1 Introduction Nothing so dif®cult as a beginning In poesy, unless perhaps the end. (Lord Byron, Don Juan, 18211 ) De®nitional issues It was Farmer Nicholas Snowe in Lorna Doone who said, with the insight given to simple rustics in Victorian novels, `virst zettle the pralimbinaries; and then us knows what be drivin' at'.2 In an enterprise such as the present one, settling the preliminaries inevitably comprises de®nitional matters, and this means the two principal objects of our concern: economics and culture. It might appear that the ®rst of these could be dispensed with quickly. There is apparently so little disagreement among contemporary economists as to the scope and content of their discipline that the introductory chapters of most modern textbooks of economics are virtually identical. The outline of the `economic problem' always emphasises scarcity, such that the decision facing actors in the economic drama is one of how to allocate limited means among competing ends. Individual consumers have wants to be satis®ed, productive enterprises have the technologies to provide the goods and services to satisfy those wants and processes of exchange link the one side of the market with the other. Much of the economics that is taught to students at universities and colleges throughout the western world nowadays is concerned with the ef®ciency of these processes of production, consumption and exchange, much less is concerned with questions of equity or fairness within the operation of economic systems. As a result, issues such as that of redistributive justice tend to play a secondary role in the thinking of many younger professional economists, if indeed such issues bother them at all. 1 2 Economics and Culture The introductory textbooks also universally make the standard distinction between the study of the micro behaviour of individual units in the economy ± consumers and ®rms ± and the macro behaviour of the economy itself. In so doing, these texts lay the foundation for the rei®cation of the economy, a process which has had profound effects on popular perceptions of economics and on the construction of public policy in the present generation. The increasing dominance of macroeconomics as the foundation stone of national and international public policy over recent decades has led to perceptions of the economy as having an identity of its own which seems to transcend its constituent elements. Ironically this view could be seen to parallel the concept of the state as having an independent existence, a concept eschewed by the model of libertarian individualism which is central to modern economics. In some cases the rei®cation of the economy in the media and elsewhere seems to extend almost to personalisation; we speak of economies as `strong' or `weak', `dynamic' or `sluggish', needing to be nursed when they are sick and requiring the administration of appropriate medicines to bring them back to health. In considering these texts as providing a de®nition of the domain and methods of contemporary economics, we should bear in mind that they mostly re¯ect the dominant neoclassical paradigm which has held sway in economics for the better part of a century and which in the last few decades has been brought to a high level of theoretical and analytical re®nement. This paradigm has provided a comprehensive and coherent framework for representing and analysing the behaviour of individuals, ®rms and markets, and it has yielded an array of testable hypotheses which have been subject to extensive empirical scrutiny. Moreover, the range of phenomena which it has embraced has been continually expanding; the model of rational utilitarian decision-making operating within competitive markets has in recent years been applied to an everwidening array of areas of human behaviour, including marriage, crime, religion, family dynamics, divorce, philanthropy, politics and law, as well as production and consumption of the arts. Yet despite its intellectual imperialism, neoclassical economics is in fact quite restrictive in its assumptions, highly constrained in its mechanics and ultimately limited in its explanatory power. It has been subject to a vigorous critique from both within and without the discipline. Furthermore, its supremacy can be challenged if a broader view of the discourse of economics is taken. In common with all great areas of intellectual endeavour, economics comprises not a single paradigm, but a number of schools of thought offering alternative or contestable ways of analysing the functioning of the economy or the actions of individual Introduction 3 economic agents. For present purposes, we are quite likely to ®nd such alternative approaches useful in thinking about cultural phenomena. But while de®ning economics and the economy may, for the time being, be disposed of relatively easily, de®ning culture is an altogether different story. Raymond Williams describes culture as `one of the two or three most complicated words in the English language'.3 Robert Borofsky suggests that attempts to de®ne culture are `akin to trying to encage the wind';4 this picturesque metaphor captures the protean nature of culture and emphasises how hard it is to be precise about what the term means. The reasons are not dif®cult to ®nd. `Culture' is a word employed in a variety of senses in everyday use but without a tangible or generally agreed core meaning. At a scholarly level it relates in some way or another to concepts and ideas which occur throughout the humanities and social sciences, but it is often deployed without precise de®nition and in ways which differ both within and between different disciplines.5 As always, an etymological analysis can throw some light on the evolution of meaning. The original connotation of the word `culture', of course, referred to the tillage of the soil. In the sixteenth century this literal meaning became transposed to the cultivation of the mind and the intellect. Such ®gurative usage is still in active service today: we refer to someone well versed in the arts and letters as a `cultured' or `cultivated' person, and the noun `culture' is often used without quali®cation to denote what, under a more restrictive de®nition, would be referred to as the products and practices of the `high' arts. But since the early nineteenth century the term `culture' has been used in a broader sense to describe the intellectual and spiritual development of civilisation as a whole. In turn, this usage became focused onto these same characteristics when evidenced in particular societies, such as nation states. In due course this humanistic interpretation of culture was supplanted by a more all-encompassing concept whereby culture was seen to embrace not just intellectual endeavour but the entire way of life of a people or society. All of these usages, and more, survive in various guises today. How, then, are we to make progress in de®ning culture in a manner that is analytically and operationally useful? Some usages are so narrow as to be restrictive of the range of phenomena that are our legitimate concern; others, such as the all-inclusive societal de®nition where culture is in effect everything, become analytically empty and operationally meaningless. Despite these dif®culties, it is possible to re®ne the range of de®nitions down to two, and indeed these will be taken to be the dual sense in which the term `culture' will be used throughout this book. The ®rst sense in which we shall use the word `culture' is in a broadly 4 Economics and Culture anthropological or sociological framework to describe a set of attitudes, beliefs, mores, customs, values and practices which are common to or shared by any group. The group may be de®ned in terms of politics, geography, religion, ethnicity or some other characteristic, making it possible to refer, for example, to Mexican culture, Basque culture, Jewish culture, Asian culture, feminist culture, corporate culture, youth culture and so on. The characteristics which de®ne the group may be substantiated in the form of signs, symbols, texts, language, artefacts, oral and written tradition and by other means. One of the critical functions of these manifestations of the group's culture is to establish, or at least to contribute to establishing, the group's distinctive identity, and thereby to provide a means by which the members of the group can differentiate themselves from members of other groups. This interpretation of culture will be especially useful for present purposes in examining the role of cultural factors in economic performance and the relationship between culture and economic development. The second de®nition of `culture' has a more functional orientation, denoting certain activities that are undertaken by people, and the products of those activities, which have to do with the intellectual, moral and artistic aspects of human life. `Culture' in this sense relates to activities drawing upon the enlightenment and education of the mind rather than the acquisition of purely technical or vocational skills. In such usage, the word is more likely to occur as an adjective than as a noun,6 as in `cultural goods', `cultural institutions', `cultural industries' or the `cultural sector of the economy'. To give this second de®nition more precision, let us propose that the connotation contained in this usage of the word `culture' can be deemed to derive from certain more or less objectively de®nable characteristics of the activities concerned. Three such characteristics are suggested. They are: . that the activities concerned involve some form of creativity in their production . that they are concerned with the generation and communication of symbolic meaning, and . that their output embodies, at least potentially, some form of intellectual property. Of course, any such list presupposes a further set of de®nitions; words such as `creativity', `symbolic meaning' and even `intellectual property' beg some further elaboration, to which in due course we shall return. For now, let us accept a standard interpretation of these terms to allow us to proceed with a working de®nition of culture in this functional sense. Generally speaking possession of all three of these characteristics Introduction 5 could be regarded as a suf®cient condition in order for this interpretation of culture to apply to a given activity. So, for example, the arts as traditionally de®ned ± music, literature, poetry, dance, drama, visual art and so on ± easily qualify. In addition, this sense of the word `culture' would include activities such as ®lm-making, story-telling, festivals, journalism, publishing, television and radio and some aspects of design, since in each case the required conditions are, to a greater or lesser degree, met. But an activity such as, say, scienti®c innovation would not be caught by this de®nition, because although it involves creativity and could lead to output capable of being copyrighted or patented, it is directed generally at a routine utilitarian end rather than at the communication of meaning.7 Similarly, road signs convey symbolic meaning in a literal sense but fail on the other criteria to qualify as cultural products. Organised sport occupies a somewhat ambiguous position. While sport possibly meets all of the three criteria, some people may still ®nd dif®culty accepting it as a cultural activity, especially if it is thought that it does not embody creativity but only technical skill. Nevertheless, there can be little doubt that sport is an element of culture in the ®rst sense de®ned above, that is as a ritual or custom expressing shared values and as a means of af®rming and consolidating group identity.8 While the three criteria listed above may be suf®cient for providing a functional de®nition of culture and cultural activities, they may not be the full story when it comes to de®ning cultural goods and services as a distinct category of commodities for purposes of economic analysis. There has been some debate among cultural economists as to whether a class of goods exists, called `cultural goods', which can be differentiated in some fundamental way from `ordinary economic goods'.9 The above criteria can be seen as a useful ®rst step towards making such a distinction, and indeed they might on their own provide a suf®ciently precise de®nition for some purposes. However, in other contexts a more rigorous speci®cation may be necessary, requiring some appeal to questions of cultural value, a matter to which we return in chapter 2. It should be noted that no universality can be claimed for these two de®nitions of culture. Some phenomena that some people may describe as culture will lie beyond their reach. Furthermore, the de®nitions are by no means mutually exclusive, but overlap in a number of important ways ± the functioning of artistic practices in de®ning group identity, for example.10 In addition, counter-examples and anomalies can doubtless be suggested. But as a basis for proceeding, the de®nitions will serve our purpose. 6 Economics and Culture Some quali®cations Three aspects of these de®nitions of culture require further elaboration. The ®rst is the fact that, although the term `culture' is used generally in a positive sense, implying virtuous and life-enhancing qualities, there is a spectre at the feast: culture, in the ®rst of the connotations de®ned above, can also be deployed as an instrument of brutality and oppression. The Soviet state culture that was imposed on artists like Shostakovitch, the cultural underpinnings of Nazism, religious wars, ethnic cleansing, the `culture of corruption' that may exist in a police force or an organisation, the gang culture that rules on the streets of large cities, ma®a culture and other such phenomena, are all examples of shared values and group identi®cation that can indeed be construed as manifestations of culture, if it is de®ned as we have above. One approach to the dark side of culture is to ignore it, to make no value judgement as to good or bad cultures, and simply to analyse all cultural phenomena at face value as they present themselves. An alternative that confronts this issue more directly is to admit the possibility of an ethical standard which would outlaw, by common consent, certain characteristics which were universally agreed to be unacceptable. Such a standard might incorporate notions such as fairness, democracy, human rights, free speech and freedom from violence, war and oppression, as basic human values. Acceptance of such a standard would disqualify all the negative examples listed above from consideration as culture, and would prevent certain barbaric and oppressive practices from being excused on the grounds that they were part of the cultural tradition of a particular group. It might be observed that a resolution of the problem of negative cultural manifestations in this way could itself be interpreted in cultural terms. Suppose a minimum ethical standard could be generally agreed upon which accepted as axiomatically desirable such concepts as individual rights, democracy, the protection of minorities, peaceful resolution of con¯ict and the promotion of civil society.11 It could be argued in such a circumstance that the values enshrined as universal could be seen to comprise the de®ning symbols of civilised human existence, and as such could be interpreted as key elements of an overarching human culture which transcends other forms of cultural differentiation. The second issue requiring some further elaboration is whether culture is a thing or a process. In the above de®nitions, we have emphasised the former, de®ning a set of characteristics which describe what culture is, rather than who makes it, or who decides how it is used. When the idea Introduction 7 of culture as process is entertained, questions are raised about power relationships between affected and affecting groups.12 Culture may in these circumstances become a contested phenomenon rather than an area of agreement and harmony. So, for example, it becomes possible to speak of a dominant culture, imposed intentionally or otherwise by an elite group in society on an unwilling or unwitting populace. It also opens up the question of de®ning `popular culture', an area seen in contemporary cultural studies as being oppositional to the hegemonic and restrictive practices of `high culture'. Furthermore, concepts of culture as transactional emphasise the fact that culture is not homogeneous and static, but an evolving, shifting, diverse and many-faceted phenomenon. The effect of these considerations is not so much to undermine or replace the concept of culture as an inventory of objects or practices, but rather to suggest that the inventory becomes unstable and its content contestable when the dynamics of cultural processes and the power relationships they imply are brought into account. The third aspect requiring clari®cation is the question of how far the de®nitions of culture as proposed above overlap with ideas about society which are the substance of sociological concern. It might be suggested that a de®nition of culture which relies on identifying distinguishing characteristics of groups might be seen to parallel a notion of such groups as societies or as social units within a society. Thus, for example, to say that traditions, customs, mores and beliefs comprise the culture of a group might simply describe a set of variables which, to a sociologist, de®ne the basis for providing social cohesion and social identity to the group. Nevertheless, while there will inevitably be some blurring of the lines between cultural and social, and between culture and society, it can be argued that there is a suf®ciently clear distinction to allow these domains to be separated, as indeed Raymond Williams was able to do in his in¯uential work whose title, Culture and Society, crystallises such a distinction.13 If culture, in both of the senses de®ned above, embraces the intellectual and artistic functions of humankind (even if these are exercised unconsciously, as for example in the use of language), its source can be differentiated from those processes of social organisation, both deliberate and spontaneous, which go towards de®ning society. Is economics culture-free? The cultural context of economics The formal precision of modern economics, with its theoretical abstraction, its mathematical analytics and its reliance on disinterested scienti®c 8 Economics and Culture method in testing hypotheses about how economic systems behave, might suggest that economics as a discipline does not have a cultural context, that it operates within a world that is not conditioned by, nor conditional upon, any cultural phenomena. But just as the radical critique of contemporary economics has argued that the sort of economics described above cannot be value-free, so also can it be suggested that economics as an intellectual endeavour cannot be culture-free. To begin with, it is apparent that the many schools of thought that go to make up the full complement of economic science as it has evolved over at least two centuries themselves comprise a series of separate cultures or subcultures, each de®ned as a set of beliefs and practices which bind the school together. Just as shared values provide the basis for cultural identity of various sorts in the world at large, so also in the restricted domain of the intellectual discourse of economics we can interpret the coalescence of schools of thought, whether they be Marxist, Austrian, Keynesian, neoclassical, new classical, old institutional, new institutional or whatever, as a cultural process. However, the impact of culture on the thinking of economists goes further, because the cultural values they inherit or learn have a profound and often unacknowledged in¯uence on their perceptions and attitudes. Of course, to argue that cultural considerations affect the way in which economists practise their trade is simply an extension of the well known argument that the ideological standpoint of the observer in¯uences the way he or she perceives the world, and that objectivity in the social sciences generally is impossible since even the choice of which phenomena to study is itself a subjective process. Recognising this in the present context, we might ask, for example, whether the apparent acceptance by the great majority of contemporary Western economists of the dominant intellectual paradigm in their discipline ± a belief in the ef®cacy of competitive markets, the foundation upon which the political system of capitalism is built ± derives from a process of intellectual persuasion or simply from an unexamined cultural predisposition shaped by the values of their profession. Furthermore the cultural context of economics as a discipline relates not only to the conditioning of its practitioners, but also to the methodology of its discourse. The processes by which economic ideas are generated, discussed, appraised and transmitted have been subject to analysis in terms which draw upon the work of theoreticians in literary and critical analysis such as Derrida and Foucault. Turning attention to the textual nature of economic knowledge and to the functioning of rhetoric in economic discourse has been seen by economists such as Deirdre McCloskey as opening up new `conversations' in the philosophy Introduction 9 of economics and in the interpretation of the history of economic thought.14 Argument, persuasion and other processes involved in conversations among economists or between economists and others have clear cultural connotations, as indicated, for example, in Arjo Klamer's writings on the growth, communication and dissemination of economic knowledge;15 it is perhaps no coincidence that Klamer occupies the world's ®rst chair in the economics of art and culture, at Erasmus University in Rotterdam. Let us turn now from the cultural context of economics as a system of thought to the cultural context of the economy as a system of social organisation. The fact that economic agents live, breathe and make decisions within a cultural environment is readily observable. So, too, is the fact that this environment has some in¯uence on shaping their preferences and regulating their behaviour, whether this behaviour is observed at the level of the individual consumer or ®rm or at the aggregated level of the macroeconomy. Yet in its formal analytics, mainstream economics has tended to disregard these in¯uences, treating human behaviour as a manifestation of universal characteristics which can be fully captured within the individualistic, rational-choice, utilitymaximising model, and seeing market equilibria as being relevant to all circumstances regardless of the historical, social or cultural context.16 Indeed, when neoclassical modelling does attempt to account for culture, it can do so only within its own terms. So, for example, Guido Cozzi interprets culture as a social asset that enters the production functions of labour ef®ciency units as a public-good input within an overlappinggenerations model.17 While such efforts may capture some of the characteristics of culture in an abstract economy, they remain remote from an engagement with the wider issues of culture and real-world economic life. At the same time it is important to note that there has long been an interest in examining the role of culture as a signi®cant in¯uence on the course of economic history, arising within several schools of economic thought. Perhaps the most celebrated contribution to the ®eld has been Max Weber's analysis of the in¯uence of the Protestant work ethic on the rise of capitalism.18 Here the cultural conditions in which economic activity occurs are linked very directly to economic outcomes. Many other speci®c illustrations of the historical in¯uence of culture on economic performance can be cited. For example, the spirit of individualism inherent in Anglo-Saxon culture, ®rst noted in Adam Smith's discussion of the division of labour, and developed further by the great nineteenth-century political economists, especially John Stuart Mill, can be seen to have provided the conditions for the spread of the industrial 10 Economics and Culture revolution in Britain and almost concurrently in the United States.19 Closer to our own time, there has been much speculation on what it is that explains the `Asian economic miracle' in the post-war years, beginning with the spectacular industrial dynamism of Japan, and followed by the phenomenal growth rates in South Korea, Taiwan, Hong Kong and Singapore. We shall return to these questions in chapter 4. Culture as economy: the economic context of culture In the same way as economic discourse and the operation of economic systems function within a cultural context, so also is the reverse true. Cultural relationships and processes can also be seen to exist within an economic environment and can themselves be interpreted in economic terms. Both of the conceptualisations of culture de®ned earlier ± the broad anthropological de®nition and the more speci®c functional intention of culture ± can be considered in this light. Let us deal with them in turn. If culture can be thought of as a system of beliefs, values, customs, etc. shared by a group, then cultural interactions among members of the group or between them and members of other groups can be modelled as transactions or exchanges of symbolic or material goods within an economising framework. Anthropologists have characterised primitive and not-so-primitive societies in these terms, where ideas of markets, exchange value, currency, price and other such phenomena take on cultural meaning. One speci®c area of interest has been built around the proposition that all cultures are adapted to, and are explicable through, their material environment. Cultures may differ, but their evolution will be determined not by the ideas that they embody but by their success in dealing with the challenges of the material world in which they are situated. Such `cultural materialism' has a clear counterpart in economics, especially in the `old' school of institutional economics, where culture underpins all economic activity. Indeed William Jackson sees cultural materialism as providing the means of reintegrating culture into the same material, natural world as economics.20 Furthermore, considerations of the role of culture in the economic development of the Third World place the cultural traditions and aspirations of poor people into an economic framework, as a means of identifying ways in which their material circumstances can be improved in a manner consistent with cultural integrity. In fact, as the UN World Commission on Culture and Development (1995) has made abundantly Introduction 11 clear, the concepts of culture and of development are inextricably intertwined in any society. Thus, for example, development projects in poor countries such as those ®nanced by international agencies, NGOs, foreign aid programmes and so on, are likely to be effective in raising living standards in such countries only if they recognise that the culture of the target community is the fundamental expression of their being, and that this culture is placed within an economic context that determines the scope and extent of material progress that is possible. We consider these issues more fully in chapter 4. Turning now to the interpretation of culture in functional terms, we can again identify the notion of culture as economy and the interpretation of culture as residing within an economic milieu. Perhaps the most obvious place to start is with the proposition that cultural production and consumption can be situated within an industrial framework, and that the goods and services produced and consumed can be regarded as commodities in the same terms as any other commodities produced within the economic system. The term `cultural industry' was coined by Max Horkheimer and Theodor Adorno of the Frankfurt School in 1947 as a despairing indictment of the commodi®cation inherent in mass culture. They saw culture as being transformed by the technology and ideology of monopoly capitalism; for them, an economic interpretation of cultural processes was an expression of disaster.21 Since then notions of cultural commodi®cation have developed along several different paths indicating different contextualisations of culture within a broadly economic domain. One such path, leading through to contemporary cultural studies, recognises the pervasiveness of cultural phenomena in everyday life, and investigates popular culture, largely but not exclusively from a left viewpoint, in terms of economic and social relationships in contemporary society.22 Another line of development might be traced to postmodern thinkers such as Jean Baudrillard who locate culture in a shifting universe of tangible and intangible social and economic phenomena. Steven Connor characterises Baudrillard as arguing that `it is no longer possible to separate the economic or productive realm from the realms of ideology or culture, since cultural artefacts, images, representations, even feelings and pyschic structures have become part of the world of the economic'.23 There is a blurring of the boundary between image or `simulation' and the reality it represents (a `hyperreality'). Thus, for example, Baudrillard suggests that Disneyland is more real than the `real' United States which it imitates.24 Yet another line of development has been pursued within cultural economics. It concentrates on the production and consumption of 12 Economics and Culture culture (mostly the arts) characterised as purely economic processes. The roots of the subdiscipline of cultural economics are ®rmly planted in economics, and it can now be seen to be established as a legitimate and distinctive area of specialisation within economics, with its own international association, congresses and scholarly journal (the Journal of Cultural Economics), and its own separate classi®cation in that arbiter of the taxonomy of economic discourse, the Journal of Economic Literature. It traces its modern origins to John Kenneth Galbraith's ®rst writings in economics and art,25 though identifying its seminal work as Baumol and Bowen's book Performing Arts: The Economic Dilemma (1966). Since then a number of good specialised books have appeared, and there is an expanding theoretical and applied literature in cultural economics in academic journals26 and elsewhere. Within this tradition, the cultural industries are interpreted using the conventional paraphernalia of economic analysis, albeit with some innovative twists and adaptations to account for the peculiarities of artistic demand and supply. So, for example, the work of artists is construed as occurring within a labour market whose operations can be analysed using concepts familiar to economists such as labour supply equations and earnings functions, but whose predictions of behaviour may diverge from the expected because of the idiosyncrasies of artists as a class of workers. When viewed in this way the cultural industries can easily be integrated into a wider model of an economy, such as an input±output model, where the relationships between culture and other industries can be spelled out. The commodi®- cation of culture involved in this approach does not crowd out other constructions of cultural production, including the view that art can be rationalised only in self-referential terms. Rather, this economic view of culture simply accepts as a descriptive fact that the activities of producing and consuming cultural goods and services within an economic system do generally involve economic transactions, that these activities can be encircled in some way and that what is contained within the circle can be called an industry and analysed accordingly. We shall return to this interpretation of culture as an industry in chapter 7. Individualism and collectivism Our overview of the domain of this book in this introductory chapter has referred to the fact that economic thought as it has evolved over two centuries is founded on individualism,27 whereas the notion of culture, at least in the senses de®ned above, is a manifestation of group or collective Introduction 13 behaviour. It is useful, in concluding this introduction, to codify this distinction between economics and culture, as a basis for our further consideration of the interrelationships between them. We do this by putting forward the following proposition: the economic impulse is individualistic, the cultural impulse is collective. This proposition asserts, ®rst, that there is behaviour which can be termed `economic' which re¯ects individual goals and which is portrayed in the standard model of an economy comprising self-interested individual consumers seeking to maximise their utility and self-interested producers seeking to maximise their pro®ts. The ®rst part of the proposition remains true in an economy where large corporations dominate the production sector, since they represent simply the means whereby their owners and managers can pursue their own economic selfinterest more effectively. In the standard neoclassical model of the economy, markets exist to enable mutually bene®cial exchange to occur, and according to the theory of general equilibrium such markets will lead under certain assumptions to the maximisation of social welfare, de®ned only in terms of the individuals who comprise the economy, and given the initial distribution of income. Of course in this economy collective action may occur. If markets fail or do not exist, voluntary or coercive collective action may be required in order for optimal social outcomes to be achieved. For example, public goods such as national defence or law and order, which cannot be ®nanced directly through individual demand, must be supplied through the state or via voluntary cooperation. Other forms of cooperative behaviour will spring up in an individualistic economy. But all these manifestations of collective action are traceable back to individual demand and, within the economic model, to the self-interest of the economic agents involved. Even altruism is identi®able in this model as an expression of individual utility maximisation. The above proposition asserts secondly that there is behaviour, distinguishable from the economic behaviour described in the previous paragraphs, which can be termed `cultural'; such behaviour re¯ects collective as distinct from individualistic goals, and derives from the nature of culture as expressing the beliefs, aspirations and identi®cation of a group as de®ned above. Thus the cultural impulse can be seen as a desire for group experience or for collective production or consumption that cannot be fully factored out to the individuals comprising the group. These desires range over many types of activities, but we might use the arts as illustration. On the production side, many artistic goods and services are produced by group activity where the outcome is a collective effort acknowledged by the participants as having a value or meaning 14 Economics and Culture beyond that which could simply be attributed to the totality of the inputs of the individuals involved.28 Similarly consumption of the arts ± for example, in theatres and concert halls ± is frequently a collective activity moved by a sense that the group experience transcends that of the sum of the individual consumer responses. Of course again seemingly contrary cases can be cited. Much art is produced as an individual, even lonely activity, and a person reading a novel or listening to music in the privacy of their own home is engaged in solely individualistic consumption. Nevertheless, artists working alone are generally doing so in the expectation that their work will communicate with others; similarly, lone consumers of the arts are likely to be making some wider human connection. Thus, whatever the artistic products produced and consumed, the processes of producing and consuming them can be seen not only as individual enterprise, but also as expressions of a collective will which transcends that of the individual participants involved. To sum up, we are suggesting that the economic impulse as speci®ed in the above proposition can be seen as expressing the individual desires of members of society on their own behalf, and the cultural impulse can be seen as gathering together the collective desires of the group or groups within a given society for the sorts of cultural expression referred to in the de®nitions of culture put forward earlier. This proposition will provide a useful basis for distinguishing the economic and the cultural throughout our subsequent discussion in this book. Outline of the book It can be suggested that at some fundamental level, the conceptual foundations upon which both economics and culture rest have to do with notions of value. Certainly theories of value have been central to the development of economic thought since Adam Smith, and whatever the disciplinary starting point for a consideration of culture, whether it be aesthetics or contemporary cultural studies, questions of value are fundamental there, too. Chapter 2 of this volume therefore lays the theoretical basis for the remainder of our work, by considering notions of value in economics and in culture, how they are codi®ed and how they are assessed. The next step then is to propose a means of representing culture in terms which may be capable of bridging whatever divide exists between it and economics, that is to propose a way of conceptualising culture in a form which captures its essential characteristics but which is also amenable to economic manipulation and analysis. This step is taken Introduction 15 in chapter 3, where the notion of `cultural capital' is put forward, as a means of representing both tangible and intangible manifestations of culture. The de®nition of cultural capital depends on our earlier consideration of economic and cultural value, and allows the characterisation of cultural goods and services, cultural activities and other phenomena in a way which recognises both their economic and their cultural importance. Given the long time-frames within which in practical terms culture needs to be evaluated, we go on to consider the intertemporal characteristics of cultural capital: how it is received as a bequest from the past, how it is dealt with in the present and how it is handed on to the future. Such an agenda can be brought together under the ubiquitous rubric of `sustainability', in a manner which parallels the treatment of natural capital in an environmental and ecological context. Chapters 4 and 5 then take up two speci®c aspects building on the discussion of value and of cultural capital and sustainability. In chapter 4 we consider culture in economic development, looking ®rst at cultural determinants of economic performance, and then broadening our consideration to the role of culture as the means of representing the whole gamut of human development in both developing and industrialised countries. Sustainability questions are important here, not just for the importance of a long-term perspective on economic development, but also because of the need to maintain cultural systems as integrating elements in the development process. Chapter 5 looks speci®cally at the economics of cultural heritage, perhaps the most obvious manifestation of cultural capital and one where sustainability principles can be most clearly articulated and applied. In chapter 6 we turn our attention to the process of creativity as a mainspring of cultural growth and development. Economists have long been interested in creativity as a motivating force in innovation and technological change, but have rarely been concerned to venture beneath the surface to speculate about the origins of the creative drive and about the ways in which economic incentives and constraints in¯uence its expression among individual creators, apart from the somewhat sweeping observation that the principal incentive to innovation is the prospect of pro®t. Again, issues of value are central to our argument; in characterising the work of creative artists, we are able to identify the production of both economic and cultural value in the generation and disposition of their work. We return in chapter 7 to the notion of the cultural industries as a means of representing cultural activity in economic terms. The approach which has developed as the core of cultural economics over the past thirty years is reviewed and discussed in the context of construing the 16 Economics and Culture creative arts as an industry. We then widen the focus to consider the cultural industries more generally, with attention to culture in urban development, tourism and trade. We also discuss the potential for the cultural industries to contribute to economic development in the developing world, using the music industry as a case study. If cultural activity in the economy is interpreted in an industrial framework, and if the cultural industries so designated lead to economic output and generate employment, governments which may be unconcerned about culture per se will begin to take an interest. This leads us directly into chapter 8 where we take up the more general issue of how the state might intervene to affect culture in some way or another. The means of such intervention is cultural policy, an area of government involvement which has emerged as a speci®c policy arena only in very recent times. Given that in the contemporary world much of public policy is concerned with economic phenomena, it is not surprising that cultural policy raises very directly the relationship between economics and culture. Cultural policy also has signi®cant political rami®cations. In our discussion we focus attention on the complementarities and con¯icts between cultural and economic policy: again, questions of economic and cultural value emerge as decisive in mapping the territory and in guiding decision-making. Finally, in the concluding chapter 9, some effort is made to draw the threads together and to point the way ahead. In a contemporary world where we can see juxtaposed the oppositional dynamics of economic globalisation and cultural differentiation, can the twin subjects of this volume, economics and culture, be seen as two organising principles for contemporary society, de®ning both the scope and limitations for civilised progress into the third millennium? Notes 1 Quotation is from Canto IV, stanza I, lines 1±2; see Byron (1986, V, p. 203). 2 See Blackmore (1869, p. 82). 3 Williams (1976, p. 76). 4 Borofsky (1998, p. 64). 5 For considerations of the changing concept of culture within anthropology, a discipline where culture is at its very core, see Marcus and Fischer (1986), the appendix to Ruttan (1988), Appadurai (1996, ch. 3), and the further references contained in notes 1 and 2 of Borofsky (1998). For theories of culture in sociology see Di Maggio (1994) and in psychology see Cooper and Denner (1998). 6 Williams (1976, p. 81) notes that the adjectival form dates only from the late nineteenth century. 7 Except insofar as basic scienti®c research ± pure rather than applied ± may be Introduction 17 aimed at a general advancement of knowledge and understanding, and as such could be seen as bearing some similarities with art. 8 For a review of the relationships between the economics of sport and of the arts, see Seaman (1999). 9 This debate has focused almost entirely on the demand side, with cultural goods being distinguished by the peculiar nature of tastes for them, a matter to which we return on pp. 114ff. below; see also the delineation of `creative goods' in Caves (2000). 10 A good illustration is pop music; see Dolfsma (1999). 11 Such a system of `global ethics' has been proposed by the World Commission on Culture and Development (1995, pp. 33±51); nevertheless, despite the apparent persuasiveness of such proposals, reaching agreement on universal ethical standards remains, both theoretically and practically, a matter of considerable controversy. 12 For discussion of these issues, see Wright (1998). 13 Williams (1958); for a consideration of the society±culture relationship, see Peterson (1976). 14 See McCloskey (1985, 1994); Amariglio (1988). 15 See, for example, Klamer's comparison between the shared values of baseball fans in conditioning their discussions (a cultural phenomenon), and the skills required of participants in economic discourse (Klamer 1988, pp. 260±2). 16 Economists outside the mainstream, however, are less narrowly focused. Institutional economists, for example, take culture as the foundation of economic processes and treat `all human behaviour [as] cultural behaviour' (Mayhew 1994, p. 117); see also North (1990), Stan®eld (1995). For a penetrating appraisal of economics as culture from an anthropological viewpoint, see Escobar (1995, ch. 3). 17 Cozzi (1998). 18 First published in 1904±5; for an English translation see Weber (1930). See further in O'Neil (1995); Armour (1996). 19 See Landes (1969); Temin (1997). 20 Jackson (1996); for the origins of cultural materialism, see Harris (1979). 21 See Adorno and Horkheimer (1947); Adorno (1991). 22 Storey, (1993, pp.6±18). 23 Connor (1997, p. 51). 24 Baudrillard (1994, pp. 12±14); see also Best and Kellner (1991, pp. 111±45), Storey (1993, pp. 162±5). 25 Of which the ®rst appears to be Galbraith (1960, Ch. 3). 26 For reviews of the scope and content of cultural economics, see Throsby (1994b); Hutter (1996); Towse (1997a, vol. I, pp. xiii-xxi); Blaug (1999). Some signi®cant books in the ®eld include Peacock and Weir (1975); Blaug (1976); Netzer (1978); Throsby and Withers (1979); Hendon, Shanahan and MacDonald (1980); Feld, O'Hare and Schuster (1983); Hendon and Shanahan (1983); Kurabayashi and Matsuda (1988); Frey and Pommerehne (1989); Grampp (1989); Feldstein (1991); Moulin (1992); Towse and Khakee 18 Economics and Culture (1992); Heilbrun and Gray (1993); Mossetto (1993); Peacock (1993); Towse (1993); Trimarchi (1993); Farchy and Sagot-Duvauroux (1994); Peacock and Rizzo (1994); Ginsburgh and Menger (1996); Klamer (1996); Hutter and Rizzo (1997); Cowen (1998); O'Hagan (1998); Benhamou (2000); Frey (2000); and many others. Towse (1997a) reprints a large number of important journal articles in this ®eld. 27 For an account of the place of methodological individualism in the history of economic thought, see Infantino (1998); for a discussion relating speci®cally to the economics of the arts, see Rushton (1999). 28 For a view of the sociological signi®cance of artistic production as collective action, see Becker (1974).

NBER WORKING PAPER SERIES DIGITAL ECONOMICS Avi Goldfarb Catherine Tucker Working Paper 23684 http://www.nber.org/papers/w23684 NATIONAL BUREAU OF ECONOMIC RESEARCH 1050 Massachusetts Avenue Cambridge, MA 02138 August 2017 We thank Andrey Fradkin and Kristina McElheran for helpful comments. We are grateful to the Sloan Foundation for its support of the NBER Digitization Initiative, which built the research community around which this review is based. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research. NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications. © 2017 by Avi Goldfarb and Catherine Tucker. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source. Digital Economics Avi Goldfarb and Catherine Tucker NBER Working Paper No. 23684 August 2017 JEL No. L81,L86,O33 ABSTRACT Digital technology is the representation of information in bits. This technology has reduced the cost of storage, computation, and transmission of data. Research on digital economics examines whether and how digital technology changes economic activity. In this review, we emphasize the reduction in five distinct economic costs associated with digital economic activity: Search costs, replication costs, transportation costs, tracking costs, and verification costs. Avi Goldfarb Rotman School of Management University of Toronto 105 St. George Street Toronto, ON M5S 3E6 CANADA and NBER agoldfarb@rotman.utoronto.ca Catherine Tucker MIT Sloan School of Management 100 Main Street, E62-533 Cambridge, MA 02142 and NBER cetucker@mit.edu Digital Economics Avi Goldfarb∗and Catherine Tucker† July 25, 2017 Abstract Digital technology is the representation of information in bits. This technology has reduced the cost of storage, computation, and transmission of data. Research on digital economics examines whether and how digital technology changes economic activity. In this review, we emphasize the reduction in five distinct economic costs associated with digital economic activity: Search costs, replication costs, transportation costs, tracking costs, and verification costs. Digital technology is the representation of information in bits. This reduces the cost of storage, computation, and transmission of data. Research on digital economics examines whether and how digital technology changes economic activity. Understanding the effects of digital technology does not require fundamentally new economic theory. However, it requires a different emphasis. Studying digital economics starts with the question of “what is different?” What is easier to do when information is represented by bits rather than atoms? Digital technology often means that costs that may ∗Avi Goldfarb is the Ellison Professor of Marketing at the Rotman School of Management, University of Toronto, and Research Associate at the NBER. We thank Andrey Fradkin and Kristina McElheran for helpful comments. We are grateful to the Sloan Foundation for its support of the NBER Digitization Initiative, which built the research community around which this review is based. †Catherine Tucker is the Sloan Distinguished Professor of Management Science at MIT Sloan School of Management, Cambridge, MA, and Research Associate at the NBER. 1 constrain economic actions. Therefore, digital economics explores how standard economic models change as certain costs fall substantially and perhaps approach zero. We emphasize how this shift in costs can be divided into five types: 1. Lower search costs 2. Lower replication costs 3. Lower transportation costs 4. Lower tracking costs 5. Lower verification costs Search costs are lower in digital environments, enlarging the potential scope and quality of search. Digital goods can be replicated at zero cost, meaning they are often non-rival. The role of geographic distance changes as the cost of transportation for digital goods and information is approximately zero. Digital technologies make it easy to track any one individual’s behavior. Last, digital verification can make it easier to verify the reputation and trustworthiness of any one individual, firm, or organization in the digital economy. Each of these cost changes draws on a different set of well-established economic models: Primarily search models, non-rival goods models, transportation cost models, price discrimination models, and reputation models. Early research tested straightforward models of lower costs. For example, the search literature of the late 1990s and early 2000s built directly on earlier models by Diamond (1971) and Varian (1980). As we detail below, empirical work emerged that found some inconsistencies with the simple models, and so richer models and empirical analysis of the cost reductions developed to take account of the subtleties of the digital context. Other authors have also emphasized the role of lower costs for digital economics (e.g. Shapiro and Varian (1998), Borenstein and Saloner (2001), Smith et al. (2001), etc.) Ellison 2 and Ellison (2005) the implications of these lower search and transportation costs for industrial organization with respect to increasing returns, distance and two-sided markets. Since their article, the digital economics literature has grown to contribute to the economics of crime, the economics of public goods, organizational economics, finance, urban economics, labor economics, development economics, health economics, political economy, media economics, public finance, and international economics. In this sense, we view digital economics as a way of thinking that touches many fields of economics. This review starts with a brief history of digital technology and the internet. It will then discuss each of the cost changes associated with digitization. In each section, we emphasize the key research questions that have driven the area and how they have evolved, and relate them to policy where applicable. After discussing each cost change, we finish by discussing the consequences of digitization for countries, regions, firms, and individuals. 1 Digital Technology: A Brief History The history of modern computing begins not with the internet but in 1945 with the commercialization of technologies developed during World War II (Ceruzzi, 2003). These first machines focused on rapid calculation with little capacity for storing and retrieving information. By the early 1950s, magnetic core memories enabled efficient digital information storage and perhaps the first real non-arithmetical benefit of representing information in bits emerged: The lower marginal cost of reproducing information. Over time, storage technology, software, and hardware improved so that information processing and reproduction became widespread. The software and hardware industries grew rapidly (Ceruzzi, 2003; Campbell-Kelly, 2004). Limited communication between computers limited their effect on the economy. It was with the rise of the internet–and with it, low-cost, commercial, computer-to-computer communication–that the representation of information in bits began to have a measurable 3 effect on multiple markets. This rise was built on key inventions developed by US military funding in the 1960s and 1970s (Hafner and Lyon, 1996; Greenstein, 2015). For example, DARPA funded the invention of packet switching which breaks down a long message into shorter messages that can be sent through the network and then reassembled upon receipt. DARPA-funded researchers also developed the particular packet switching standards that define internet communication: the Transmission Control Protocol/Internet Protocol or TCP/IP. The NSF began managing a network using that protocol in the 1980s, building a reliable infrastructure that was relatively easy to adopt but also restricted to researchers. Privatization occurred between 1990 and 1995, leading to the modern commercial internet. The commercial internet diffused quickly, with universities playing a key role in the diffusion process (Goldfarb, 2006). There was near-universal availability and widespread adoption in the United States by 2000 (Greenstein, 2000).1 Over time, new technologies have been layered on top of the basic TCP/IP-based internet, including browsers, search engines, online shopping, social networks, mobile communications protocols, security standards, customer relationship management systems, and many others. These technologies and others have enabled increased collection and use of data. During this process there has been an open question of who should control various aspects of commercial internet activity given this historical context of decentralization. Standards are often agreed upon through committees with representatives from industry and academia. Such standards have an influence on which technologies are widely adopted (Rysman and Simcoe, 2008). Therefore, standards setting creates winners and losers. Simcoe (2012) examines the incentives in standards development for one such standard setting organization, the Internet Engineering Task Force, demonstrating that the commercialization of the internet slowed standards development due to competing commercial interests. Given their 1This rapid speed of diffusion proved useful for identification in the empirical papers examining the impact of the internet on regions, firms, and individuals that we discuss in the penultimate section. 4 importance, control of hardware and software standards has been controversial. Echoing this question of control, the earlier literature on the economics of the internet focused on pricing the sending of information and how it varies with interconnection, competition, and the nature of the content (MacKie-Mason and Varian, 1994). In other words, there is a question about the role of internet service providers in controlling access. Laffont et al. (2003) emphasized how the need for interconnection can affect prices and welfare. This literature emphasized network effects and the challenges of interconnection (Cremer et al., 2000; Besen et al., 2001; Laffont et al., 2001; Caillaud and Jullien, 2003). As data transmission became a key aspect of digital technology, the question of net neutrality has become a central research and policy focus. Net neutrality means that an internet service provider should treat all data in the same way; regardless of the content provider or content, companies cannot pay an internet service provider to have faster speeds. The net neutrality debate asks whether internet service providers should exercise control over content. Put differently, net neutrality is the norm that Netflix pays the same to send a gigabyte of data to one of their customers as a small startup would pay to send data to the same customer. Internet services have had a historic norm of net neutrality, though this has been challenged in recent years by internet service providers and policy makers in the US and globally. The net neutrality literature therefore emphasizes the role of the connection intermediary (Economides and Hermalin, 2012; Bourreau et al., 2015; Choi et al., 2015; Goetz, 2017). As shown by Lee and Wu (2009) and Greenstein et al. (2016) the particulars of the model matter, and the costs and benefits of net neutrality depend on the specific setting. Thus, a key theme in the history of digital technology is a tension between openness and control. As we discuss below, this tension is at the center of much of the digital policy literature with respect to copyright, privacy, and discrimination. 5 2 Reduction in Search Costs Search costs are the costs of looking for information. Every information gathering activity therefore involves search costs. The basic idea with respect to digital economic activity is that it is easier to find and compare information about potential economic transactions online than offline. At the beginning of the commercial internet, there was much discussion among economics researchers around how a dramatic reduction in search costs might transform the economy by reducing prices, price dispersion, unemployment, vacancies, and inventories. Allan Greenspan argued that the ICT revolution would reduce the severity of business cycles.2 The consequences of low search costs were discussed in financial markets (Barber and Odean, 2001), labor markets (Autor, 2001) and retail markets (Borenstein and Saloner, 2001; Bakos, 2001). The ideas in these papers have their roots in the early search literature which modeled search costs as the costs of gathering information (Stigler, 1961; Diamond, 1971; Varian, 1980). Reflecting this early focus and solid base of economic understanding, the literature on the effects of lower digital search costs is more established than the other parts of the digital economics literature. 2.1 Are prices and price dispersion lower online? Low search costs make it easier for consumers to compare prices, putting downward pressure on prices for similar products. This should reduce both prices and price dispersion. Brynjolfsson and Smith (2000) compares prices of books and CDs at four internet-only retailers, four offline retailers, and four ‘hybrid’ retailers who had both online and offline stores. They identified 20 books and 20 CDs, half of which were bestsellers, and half of which were randomly selected among titles popular enough to be sold in most offline stores. They showed 2 ’Information technology has doubtless enhanced the stability of business operations,” Federal Reserve Chairman Alan Greenspan, Feb. 26 1997 testimony before Congress. https://www.federalreserve.gov/ boarddocs/hh/1997/february/testimony.htm 6 that online prices for these items were substantially lower than offline prices. Relatively low online prices have been shown in a variety of other settings, including insurance (Brown and Goolsbee, 2002), automotive products (Zettelmeyer and Silva-Risso, 2001), and airlines (Orlov, 2011). However, though prices may be lower, substantial price dispersion remains. Brynjolfsson and Smith (2000) shows this in their online-offline retail study. Baye et al. (2004a) and Baye et al. (2004b) use evidence from thousands of products and prices to document large and persistent online price dispersion. Orlov (2011) finds that the internet increases the intrafirm dispersion of airline prices, but had no impact on interfirm price dispersion. By contrast, the development economics literature measuring the effect of mobile phones on commodity prices suggests that lower search costs reduced price dispersion (Jensen, 2007; Aker, 2010; Parker et al., 2016). Given evidence of the persistence of price dispersion online, research turned to explore why price dispersion does not disappear. Of course comparison of online products does not always compare apples-to-apples. In comparing book prices, the book may be the same, but the retailer is different. Different retailers offer different quality, different shopping experiences, and different shipping policies. Firms with higher quality may develop stronger brands, and therefore command higher prices (Waldfogel and Chen, 2006). Firms selling products can also shape the search process. When consumers search, they assess multiple dimensions of information: price, quality, reputation, shipping fees, time to delivery, color, etc. Lynch and Ariely (2000) demonstrates this for online wine purchasing in a laboratory. If price was available on the first page, consumers focused on price. If consumers needed to click further to learn the price, other attributes became more important for purchase decisions. Fradkin (2017) shows that the details of the search process matter in the context of short term accommodation platform Airbnb. Structural estimates of the cost of an extra click in the consumer search process suggest they are larger than might 7 be supposed (Honka, 2014; De Los Santos et al., 2012). This means that consumers stop searching sooner than predicted by models that assume search costs close to zero. In the presence of search costs, and multiple dimensions of information, firms can partly choose which information has the lowest search costs. Ellison and Ellison (2009a) demonstrates that computer memory chip retailers attract customers with low prices at an online price comparison website, and then show customers other (typically higher quality and higher margin) products once they arrive. Using data from Ebay, Dinerstein et al. (2017) emphasizes how the design of the search algorithm on eBay affects markups charged by eBay sellers. More directly, Hossain and Morgan (2006) shows that online sellers often hide shipping fees until the final purchase page. Moshary et al. (2017) shows a similar phenomenon in the information revealed in ticket prices at an online ticket platform. Therefore, while prices have fallen, price dispersion has persisted. The initial predictions of low price dispersion missed the point that search costs are endogenous, and so firms can manipulate the search process in order to sustain higher margins and prices. 2.2 How do low search costs affect variety? Low search costs may mean that it is easier to find rare and niche products (Yang, 2013). In this case, digital search might lead to an increase in the proportion of sales going to products that are relatively rarely purchased, a phenomenon dubbed ‘the long tail’ by Anderson (2006). Using data from a retailer with both online and offline channels, Brynjolfsson et al. (2011) documents that the variety of products available, and purchased, online is higher than offline. Low search costs may facilitate discovery of relatively unknown products (Zhang, 2016).3 Low search costs could also generate superstar effects (Rosen, 1981). If there are verti3 In addition to search costs, variety may increase because digital technologies can make inventory systems more efficient, meaning firms can hold millions of products, especially for digital goods that have no physical presence. People may also be less inhibited from purchasing non-standard items when purchasing on a screen rather than from a human (Goldfarb et al., 2015). 8 cally differentiated products, and the marginal cost of production is zero then homogeneous consumers will all agree which product is best and buy it. Consistent with this, Goldmanis et al. (2010) shows that the internet initially led to a relative increase in the number of large offline bookstores and travel agencies. Bar-Isaac et al. (2012) explains how both superstar and long tail effects may both result from a reduction in search costs. If products are both vertically and horizontally differentiated, a reduction in search costs may lead to an equilibrium where the most popular and highest quality products are produced in high enough quantity to be sold to everyone while niche products are sold through long tail retailers. The increase in tails at the right and left of the distribution comes at the cost of products in the middle. The degree to which search costs generate more or less variety depends on the search process endogenously chosen by the firm. Recommendation engines are a key aspect of the online search process. Fleder and Hosanagar (2009) demonstrates this, showing that algorithms that emphasize ‘people who bought this also bought’ move the sales distribution toward superstars. If many people buy Harry Potter, this recommendation engine will recommend Harry Potter to everyone else. In contrast, if the algorithm emphasizes ‘people who bought this disproportionately bought’, relatively unusual items that demonstrate niche tastes will be sold. Empirically, Tucker and Zhang (2011) documents that popularity information has asymmetrically large effects for niche products. Popularity information affects sales in general. Many online platforms sort items by popularity and feature popularity prominently, reducing search costs for this type of information. Showing such popularity information affects purchase behavior not only in retail but also online lending (Zhang and Liu, 2012), and online investing (Agrawal et al., 2015). The effect on welfare of this change in variety is not obvious, and so it has been the subject of a rich discussion in the literature. Lower search costs that lead people to buy the products that more closely match their preferences should increase welfare. Consistent with 9 this, Brynjolfsson et al. (2003) shows that increased variety increases consumer surplus. At the same time, improvements in welfare may be small. The increase in matching of products to preferences is, by definition, marginal. The new products offered are the products on the margin of being produced. The superstar effects may be marginal relative to the consumers who bought products in the middle because they were unwilling to pay search costs. For example, Ershov (2017) shows that a reduction in search costs in the mobile app market reduced average product quality. On balance, however, it also shows that the increase in variety led to a substantial increase in overall welfare despite the incremental nature of the new products. Aguiar and Waldfogel (2016) argues that this marginal argument misses the substantial uncertainty about product quality for many information goods. In the context of music, they show that several songs and musicians that seem marginal ex ante ended up having substantial sales. Thus, by enabling such music to get produced, digital markets led to a large change in the relative sales of products. Uncertainty in the process meant better and more music was created. A great deal of attention has focused on the increase in variety in consumption of media in particular. The internet might also enable people to only read information of that reflects their narrow viewpoint; despite the variety, there is no need to search widely. The latter idea has been emphasized by Cass Sunstein as an ‘echo chamber’ (Sunstein, 2001). Consistent with the idea of wide variety available but consumption in echo chambers, Greenstein and Zhu (2012) examines the bias of Wikipedia and show that, while, on aggregate, Wikipedia has become less politically biased (towards Democrats) over time, the bias of articles has not changed much. Instead, the political bias has mainly dropped because of the arrival of new, relatively right-wing articles. By contrast, Gentzkow and Shapiro (2011) shows that internet media consumption is more varied than offline media consumption. Thus, in this context, low search costs lead to 10 increased variety. Boxell et al. (2017) argues that the internet is unlikely to be responsible for increased polarization of digital content because the increase in polarization is largest for demographic groups with the least internet usage. Polarized media may be less concentrated, generating incentives for niche sources to intentionally mislead. Allcott and Gentzkow (2017) show that false news stories about the 2016 presidential election were shared tens of millions of times, though they demonstrate the fake news was unlikely to have changed the election outcome. Long before the attention to fake news in the 2016 election, Antweiler and Frank (2004) examines how anonymous, and potentially misleading, online investing advice affects stock prices. Low search costs–in the absence of a reliable quality filter–meant that this information could be more easily found and shared. Low online search costs have also transformed the way academic research is consumed. McCabe and Snyder (2015) shows that JSTOR led to an increase in citations of included articles at the expense of others. Search costs fell, but because they fell more for some articles than others, it changes the nature of attention to specific articles and ideas. More starkly, Ellison (2011) argues that peer review may be in decline because of low online search costs. In particular, he shows that high-profile researchers do not need to rely on academic journals to disseminate their ideas. They can post online and people will find their work. In other words, similar to the superstar effect in products, low search costs combined with thousands of research articles benefit the superstar researchers. 2.3 How do low search costs affect matching? Reduced search costs facilitate exchange more generally, often enabled by large digital platforms. Dana and Orlov (2014) shows that airlines are better able to fill capacity. Ellison et al. (2014) shows that online buyers are better able to find the specific books they want. Kroft and Pope (2014) finds online search through Craigslist decreased rental apartment 11 and home vacancies (though they measure no effect on unemployment). Anenberg and Kung (2015) shows that online search enabled the rise of a market for truck-based mobile restaurants (“food trucks”). To the extent that the literature emphasizing matching is distinct from search, the matching literature emphasizes that both sides of the market engage in the search process. Related to the above ideas, low search costs are likely to increase the quality of matches between buyers and sellers, firms and workers, etc. The labor economics literature has emphasized that the internet should reduce unemployment and vacancies. Kuhn and Skuterud (2004) finds no effect of internet job search on employment. Kuhn and Mansour (2014) revisits the analysis several years later with updated data and finds that job searchers that used the internet in job search were indeed more likely to match to an employer. The reduced costs of search have led to the development of online ‘peer-to-peer’ platforms dedicated to facilitate matching. The variety of such online matching markets is extraordinary: Workers and firms, buyers and sellers, investors and entrepreneurs, vacant rooms and travelers, charities and donors, dog walkers and dog owners, etc. Several of these markets have been dubbed the ‘sharing economy’ because people are able to use unused objects or skills better. Most ‘sharing economy’ platforms are not sharing in the sense learned by kindergarteners: Customers typically pay for the ‘shared’ services. Horton and Zeckhauser (2016) emphasizes that many of these markets are driven by an unused capacity for durable goods. Low search costs enable such unused capacity to be filled more efficiently. In a review of the peer-to-peer markets literature, Einav et al. (2016) notes that much of the research takes a market design perspective. For example, Cullen and Farronato (2016) examines an online marketplace that matches buyers and sellers of domestic tasks, such as cleaning, moving, and simple home repair. They emphasize the challenges in growing both the demand and supply sides with respect to variation in the quantity of buyers and sellers over time, economies of scale in matching, and geographic density. A key result is that 12 demand fluctuations in this two-sided market lead to changes in quantity supplied rather than changes in prices. Similarly, Hall et al. (2016), Fradkin and Farronato (2016) and Zervas et al. (2016) also show that the responsiveness in quantity supplied to changes in demand conditions is a key aspect of peer-to-peer platforms (specifically, Uber and Airbnb). Low search costs provide market demand information that enables supply to enter the market when needed. 2.4 Why are digital platform-based businesses so prevalent? Platforms are intermediaries that enable exchange between other players. Digitization has led to an increase in the prevalence of platform businesses, even beyond the peer-to-peer platforms discussed above. Most of the major technology firms can be seen as platformbased businesses. For example, Apple provides hardware and software platforms for others to build applications around. Google provides platforms for bringing together advertisers and potential buyers. As highlighted in Jullien (2012), there are two main reasons digital markets give rise to platforms. First, platforms facilitate matching. In particular, as in the sharing economy platforms, they provide a structure that can take advantage of low search costs to create efficient matches. Often platforms serve as intermediaries between buyers and sellers, as highlighted in Nocke et al. (2007) and Jullien (2012). In the context of a central role of matching, a rich theory literature has arisen that examines competition and pricing strategy in such platform businesses, with an emphasis on the importance of indirect network effects (for example Baye and Morgan (2001a); Caillaud and Jullien (2003); Weyl (2010); Hagiu and Jullien (2011) and, de Corniere (2016)). Second, platforms increase the efficiency of trade. They do this through lower search costs as well as other aspects of digitization that we discuss below: Low reproduction costs and low verification costs. Hagiu (2012) emphasizes how software platforms enable application 13 providers to serve a large number of customers quickly, with the only requirement that the application serve some particular customer need, reproduce at zero cost, and rely on the platform and the other applications to serve other needs. Interoperability is therefore a key aspect of platforms. There is a large literature on the topic, as reviewed in Farrell and Simcoe (2012). A key contribution of this literature is the emphasis on the strategic nature of decisions on interoperability and standards (Rysman and Simcoe, 2008; Simcoe, 2012). A related set of questions examines whether market participants will ‘multi-home’ and use multiple platforms (Rochet and Tirole, 2003; Rysman, 2007; Halaburda and Yehezkel, 2013). 2.5 How do low search costs affect the organization of the firm? Lucking-Reiley and Spulber (2001) discusses several hypotheses with respect to the impact of the internet on firm structure in terms of the role of online intermediaries and vertical integration. This literature emphasizes information flow generally, in which search is one key type of information flow. Garicano (2000) shows that low-cost digital information flow could increase centralization, by enabling headquarters, and organizational leaders, to understand better what is happening at a distance. On the other hand, Garicano (2000) also shows that low-cost communication could decrease centralization, by enabling front-line employees to access information previously only available to senior employees at headquarters. A variety of papers have explored nuances in this tradeoff within organizations, emphasizing the importance of the particular technology studied. Bloom et al. (2014) tests this theory directly, using data on European and American manufacturing firms to show that information technology is a centralizing force and communication technology is a decentralizing force. Acemoglu et al. (2007) also discusses the decentralizing role of information technology. For example, Forman and van Zeebroeck (2012) shows that digital communication increases in research collaboration across establishments within an organization. Baker and Hubbard (2003) examines the impact of on-board com14 puters on asset ownership in the trucking industry. They emphasize tracking costs more than search costs and find that aspects of on-board computing that improve monitoring pushed trucking firms to more ownership of trucks while aspects of on-board computing that improve real-time location information pushed trucking firms to less ownership of trucks. Thus, while adoption of digital technology led to improved efficiency, the impact on organization of the firm in equilibrium depends on the nature of the technology and how its specific features affect tradeoffs between competing tensions at the boundary of the firm. McElheran (2014) examines the decision to centralize or delegate IT adoption decisions within firms. Firms with a greater need for integrated processes (digital or otherwise) delegate less. Forman and McElheran (2013) shows that this tendency is mitigated by the ease with which IT enables coordination across firms, so that disintegration of the firm boundary can be seen as an extreme form of delegation. In addition to the impact on the domestic boundaries of the firm, the reduction in search costs (combined with the reduction in verification costs discussed below) has also led to an increase in international hiring and outsourcing. While international outsourcing is not a new phenomenon (Leamer, 2007), the recent rise of digital international labor market platforms suggests a different avenue for international hiring. Agrawal et al. (2016) shows that online platforms with standardized information disproportionately benefit workers from developing countries. The objective information available online, combined with the ability to send the output of the work (typically information such as data or software code) for free over long distance helps workers who are far from the buyer. Such online labor markets have several important challenges. Using data from online labor markets, Lyons (2017) shows that crosscultural international teams can be less productive because of communication challenges. Relatedly, Ghani et al. (2014) shows that employers in the Indian diaspora are more likely to hire Indians online. 15 3 The Replication Cost of Digital Goods is Zero The key shift in the production function is not that digital goods have a marginal cost of zero. Simple microeconomic models with zero marginal cost are not so different from models with positive marginal cost. The demand curve slopes downward and firms price where marginal revenue equals zero. Instead, a key distinction between goods made of atoms and goods made of bits is that bits are non-rival, meaning that they can be consumed by one person without reducing the amount or quality available to others. A common analogy for non-rival goods is that just as one person can start a fire without diminishing another’s fire, information can be shared without diminishing the original information. In the absence of deliberate legal or technological effort to exclude, bits can be reproduced by anyone–not just the producing firm–at near zero cost without degrading the quality of the initial good. As Shapiro and Varian (1999, p. 83) put it, the internet can be seen as a “giant, out of control copying machine.” Nevertheless, the economics of zero marginal cost, non-rival goods can shift things in favor of producers, consumers, or both. In a static model, as marginal costs fall the potential surplus rises and so the welfare impact depends on the final price and associated deadweight loss. The final price and deadweight loss depend on legal and technological tools for exclusion (Cornes and Sandler, 1986), which relate to the ability to track behavior – the subject of the next section. In this section, we emphasize that the underlying technology enables firms and governments to make a choice not to exclude. This can allow individuals to enjoy the full benefits of the non-rival nature of information-based goods. 3.1 How can non-rival digital goods be priced profitably? The non-rival nature of digital goods has led to questions of how to structure pricing of a large variety of non-rival zero-cost goods, should a producer choose to charge. Bundling 16 occurs when two or more products are sold together at a single price (Shapiro and Varian, 1998; Choi, 2012). Bundling models have a long history in economics. Stigler (1963) and Adams and Yellen (1976) note that the price discrimination benefit of bundling arises when consumers have negatively correlated preferences. Some people may value an action movie at $10 and a romance at $2. Others may value the romance at $10 and the action movie at $2. Selling the bundle at $12 yields higher profits than selling the action and romance movies separately. The challenge for firms is to identify such negative correlations in preferences to identify when bundling will increase profits. Bakos (1999) and Bakos and Brynjolfsson (2000) recognize that, under certain assumptions, this challenge is overcome when many products can be bundled, due to the law of large numbers. Furthermore, the non-rival nature of information goods means that large numbers of information goods can be bundled without substantially increasing costs. Therefore, a simple and useful insight on the economics of non-rival information goods is that it will sometimes be optimal to bundle thousands of digital products together. Chu et al. (2011) uses an empirical example to show that the intuition of Bakos (1999) applies to relatively small numbers of goods in the bundle. There are also strategic reasons to bundle because it can reduce competition (Carbajo et al., 1990). When bundling has zero marginal cost, such strategic considerations can become particularly relevant (Carlton et al., 2010; Choi, 2012). Despite the extensive theory work, it is only recently that empirical examples of such massive bundles appeared in the literature, in the form of subscription services for video such as Netflix and music such as Spotify and Apple Music. Aguiar and Waldfogel (2015) shows the Spotify displaces sales but it also displaces ‘piracy’ or the downloading of music without permission from the copyright holder. They estimate that the reduction in sales and the increase in legal music consumption balance each other so that Spotify appears to be revenue neutral in the 2013-15 time period. 17 3.2 What are the motivations for providing digital public goods? Information providers can deliberately decide not to exclude. It is somewhat of a puzzle why private actors would choose to create public goods. Two prominent examples of non-rival public digital goods are open source software and Wikipedia. Both cases involve a deliberate decision not to exclude, and applying established models is somewhat less straightforward than the bundling models highlighted in the preceding subsection. Lerner and Tirole (2002) asks why software developers would freely share their code with no direct payment. They emphasize two core benefits from open source that do not appear in standard models of public goods. For individual developers, providing high quality open source code is a way to signal their skills to potential employers. For companies, improving the quality of open source software may allow them to sell other services, that are complementary to open source software (such as hardware or consulting services), at a premium. Underlying these core benefits is the non-rival nature of the code: Digital distribution through the internet means that (high quality) open source contributions can be widely adopted. The literature on the economics of open source that followed has largely supported their hypotheses of career concerns and complementarity (Johnson, 2002; Bitzer and Schroder, 2005; Mustonen, 2005; Lerner et al., 2006; Henkel, 2009; Xu et al., 2016). Wikipedia represents a different important context for the puzzle of why people contribute to digital public goods. Zhang and Zhu (2011) emphasizes social benefits related to breadth of readership. In the context of Chinese language Wikipedia, they show that users care about audience size, and decrease contributions when part of the audience is blocked due to Chinese government policy. Consistent with this idea of a social benefit, Aaltonen and Seiler (2016) and Kummer et al. (2015) together provide evidence for a virtuous circle in which more editing leads to more views and more views lead to more editing. Contributions are likely related to the interests of the contributors: Wikipedia leaned sharply Democratic 18 early on and has gradually become more neutral (Greenstein and Zhu, 2012). Nagaraj (2016) suggests the potential for government sponsorship of digital public goods. He finds that open mapping information led to a substantial increase in mining activity, particularly for smaller firms with fewer resources. Therefore, open data enabled a wider set of participants to succeed. More generally, the non-rivalrous nature of digital technology could enable consumers and workers in developing countries to access the same information as people in developed countries, conditional on having access to the internet. In the context of education, Kremer et al. (2013) argues that information technology can improve pedagogy in the developing world. Underlying their argument is an emphasis on non-rival non-excludable digital information, and the public internet-based posting of educational materials. Correspondingly, Acemoglu et al. (2014) emphasizes that digital education will lead to a more equal distribution of educational resources. There are, however, situations in which welfare may decrease because of a decision not to exclude digital goods from widespread copying. The decision not to exclude non-rival goods can reduce the incentives to produce information goods, a subject we discuss below in the context of copyright policy. It can also create negative externalities. For example, Acquisti and Tucker (2014) shows that policies that mandate ‘Open Data’ by government may lead to data leakages (or privacy breaches) that affect individuals’ welfare offline. Openness, almost by definition, implies a reduction in privacy. Relatedly, Acquisti and Gross (2009) shows that using public data online makes it possible to predict an individual’s social security number. This feeds back in general to the idea that while non-excludability may be attractive in principle, it can lead to questions of appropriate data security practices (Gordon and Loeb, 2002; Gal-Or and Ghose, 2005), especially if costly investments in data security also are a public good. 19 3.3 How do digital markets affect copyright policy? While digital technology creates public goods, zero marginal cost of production can also create public bads, such as spam (Rao and Reiley, 2012) and online crime (Moore et al., 2009). These have led to policy responses such as the US CANSPAM act. Another example, of digital spam is junk telephone calls, the automation of which has been enabled by digital technologies. Petty (2000) and Varian et al. (2005) evaluate the role of the federally sponsored ‘Do Not Call’ list in preventing potentially intrusive direct sales calls and find positive effects. That said, the economics of such bads are relatively straightforward. In contrast, the more challenging policy question for non-rival digital goods is whether the government should intervene through copyright policy to enforce excludability despite the non-rival nature of the goods. As the internet first diffused in the late 1990s, copyright of music (and text) was often ignored as people freely posted copyrighted goods online. Because of the non-rival nature of digital information, one posted copyrighted item could be useful to millions of people, potentially replacing sales. At the same time, music industry revenue began to fall (Waldfogel, 2012) and this was widely blamed on changes brought by the internet. Optimal enforcement of copyright has therefore been a key focus of the digital economics literature. The early work focused on the revenue consequences of free online copying. This was referred to as ‘file-sharing’ to those who believe it should be allowed, and as ‘piracy’ by those who didn’t. The direct effect of free online copying of media is that revenues from the sale of copies of that media fall. At the same time, revenues could rise if the free copies are merely sampled and consumers buy what they like (Peitz and Waelbroeck, 2006). Revenues could also rise for complementary goods like live performances (Mortimer et al., 2012). Finally, revenues could rise if the free copies are limited to developing markets for products with network effects (Takayama, 1994). Empirically, though a small number of 20 studies have found positive effects (Oberholzer-Gee and Strumpf, 2007), most studies have found that free online copying reduces revenues in music (Rob and Waldfogel, 2006; Zentner, 2006; Liebowitz, 2008; Waldfogel, 2010), in video (Rob and Waldfogel, 2007; Liebowitz and Zentner, 2012; Danaher et al., 2014; Danaher and Smith, 2014; Reis et al., 2015; Peukert et al., 2017), and in books (Reimers, 2016). This echoes a non-digital historical literature (Li et al., 2015; MacGarvie and Moser, 2015) suggesting a continuity between policy governing digital technologies and earlier policies. How does copyright affect the creation of new works? This is a more difficult research question as it requires some attempt to measure counterfactual quality and quantity of goods had copyright law not existed (Varian, 2005; Peitz and Waldfogel, 2012; Danaher et al., 2013). Waldfogel (2012) addresses this challenge using two measures of music quality: Historical ‘best albums’ lists and usage information over time. In both cases, he shows that the quality of music began to decline in the early 1990s and stopped declining after the arrival of free online copying in 1999. Why did quality rise despite declining revenue? He argues that simultaneously with the decline in revenue came a decline in the cost of producing and distributing music. Digitization affected the supply side as well as the demand side, and so quality rose. Results are similar in movies (Waldfogel, 2016) and books (Waldfogel and Reimers, 2015). This contrasts with the economic history literature, which suggested that copyright alone could increase the quality of creative output (Giorcelli and Moser, 2016). In addition to affecting incentives to innovate, digital challenges to copyright protection may affect incentives to build on prior work. Williams (2013) demonstrates this point in a different intellectual property context and shows that intellectual property protections limit follow-on innovation in gene sequencing. Heald (2009) shows that copyrighted music is less used in the movies than non-copyrighted music. Nagaraj (2017) shows that copyright protection of old sports magazines reduces the quality of Wikipedia pages decades later. This phenomenon is not unique to the digital context. Biasi and Moser (2016) shows that 21 eliminating copyright of German books during World War II led to a substantial increase in US scientific output, measured by PhDs in mathematics and patents that cited the German books. Another challenge for copyright policy driven by the shift in costs of replication is that it has made it easier for other firms to replicate digital content and attempt to aggregate it. This practice has been particularly prevalent in the news media, where policy makers have been encouraged to take action to protect the interest of the newspapers that actually originated this news content. However, in general the work in economics which has evaluated the effect of these aggregators has been to emphasize that such aggregation promotes more exploration rather than necessarily cannibalizing content (Calzada and Gil, 2016; Chiou and Tucker, 2017; Athey et al., 2017). Overall, copyright law is more important in digital markets because goods can be copied at zero cost. Stricter enforcement of copyright appears to increase revenue to the copyright holder, increase some incentives by potential copyright holders to innovate, but reduce incentives by others to build on copyrighted work. Nevertheless, the literature also shows that, despite ease of copying, digitization has not killed creative industries because production and distribution costs have fallen and because the technology has caught up to facilitate copyright enforcement. 4 Lower Transportation Costs Related to replication being costless, the cost of transporting information stored in bits over the internet is near zero.4 Put differently, the cost of distribution for digital goods approaches zero and the difference in the cost of nearby and distant communication approaches zero. 4While transportation costs could be positive and even high due to network congestion, in practice this has not been an issue. Early on, such network congestion was a key focus of the literature. For example, one of the first volumes on internet economics, Mcknight and Bailey (1998), has several articles on congestion pricing. This early literature on backbone competition and congestion ended up influencing our understanding the economics of net neutrality discussed above (Cremer et al., 2000; Laffont et al., 2001; Besen et al., 2001; Laffont et al., 2003). 22 In addition, digital purchasing technologies have reduced transportation costs. Consumers buy physical goods online, particularly when offline purchasing is costly or difficult (Goolsbee, 2000; Forman et al., 2009; Brynjolfsson et al., 2009). Furthermore, Pozzi (2013) shows that consumers also use online shopping to overcome the transportation costs of carrying things from the store. In this way, the internet facilitates stockpiling, allowing people to buy in bulk when a discount appears because delivery means there is no need to carry the large quantity of items purchased. Thus, for information, for digital goods, and for physical goods, transportation costs are lower online. 4.1 Does distance still matter if transportation costs are near zero? Low transportation costs for information mean that the cost of distribution for digital goods approaches zero and that the difference in cost of nearby and distant communication approaches zero. The potential implications of low transportation costs have been explored in the popular press. Cairncross (1997) suggests that this fall in the costs of transporting information would lead to a “death of distance”. Isolated individuals and companies would be able to plug into the global economy. Rural consumers would benefit by having access to the same set of digital products and services as everyone else. There would be a global diffusion of knowledge. Friedman (2005) identifies several of the same themes in predicting a “flat world” in which businesses anywhere could plug into the global supply chain and produce. Being in the United States would not confer a meaningful advantage relative to India. Both Cairncross and Friedman suggested the potential arrival of a global culture, in which everyone everywhere would consume the same information, an idea with its roots in McLuhan (1964). This idea is implicit in the trade model of Krugman (1979): Countries consume the same goods as transport costs approach zero. Rosenblat and Mobius (2004) formalizes some of 23 these ideas in a different context, using network model of collaboration in which long distance collaboration rises but coauthor similarity in other dimensions (such as field of research) also rises. A less extreme question than “Is distance dead?” is “Does distance matter more or less than it used to?” The most definitive answer to that question comes from Lendle et al. (2016). They compare cross-border sales on Ebay with international trade data. They demonstrate that, while distance predicts both online and offline trade flows, distance matters substantially less on Ebay. The digital economic literature has emphasized what factors influence the extent to which distance still matters. As Lemley (2003) notes, “No one is ‘in’ cyberspace.” (p. 523). Therefore, offline options matter. Balasubramanian (1998) examines the importance of offline options using a circular city/Salop (1979) model with the cost of using the direct retailer as constant for all locations, but the cost of using the stores located around the circle depends on transportation costs. The model shows that the benefit of a direct (online) retailer will be largest for those who live far from an offline retailer. Forman et al. (2009) provides evidence to support this model, demonstrating that when a Walmart or Barnes & Noble opens offline, people substitute away from purchasing books on Amazon. A number of other studies also demonstrate how offline retail affects online purchasing. Related models include Loginova (2001) and Dinlersoz and Pereira (2007) which examine the role of loyalty to the offline store in driving the more price sensitive customers online. Empirically, Brynjolfsson et al. (2009) shows that online sales at a women’s clothing retailer are lower from places with many offline women’s clothing stores. This impact is driven by the more popular products that are likely to be available in a typical offline store. Choi and Bell (2011) shows that online sales of niche diaper brands are higher in places where they are unlikely to be available offline. Goolsbee (2001), Prince (2007), and Duch-Brown et al. (2017) all show substitution between online and 24 offline sales of personal computers. Gentzkow (2007) demonstrates substitution between the online and offline news in Washington DC. Seamans and Zhu (2013), Goldfarb and Tucker (2011a), and Goldfarb and Tucker (2011d) demonstrate substitution between online and offline advertising. Gertner and Stillman (2001) shows how channel conflict interacts with vertical integration and show that vertically integrated apparel retailers went online first. In their review of the literature on online-offline competition, Lieber and Syverson (2012) provides some additional evidence that offline options affect online purchasing. Similarly, in the digital media context, evidence suggests that online media consumption substitutes for, and is replacing, offline media consumption (Wallsten, 2013; Gentzkow, 2007). In addition to the offline option, the fact that tastes are spatially correlated also matters for the persistent role of distance. Blum and Goldfarb (2006) examines the international internet surfing behavior of about 2600 American internet users, and demonstrate that internet surfing behavior is consistent with the well-established empirical finding in the trade literature that bilateral trade decreases with distance (Overman et al., 2003; Anderson and van Wincoop, 2004; Disdier and Head, 2008). In other words, even for a product with zero shipping costs (visiting websites), people are more likely to visit websites from nearby countries than from faraway countries. This relationship between distance and website visits is much higher in taste dependent categories (and loses statistical significance in the non-taste dependent categories). Distance matters because it proxies for taste similarity. Alaveras and Martens (2015) replicates this core result using much richer data on website visits by users in a large number of countries. Sinai and Waldfogel (2004) also shows that highly populated areas produce more content, and that because tastes are spatially correlated in the sense that people are more likely to consume local media than distant media, people in highly populated areas are particularly likely to go online. This geographically specific nature of tastes is also reflected in the consumption of digital goods such as music (Ferreira and Waldfogel, 2013) and content (Gandal, 2006). Quan and Williams (2017) demonstrates 25 that accounting for spatial correlation in tastes reduces the estimated consumer surplus from increased online variety by 30 percent. In addition to offline choices and spatially correlated tastes, another factor which explains the continuing role of distance is the presence of social networks. Much online behavior is social, and social networks are highly local (Hampton and Wellman, 2003). Thus, while zero transportation costs of information mean that you can communicate with anyone anywhere in the world for the same price, the vast majority of most people’s email comes from those who either live at the same home or work in the same building. Gaspar and Glaeser (1998) speculates that because of the spatial correlation of social networks, the internet may be a complement to cities. More efficient communication would be especially important for those who communicate frequently. In other words, though the relative costs of communication fall more for distant communication, the overall importance of local communication might mean that cities benefit most. Agrawal and Goldfarb (2008) provides some evidence in support of this hypothesis by showing that as new universities connected to a 1980s internet-like network, they increased their collaboration rate with those already connected. The biggest change in collaboration rates were for co-located universities in different quality tiers. The paper emphasizes the likely local social networks of researchers in the same city. Looking at online ‘crowdfunding’ of music, Agrawal et al. (2015) provides further evidence of the importance of local social networks by showing that musicians’ early funding tends to come from local supporters who the musicians knew prior to joining the crowdfunding platform. As the musician gains prominence on the website, the later funding often comes from distant strangers. Finally, in the absence of the improvements in verification discussed below, trust is easier locally. Hortacsu et al. (2009) shows that same-city sales on Ebay and MercadoLibre (a Brazilian electronic commerce platform) are disproportionately high, likely because some products are observed and delivered in person. Furthermore, Forman et al. (2009) shows 26 that Americans follow the online product recommendations of others who live near them. 4.2 Can policy constrained by geographic boundaries shape digital behavior? Early work worried that the internet could undermine local regulation and national sovereignty (Castells, 2001). Some research is consistent with this idea: Online sales have been higher where the difference between online and offline tax rates is highest (Goolsbee, 2000; Ellison and Ellison, 2009b; Anderson et al., 2010; Einav et al., 2014). When local regulation prohibits offline advertising, similar online advertising is more expensive (Goldfarb and Tucker, 2011d) and more effective (Goldfarb and Tucker, 2011a). This substitution suggests that online and offline markets should be considered together in the context of antitrust (Goldfarb and Tucker, 2011e; Brand et al., 2014). At the same time, regulation can mean that users experience the internet differently in different locations. At the extreme, regulation can prohibit certain content, making the experience of using the internet different across locations. Zhang and Zhu (2011) examines the impact of the blocking of Wikipedia in China in October 2005 on the motivations of others outside China to contribute. Thus, a key online website was available in some places and not others. More generally, some countries regularly block access to certain websites changing the nature of the internet across locations. Regulation can also change what users find available across locations. Copyright policy leads to variation in the availability and consumption of media across locations (GomezHerrera et al., 2014; Chiou and Tucker, 2017; Athey et al., 2017; Calzada and Gil, 2016). Privacy policy leads to different advertising and different website success (Goldfarb and Tucker, 2011f; Tucker, 2015). Trademark policy leads to different search experiences (Chiou and Tucker, 2012; Bechtold and Tucker, 2014). Thus, when regulation does not reach into the online sphere, the zero transportation costs of information in the online channel generate a disproportionate benefit of online information 27 in regulated contexts. However, when regulation does reach the online sphere, it can have a substantial impact on the nature of the internet across locations. 5 Lower Tracking Costs The first three drops in costs, those associated with search, replication, and distance, were well discussed in the early digital economics literature. However, the importance of the lowering of the next two costs we discuss, tracking and verification, has only become clear in the last decade. Digital activity is easily recorded and stored. In fact, the technology typically stores all information automatically, and firms and consumers have to make a deliberate decision to discard data. Reductions in tracking costs enable personalization and the creation of one-to-one markets, leading to renewed interest in established economic models with asymmetric information and differentiated products such as price discrimination, auctions, and advertising models. 5.1 Do lower tracking costs enable novel forms of price discrimination? The ability to use digital technologies to track individuals enables personalized markets. Several economists recognized this potential for digital price discrimination as the internet commercialized in the late 1990s (Shapiro and Varian, 1998; Smith et al., 2001; Bakos, 2001). Even first-degree price discrimination seemed like it might become more than a theoretical curiosity. One form of price discrimination that has received a great deal of attention in the theory literature on digital markets is behavioral price discrimination (see Fudenberg and VillasBoas (2012) and Fudenberg and Villas-Boas (2007) for reviews). This literature emphasizes that the low cost of collecting digital information makes it easier for companies to pricediscriminate based on an individual’s past behavior. The research builds on a large price discrimination literature that does not specifically emphasize digital markets (Hart and Ti28 role, 1988; Chen, 1997; Fudenberg and Tirole, 2000). Broadly, the research explores the benefits and costs of identifying previous customers for monopolies (Villas-Boas, 2004) and competing firms (Shin and Sudhir, 2010; Chen and Zhang, 2011). Fudenberg and Villas-Boas (2012) summarize this literature to conclude that under monopoly, firms benefit from the additional information but under competition the information may increase the intensity of competition. Furthermore, the benefits of the information to a monopoly may lead consumers to strategically withhold information. In other words, consumers become privacy-sensitive (Taylor, 2004; Acquisti and Varian, 2005; Hermalin and Katz, 2006). In the opposite direction, rules that restrict the flow of information hurt firms’ ability to price-discriminate and therefore may leave some consumers unwilling to buy at the offered prices (Taylor and Wagman, 2014; Kim and Wagman, 2015). Another form of price discrimination that has received attention in the digital economics literature is versioning. Bhargava and Choudhary (2008) provide a model of versioning when variable costs are zero. Fay and Xie (2008) explore versioning based on probabilistic selling. For example, airlines and hotels offer low price versions of their products on Priceline.com, in which there is buyer uncertainty about the specific product being bought. Empirical support for digital price discrimination is limited, despite the rich theoretical discussion of the potential for personalized pricing. For example, versioning is a basic form of third-degree price discrimination that precedes most digital markets (Maskin and Riley, 1984; Deneckere and McAfee, 1996; Corts, 1998; Fudenberg and Tirole, 1998). Rao (2015) provides experimental support for the value of versioning digital products, demonstrating that online limited time ‘rentals’ can increase profits by segmenting high and low value consumers. Despite the ease of even this most straightforward form of price discrimination, Shiller and Waldfogel (2011) argue digital firms may not be versioning, or more generally price-discriminating, as much as would be optimal. In particular, they puzzle over the surprisingly uniform nature of pricing for digital music. They argue that uniform pricing 29 of music appears to lead to lower-than-optimal profits for firms, but do not provide a clear answer to this puzzle. While there is evidence of broad versioning of online media (Chiou and Tucker, 2013; Lambrecht and Misra, 2017), the theoretical literature on digital price discrimination seems to be ahead of the empirical work and of firm practices. While there is evidence of first degree price discrimination in higher education (Waldfogel, 2015), the only online research example we found is Dube and Misra (2017), who demonstrate the feasibility and profitability of targeting many prices to different customers of an online service based on a large number of characteristics. 5.2 Why has there been a shift from personalized pricing to personalized advertising? Given the emphasis of the theoretical literature on the ease and practicality of behavioral price discrimination and the potential for personalized pricing of goods online, it is perhaps a surprise that for many of these goods consumers face a price of zero (Evans, 2009). Thus, perhaps the most striking effect of the creation of low online tracking costs has not been to use personalized profiles to charge different consumers different prices, but instead to show these different consumers more appropriate, relevant, and profitable advertising. Variants of these ideas appear in a rich theory literature on two-sided markets, emphasizing the digital context (Baye and Morgan, 2001b; Anderson and de Palma, 2009, 2013; White, 2013; Athey et al., 2014). Baye and Morgan (2001b) demonstrates that an information intermediary will price low to consumers, while charging advertisers a high enough price that some choose not to participate. Anderson and de Palma (2009) and Athey et al. (2014) each model consumer attention as scarce and explore advertiser competition for that attention. Athey et al. (2014) emphasizes that if an advertiser wants to send a message to a customer offline, they need to rely on noisy signals based on media demographics. In contrast, online targeting technology is such that an advertisers can target a particular consumer. In the 30 presence of multiple media outlets and multi-homing by consumers, the equilibrium outcome is that online advertising prices can be much lower than offline advertising prices even though the online advertising is in fact more useful to the advertiser. However, Gentzkow (2014) argues that the price of attention is not lower online than offline, which challenges this prediction. Perhaps because of these forces, many of the largest online companies–in terms of revenues, profits, and users–are advertising-supported. Low-cost tracking means that what distinguishes online advertising from offline advertising is that it is targeted (Goldfarb and Tucker, 2011b; Goldfarb, 2014). This difference is highlighted in models that explore competition between online and offline advertising (Athey and Gans, 2010; Bergemann and Bonatti, 2011; Johnson, 2013). Athey et al. (2014) and Levin and Milgrom (2010) use very different models to demonstrate that better targeting may not help online media. Athey et al. (2014) show that improved tracking can increase competition between media outlets. Levin and Milgrom (2010) show that too much targeting can lead to insufficient competition among advertisers for the user attention sold by a monopolist media firm. This better targeting has led to a thriving literature that measures advertising effectiveness. Because ad messages are sent to individuals in bits (rather than broadcast through billboards and newspapers), it is relatively easy to identify consumers that see ads, to randomize which consumers see ads, and even to track those consumers through purchase. Until recently, this was very difficult and so there were few well-identified studies of advertising effectiveness. Low tracking costs make it relatively easy to run field experiments online, and large scale field experiments have been the focus on the recent literature. Research on online advertising effectiveness has been largely conducted by research economists working with industry. For example, Lewis and Reiley (2014) uses a field experiment on 1.6 million Yahoo customers that connects online advertising to offline department store sales. They find that online advertising increases offline sales in a department store. 31 Blake et al. (2015) shows that in many cases search engine advertising–the key revenue generator for Google–does not work. In particular, they demonstrate with a large field experiment at Ebay that consumers will often click on the ‘organic’ link anyway and navigate to the advertiser’s page. They argue that much search engine advertising is wasted. Simonov et al. (2017) uses data from Microsoft’s Bing search engine to show that the results for Ebay may be driven by the strength of Ebay as a particularly well-known brand. Less well-known advertisers seem to benefit from search advertising. While much better than prior ways to measure advertising effectiveness, there are still substantial challenges. Correlational research, even with detailed data, typically yields inaccurate measures of advertising effects because the signal-to-noise ratio for advertising’s effect on sales is low (Lewis et al., 2015; Gordon et al., 2016). Furthermore, even with experiments, advertising effects are subtle relative to the variance in purchase behavior and so studies need to be highly powered (Lewis and Rao, 2015). A large literature also emphasizes the role of targeting as a distinct and important feature of online advertising. Goldfarb and Tucker (2011c) shows that targeted banner advertising is effective, but only as long as it does not take over the screen too much. Targeting works when subtle, in the sense that it has the biggest impact on plain banner ads, relative to how it increases the effectiveness of other types of ads. Lambrecht and Tucker (2013) and Tucker (2012b) demonstrate the effectiveness of other types of online advertising targeting. As noted above, online media support their business by selling scarce consumer attention to advertisers. New technologies are emerging that allow consumers to block advertising online. Such ad blocking may reduce revenues and, perhaps counterintuitively, increase the quantity of ads shown to those without ad blockers (Anderson and Gans, 2011). In a test of these ideas, Shiller and Waldfogel (2017) uses data on ad blocking and website visits to show that widespread use of ad blockers may decrease the quality of websites on the advertising-supported internet. 32 5.3 Why are online goods and services often sold by auction? The rise of online advertising, along with individual-level tracking technologies, has created a difficult pricing problem: How can a firm choose prices for thousands of advertisements that might be priced differentially to millions or even billions of customers? As economists have long-recognized, auctions are a particularly useful tool for price discovery. Consequently, digital markets typically use auctions to determine prices for advertising. Auctions are also used to price some other goods. Originally, advertising on Yahoo!’s search page in the 1990s was priced according to a standard rate. Goto.com’s insight–that an auction could leverage the fact that the value of advertising depended on the search term–led to a new way to price discriminate in advertising. Rather than price for the search page, price could be at the level of the search term. Google and Bing’s ad auctions run on this insight. A large literature has arisen to develop auction formats for this context (Varian, 2007; Edelman et al., 2007; Levin and Milgrom, 2010; Arnosti et al., 2016). Today, advertising auctions, particularly for display advertising, often take into account addition information provided by online tracking technologies, such as websites visited in the past and products observed. Less related to tracking costs, online auctions have also been used for price discovery for goods, most notably on Ebay. An early review of the auction literature is provided in Ockenfels et al. (2006). They emphasize that the transactions costs of conducting and participating in auctions are lower in the digital context. Furthermore, many digital goods are not standardized in the sense that buyer valuations vary over time and location, and so the price discovery function of the auction is particularly useful. This idea also appears in Varian (2010) which describes the benefits of computer-mediated transactions with respect to decentralized price discovery, and therefore more finely based price discrimination. While auctions for goods (rather than advertising) still exist online, Einav et al. (2017) shows 33 that goods auctions are in decline as online markets have matured. The prominent role of auctions in economic theory means that a separate literature has used the digital setting as a context to test long-established theory. This research, pioneered by Lucking-Reiley (1999), is not about digital markets per se, but uses the digital context to inform a broader theory literature (Roth and Ockenfels, 2002; Bajari and Hortacsu, 2003; Einav et al., 2016). 5.4 How do digital markets affect privacy policy? Low tracking costs have led to a renewed interest in the economics of privacy, as highlighted by a recent review in this journal (Acquisti et al., 2016). In general the economics literature on privacy, both offline and online, grapples with the question of how privacy should be treated in terms of the consumers’ utility function. Should economists treat privacy as an intermediate good, that is a good whose value simply lies in the way it can moderate the achievement of another good, or as a final good, that is, a good that should be enjoyed and valued for its own sake (Farrell, 2012)? Much policymaking is grounded on the idea that privacy is a final good where a distaste for others intruding on or gathering knowledge about an individual’s personal domain is valid as a driver of an individual’s utility. However, much the theoretical literature analyzes privacy as an intermediate good, because of the implications for personalized pricing that are discussed above (Taylor, 2004; Acquisti and Varian, 2005; Hermalin and Katz, 2006). Privacy regulation can affect the nature and distribution of economic outcomes (Goldfarb and Tucker, 2012a). (Edelman, 2009) and (Lenard and Rubin, 2009) emphasize that there is a trade-off between the use of online customer data to subsidize zero-price goods and advertising performance. Goldfarb and Tucker (2011f) show that European privacy regulation that restricted online tracking led to a substantial decline in the effectiveness of online advertising in Europe. Johnson (2014) estimates the financial effect of privacy policies on the online display ad industry, suggesting an opt-in policy or a tracking ban would reduce 34 welfare substantially, though an opt-out policy would have little effect. Johnson’s paper is very useful for understanding the effect on publishers (rather than advertisers) of privacy regulation. Kim and Wagman (2015) shows that regulation of sharing financial information increased defaults on loans during the financial crisis. Miller and Tucker (2009, 2011) show that US healthcare privacy regulation reduced hospital adoption of electronic medical records, leading to worse health outcomes. On a more positive note in favor of privacy, Tucker (2014) shows that firm-implemented privacy controls designed to encourage con-sumers’ perceptions of control can actually enhance the performance of online advertising. Tucker (2012a) compares this result with work that suggests there may be benefits from addressing consumer privacy concerns, building on research that illustrates how perceptions of control influence privacy concerns in general (Brandimarte et al., 2012). In general, the precise nature of privacy protection can be expected to matter a lot for the direction of innovation: It is not a matter of a simple binary choice to have privacy protection or not. This is emphasized in Miller and Tucker (2014), which shows that different types of privacy protections had very different effects on the adoption of personalized medicine technologies: Regulations that gave consumers control over disclosures enhanced adoption, but regulations that imposed consent requirements decreased adoption. Privacy regulation puts a cost on tracking information flows. The welfare effects of these costs may be ambiguous. First, there may be knock-on effects to industry structure from privacy regulation. Campbell et al. (2015) shows that because privacy regulations typically require firms to persuade their consumers to give consent, which in turn imposes a cost on the consumer, small firms and new firms are disproportionately affected, because it is harder for them to obtain consent under the regulation. Second, welfare complications of privacy policies are also hard to assess due to a privacy 35 paradox, where consumers state an affinity for privacy, but then act in ways which is not consistent with this stated preference. Athey et al. (2017) provides some evidence about the extent to which small incentives, distracting information, and small navigation costs can lead to a gap between stated privacy preferences and actual behavior. Furthermore, assessing the value of privacy is complicated for many reasons, including that privacy preferences for the same individual change over time Goldfarb and Tucker (2012b). Third, much of the work in the economics of privacy has understandably focused on questions relating to industrial organization, there are also implications of digital technologies and privacy for the economics of national security. In addition to improving the ability of firms to track consumers, digital technology allows government crime-fighting agencies to track a broad swathe of the population. Marthews and Tucker (2014) shows that increasing consumer awareness of government data use leads to increased privacy-protecting behavior among consumers in their interactions with firms. 6 Reduction in Verification Costs The reduction in tracking costs has also led to a reduction in costs associated with the verification of identity and reputation. This was not anticipated by the early literature in economics because earliest reporting on the internet suggested that it would be a vehicle for anonymity - ”On the Internet, nobody knows you’re a dog.”5 Furthermore, in addition to tracking cost falling, digital technologies have also made it easier to verify identity and also create a digital reputation. In the absence of such technologies, a long-standing solution for firms to provide credible information about quality was to develop a reputation in the form of a brand (Tadelis, 1999; Smith and Brynjolfsson, 2001; Waldfogel and Chen, 2006). However, digital markets involve thousands of small players. Furthermore, these small players can be unfamiliar to potential 5The New Yorker on July 5, 1993 36 customers. Einav et al. (2017) estimates the 88% of online Visa transactions are with a merchant that the customer does not visit offline. Alternative mechanisms to brand-based reputations are needed. The literature on verification costs builds on economic models of reputation, exploring when the experiences of previous buyers and sellers can enable market exchange in the presence of asymmetric information about quality and trustworthiness. This emphasis on reputation models distinguishes the literature on verification costs from the literature on tracking costs, with emphasizes price discrimination, advertisement targeting, and other forms of personalization. 6.1 How do online reputation systems facilitate trust? The most common such mechanism is an online rating systems in which ratings from past buyers and sellers are posted for future market participants to see. The marketplace that has received the most attention in the literature is Ebay. As mentioned above, one reason Ebay has received so much attention by economists is that it provided a useful setting to test auction theory. Another reason relates to reputation mechanisms. Ebay recognized the challenges of getting people to buy from strangers who they will not meet in person (Resnick and Zeckhauser, 2002; Livingston, 2005). To address this issue, they built, and continually adapted and improved, a ratings system. The effectiveness and development of this ratings system has been the subject of hundreds of papers in economics and management. For example, Ba and Pavlou (2002) shows how a ratings system can enable trust in the absence of repeated interactions. A number of papers empirically demonstrate that better-rated sellers have higher prices and higher revenues (Melnik and Alm, 2002; Livingston, 2005; Houser and Wooders, 2005; Lucking-Reiley et al., 2007). Cabral and Hortacsu (2010) demonstrates differences between positive and negative feedback, emphasizing how the ratings system acts as a disciplining force in the marketplace in which sellers with low ratings exit from Ebay’s platform. 37 Therefore, the original emphasis of the reputation literature was as a platform for overcoming trust in long-distance transactions. Dellarocas (2003) recognizes early on that the application of these feedback mechanisms was not limited to online exchange. Instead, Dellarocas argued that such mechanisms would enable a variety of market activities, both online and offline. As long as incentives to deviate are not too high, such systems can provide credible quality signals in a variety of settings (Dellarocas, 2003; Cabral, 2012). One key application is to provide information on product quality. Rather than enhance information about a particular seller, ratings can inform consumers about the best products available within a platform. It might be in the platform’s interest to provide such information so that consumers are directed to the highest quality products. Comparing changes in reviews on Amazon relative to Barnes & Noble, Chevalier and Mayzlin (2006) demonstrates that positive reviews lead to higher sales. More recently, the literature has focused on how online tools reduce verification costs in offline settings. Luca (2011) shows how online restaurant reviews on Yelp impact restaurant demand, particularly for independent restaurants. Overall, his results suggest that Yelp led to a decrease in the share of chain restaurants relative to independents. Hollenbeck (2016) finds a similar result for hotels. It is easier to establish an online reputation using online reputation mechanisms, but the mechanisms for damaging that reputation in the form of consumer complaints have also become easier. Historically, complaints were registered with letters, and then calls into call centers. Social media enables rapid widespread communication of complaints to both the firm and a wider audience. Gans et al. (2016) uses data from Twitter to explore ideas on the relationship between market power and consumer voice first sketched out in Hirschman (1970). They show that consumers are more likely to voice their complaints via Twitter in locations where airlines have a higher share of flights. In turn, airlines are more likely to respond to consumers in these markets. Tucker and Yu (2017) shows some positive effects 38 of digital technologies, in that the use of mobile apps to receive complaints can actually advantage less educated consumers who are more likely to suffer from employee-consumer discrimination in the treatment of their complaints. A benefit of improved verification procedures online for individuals has been the ability to more securely and easily make payments. This is demonstrated by Economides and Jeziorski (2017), which shows the power of using mobile devices to digitally verify identity in Tanzania. They show that this power enables the use of mobile payments networks to transfer money to others, but also, equally importantly, to transport money over short distances. People appear to deposit cash after work, walk home, and then pick up the cash at home. The verification system enables easy deposits and withdrawals, thereby reducing the risk of robbery. Digital verification, in the form of DNA databases, has also been shown to reduce crime (Doleac, 2017). As technology improves, verification may continue to become easier. Researchers have speculated that the blockchain is a promising technology for reducing verification costs further (Catalini and Gans, 2016). Currently most of the literature on blockchain technologies focuses on specific applications of the technology such as cryptocurrencies (B¨ohme et al., 2015; Catalini and Tucker, 2016). However, if blockchain technologies achieve the promise highlighted in Catalini and Gans (2016), then we might see a diverse literature emerge over the next few years on the consequences of low-cost verification across a variety of empirical setting. 6.2 Is there a role for policy in reducing reputation system failures? Given the important role of such systems in generating demand, it is perhaps unsurprising that the economics literature has focused on questioning when reputation systems fail. Often the failures relate to incomplete ability to verify the person doing the rating online. One type of failure relates to a selection bias: Not all consumers provide ratings. Nosko and Tadelis 39 (2017) shows evidence of such a selection bias, in which buyers with a bad experience do not bother to rate the seller. They instead stop buying from any sellers on the platform into the future. Poor service by a seller therefore creates an externality. The failure of the reputation systems hurts the platform rather than the individual seller. Another type of failure relates to direct manipulation of the ratings by the firms or their competitors. Mayzlin et al. (2014) and Luca and Zervas (2016) show evidence of manipulation, in which firms seem to give themselves high ratings while giving low ratings to their competitors. This evidence of manipulation suggests that ratings systems alone are insufficient. The challenges of ratings systems were recognized relatively early in the digital economics literature. Consider the market for collectible baseball cards. When buyer and seller are in the same place, the buyer can inspect the quality of the card in the store. They can look for rips, folds, or frayed edges. Online, quality is hard to assess. Jin and Kato (2006) provides evidence of fraud in these markets. They show that the online reputation system is insufficient in many ways. In a companion paper (Jin and Kato, 2007), they show how a professional grading industry grew to help solve the information asymmetry between buyers and sellers online. Stanton and Thomas (2016) shows the value of online intermediaries in providing information beyond platform ratings by examining worker and firm behavior on an online labor market. They show that new workers benefit from affiliating themselves with an agency. The platforms also work to improve their reputation systems. Fradkin et al. (2017) documents two experiments made at Airbnb to improve their reputation system: Offering monetary incentives to submit reviews and implementing a simultaneous review process to reduce strategic reciprocity. Hui et al. (2016) shows, in the context of Ebay, that platforms benefit by having both reputation systems and regulations to expel bad actors. In each of these cases, it has been the private sector that has reduced these reputation system failures. To the extent that there has been a role for policy, it has been in the 40 enforcement of contracts and prevention of fraud. At this point, the literature does not point to a specific digital policy with respect to reputation systems failures. One aspect of policy related to verification is the nature of intellectual property tools such as trademarks. Trademarks allow customers to verify whether a brand is indeed the brand it claims to be. Chiou and Tucker (2012) and Bechtold and Tucker (2014) document that, online, consumers use trademarks to search pro-actively. The trademark therefore serves two purposes: It verifies identity and it provides a path to search for related products. Trademark policy needs to be narrow enough to facilitate search related to trademarks, but broad enough to ensure that such search does not sow confusion on brand identity. 6.3 How do digital markets affect anti-discrimination policy? A second policy issue driven by changes in verification relates to discrimination. If people were indeed truly anonymous on the internet then there could be no direct discrimination. However, the drop in verification costs and the ability to identify an individual and also their characteristics makes discrimination possible (and potentially low cost) in a digital environment. The question then for policy makers is whether there is something unique to the online setting which requires additional regulation beyond existing anti-discrimination law. One area this is hotly debated is in the use of algorithms to parse data and automate the allocation of resources and decision making. This is investigated in Lambrecht and Tucker (2016), which shows that algorithms may lead to apparently discriminatory outcomes for innocent reasons. In particular, they show that ads for STEM education are disproportionately shown to men by online algorithms because advertising to men is less expensive overall than advertising to women, and so advertisers who are indifferent to gender end up showing their ads to men more often. Broadly, on the one hand, while tracking is easier, such tracking may focus on dimensions 41 that are legally and morally less controversial, such as preferences rather than race. If digital transactions mean that gender and race information is not revealed, then discrimination may fall. Morton et al. (2003) shows that internet car purchasing reduces gender- and race-based price discrimination. Cullen and Pakzad-Hurson (2017) shows that a reduction of privacy of wages in online platforms decreases pay differences across workers (though it also reduces average pay). On the other hand, if gender or race or other sensitive information are revealed, it is possible that, in the absence of other information, discrimination is high. For example, Ayres et al. (2015) and Doleac and Stein (2013) show that sellers receive lower prices when a black hand is shown with the item than when a white hand is shown. Acquisti and Fong (2013) presents the results of a field experiment to study how employers use information on social networks to filter the suitability of employees. They find considerable use of social networking sites for potentially discriminatory purposes. Similar results have been found in a variety of other online contexts (Pope and Sydnor, 2011; Edelman and Luca, 2014). Both online and offline, discrimination is prevalent. Open questions remain as to whether discrimination is more prevalent online or offline, and as to whether policies aimed at reducing online discrimination specifically will reduce discrimination overall, or simply push discrimination into another setting. 7 Consequences of Digitization for Economic Actors As people spend more time consuming digital media and buying products online, and as business and government increasingly use digital technology, it suggests a broader question: How does storing information in bits rather than atoms affect welfare? As search, reproduction, transportation, tracking, and verification costs fall, has that had an effect on the economy? Broadly, the literature has tackled this question in four different ways: Country-level 42 effects, region-level effects, firm-level effects, and consumer-level effects. 7.1 Country-level effects The macroeconomic productivity literature with respect to internet technology has its roots in the Solow (1987) claims that “you can see the computer age everywhere but in the productivity statistics.” This productivity puzzle persisted for many years. A large growth accounting literature has arisen to examine this ‘productivity puzzle’ and measure the overall impact of digital technologies on the economy. While we view this literature as beyond the scope of this article, Jorgenson et al. (2008) and Van Reenen et al. (2010) both summarize it to suggest that there was a post-1995 productivity surge that was largely driven by digital technology investment and usage. Still, measuring the productivity shifts is difficult. Haltiwanger and Jarmin (2000) lays out several of the anticipated challenges in measuring the effect of the digital economy: Service industry output, data on digital technology spending, price deflators, etc. A key challenge relates to intangible capital (Corrado and Hulten, 2010) which has been found to affect productivity measurement in both the United States and the United Kingdom (Corrado et al., 2009; Marrano et al., 2009). Soloveichik (2010) takes on this measurement challenge and identifies about 65 billion dollars in intangible capital related to books, movies, music, and television. A different stream of work on country-level effects examines how digital communication may affect trade flows for digital and physical goods. Freund and Weinhold (2004) provides suggestive evidence that the internet increased trade in physical goods due to a reduction in the cost of international communication. The asynchronous nature of email communication may be particularly important for reducing the cost of communication across many time zones (Borenstein and Saloner, 2001). Gomez-Herrera et al. (2014) suggests, however, that this increase may disproportionately benefit English-language countries. Several of the 43 papers highlighted earlier in this review demonstrate that the internet facilitated trade in digital services (Blum and Goldfarb, 2006; Alaveras and Martens, 2015; Lendle et al., 2016), and this might lead to offshoring of certain jobs (Tambe and Hitt, 2012). While there is some debate about whether distance matters less overall than it did prior to the diffusion of the internet (Leamer, 2007; Cristea, 2011; Krautheim, 2012), our reading of the literature is that those papers that focus on the direct impact of the internet find a decrease in the role of distance in trade (Freund and Weinhold, 2004; Clarke, 2008; Lendle et al., 2016; Hui, 2017) while other papers identify other weaker forces moving in the opposite direction. Consistent with an impact of easy international communication on trade, Gorodnichenko and Talavera (2017) shows that exchange rate pass-through is faster online. 7.2 Region-level effects Another question is the extent to which the internet has led to redistribution of economic benefits within countries and in particular between cities and rural areas. Gaspar and Glaeser (1998) notes that digital communication could be a substitute or a complement to cities. Overall, the literature suggests that the biggest beneficiaries of digital technologies and data have been in large urban areas. The prime early beneficiaries of online media were in urban areas because the highest quality online content was produced in urban areas. This might be one reason why Savage and Waldman (2009) finds that urbanites have higher willingness to pay for broadband. Eichengreen et al. (2016) shows that efficient electronic communication in foreign exchange markets led to an increase in offshore currency trading and the consequent agglomeration of currency markets in London and a small number of other major financial centers. Forman et al. (2012) shows that wealthy cities were the primary beneficiaries of the business internet. The mechanism through which cities appear to have benefited has been shown to depend on agglomeration effects, particularly with respect to skilled workers in local labor markets. 44 Forman et al. (2005, 2008) show that internet adoption by businesses is higher in cities and in large companies but the advantage associated with being in a city or a large company are substitutes for each other. This indicates the importance of agglomeration effects. Dranove et al. (2014) finds similar results for hospitals. In contrast to the above work, there is some evidence that internet adoption has some benefits for isolated individuals and rural areas. Autor (2001) and Gaspar and Glaeser (1998) speculated that the internet might reduce the need for task-specific workspace, thereby increasing the prevalence of ‘telecommuting’ and reducing the need for home and work to be nearby. Kolko (2012) shows that broadband disproportionately benefited people in low density areas in terms of employment, though the overall effect is small. Furthermore, while the primary result in the Sinai and Waldfogel (2004) study cited above is that urban areas have higher quality internet content, they also show that isolated individuals consume disproportionately more internet news. For example, blacks in white neighborhoods consume more internet news. Finally, Forman et al. (2005) shows that basic internet technologies have (perhaps disproportionately) benefited rural and isolated cities. Overall, two forces are at play. Agglomeration effects mean that cities disproportionately benefit. Low cost communication, however, can benefit the geographically isolated. In any particular context, the overall result depends on the balance between these forces. Generally, the more difficult the technology is to use, the more likely that agglomeration effects dominate. 7.3 Firm-level effects As noted above, the growth accounting literature has suggested a compelling link between digital technology investments and productivity growth at the country level; however, causal inference is difficult with macro-level measurement. There is a large and growing literature that documents a direct link from digital technology adoption and usage to productivity 45 growth at the firm level. By using micro data and various econometric techniques to address selection, omitted variables bias, and simultaneity, this literature has found that digital technology adoption and usage does enhance productivity. However, the story is not as simple as it seems at first. Only some types of firms experience improved productivity. Various factors enhance or mitigate this relationship, including organizational change, skills, geography, regulation, firm size and age, and the potential for spillovers and/or network externalities. Reviews by Brynjolfsson and Saunders (2010) and Draca et al. (2006) conclude that ICT adoption and usage increase firm performance. This conclusion is driven by a large number of papers and a variety of settings. The correlation between IT and productivity is even stronger when ICT investment is modeled with a lag (Brynjolfsson and Hitt, 2003). There are also specific case studies on the effects of ICT on productivity. Baker and Hubbard (2004) show that ICT improved productivity in trucking. McElheran and Jin (2017) show improved productivity in manufacturing. Agrawal and Goldfarb (2008) show that BITNET increased academic productivity at middle-tier universities. In healthcare, Athey and Stern (2002) show that ICT, in the form of Enhanced 911, improved emergency response, Miller and Tucker (2011) and McCullough et al. (2016) show that Electronic Medical Records improve patient outcomes, Dranove et al. (2014) show that EMR reduces hospital costs in the presence of complementary skills but not otherwise, and Lee et al. (2013) show that electronic medical records (EMR) increase hospital productivity. Bloom et al. (2012) use a large-scale multi-country firm-level panel database on ICT and productivity. Their database contains 19,000 firms in 13 EU countries over 11 years, plus a smaller panel of US firms over the same time period. They conclude that ICT does increase productivity, though they find considerable heterogeneity in this effect across countries and type of firm. They emphasize the importance of organizational capital, showing that US multinationals operating in the UK experienced the same productivity miracle as 46 US-based establishments. In contrast, other multinationals (and other firms) in the UK did not. The title communicates the idea well: Americans do I.T. better. They argue that US firms are organized in way that allows them to use ICT more efficiently. This essential role of organizational capital and organizational structure in making productive use of ICT investments is a recurring theme elsewhere in the literature (Bresnahan et al., 2002; Brynjolfsson and Saunders, 2010; Garicano, 2010; Tambe et al., 2012; Brynjolfsson and McElheran, 2016). In addition to change in the organizational structure, the most effective use of advanced ICT also involves ‘co-invention’, the process of adapting ICT to the organization’s needs (Bresnahan and Greenstein, 1996). Such process innovation is easiest for firms in places that have a pool of local ICT expertise to draw on (Forman et al., 2008; Dranove et al., 2014). This of course reflects the extensive literature on skill-biased technological change, which is long and beyond the scope of this review. As reviewed in Acemoglu and Autor (2012), given that prior generations of IT are skill-biased, it is perhaps unsurprising that use of the internet to enhance productivity is also skill-biased. Correspondingly, in the context of the internet, Akerman et al. (2015) provide evidence that broadband diffusion in Norway disproportionately benefited skilled workers. 7.4 Consumer-level effects Measurements that focus on productivity or national income accounts do not measure consumer surplus. To the extent that much of the most valuable content online is free, measures of productivity and GDP may miss a potential increase in consumer surplus driven by the internet (Scott and Varian, 2015; Brynjolfsson et al., 2017; Greenstein and McDevitt, 2011; Goolsbee and Klenow, 2006). With time use data, Wallsten (2013) demonstrates that we are spending an increasing proportion of our leisure time online, substituting for offline leisure (including television), and to a lesser extent work and sleep. Also with time use data, 47 Goolsbee and Klenow (2006) estimates a consumer surplus of $3000 per person-year in 2005. Goldfarb and Prince (2008) shows that this effect is heterogeneous. Overall, rich educated Americans are more likely to adopt and therefore overall consumer surplus disproportionately goes to the wealthy. At the same time, conditional on adoption, lower-income people spend more time online. Therefore, among adopters, consumer surplus (at least relative to overall consumption) is higher for lower-income people. Many studies arrive at specific estimates of the consumer surplus from internet-related technologies. Greenstein and McDevitt (2011) measures the consumer surplus associated with broadband diffusion at $4.8 to $6.7 billion between 1999 and 2006. Brynjolfsson and Oh (2012) estimates the consumer surplus from free online services to be close to $100 billion. Cohen et al. (2016) estimates billions of dollars in consumer surplus from the UberX car service alone.6 Brynjolfsson et al. (2017) provides perhaps the most comprehensive estimate of the consumer surplus of the internet by using (incentive compatible) choice experiments. For example, in one study, they asked people how much they would need to be paid in order to not have access to Facebook for a month. They then implemented the result by actually blocking their respondents’ access to Facebook in exchange for payment. They estimate a value of Facebook of about $750 per user per year, or $18 billion for the United States. They also generated user-level survey estimates of the consumer surplus from other free online services such as search engines ($16,000 per user per year) and online video ($900 per user per year). Before concluding, it is important to recognize that there are other, perhaps negative, changes to overall welfare that may result from shifts in internet consumption that are not captured by these surplus measures. Belo et al. (2013) shows a reduction in grades associated with school’s adopting broadband, perhaps because online games distracted students. 6Greenstein and Nagle (2014) estimates an intangible benefit of digitization distinct from consumer surplus: The value of open source. It shows that open source software Apache generates at least $2 billion in unmeasured benefits to the US economy. 48 Bhuller et al. (2013) argues that internet diffusion may have increased sex crime, likely due to increased consumption of pornography (not because of reporting or matching between offenders and victims). Similarly, Chan et al. (2015) suggests an increase in racial hate crimes associated with the internet, and Falck et al. (2014) suggests that internet availability reduces voter turnout in elections. 8 Conclusions Across a variety of fields, economists examine how digital technologies change economic activity. While these papers often have different perspectives and cite different literatures, a core theme is that digitization has reduced a number of specific economic costs. We have identified five such costs: Search, reproduction, transportation, tracking, and verification. These themes inform our understanding of the nature of digital economic activity, and of the interaction between digital and non-digital settings. In defining the scope of this article, we drew boundaries. For example, we did not discuss work on skill-biased technical change. Because skill-bias is not primarily driven by the storage of information in bits, and because there are several other reviews of that literature, we instead refer to Katz and Autor (1999), Acemoglu (2002), Goldin and Katz (2008), and Acemoglu and Autor (2012). Similarly, we limit the discussion of the digital technology growth accounting literature, referring the reader to Jorgenson et al. (2008) and Van Reenen et al. (2010). We also limited our discussion on three topics that have already received reviews in the Journal of Economic Literature: privacy (Acquisti et al., 2016), online auctions (Bajari and Hortacsu, 2004), and telecommunications pricing and universal service Vogelsang (2003). This overview highlights that changes to economic behavior that result from the change of costs inherent in the digital context are not as obvious as basic economic models might imply. Key open questions remain with respect to each of the cost changes highlighted. 49 Further, other categories of costs may shift downwards as digital technology evolves. 50 References Aaltonen, A. and S. Seiler (2016). Cumulative growth in user-generated content production: Evidence from wikipedia. Management Science 62 (7), 2054–2069. Acemoglu, D. (2002). Technical change, inequality, and the labor market. Journal of Economic Literature 40 (1), 7–72. Acemoglu, D., P. Aghion, C. LeLarge, J. V. Reenan, and F. Zilibotti (2007). Technology, information and the decentralization of the firm. Quarterly Journal of Economics 122, 1759–1799. Acemoglu, D. and D. Autor (2012). Handbook of Labor Economics, Volume 4, Chapter 12, pp. 1043–1171. Elsevier. Acemoglu, D., D. Laibson, and J. A. List (2014, May). Equalizing superstars: The internet and the democratization of education. American Economic Review 104 (5), 523–27. Acquisti, A. and C. M. Fong (2013). An experiment in hiring discrimination via online social networks. Mimeo, CMU . Acquisti, A. and R. Gross (2009). Predicting social security numbers from public data. Proceedings of the National Academy of Sciences. Acquisti, A., C. R. Taylor, and L. Wagman (2016). The economics of privacy. Journal of Economic Literature. Acquisti, A. and C. Tucker (2014). Guns, privacy and crime. Mimeo, MIT. Acquisti, A. and H. R. Varian (2005). Conditioning prices on purchase history. Marketing Science 24 (3), 367–381. 51 Adams, W. and J. L. Yellen (1976). Commodity bundling and the burden of monopoly. The Quarterly Journal of Economics 90 (3), 475–498. Agrawal, A., C. Catalini, and A. Goldfarb (2015). Crowdfunding: Geography, social networks, and the timing of investment decisions. Journal of Economics & Management Strategy 24 (2), 253–274. Agrawal, A. and A. Goldfarb (2008). Restructuring research: Communication costs and the democratization of university innovation. American Economic Review 98 (4), 1578–90. Agrawal, A., N. Lacetera, and E. Lyons (2016). Does standardized information in online markets disproportionately benefit job applicants from less developed countries? Journal of International Economics 103, 1 – 12. Aguiar, L. and J. Waldfogel (2015). Streaming reaches flood stage: Does spotify stimulate or depress music sales? Working Paper, University of Minnesota. Aguiar, L. and J. Waldfogel (2016). Digitization, copyright, and the welfare effects of music trade. Working Paper, University of Minnesota. Aker, J. C. (2010). Information from markets near and far: Mobile phones and agricultural markets in niger. American Economic Journal: Applied Economics 2 (3), 46–59. Akerman, A., I. Gaarder, and M. Mogstad (2015). The skill complementarity of broadband internet. The Quarterly Journal of Economics 130 (4), 1781. Alaveras, G. and B. Martens (2015, August). International Trade in Online Services. JRC Working Papers on Digital Economy 2015-08, Joint Research Centre (Seville site). Allcott, H. and M. Gentzkow (2017, May). Social media and fake news in the 2016 election. Journal of Economic Perspectives 31 (2), 211–36. 52 Anderson, C. (2006). The Long Tail. Hyperion. Anderson, E. T., N. M. Fong, D. I. Simester, and C. E. Tucker (2010). How sales taxes affect customer and firm behavior: The role of search on the internet. Journal of Marketing Research 47 (2), 229–239. Anderson, J. E. and E. van Wincoop (2004). Trade costs. Journal of Economic Literature 42 (3), 691–751. Anderson, S. P. and A. de Palma (2009). Information congestion. RAND Journal of Economics 40 (4), 688–709. Anderson, S. P. and A. de Palma (2013, July). Shouting to Be Heard in Advertising. Management Science 59 (7), 1545–1556. Anderson, S. P. and J. S. Gans (2011, November). Platform siphoning: Ad-avoidance and media content. American Economic Journal: Microeconomics 3 (4), 1–34. Anenberg, E. and E. Kung (2015). Information technology and product variety in the city: The case of food trucks. Journal of Urban Economics 90, 60 – 78. Antweiler, W. and M. Z. Frank (2004). Is all that talk just noise? The information content of internet stock message boards. Journal of Finance 59 (3), 1259–1294. Arnosti, N., M. Beck, and P. Milgrom (2016). Adverse selection and auction design for internet display advertising. American Economic Review 106 (10), 2852–66. Athey, S., E. Calvano, and J. S. Gans (2014). The impact of the internet on advertising markets for news media. Rotman School of Management Working Paper (2180851). Athey, S., C. Catalini, and C. E. Tucker (2017). The digital privacy paradox: Small money, small costs, small talk. Working Paper, MIT. 53 Athey, S. and J. S. Gans (2010, May). The impact of targeting technology on advertising markets and media competition. American Economic Review 100 (2), 608–13. Athey, S., M. Mobius, and J. Pal (2017, January). The impact of news aggregators on internet news consumption: The case of localization. Technical Report 3353. Athey, S. and S. Stern (2002, Autumn). The impact of information technology on emergency health care outcomes. RAND Journal of Economics 33 (3), 399–432. Autor, D. H. (2001, March). Wiring the labor market. Journal of Economic Perspectives 15 (1), 25–40. Ayres, I., M. Banaji, and C. Jolls (2015). Race effects on ebay. RAND Journal of Economics 46 (4), 891–917. Ba, S. and P. Pavlou (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. MIS Quarterly 26, 243–268. Bajari, P. and A. Hortacsu (2003). The winner’s curse, reserve prices, and endogenous entry: Empirical insights from ebay auctions. RAND Journal of Economics 34 (2), 329–355. Bajari, P. and A. Hortacsu (2004). Economic insights from internet auctions. Journal of Economic Literature 42 (2), 457–486. Baker, G. P. and T. N. Hubbard (2003). Make versus buy in trucking: Asset ownership, job design, and information. American Economic Review 93 (3), 551–572. Baker, G. P. and T. N. Hubbard (2004). Contractibility and Asset Ownership: OnBoard Computers and Governance in U. S. Trucking. The Quarterly Journal of Economics 119 (4), 1443–1479. 54 Bakos, Yannis & Brynjolfsson, E. (1999, 12). Bundling information goods: Price, profits, and efficiency. Management Science 45 (12), 1613–1630. Bakos, Y. (2001, March). The emerging landscape for retail e-commerce. Journal of Economic Perspectives 15 (1), 69–80. Bakos, Y. and E. Brynjolfsson (2000). Bundling and competition on the internet. Marketing Science 19 (1), 63–82. Balasubramanian, S. (1998). Mail versus mall: A strategic analysis of competition between direct marketers and conventional retailers. Marketing Science 17 (3), 181–195. Bar-Isaac, H., G. Caruana, and V. Cunat (2012). Search, design, and market structure. American Economic Review 102 (2), 1140–60. Barber, B. M. and T. Odean (2001, March). The internet and the investor. Journal of Economic Perspectives 15 (1), 41–54. Baye, M. R. and J. Morgan (2001a, June). Information gatekeepers on the internet and the competitiveness of homogeneous product markets. American Economic Review 91 (3), 454–474. Baye, M. R. and J. Morgan (2001b). Information gatekeepers on the internet and the competitiveness of homogenous product markets. American Economic Review 91 (3), 454– 474. Baye, M. R., J. Morgan, and P. Scholten (2004a). Price dispersion in the small and in the large: Evidence from an internet price comparison site. Journal of Industrial Economics 52 (4), 463–496. 55 Baye, M. R., J. Morgan, and P. Scholten (2004b). Price dispersion in the small and the large: Evidence from an internet price comparison site. Journal of Industrial Economics 52 (4), 463–496. Bechtold, S. and C. Tucker (2014, December). Trademarks, triggers, and online search. Journal of Empirical Legal Studies 11 (4). Belo, R., P. Ferreira, and R. Telang (2013). Broadband in school: Impact on student performance. Management Science 60 (2), 265–282. Bergemann, D. and A. Bonatti (2011). Targeting in advertising markets: implications for offline versus online media. RAND Journal of Economics 42 (3), 417–443. Besen, S., P. Milgrom, B. Mitchell, and P. Srinagesh (2001). Advances in routing technologies and internet peering agreements. The American Economic Review 91 (2), 292–296. Bhargava, H. K. and V. Choudhary (2008). Research note: when is versioning optimal for information goods? Management Science 54 (5), 1029–1035. Bhuller, M., T. Havnes, and M. Leuven, Edwin; Mogstad (2013). Broadband internet: An information superhighway to sex crime? Review of Economic Studies 80 (4), 1237–66. Biasi, B. and P. Moser (2016, October). Effects of copyright on science: Evidence from the world war ii book republication program. Technical report. Bitzer, J. and P. J. Schroder (2005, July). Bug-fixing and code-writing: The private provision of open source software. Information Economics and Policy 17 (3), 389–406. Blake, T., C. Nosko, and S. Tadelis (2015). Consumer heterogeneity and paid search effectiveness: A large scale field experiment. Econometrica 83 (1), 155–174. 56 Bloom, N., L. Garicano, R. Sadun, and J. Van Reenen (2014). The distinct effects of information technology and communication technology on firm organization. Management Science 60 (12), 2859–2885. Bloom, N., R. Sadun, and J. Van Reenen (2012, February). Americans do it better: Us multinationals and the productivity miracle. American Economic Review 102 (1), 167– 201. Blum, B. S. and A. Goldfarb (2006). Does the internet defy the law of gravity? Journal of International Economics 70 (2), 384–405. B¨ohme, R., N. Christin, B. Edelman, and T. Moore (2015). Bitcoin: Economics, technology, and governance. Journal of Economic Perspectives 29 (2), 213–238. Borenstein, S. and G. Saloner (2001, March). Economics and electronic commerce. Journal of Economic Perspectives 15 (1), 3–12. Bourreau, M., F. Kourandi, and T. Valletti (2015). Net neutrality with competing internet platforms. The Journal of Industrial Economics 63 (1), 30–73. Boxell, L., M. Gentzkow, and J. M. Shapiro (2017, March). Is the internet causing political polarization? evidence from demographics. Working paper, Brown University. Brand, K., M. Gaynor, P. McAlvanah, D. Schmidt, and E. Schneirov (2014). Economics at the FTC: Office Supply Retailers Redux, Healthcare Quality Efficiencies Analysis, and Litigation of an Alleged Get-Rich-Quick Scheme. Review of Industrial Organization 45 (4), 325–344. Brandimarte, L., A. Acquisti, and G. Loewenstein (2012). Misplaced confidences: Privacy and the control paradox. Social Psychological and Personality Science. 57 Bresnahan, T. F., E. Brynjolfsson, and L. M. Hitt (2002, February). Information technology, workplace organization and the demand for skilled labor: Firm-level evidence. The Quarterly Journal of Economics 117 (1), 339–76. Bresnahan, T. F. and S. Greenstein (1996). Technical progress and co-invention in computing and in teh uses of computers. Brookings Papers on Economic Activity. Microeconomics, 1–78. Brown, J. and A. Goolsbee (2002). Does the internet make markets more competitive? Evidence from the life insurance industry. Journal of Political Economy 110 (3), 481–507. Brynjolfsson, E., F. Eggers, and A. Gannamaneni2017 (2017). Using massive online choice experiments to measure changes in well-being. Working paper, MIT. Brynjolfsson, E. and L. Hitt (2003, November). Computing productivity: Firm-level evidence. Review of Economics and Statistics 85 (4), 793–808. Brynjolfsson, E., Y. Hu, and M. S. Rahman (2009). Battle of the retail channels: How product selection and geography drive cross-channel competition. Management Science 55 (11), 1755–1765. Brynjolfsson, E., Y. Hu, and D. Simester (2011). Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. Management Science 57 (8), 1373–1386. Brynjolfsson, E., Y. J. Hu, and M. D. Smith (2003, November). Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. Management Science 49 (11), 1580–1596. Brynjolfsson, E. and K. McElheran (2016, May). The rapid adoption of data-driven decisionmaking. American Economic Review 106 (5), 133–39. 58 Brynjolfsson, E. and J. Oh (2012). The attention economy: Measuring the value of free digital services on the internet. In ICIS. Association for Information Systems. Brynjolfsson, E. and A. Saunders (2010). Wired for Innovation: How Information Technology is Reshaping the Economy. The MIT Press. Brynjolfsson, E. and M. Smith (2000). Frictionless commerce? A comparison of internet and conventional retailers. Management Science 46 (4), 563–585. Cabral, L. (2012). Reputation on the internet. Oxford Handbook of the Digital Economy, 343–354. Cabral, L. and A. Hortacsu (2010). The dynamics of seller reputation: Evidence from ebay\*. Journal of Industrial Economics 58 (1), 54–78. Caillaud, B. and B. Jullien (2003). Chicken & egg: Competition among intermediation service providers. The RAND Journal of Economics 34 (2), 309–328. Cairncross, F. (1997). The Death of Distance. Harvard University Press. Cambridge, MA. Calzada, J. and R. Gil (2016, September). What do news aggregators do? evidence from google news in spain and germany. Technical report. Campbell, J., A. Goldfarb, and C. Tucker (2015). Privacy regulation and market structure. Journal of Economics & Management Strategy 24 (1), 47–73. Campbell-Kelly, M. (2004). From Airline Reservations to Sonic the Hedgehog: A history of the software industry. MIT Press. Carbajo, J., D. de Meza, and D. J. Seidmann (1990). A strategic motivation for commodity bundling. The Journal of Industrial Economics 38 (3), 283–298. 59 Carlton, D. W., J. S. Gans, and M. Waldman (2010, August). Why tie a product consumers do not use? American Economic Journal: Microeconomics 2 (3), 85–105. Castells, M. (2001). The Internet Galaxy: Reflections on the Internet, Business and Society. Oxford University Press, London. Catalini, C. and J. S. Gans (2016). Some simple economics of the blockchain. SSRN Working Paper 2874598, https://papers.ssrn.com/sol3/papers.cfm?abstract id=2874598 . Catalini, C. and C. Tucker (2016). Seeding the s-curve? the role of early adopters in diffusion. Ceruzzi, P. E. (2003). A History of Modern Computing, Second edition. MIT Press. Chan, J., A. Ghose, and R. Seamans (2015). The internet and racial hate crime: Offline spillovers from online access. Chen, Y. (1997). Paying customers to switch. Journal of Economics & Management Strategy 6 (4), 877–897. Chen, Y. and T. Zhang (2011, November). Equilibrium price dispersion with heterogeneous searchers. International Journal of Industrial Organization 29 (6), 645–54. Chevalier, J. and D. Mayzlin (2006). The effect of word of mouth online: Online book reviews. Journal of Marketing Research 43 (345-354). Chiou, L. and C. Tucker (2012). How does the use of trademarks by third-party sellers affect online search? Marketing Science 31 (5), 819–837. Chiou, L. and C. Tucker (2013). Paywalls and the demand for news. Information Economics and Policy 25 (2), 61–69. Chiou, L. and C. Tucker (2017). Content aggregation by platforms: The case of the news media. Journal of Economics & Management Strategy, n/a–n/a. 60 Choi, J. and D. Bell (2011). Preference minorities and the internet. Journal of Marketing Research 58 (3), 670–682. Choi, J. P., D.-S. Jeon, and B.-C. Kim (2015, August). Net neutrality, business models, and internet interconnection. American Economic Journal: Microeconomics 7 (3), 104–41. Choi, T.-M. (2012). Handbook of Newsvendor Problems: Models, Extensions, and Applications, Volume 176 of International Series in Operations Research & Management Science. Springer Science and Business Media. Chu, C. S., P. Leslie, and A. Sorensen (2011, February). Bundle-size pricing as an approximation to mixed bundling. American Economic Review 101 (1), 263–303. Clarke, G. R. (2008). Has the internet increased exports for firms from low and middleincome countries? Information Economics and Policy 20 (1), 16 – 37. Cohen, P., R. Hahn, J. Hall, S. Levitt, and R. Metcalfe (2016, September). Using big data to estimate consumer surplus: The case of uber. Working Paper 22627, National Bureau of Economic Research. Cornes, R. and T. Sandler (1986). The theory of externalities, public goods, and club goods. Cambridge University Press. Corrado, C., C. Hulten, and D. Sichel (2009). Intangible capital and u.s. economic growth. Review of Income and Wealth 55 (3), 661–685. Corrado, C. A. and C. R. Hulten (2010). How do you measure a ”technological revolution”? The American Economic Review 100 (2), 99–104. Corts, K. (1998). Third-degree price discrimination in oligopoly: All-out competition and stretegic commitment. RAND Journal of Economics 29 (2), 306–23. 61 Cremer, J., P. Rey, and J. Tirole (2000). Connectivity in the commercial internet. Journal of Industrial Economics 48 (4), 433–472. Cristea, A. D. (2011). Buyer-seller relationships in international trade: Evidence from u.s. states’ exports and business-class travel. Journal of International Economics 84 (2), 207 – 220. Cullen, Z. and C. Farronato (2016). Outsourcing tasks online: Matching supply and demand on peer-to-peer internet platforms. HBS Working Paper . Cullen, Z. and B. Pakzad-Hurson (2017). Equilibrium effects of pay transparency. Working paper, Harvard University. Dana, James D., J. and E. Orlov (2014, November). Internet penetration and capacity utilization in the us airline industry. American Economic Journal: Microeconomics 6 (4), 106–37. Danaher, B. and M. D. Smith (2014). Gone in 60 seconds: The impact of the megaupload shutdown on movie sales. International Journal of Industrial Organization 33, 1 – 8. Danaher, B., M. D. Smith, and R. Telang (2013, June). Piracy and Copyright Enforcement Mechanisms, pp. 25–61. University of Chicago Press. Danaher, B., M. D. Smith, R. Telang, and S. Chen (2014). The effect of graduated response anti-piracy laws on music sales: Evidence from an event study in france. The Journal of Industrial Economics 62 (3), 541–553. de Corniere, A. (2016). Search advertising. American Economic Journal: Microeconomics 8 (3), 156–188. 62 De Los Santos, B., A. Hortasu, and M. R. Wildenbeest (2012, May). Testing models of consumer search using data on web browsing and purchasing behavior. American Economic Review 102 (6), 2955–80. Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. Management Science 49 (10), 1407–1424. Deneckere, R. J. and R. P. McAfee (1996). Damaged goods. Journal of Economics and Management Strategy 5 (2), 149–74. Diamond, P. (1971). A model of price adjustment. Journal of Economic Theory 3 (2), 156–168. Dinerstein, M., L. Einav, J. Levin, and N. Sundaresan (2017). Consumer price search and platform design in internet commerce. Working paper, Stanford University. Dinlersoz, E. and P. Pereira (2007). On the diffusion of electronic commerce. International Journal of Industrial Organization 25 (3), 541–574. Disdier, A.-C. and K. Head (2008). The puzzling persistence of the distance effect on bilateral trade. Review of Economics and Statistics 90 (1), 37–48. Doleac, J. L. (2017, January). The effects of dna databases on crime. American Economic Journal: Applied Economics 9 (1), 165–201. Doleac, J. L. and L. C. Stein (2013). The visible hand: Race and online market outcomes. The Economic Journal 123 (572), F469–F492. Draca, M., R. Sadun, and J. Van Reenen (2006). Productivity and ict: A review of the evidence. 63 Dranove, D., C. Forman, A. Goldfarb, and S. Greenstein (2014). The trillion dollar conundrum: Complementarities and health information technology. American Economic Journal: Economic Policy 6 (4), 239–270. Dube, J.-P. and S. Misra (2017). Scalable price targeting. Working paper, University of Chicago. Duch-Brown, N., L. Grzybowski, A. Romahn, and F. Verboven (2017). The impact of online sales on consumers and firms. evidence from consumer electronics. International Journal of Industrial Organization 52, 30 – 62. Economides, N. and B. E. Hermalin (2012). The economics of network neutrality. The RAND Journal of Economics 43 (4), 602–629. Economides, N. and P. Jeziorski (2017). Mobile money in tanzania. Marketing Science forthcoming. Edelman, B. (2009). Prices and unpriced online markets. Journal of Economic Perspectives 23 (3), 21–36. Edelman, B. and M. Luca (2014, January). Digital discrimination: The case of airbnb.com. Edelman, B., M. Ostrovsky, and M. Schwarz (2007, March). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. American Economic Review 97 (1), 242–259. Eichengreen, B., R. Lafarguette, and A.Mehl (2016). Cables, sharks and servers: Technology and the geography of the foreign exchange market. NBER Working Paper 21884 . Einav, L., C. Farronato, J. Levin, and N. Sundaresan (2016, January). Auctions versus posted prices in online markets. Journal of Political Economy - forthcoming. 64 Einav, L., C. Farronato, J. Levin, and N. Sundaresan (2017). Auctions versus posted prices in online markets. Journal of Political Economy. Einav, L., P. Klenow, B. Klopack, J. Levin, L. Levin, and W. Best (2017). Assessing the gains from e-commerce. Working paper, Stanford University. Einav, L., D. Knoepfle, J. Levin, and N. Sundaresan (2014). Sales taxes and internet commerce. American Economic Review 104 (1), 1–26. Ellison, G. (2011). Is peer review in decline? Economic Inquiry 49 (3), 635–657. Ellison, G. and S. F. Ellison (2005, June). Lessons about markets from the internet. Journal of Economic Perspectives 19 (2), 139–158. Ellison, G. and S. F. Ellison (2009a). Search, obfuscation, and price elasticities on the internet. Econometrica 77 (2), 427–452. Ellison, G. and S. F. Ellison (2009b, August). Tax sensitivity and home state preferences in internet purchasing. American Economic Journal: Economic Policy 1 (2), 53–71. Ellison, G., S. F. Ellison, D. Liu, H. Zhang, and V. Bhattacharya (2014, September). Match quality, search, and the internet market for used books. Massachusetts Institute of Technology. Ershov, D. (2017). The effect of consumer search costs on entry and quality in the mobile app market. Working Paper, University of Toronto. Evans, D. S. (2009). The online advertising industry: Economics, evolution, and privacy. Journal of Economic Perspectives 23 (3), 37–60. Falck, O., R. Gold, and S. Heblich (2014, July). E-lections: Voting behavior and the internet. American Economic Review 104 (7), 2238–65. 65 Farrell, J. (2012). Can privacy be just another good. J. on Telecomm. & High Tech. L. 10, 251. Farrell, J. and T. Simcoe (2012). Oxford Handbook of the Digital Economy, Chapter Four Paths to Compability, pp. 34–58. Oxford University Press. Fay, S. and J. Xie (2008, July-August). Probabilistic goods: A creative way of selling products and services. Marketing Science 27 (4), 674–90. Ferreira, F. and J. Waldfogel (2013). Pop internationalism: has half a century of world music trade displaced local culture? Economic Journal 123 (569), 634–664. Fleder, D. M. and K. Hosanagar (2009, May). Blockbuster culture’s next rise or fall: The impact of recommender systems on sales diversity. Managment Science 55 (5), 697–712. Forman, C., A. Ghose, and A. Goldfarb (2009). Competition between local and electronic markets: How the benefit of buying online depends on where you live. Management Science 55 (1), 47–57. Forman, C., A. Ghose, and B. Wiesenfeld (2009). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. Information Systems Research 19 (3). Special issue on the interplay between digital and social networks. Forman, C., A. Goldfarb, and S. Greenstein (2005). How did location affect adoption of the commercial internet? Global village vs. urban leadership. Journal of Urban Economics 58 (3), 389–420. Forman, C., A. Goldfarb, and S. Greenstein (2008). Understanding the inputs into innovation: Do cities substitute for internal firm resources? Journal of Economics & Management Strategy 17 (2), 295–316. 66 Forman, C., A. Goldfarb, and S. Greenstein (2012). The internet and local wages: A puzzle. American Economic Review 102 (1), 556–575. Forman, C. and K. McElheran (2013). The digital reorganization of firm boundaries. Working paper, University of Toronto. Forman, C. and N. van Zeebroeck (2012). From wires to partners: How the internet has fostered r&d collaborations within firms. Management Science 58 (8), 1549–1568. Fradkin, A. (2017, March). Search, matching, and the role of digital marketplace design in enabling trade: Evidence from airbnb. Mimeo, MIT Sloan School of Business. Fradkin, A. and C. Farronato (2016, January). Market structure with the entry of peer-topeer platforms: The case of hotels and airbnb. Fradkin, A., E. Grewal, and D. Holtz (2017). The determinants of online review informativeness: Evidence from field experiments on airbnb. Working paper, MIT Sloan School of Management. Freund, C. L. and D. Weinhold (2004). The effect of the internet on international trade. Journal of International Economics 62 (1), 171–189. Friedman, T. (2005). The World is Flat: A Brief History of the Twenty-First Century. New York: Farrar, Straus, and Giroux. Fudenberg, D. and J. Tirole (1998). Upgrades, tradeins, and buybacks. The RAND Journal of Economics 29 (2), 235–258. Fudenberg, D. and J. Tirole (2000). Customer poaching and brand switching. The RAND Journal of Economics 31 (4), 634–657. 67 Fudenberg, D. and J. M. Villas-Boas (2007). Behaviour-Based Price Discrimination and Customer Recognition. Oxford: Elsevier Science. Fudenberg, D. and J. M. Villas-Boas (2012, August). Price discrimination in the digital economy. Oxford Handbook of the Digital Economy. Gal-Or, E. and A. Ghose (2005). The economic incentives for sharing security information. Information Systems Research 16 (2), 186–208. Gandal, N. (2006). The effect of native language on internet usage. International Journal of the Sociology of Language 182, 25–40. Gans, J. S., A. Goldfarb, and M. Lederman (2016, August). Exit, tweets, and loyalty. (2889388). Garicano, L. (2000). Hierarchies and the organization of knowledge in production. Journal of Political Economy 108 (5), 874–904. Garicano, L. (2010, July). Policemen, managers, lawyers: New results on complementarities between organization and information and communication technology. International Journal of Industrial Organization 28 (4), 355–358. Gaspar, J. and E. L. Glaeser (1998). Information technology and the future of cities. Journal of Urban Economics 43 (1), 136–156. Gentzkow, M. (2007, June). Valuing new goods in a model with complementarity: Online newspapers. American Economic Review 97, 713–744(32). Gentzkow, M. (2014, May). Trading dollars for dollars: The price of attention online and offline. American Economic Review 104 (5), 481–88. 68 Gentzkow, M. and J. M. Shapiro (2011). Ideological segregation online and offline. Quarterly Journal of Economics 126 (4), 1799–1839. Gertner, R. H. and R. S. Stillman (2001). Vertical integration and internet strategies in the apparel industry. Journal of Industrial Economics 49 (4), 441–461. Ghani, E., W. R. Kerr, and C. Stanton (2014). Diasporas and outsourcing: Evidence from odesk and india. Management Science 60 (7), 1677–1697. Giorcelli, M. and P. Moser (2016, December). Copyrights and creativity: Evidence from italian operas. Journal of Economic Literature (O3, O33, O34, K11, N3). Goetz, D. (2017). Competition and dynamic bargaining in the broadband industry. Working paper, University of Toronto. Goldfarb, A. (2006). The (teaching) role of universities in the diffusion of the internet. International Journal of Industrial Organization 24 (2), 203–225. Goldfarb, A. (2014). What is different about online advertising? Review of Industrial Organization 44 (2), 115–129. Goldfarb, A., R. C. McDevitt, S. Samila, and B. S. Silverman (2015). The effect of social interaction on economic transactions: Evidence from changes in two retail formats. Management Science 61 (12), 2963–2981. Goldfarb, A. and J. Prince (2008). Internet adoption and usage patterns are different: Implications for the digital divide. Information Economics and Policy 20 (1), 2–15. Goldfarb, A. and C. Tucker (2011a, April). Advertising Bans and the Substitutability of Online and Offline Advertising. Journal of Marketing Research 48 (2), 207–227. 69 Goldfarb, A. and C. Tucker (2011b). Chapter 6 - online advertising. Volume 81 of Advances in Computers, pp. 289 – 315. Elsevier. Goldfarb, A. and C. Tucker (2011c, May). Online display advertising: Targeting and obtrusiveness. Marketing Science 30, 389–404. Goldfarb, A. and C. Tucker (2011d). Search engine advertising: Channel substitution when pricing ads to context. Management Science 57 (3), 458–470. Goldfarb, A. and C. Tucker (2011e). Substitution between offline and online advertising markets. Journal of Competition Law & Economics 7 (1), 37–44. Goldfarb, A. and C. Tucker (2012a). Privacy and innovation. In Innovation Policy and the Economy, Volume 12, NBER Chapters. National Bureau of Economic Research, Inc. Goldfarb, A. and C. Tucker (2012b). Shifts in privacy concerns. American Economic Review: Papers and Proceedings 102 (3), 349–53. Goldfarb, A. and C. E. Tucker (2011f). Privacy regulation and online advertising. Management Science 57 (1), 57–71. Goldin, C. and L. F. Katz (2008). The Race Between Education and Technology. Harvard University Press. Goldmanis, M., A. Hortasu, C. Syverson, and n. Emre (2010). E-commerce and the market structure of retail industries\*. The Economic Journal 120 (545), 651–682. Gomez-Herrera, E., B. Martens, and G. Turlea (2014). The drivers and impediments for cross-border e-commerce in the eu. Information Economics and Policy 28 (C), 83–96. Goolsbee, A. (2000, May). In a world without borders: The impact of taxes on internet commerce. Quarterly Journal of Economics 115 (2), 561–576. 70 Goolsbee, A. (2001). Competition in the computer industry: Online versus retail. Journal of Industrial Economics 49 (4), 487–99. Goolsbee, A. and P. J. Klenow (2006, May). Valuing consumer products by the time spent using them: An application to the internet. American Economic Review 96 (2), 108–113. Gordon, B., F. Zettelmeyer, N. Bhargava, and D. Chapsky (2016, 7). A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. Gordon, L. A. and M. P. Loeb (2002). The economics of information security investment. ACM Transactions on Information and System Security (TISSEC) 5 (4), 438–457. Gorodnichenko, Y. and O. Talavera (2017). Price setting in online markets: Basic facts, international comparisons, and cross-border integration. American Economic Review 107 (1), 249–282. Greenstein, D. (2000). Building and delivering the virtual world: Commercializing services for internet access. The Journal of Industrial Economics 48, 391–411. Greenstein, S. (2015). How the Internet Became Commercial. Princeton University Press. Greenstein, S. and R. McDevitt (2011). The broadband bonus: Estimating broadband internet’s economic value. Telecommunications Policy 35 (7), 617–632. Greenstein, S. and F. Nagle (2014). Digital dark matter and the economic contribution of Apache. Research Policy 43 (4), 623–631. Greenstein, S., M. Peitz, and T. Valletti (2016, May). Net neutrality: A fast lane to understanding the trade-offs. Journal of Economic Perspectives 30 (2), 127–50. Greenstein, S. and F. Zhu (2012). Is Wikipedia biased? American Economic Review 102 (3), 343–348. 71 Hafner, K. and M. Lyon (1996). where wizards stay up late. Simon and Schuster Paperbacks. Hagiu, A. (2012). Oxford Handbook of the Digital Economy, Chapter Software Platforms, pp. 59–82. Oxford University Press. Hagiu, A. and B. Jullien (2011). Why do intermediaries divert search? The RAND Journal of Economics 42 (2), 337–362. Halaburda, H. and Y. Yehezkel (2013, August). Platform competition under asymmetric information. American Economic Journal: Microeconomics 5 (3), 22–68. Hall, J., C. Kendrick, and C. Nosko (2016). The effects of uber’s surge pricing: A case study. Technical report. Haltiwanger, J. and R. S. Jarmin (2000). Understanding the Digital Economy: Data, Tools, and Research, Chapter Measuring the Digital Economy, pp. 13–33. MIT Press. Hampton, K. and B. Wellman (2003). Neighboring in netville: How the internet supports community and social capital in a wired suburb. City & Community 2 (4), 277–311. Hart, O. D. and J. Tirole (1988). Contract renegotiation and coasian dynamics. The Review of Economic Studies 55 (4), 509–540. Heald, P. (2009). Does the song remain the same? : an empirical study of bestselling musical compositions (1913-1932) and their use in cinema (1968-2007). Case Western Reserve Law Review. Henkel, J. (2009). Champions of revealingthe role of open source developers in commercial firms. Industrial and Corporate Change 18 (3), 435. Hermalin, B. and M. Katz (2006, September). Privacy, property rights and efficiency: The economics of privacy as secrecy. Quantitative Marketing and Economics 4 (3), 209–239. 72 Hirschman, A. O. (1970, January). Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States. Harvard University Press. Hollenbeck, B. (2016). Online reputation mechanisms and the decreasing value of brands. UCLA Working paper . Honka, E. (2014). Quantifying search and switching costs in the us auto insurance industry. The RAND Journal of Economics 45 (4), 847–884. Hortacsu, A., F. A. Martinez-Jerez, and J. Douglas (2009). The geography of trade in online transactions: Evidence from eBay and Mercadolibre. American Economic Journal: Microeconomics 1 (1), 53–74. Horton, J. J. and R. J. Zeckhauser (2016). Owning, using and renting: Some simple economics of the ‘sharing economy’. NBER Working Paper No. 22029 . Hossain, T. and J. Morgan (2006). ... plus shipping and handling: Revenue (non) equivalence in field experiments on ebay. Advances in Economic Analysis & Policy 5 (2). Houser, D. and J. Wooders (2005). Reputation in auctions: Theory and evidence from ebay. Journal of Economics and Management Strategy 15, 353–369. Hui, X. (2017). E-comemrce platforms and international trade: A large-scale field experiment. Working paper, MIT-Sloan School of Management. Hui, X., M. Saeedi, Z. Shen, and N. Sundaresan (2016). Reputation and regulations: Evidence from ebay. Management Science 62 (12), 3604–3616. Jensen, R. (2007). The digital provide: Information (technology), market performance, and welfare in the South Indian fisheries sector. Quarterly Journal of Economics, 879–924. 73 Jin, G. Z. and A. Kato (2006). Price, quality, and reputation: Evidence from an online field experiment. RAND Journal of Economics 37 (4), 983–1005. Jin, G. Z. and A. Kato (2007). Dividing online and offline: A case study. Review of Economic Studies 74 (3), 981–1004. Johnson, G. (2014). The impact of privacy policy on the auction market for online display advertising. Available at SSRN 2333193 . Johnson, J. P. (2002). Open source software: Private provision of a public good. Journal of Economics & Management Strategy 11 (4), 637–662. Johnson, J. P. (2013). Targeted advertising and advertising avoidance. RAND Journal of Economics 44 (1), 128–144. Jorgenson, D. W., M. S. Ho, and K. J. Stiroh (2008, March). A retrospective look at the u.s. productivity growth resurgence. Journal of Economic Perspectives 22 (1), 3–24. Jullien, B. (2012). Oxford Handbook of the Digital Economy, Chapter Two-Sided B to B Platforms, pp. 161–185. Oxford University Press. Katz, L. F. and D. Autor (1999). Handbook of Labor Economics, Volume 3, Chapter 26, pp. 1463–1555. Elsevier. Kim, J.-H. and L. Wagman (2015). Screening incentives and privacy protection in financial markets: a theoretical and empirical analysis. The RAND Journal of Economics 46 (1), 1–22. Kolko, J. (2012). Broadband and local growth. Journal of Urban Economics 71 (1), 100–113. Krautheim, S. (2012). Heterogeneous firms, exporter networks and the effect of distance on international trade. Journal of International Economics 87 (1), 27 – 35. Symposium on the Global Dimensions of the Financial Crisis. 74 Kremer, M., C. Brannen, and R. Glennerster (2013). The challenge of education and learning in the developing world. Science 340 (6130), 297–300. Kroft, K. and D. G. Pope (2014). Does Online Search Crowd Out Traditional Search and Improve Matching Efficiency? Evidence from Craigslist. Journal of Labor Economics 32 (2), 259 – 303. Krugman, P. (1979). A model of innovation, technology transfer, and the world distribution of income. Journal of Political Economy, 253–266. Kuhn, P. and H. Mansour (2014). Is internet job search still ineffective? Economic Journal 124 (581), 1213–1233. Kuhn, P. and M. Skuterud (2004, March). Internet job search and unemployment durations. American Economic Review 94 (1), 218–232. Kummer, M. E., O. Slivko, and X. M. Zhang (2015, November). Economic downturn and volunteering: Do economic crises affect content generation on wikipedia? Technical Report 15-078. Laffont, J.-J., S. Marcus, P. Rey, and J. Tirole (2001, May). Internet peering. American Economic Review 91 (2), 287–291. Laffont, J.-J., S. Marcus, P. Rey, and J. Tirole (2003). Internet interconnection and the off-net-cost pricing principle. The RAND Journal of Economics 34 (2), 370–390. Lambrecht, A. and K. Misra (2017). Fee or free: When should firms charge for online content? Management Science 63 (4), 1150–1165. Lambrecht, A. and C. Tucker (2013). When does retargeting work? Information specificity in online advertising. Journal of Marketing Research 50 (5), 561–576. 75 Lambrecht, A. and C. E. Tucker (2016). Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of stem career ads. Leamer, E. E. (2007, March). A flat world, a level playing field, a small world after all, or none of the above? a review of thomas l friedman’s the world is flat. Journal of Economic Literature 45 (1), 83–126. Lee, J., J. S. McCullough, and R. J. Town (2013). The impact of health information technology on hospital productivity. RAND Journal of Economics 44 (3), 545–568. Lee, R. S. and T. Wu (2009, September). Subsidizing creativity through network design: Zero-pricing and net neutrality. Journal of Economic Perspectives 23 (3), 61–76. Lemley, M. (2003). Place and cyberspace. California Law Review 91 (2), 521–542. Lenard, T. M. and P. H. Rubin (2009). In Defense of Data: Information and the Costs of Privacy. Technology Policy Institute Working Paper . Lendle, A., M. Olarreaga, S. Schropp, and P. Vzina (2016, 03). There Goes Gravity: eBay and the Death of Distance. Economic Journal 126 (591), 406–441. Lerner, J., P. A. Pathak, and J. Tirole (2006, May). The dynamics of open-source contributors. American Economic Review 96 (2), 114–118. Lerner, J. and J. Tirole (2002). Some simple economics of open source. The Journal of Industrial Economics 50 (2), 197–234. Levin, J. and P. Milgrom (2010, May). Online advertising: Heterogeneity and conflation in market design. American Economic Review 100 (2), 603–07. Lewis, R., J. M. Rao, and D. H. Reiley (2015, April). Measuring the Effects of Advertising: The Digital Frontier, pp. 191–218. University of Chicago Press. 76 Lewis, R. A. and J. M. Rao (2015). The unfavorable economics of measuring the returns to advertising. The Quarterly Journal of Economics 130 (4), 1941. Lewis, R. A. and D. H. Reiley (2014). Online ads and offline sales: measuring the effect of retail advertising via a controlled experiment on yahoo! Quantitative Marketing and Economics 12 (3), 235–266. Li, X., M. MacGarvie, and P. Moser (2015, September). Dead poet’s property - how does copyright influence price? Working Paper 21522, National Bureau of Economic Research. Lieber, E. and C. Syverson (2012). Oxford Handbook of the Digital Economy, Chapter Online versus offline competition, pp. 189–223. Oxford University Press. Liebowitz, S. J. (2008). Research notetesting file sharing’s impact on music album sales in cities. Management Science 54 (4), 852–859. Liebowitz, S. J. and A. Zentner (2012). Clash of the titans: Does internet use reduce television viewing? Review of Economics and Statistics 94 (1), 234–245. Livingston, J. (2005). How valuable is a good reputation? a sample selection model of internet auctions. The Review of Economics and Statistics 87 (3), 453–465. Loginova, O. (2001). Real and virtual competition. Journal of Industrial Economics 57 (2), 319–342. Luca, M. (2011). Reviews, reputation, and revenue: The case of yelp. com. Harvard Business School NOM Unit Working Paper (12-016). Luca, M. and G. Zervas (2016, December). Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. Management Science 62 (12), 3412–3427. 77 Lucking-Reiley, D. (1999, December). Using field experiments to test equivalence between auction formats: Magic on the internet. American Economic Review 89 (5), 1063–1080. Lucking-Reiley, D., D. Bryan, N. Prasad, and D. Reeves (2007). Pennies from ebay: The determinants of price in online auctions. Journal of Industrial Economics 55 (2), 223–233. Lucking-Reiley, D. and D. F. Spulber (2001, March). Business-to-business electronic commerce. Journal of Economic Perspectives 15 (1), 55–68. Lynch, J. G. and D. Ariely (2000). Wine online: Search costs affect competition on price, quality, and distribution. Marketing Science 19 (1), 83–103. Lyons, E. (2017). Team production in international labor markets: Experimental evidence from the field. American Economic Journal: Applied Economics. MacGarvie, M. and P. Moser (2015). Copyright and the profitability of authorship. Economic Analysis of the Digital Economy, 357. MacKie-Mason, J. K. and H. Varian (1994, September). Economic faqs about the internet. Journal of Economic Perspectives 8 (3), 75–96. Marrano, M. G., J. Haskel, and G. Wallis (2009). What happened to the knowledge economy? ict, intangible investment, and britain’s productivity record revisited. Review of Income and Wealth 55 (3), 686–716. Marthews, A. and C. Tucker (2014). Government surveillance and internet search behavior. Available at SSRN 2412564 . Maskin, E. and J. Riley (1984). Monopoly with incomplete information. The RAND Journal of Economics 15 (2), 171–96. 78 Mayzlin, D., Y. Dover, and J. Chevalier (2014). Promotional reviews: An empirical investigation of online review manipulation. American Economic Review 104 (8), 2421–55. McCabe, M. J. and C. M. Snyder (2015). Does online availability increase citations? theory and evidence from a panel of economics and business journals. The Review of Economics and Statistics 97 (1), 144–165. McCullough, J. S., S. T. Parente, and R. Town (2016). Health information technology and patient outcomes: The role of information and labor coordination. RAND Journal of Economics 47 (1), 207–236. McElheran, K. (2014). Delegation in multi-establishment firms: Evidence from I.T. purchasing. Journal of Economics & Management Strategy 23 (2), 225–258. McElheran, K. and W. Jin (2017). Economies before scale: I.t. investment and performance in young firms. Working paper, University of Toronto. Mcknight, L. W. and J. P. Bailey (1998). Internet Economics. MIT Press. McLuhan, M. (1964). Understanding Media: The Extensions of Man. New York: McGrawHill. Melnik, M. I. and J. Alm (2002). Does a sellers ecommerce reputation matter? evidence from ebay auctions. The Journal of Industrial Economics 50 (3), 337–349. Miller, A. and C. Tucker (2009, July). Privacy protection and technology adoption: The case of electronic medical records. Management Science 55 (7), 1077–1093. Miller, A. and C. Tucker (2011). Can healthcare information technology save babies? Journal of Political Economy (2), 289–324. 79 Miller, A. and C. Tucker (2014). Privacy protection, personalized medicine and genetic testing. Mimeo, MIT. Moore, T., R. Clayton, and R. Anderson (2009, September). The economics of online crime. Journal of Economic Perspectives 23 (3), 3–20. Mortimer, J. H., C. Nosko, and A. Sorensen (2012). Supply responses to digital distribution: Recorded music and live performances. Information Economics and Policy 24 (1), 3–14. Morton, F. S., F. Zettelmeyer, and J. Silva-Risso (2003). Consumer information and discrimination: Does the internet affect the pricing of new cars to women and minorities? Quantitative Marketing and Economics 1 (1), 65–92. Moshary, S., T. Blake, K. Sweeney, and S. Tadelis (2017, April). Price salience and product choice. Mustonen, M. (2005). When does a firm support substitute open source programming? Journal of Economics & Management Strategy 14 (1), 121–139. Nagaraj, A. (2016, July). The private impact of public maps: Landsat satellite imagery and gold exploration. Technical report. Nagaraj, A. (2017). Does copyright affect reuse? evidence from the google books digitization project. Management Science, forthcoming. Nocke, V., M. Peitz, and K. Stahl (2007). Platform ownership. Journal of the European Economic Association 5 (6), 1130–1160. Nosko, C. and S. Tadelis (2017). The limits of reputation in platform markets: An empirical analysis and field experiment. Working Paper, University of California Berkeley. 80 Oberholzer-Gee, F. and K. Strumpf (2007). The effect of file sharing on record sales: An empirical analysis. Journal of Political Economy 115, 1–42. Ockenfels, A., D. Reiley, and A. Sadrieh (2006, December). Online auctions. Orlov, E. (2011). How does the internet influence price dispersion? evidence from the airline industry. Journal of Industrial Economics 59 (1), 21–37. Overman, H. G., S. Redding, and A. J. Venables (2003). Handbook of International Trade, Chapter The Economic Geography of Trade, Production, and Income: A Survey of Empirics, pp. 353387. Blackwell Publishing. Parker, C., K. Ramdas, and N. Savva (2016). Is it enough? evidence from a natural experiment in indias agriculture markets. Management Science 62 (9), 2481 – 2503. Peitz, M. and P. Waelbroeck (2006). Why the music industry may gain from free downloading – the role of sampling. International Journal of Industrial Organization 24 (5), 907–913. Peitz, M. and J. Waldfogel (2012). Oxford Digital Handbooks, Chapter Digital Piracy: Empirics. Oxford University Press. Petty, R. D. (2000). Marketing without consent: Consumer choice and costs, privacy, and public policy. Journal of Public Policy & Marketing 19 (1), 42–53. Peukert, C., J. Claussen, and T. Kretschmer (2017). Piracy and box office movie revenues: Evidence from megaupload. International Journal of Industrial Organization 52, 188 – 215. Pope, D. G. and J. R. Sydnor (2011). Whats in a picture?: Evidence of discrimination from prosper.com. Journal of Human Resources 46 (1), 53–92. 81 Pozzi, A. (2013). E-commerce as a stockpiling technology: Implications for consumer savings. International Journal of Industrial Organization 31 (6), 677 – 689. The Economics of Information and Communication Technologies (ICT). Prince, J. (2007). The beginning of online/retail competition and its origins: An application to personal computers. International Journal of Industrial Organization 25 (1), 139–156. Quan, T. W. and K. R. Williams (2017). Product variety, across-market demand heterogeneity, and the value of online retail. Working paper, Yale University. Rao, A. (2015). Online content pricing: Purchase and rental markets. Marketing Science 34 (3), 430–451. Rao, J. M. and D. H. Reiley (2012, April). The economics of spam. Journal of Economic Perspectives 26 (3), 87–110. Reimers, I. (2016). Can private copyright protection be effective? evidence from book publishing. The Journal of Law and Economics 59 (2), 411–440. Reis, F., P. Ferreira, M. G. de Matos, and M. Smith (2015). Patterns of substitution between internet and television in the era of media streaming-evidence from a randomized experiment. Technical report, Working paper. Carnegie Mellon University. Resnick, P. and R. Zeckhauser (2002). Trust Among Strangers in Internet Transactions: Empirical Analysis form eBay Auctions, pp. 667–719. Amsterdam: Elsevier Science. Rob, R. and J. Waldfogel (2006). Piracy on the high c’s: Music downloading, sales displacement, and social welfare in a sample of college students. Journal of Law and Economics 49 (1), 29–62. Rob, R. and J. Waldfogel (2007). Piracy on the silver screen\*. The Journal of Industrial Economics 55 (3), 379–395. 82 Rochet, J.-C. and J. Tirole (2003). Platform competition in two-sided markets. Journal of the European Economic Association 1 (4), 990–1029. Rosen, S. (1981). The economics of superstars. The American Economic Review 71 (5), 845–858. Rosenblat, T. and M. Mobius (2004). Getting closer or drifting apart? Quarterly Journal of Economics 119 (3), 971–1009. Roth, A. E. and A. Ockenfels (2002, September). Last-minute bidding and the rules for ending second-price auctions: Evidence from ebay and amazon auctions on the internet. American Economic Review 92 (4), 1093–1103. Rysman, M. (2007). An empirical analysis of payment card usage. The Journal of Industrial Economics 55 (1), 1–36. Rysman, M. and T. Simcoe (2008). Patents and the performance of voluntary standardsetting organizations. Management Science 54 (11), 1920–1934. Salop, S. (1979). Monopolistic competition with outside goods. Bell Journal of Economics 10, 141–146. Savage, S. J. and D. M. Waldman (2009). Ability, location and household demand for internet bandwidth. International Journal of Industrial Organization 27 (2), 166 – 174. Scott, S. L. and H. R. Varian (2015, April). Bayesian Variable Selection for Nowcasting Economic Time Series, pp. 119–135. University of Chicago Press. Seamans, R. and F. Zhu (2013). Responses to entry in multi-sided markets: The impact of craigslist on local newspapers. Management Science 60 (2), 476–493. 83 Shapiro, C. and H. R. Varian (1998). Information rules: A strategic guide to the network economy. Harvard Business Press. Shiller, B. and J. Waldfogel (2011). Music for a song: An empirical look at uniform pricing and its alternatives. The Journal of Industrial Economics 59 (4), 630–660. Shiller, B. and J. Waldfogel (2017). Will ad blocking break the internet? Working paper, University of Minnesota. Shin, J. and K. Sudhir (2010, July-August). A customer management dilemma: When is it profitable to reward one’s own customers? Marketing Science 21 (4), 671–89. Simcoe, T. (2012, February). Standard setting committees: Consensus governance for shared technology platforms. American Economic Review 102 (1), 305–36. Simonov, A., C. Nosko, and J. M. Rao (2017, April). Competition and crowd-out for brand keywords in sponsored search. Journal of Economic Literature. Sinai, T. and J. Waldfogel (2004). Geography and the internet: Is the internet a substitute or a complement for cities? Journal of Urban Economics 56 (1), 1–24. Smith, M. D., J. Bailey, and E. Brynjolfsson (2001). Understanding the Digital Economy: Data, Tools, and Research, Chapter Understanding Digital Markets: Review and Assessment, pp. 99–136. MIT Press. Smith, M. D. and E. Brynjolfsson (2001). Consumer decision-making at an internet shopbot: Brand still matters. The Journal of Industrial Economics 49 (4), 541–558. Soloveichik, R. (2010, May). Artistic originals as a capital asset. American Economic Review 100 (2), 110–14. Solow, R. M. (1987, July). We’d better watch out. The New York Times, 36. 84 Stanton, C. and C. Thomas (2016). Landing the first job: The value of intermediaries in online hiring. Review of Economic Studies 83 (2). Stigler, G. (1961). The economics of information. Journal of Political Economy 69. Stigler, G. J. (1963). Public regulation of the securities markets. The Journal of Business 37, 117–117. Sunstein, C. (2001). Echo Chambers: Bush v. Gore, Impeachement, and Beyond. Princeton University Press. Tadelis, S. (1999). What’s in a name? reputation as a tradeable asset. The American Economic Review 89 (3), 548–563. Takayama, L. (1994). The welfare implications of unauthorized reproduction of intellectual property in presence of demand network externalities. Journal of Industrial Economics 42 (2). Tambe, P. and L. M. Hitt (2012, April). Now it’s personal: Offshoring and the shifting skill composition of the u.s. information technology workforce. Management Science 58 (4), 678–695. Tambe, P., L. M. Hitt, and E. Brynjolfsson (2012, May). The extroverted firm: How external information practices affect innovation and productivity. Management Science 58 (5), 843– 59. Taylor, C. and L. Wagman (2014). Consumer privacy in oligopolistic markets: Winners, losers, and welfare. International Journal of Industrial Organization 34, 80 – 84. Taylor, C. R. (2004, Winter). Consumer privacy and the market for customer information. RAND Journal of Economics 35 (4), 631–650. 85 Tucker, C. (2012a, May). The economics of advertising and privacy. International Journal of Industrial Organization 30 (3), 326–329. Tucker, C. (2012b). Patent trolls and technology diffusion. In Standards, Patents and Innovations. National Bureau of Economic Research, Inc. Tucker, C. (2014, October). Social networks, personalized advertising, and privacy controls. Journal of Marketing Research 51 (5), 546–562. Tucker, C. (2015). Handbook of Media Economics Volume 1A, Chapter Privacy and the Internet. North Holland. Tucker, C. and S. Yu (2017). Technology and inequality: The role of mobile applications in the inequality of treatment of customer complaints. Mimeo, MIT. Tucker, C. and J. Zhang (2011). How does popularity information affect choices? a field experiment. Management Science 57 (5), 828–842. Van Reenen, J., N. Bloom, M. Draca, T. Kretschmer, R. Sadun, H. Overman, and M. Schankerman (2010). The economic impact of ict. Final report. Varian, H. (2007). Position auctions. International Journal of Industrial Organization 25 (6), 1163–1178. Varian, H., F. Wallenberg, and G. Woroch (2005). The demographics of the do-not-call list. IEEE Security and Privacy 3 (1), 34–39. Varian, H. R. (1980). A model of sales. The American Economic Review 70 (4), 651–659. Varian, H. R. (2005, June). Copying and copyright. Journal of Economic Perspectives 19 (2), 121–138. 86 Varian, H. R. (2010, May). Computer mediated transactions. American Economic Review 100 (2), 1–10. Villas-Boas, J. M. (2004). Price cycles in markets with customer recognition. The RAND Journal of Economics 35 (3), 486–501. Vogelsang, I. (2003, September). Price regulation of access to telecommunications networks. Journal of Economic Literature 41 (3), 830–862. Waldfogel, J. (2010). Music file sharing and sales displacement in the itunes era. Information Economics and Policy 22 (4), 306–314. Waldfogel, J. (2012). Copyright research in the digital age: Moving from piracy to the supply of new products. American Economic Review 102 (3), 337–342. Waldfogel, J. (2015). First degree price discrimination goes to school. Journal of Industrial Economics 63 (4), 569–597. Waldfogel, J. (2016). Cinematic explosion: New products, unpredictabilty and realized quality in the digital era. The Journal of Industrial Economics 64 (4), 755–772. Waldfogel, J. and L. Chen (2006). Does information undermine brand? Information intermediary use and preference for branded web retailers. Journal of Industrial Economics 54 (4), 425–449. Waldfogel, J. and I. Reimers (2015). Storming the gatekeepers: Digital disintermediation in the market for books. Information Economics and Policy 31 (C), 47–58. Wallsten, S. (2013, October). What are we not doing when we’re online. Working Paper 19549, National Bureau of Economic Research. 87 Weyl, E. G. (2010, September). A price theory of multi-sided platforms. American Economic Review 100 (4), 1642–72. White, A. (2013). Search engines: Left side quality versus right side profits. International Journal of Industrial Organization 31 (6), 690 – 701. The Economics of Information and Communication Technologies (ICT). Williams, H. L. (2013). Intellectual property rights and innovation: Evidence from the human genome. Journal of Political Economy 121 (1), 1–27. Xu, L., T. Nian, and L. Cabral (2016, April). What makes geeks tick? a study of stack overflow careers. Technical report. Yang, H. (2013). Targeted search and the long tail effect. The RAND Journal of Economics 44 (4), 733–756. Zentner, A. (2006). Measuring the effect of file sharing on music purchases. The Journal of Law & Economics 49 (1), 63–90. Zervas, G., D. Proserpio, and J. W. Byers (2016). The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. Journal of Marketing Research 0 (ja), null. Zettelmeyer, Florian, F. S. M. and J. Silva-Risso (2001). Internet car retailing. Journal of Industrial Economics 49 (4), 501–519. Zhang, J. and P. Liu (2012). Rational herding in microloan markets. Management Science 58 (5), 892–912. Zhang, L. (2016). Intellectual property strategy and the long tail: Evidence from the recorded music industry. Management Science 0 (0), null. 88 Zhang, X. M. and F. Zhu (2011, June). Group size and incentives to contribute: A natural experiment at chinese wikipedia. American Economic Review 101 (4), 1601–15. 89

NBER WORKING PAPER SERIES THE NEW COMPARATIVE ECONOMICS Simeon Djankov Edward L. Glaeser Rafael La Porta Florencio Lopez-de-Silanes Andrei Shleifer Working Paper 9608 http://www.nber.org/papers/w9608 NATIONAL BUREAU OF ECONOMIC RESEARCH 1050 Massachusetts Avenue Cambridge, MA 02138 April 2003 We are grateful to Daron Acemoglu, Olivier Blanchard, Rafael DiTella, Simon Johnson, Peter Murrell, Katharina Pistor, Richard Posner, Dani Rodrik, Gerard Roland, Lawrence Summers, and Daniel Treisman for helpful comments. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research. ©2003 by Simeon Djankov, Edward L. Glaeser, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. All rights reserved. Short sections of text not to exceed two paragraphs, may be quoted without explicit permission provided that full credit including ©notice, is given to the source. The New Comparative Economics Simeon Djankov, Edward L. Glaeser, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer NBER Working Paper No. 9608 April 2003 JEL No. H1, K1, P1, P14, P16, P37, P5, P51 ABSTRACT In recent years, comparative economics experienced a revival, with a new focus on comparing capitalist economies. The theme of the new research is that institutions exert a profound influence on economic development. We argue that, to understand capitalist institutions, one needs to understand the basic tradeoff between the costs of disorder and those of dictatorship. We then apply this logic to study the structure of efficient institutions, the consequences of colonial transplantation, and the politics of institutional choice. Simeon Djankov Edward L. Glaeser The World Bank Harvard University 1818 H Street, NW 315a Littauer Center Washington, DC 20433 Cambridge, MA 02138 Sdjankov@worldbank.org and NBER Eglaeser@harvard.edu Rafael La Porta Florencio Lopez-de-Silanes Harvard University Yale School of Management M7 Littauer Center 135 Prospect Street Cambridge, MA 02138 New Haven, CT 06520 and NBER and NBER Rafael\_Laporta@harvard.edu florencio.lopezdesilanes@yale.edu Andrei Shleifer Harvard University M9 Littauer Center Cambridge, MA 02138 and NBER ashleifer@harvard.edu 2 This field has its own category in the Journal of Economic Literature, called Economic Systems. The subcategories are capitalist systems, socialist systems, socialist institutions, other economic systems and comparative economic systems. 1 I. Introduction. The traditional field of comparative economics deals mostly with the comparison of socialism and capitalism2 . Under socialism, the principal mechanism of resource allocation is central planning. Under capitalism, this mechanism is the market. Comparative economics, which dates back at least to the discussions of market socialism in the 1930s, asks under what circumstances either the plan or the market delivers greater economic efficiency and equality. By the time socialism collapsed in Eastern Europe and the Soviet Union, this question lost much of its appeal. It was clear that socialism produced misery and inefficiency – not to mention mass murder by the communist dictators who practiced it. Capitalism, in contrast, produced growth and wealth. With capitalism triumphant, is comparative economics dead? The answer, we argue in this paper, is NO. Traditional comparative economics has evolved into a new field. This field shares with its predecessor the notion that by comparing alternative economic systems, we can understand better what makes each of them work. But it sees the key comparisons as being those of alternative capitalist models that prevail in different countries. Each capitalist economy has many public and private institutions. These institutions function to choose political leaders, to secure property rights, to redistribute wealth, to resolve disputes, to govern firms, to allocate credit, and so on. Political economy over the last two centuries, as well as recent empirical research, demonstrate that these institutions differ tremendously and systematically among countries, with significant consequences for economic performance. The analysis of these differences is the subject of the new comparative economics. 2 In thinking about these issues, it is best to start from first principles. Since the days of the Enlightenment, economists agreed that good economic institutions must secure property rights, enabling people to keep the returns on their investment, make contracts, and resolve disputes. Such security encourages people to invest in themselves and in physical capital, and thereby fosters economic growth. As Smith [1776] wrote, “in all countries where there is tolerable security [of property], every man of common understanding will endeavor to employ whatever [capital] stock he can command... In those unfortunate countries... where men are continually afraid of the violence of their superiors, they frequently bury and conceal a great part of their [capital] stock ... in case of their being threatened with any of those disasters to which they consider themselves as at all times exposed.” But there are two sides to the security of property rights. On the one hand, investment must be secured – typically by the government – from the expropriation by one’s neighbors: thieves, competitors, or tort-feasors. Hobbes (1651) considered controlling private disorder – the war of all against all – to be the central concern of the state. The problems of disorder – crime, ethnic violence, squatter takings, bribery, investor expropriation, and so on – continue to plague modern developing countries. On the other hand, a government capable of protecting property against private infringement can itself become the violator and thief. Smith refers to the violence from “superiors;” Montesquieu (1748) is even more explicit: “Great enterprises in commerce are not found in monarchical, but republican governments... An opinion of greater certainty as to the possession of property in these [republican] states makes [merchants] undertake everything... Thinking themselves sure of what they have already acquired, they boldly expose it in order to acquire more... A general rule: A nation in slavery labors more to preserve than to acquire; a free nation, more to 3 acquire than to preserve.” In Enlightenment thinking, a crucial aspect of securing property rights is the control of dictatorship. A fundamental problem of institutional design is the conflict between the twin goals of controlling disorder and dictatorship. Hobbes, most fearful of disorder, favored absolutism, but subsequent writers recognized that extreme solutions are generally suboptimal. The framers of the U.S. constitution realized that dealing with disorder and localism through a more powerful central government directly conflicts with the objective of restraining the sovereign (Hamilton, Madison, and Jay 1788). This concern with institutional design continues in modern writing, most importantly in institutional economics (Hayek 1960, Olson 1965, 2000, Demsetz 1967, North 1981, 1990, North and Weingast 1989) and public choice (Buchanan and Tullock 1962). In addition, empirical studies confirm the close relationship between good institutions and economic development (De Soto 1989, De Long and Shleifer 1993, Besley 1995, Easterly and Levine 1997, 2003, Knack and Keefer 1995, Acemoglu et al. 2001, Rodrik et al. 2002). The interest in institutions revived with the collapse of socialism, and the transition of the economies in Eastern Europe, the former Soviet Union, and China to capitalism. The transition experience has been diverse. Many countries of Eastern and Central Europe, especially Poland, Slovenia, Hungary, and the Czech Republic, successfully established both secure democracies and many of the legal and regulatory institutions of capitalism during the 1990s. They grew rapidly, and are expected to fully integrate into Europe over the next several years. Countries further East, such as Romania and Russia, also moved to establish democracies and market institutions, but their experience has been more complex. Some of the Asian countries, from Kazakhstan to China, did not embrace democracy, but undertook significant economic reforms and grew, in China’s case 4 spectacularly. Finally, several transition economies, including Cuba, Belarus and many countries of Central Asia, did not reform and stagnated. Early discussions of the transition experiences focused on the speed of reforms – big bang versus gradualism – as a crucial determinant of performance. Although it is now clear that the absence of reform – as in Belarus and Cuba – is associated with both economic and political stagnation, the emphasis on speed turned out to be excessive. The important differences among countries had more to do with the effectiveness of the newly created institutions (Murrell 1995). The countries of Central Europe succeeded in creating successful institutions of both democracy and a market economy. Russia – having moved as fast or faster on many of its reforms – also established a successful democratic government, but faced greater problems of corruption and capture, and began growing only recently. These divergent experiences raised many questions. Some of them dealt with controlling disorder. How much government ownership is desirable? Poland, the Czech Republic, and Russia privatized extensively; China retained large state industries, yet grew the fastest. How much should governments regulate? Transition saw both successful regulation and degeneration of regulation into corruption and selective abuse of new business (Frye and Shleifer 1997, Hellman, Jones, and Kaufmann 2003, McKinsey 1999). How much should the government fight disorder at all? Many writers saw public institutions as having limited use, and stressed private orderings as the means of securing property and contract (McMillan and Woodruff 1999, Johnson, McMillan, and Woodruff 2002, Allen, Qian, and Qian 2003, Murrell 2003). Other questions focused on dictatorship. For starters, is democracy the best political system for economic reform or is dictatorship efficient when radical change is required? China’s economic 5 success under communism, contrasted with the difficulties of Yeltsin’s democracy in Russia, animated the advocates of one party rule; the successes of democracies of Central Europe pointed in the opposite direction. Within democracies, do reforms proceed better under divided or consolidated governments? Many economists assumed that consolidated government is better for reforms, yet here again, the deeply divided governments of Central Europe had the most success (Hellman 1998). Is a federal structure desirable from the viewpoint of economic transformation? Scholars of China credited its federalism and resulting competition among regions with the success of reforms (Jin and Qian 1998, Roland 2000); scholars of Russia saw its federalism and the resulting conflict between the regions and the center as a key obstacle to stability (Shleifer and Treisman 2000, Blanchard and Shleifer 2001, Zhuravskaya 2000). What can one make of these questions? First, the standard economic questions of market failure, and inefficiencies associated with it, have played virtually no role in the central debates of transition. Rather, the central issues have all dealt with property rights: how these rights can best be secured against both public and private expropriation? Second, we see in all these questions a common tradeoff. On the one hand, there is the objective of controlling disorder, that pushes toward greater state intervention. On the other hand, there is the goal of controlling dictatorship, that pushes against state power. In the rest of the paper, we explore this tradeoff. To this end, we present a framework describing the tradeoff between dictatorship and disorder, and apply it to the problem of social control of business. We argue that the four common strategies of such control: private orderings, private litigation, regulation, and state ownership, can be thought of as points on the institutional possibility frontier, ranked in terms of increasing powers of the state. These strategies are associated with progressively diminishing social costs of disorder, 6 and progressively rising social costs of dictatorship. We use this framework to analyze efficient institutional choice, and to argue that it provides useful empirical predictions as well as guidance for reform. But the efficiency perspective is not the only way to think about institutional choice. A key reason for institutional inefficiency is the transplantation of institutions through conquest and colonization. Many countries have inherited their legal systems. An institution which in the origin country respects the delicate tradeoff between dictatorship and disorder may not, once transplanted, remain efficient. This view of colonial transplantation may shed light on the amazing consistency with which a given country regulates different activities, as well as on some institutional pathologies. In addition, as both the Marxist and the public choice literature have long recognized, governments choose policies and institutions to benefits themselves – to stay in power and to get rich. The politics of institutional choice may also explain the inefficiencies. We describe our framework in Section II and illustrate it in Section III. Section IV focuses on transplantation, and Section V on politics. Section VI concludes. II. Basic Framework. The two central dangers that any society faces are disorder and dictatorship. By disorder we mean the risk to individuals and their property of private expropriation in the form of murder, theft, violation of agreements, torts, monopoly pricing, and so on. Disorder, in this framework, is also reflected in private subversion of public institutions, such as courts, through bribes and threats, which allows private violators to escape penalties. By dictatorship we mean the risk to individuals and their property of expropriation by the state and its agents in the form of murder, taxation, 3 Our discussion of dictatorship and disorder is related to the literature on rent-seeking and corruption (Tullock 1967, Posner 1974, Shleifer and Vishny 1993, Ades and DiTella 1997). 7 violation of property, and so on. Dictatorship, in this framework, is also reflected in expropriation through – rather than just by – the state, as with the use of regulators to eliminate entry by competitors. Some phenomena, such as corruption, are reflections of both disorder and dictatorship: in so far as individuals pay bribes to avoid penalties for harmful conduct, corruption is a reflection of disorder, but in so far as officials create harmful rules to collect bribes from individuals seeking to get around them, corruption is a cost of dictatorship.3 Institutions function to control the twin dangers of dictatorship and disorder. We focus on a fundamental tradeoff inherent in such control: a state that has enough powers to control disorder also has enough for dictatorial abuse. Figure 1 – the basic building block of our analysis – depicts what we call the Institutional Possibility Frontier (IPF) for a society or a sector within it. On the x-axis are the social losses from a higher level of dictatorship (as opposed to gross amounts of such activities as taxation and government expropriation), measured relative to a world with perfect property rights. On the y-axis are the social losses from a higher level of disorder, again measured relative to a perfect property rights benchmark. We measure the costs of dictatorship and disorder in the same units of “social losses” to think about the tradeoff. The IPF reflects the institutional possibilities of the society: how much disorder can be reduced with an incremental increase in the power of the state. As in all standard neoclassical theory, the IPF is convex to the origin. In our framework, an institution – such as a legal or a regulatory system – is a point on the IPF. In thinking about institutions, economists usually distinguish between written rules and their 8 Independent judges Private orderings Regulatory state Socialism 45o Total loss minimization Social losses due to state expropriation (Dictatorship) Social losses due to private expropriation (Disorder) Figure 1: Institutional possibilities Institutional possibility frontier (IPF) enforcement. In our framework, this separation disappears: a rule comes with its own enforcement properties, reflected in the equilibrium degree of dictatorship and disorder arising when this rule is used. If Russia regulates its monopolies through an anti-trust agency, there will be a certain residual amount of market failure, the equilibrium disorder, given by the waste from actual monopolists escaping the law and exercising their market power, as well as from their bribing the regulators to let them do so. There will also be a certain amount of public abuse of the private sector, the equilibrium dictatorship, associated with monopolists using the regulator to restrict entry by competitors, as well as from the agency’s officials charging non-monopolists bribes in exchange for agreeing not to harass them. The downward sloping 45 degree line in Figure 1 holds constant the total social costs of dictatorship and disorder. Its point of tangency with the IPF is the efficient institutional choice for 4 Empirical growth studies generally confound the location of the IPF and the choice of a point on it under the rubric of “institutions,” but for us it is important to draw the distinction. In these empirical studies, “institutions” are measured with a variety of objective and subjective assessments of institutional quality of a country, such as law and order, risk of government expropriation, rule of law, corruption, efficiency of the judiciary, or some combination of these variables (Easterly and Levine 1997, 2003, Hall and Jones 1999, Kaufmann, Kraay, and ZoidoLobaton 2002). These variables are highly correlated with each other, and have proven to be strong predictors of per capita income, economic growth, and many other “good” outcomes, but it is not entirely clear what they measure conceptually. 9 a given society or a sector within a society. In much of our discussion, we focus on the efficient institutional choice for a given IPF. Efficient institutions could evolve from democratic pressures (Wittman 1989), from the influence of growth-seeking interest groups such as merchants (DeLong and Shleifer 1993), from a Coasian negotiation among the members of the elite, such as the Magna Carta or the American Constitutional bargain (Becker 1983), or from a long term evolutionary process described by Hayek (1960). Moreover, looking at efficient institutional choices does not mean that, in equilibrium, the society eliminates the problems of dictatorship and disorder. It does not. Consistent with Coase (1960), even the most efficient institutional structure retains residual levels of both dictatorship and disorder. The shape and the location of the IPF -- and hence the efficient choice -- varies across activities within a society, as well as across societies.4 An activity that involves repeated interactions among participants of roughly similar resources, and with little technological change, such as diamond trading, can achieve order with little dictatorship (Bernstein 1992). In contrast, an activity like security issuance, involving players with few repeated interactions and massive inequalities of power, is vulnerable to much more disorder for a given level of police. Looking across societies, the differences loom even larger. The institutional possibilities of modern Sweden, or even China, are far superior to those of Albania or Congo. Sweden can pursue 10 either extreme laissez-faire policies or toy with socialism and still achieve decent outcomes, Albania can choose a perfect balance of dictatorship and disorder for its IPF, but property rights would remain insecure. For lack of a better term, we refer to location of the IPF as “civic capital,” with the idea that societies with more such capital, and an IPF closer to the origin, are better capable of achieving cooperation among their members. We use the term “civic capital” rather than social capital because we have something similar but broader in mind. We stick with the idea of capital because investments in civic capital pay off in the medium run. Recent historical research has made progress in understanding the determinants of civic capital. Landes (1998) continues the great tradition of stressing the influence of culture. Easterly and Levine (1997) and Alesina, Baqir and Easterly (1999) show how ethnic heterogeneity, and the resulting ethnic strife, reduce institutional quality. Diamond (1997) and Engerman and Sokoloff (1997, 2002) stress the role of factor endowments and the physical environment in shaping – and limiting – the institutional opportunities of a society. For instance, the environment of Latin America, unlike that of North America, was most hospitable to large scale agricultural technologies that bred significant economic and social inequalities, causing a long-lasting damage to the regions’s institutions. Acemoglu, Johnson, and Robinson (2001) argue that the mortality of European settlers shaped settlement patterns and institutional outcomes. Putnam (1993) maintains that the long run history of cooperation in a community -- its social capital -- determines the ability of its members to cooperate in the production of public goods. But there are some other, more prosaic, determinants of civic capital. Technology of production matters. When the scale of production and the pace of interaction among individuals rise, the opportunities for private expropriation expand, moving the IPF out. The technology of 11 government repression – the efficiency of tax extraction and the monopoly on power – influence both the location and the shape of the IPF, because they determine how much dictatorship is needed to reduce disorder both on average and at the margin. Last but not least, the level of human capital in the society is itself likely to determine the location of the IPF, as better educated and informed people may be more likely to solve problems without violence. Admittedly, civic capital is a somewhat vague concept, but many aspects of it have been measured. We have already referred to the measurement of historical influences, which are predetermined for many policy choices. Other aspects of civic capital, such as human capital, technology, or inequality, can be changed in the medium run through public investments or large interventions, such as land reform. For many reforms that economists focus on, however, including social control of business, the location of the IPF is fixed in the short run. For the analysis of such reforms, civic capital is a constraint rather than a choice. To think about the IPF and institutional choices in more concrete terms, consider the problem of social control of business. Suppose that “the society” wishes to reduce disorder from monopoly pricing, torts, predatory tactics, etc. There are four distinct strategies of such control, involving ever growing powers of the officials vis-a-vis private individuals: market discipline, private legal action through courts, public enforcement through regulation, and state ownership. They are shown as points on the IPF in Figure 1. These four strategies are not mutually exclusive: competition and regulation often operate in the same market, as do private litigation and public regulation. Moreover, intermediate strategies of social control of business, such as private litigation to enforce public rules, are available. Nonetheless, these four categories provide a useful analytical classification. 12 To illustrate these categories, take the example of social control of securities issues. Suppose that the society – through its institutions – has an interest in having broad and liquid securities markets and, to this end, deems it desirable that firms issuing equity disclose accurate information about their circumstances. The society has four choices. First, it can rely on the incentives of issuers themselves, or of their underwriters, to disclose the truth about the securities because, to raise funds in the future, they need to establish a reputation for credibility. This is the market discipline solution. Second, the society can rely on private suits by buyers of securities who feel that they have been cheated by the issuers, under the general doctrines of contract or tort. For this, the society needs a court and a judge. The question for the court is whether the issuer disclosed inaccurate information or, alternatively, negligently failed to provide material information. Third, the society can designate a public regulatory agency, which mandates what should be disclosed by security issuers, inspects their books and disclosures, and penalizes issuers and underwriters who break the regulations. Between private litigation and full-scale regulation, the regulatory agency can establish the rules for security issuance, but leave the enforcement of these rules to private litigation by the wronged investors. Fourth, the society can nationalize security issuance. A company wishing to raise capital would turn over the inspection, disclosure, and sale of securities to the state. These are the four basic institutional strategies for the enforcement of good conduct. These four basic strategies differ in the degree of public control. With competition and private orderings, there is basically no public involvement. With courts, there is a role for impartial judges enforcing the rules of good behavior. These rules do not even need to come from legislation, 13 but may instead derive from custom or from judge-made common law. Even so, there is a public agent – the judge – who has at least some decision-making authority. With regulators, control by the state rises sharply. The state now writes the rules, inspects the product before it is sold, and possibly penalizes sellers for delivering a bad product. Both the scope of government activity, and its centralization, are greatly increased relative to the judges. Finally, with state ownership, government takes complete control over an activity. From the perspective of the tradeoff between dictatorship and disorder, the principal strength of market discipline as a method of enforcing good conduct is that it is free of public enforcers. There is no possibility of politicization of rules of conduct, of corruption, of costly and delayed enforcement of rules, of random or compromised choice of one competitor over another. But market discipline often fails to control disorder. Market participants use their economic, political, or social resources to damage their customers and rivals, relying on methods ranging from deception to predatory pricing to monopoly pricing to social exclusion to outright theft or violence. One man’s peaceful private orderings become another man’s death in the hands of the mafia. When market discipline can successfully control disorder and avoid Hobbesian anarchy, it works best because it minimizes the social costs of dictatorship. As a corollary, any case for public intervention relies crucially on the presumptive failure of market discipline to control disorder. This case for the effectiveness of private orderings and market discipline is often compelling. Neighbors resolve disputes among themselves, without any government intervention, because they need to get along with each other over long stretches of time (Ellickson 1991). Industries form associations that assure quality for consumers, and penalize cheaters among themselves to assure that consumers continue to patronize the industry (Greif 1989, 1993, 1994, Bernstein 1992). 14 Families, cities, and ethnic groups establish reputations in the marketplace, and penalize reputationthreatening misconduct by their members. To the extent that market discipline can control disorder, regulation, or even courts, are unnecessary. But this is far from always being the case. Market pressures may not get rid of all monopolies. Employers may under-invest in safety and then blame accidents on an injured worker’s own carelessness (Fishback and Kantor 2000). A fraudulent stock issue can separate investors from their money very quickly, undermine confidence in markets, and run off with the cash. In these instances, to control disorder, societies may efficiently accept a higher level of government intervention and dictatorship. The traditional libertarian response to these problems is to move one notch toward more dictatorship by turning to the enforcement of good conduct through private litigation over contracts and torts. Injured employees can sue their employers for harm. Investors can sue issuers and underwriters for damages when they believe that representations about securities were false or incomplete. Ideally, a judge would recognize quickly whether investors have been misled, and compel the issuer to compensate investors for their losses. Such judicial enforcement of contracts and torts is seen by libertarians as a sufficient guarantee of security of property. Private litigation has many advantages. In principle, such litigation is of no special interest to the government, and hence disputes can be resolved apolitically, with no favors to influential parties. Judges may also develop expertise in contract enforcement (as well as in handling tort cases), and hence address problems efficiently and expeditiously. This is what Coase (1960), Nozick (1974), and Posner (1985) have in mind in making the case for courts. The reality of litigation is, unfortunately, not so perfect, and the tradeoff between dictatorship 15 and disorder is helpful for thinking about courts as well. As with private orderings, the powerful and not the just often get their way in court (Galanter 1974). Some mechanisms of influencing courts, such as hiring good lawyers, delaying proceedings, and (in some instances) seeking political help, are entirely legal. But judges are also bribed with cash, benefits, or promises of promotion, as well as threatened with violence if they do not favor the strong (Dal Bó, Dal Bó, and Di Tella 2002). When the rich and the politically connected influence the path of justice, litigation over contracts and torts cannot be counted on for enforcing socially desirable conduct (Glaeser, Scheinkman, and Shleifer 2003). A common way of protecting judges from influence is to formalize laws and procedures through codes, so as to minimize judicial discretion and the potential for subversion. Such strategies can be thought of as a move along the IPF, as they reduce disorder but also offer the state more control over the outcomes of litigation. A related mechanism for controlling private subversion of courts is to make judges employees of the state, whose career concerns protect them from succumbing to outside influence. But as judges become more dependent on the state, the risk of politicization of their decisions rises, and so do the social losses from dictatorship. This brings us to the third strategy of enforcing good conduct, government regulation. Regulation has been an anathema to libertarians such as Nozick (1974), who see a sharp contrast between the enforcement of rules by judges and that by regulators. But in fact, the change to a higher level of dictatorship is incremental, and there is no profound conceptual distinction between litigation and regulation. The libertarian distinctions miss Coase’s (1960) realization that the costs of enforcement shape the optimal institutional choice. To make this continuity clear, note that a step short of full-fledged public enforcement, there 5 La Porta, Lopez-de-Silanes, and Shleifer (2002b) show that this is a key strategy for enforcing good conduct in security issuance in many countries. Barth, Caprio, and Levine (2003) show that private enforcement of public disclosure rules is a key strategy of bank supervision. 16 is an important intermediate strategy, namely private enforcement of public rules. The government can create rules governing private conduct and then leave their enforcement to private litigation. Private enforcement of such specific statutes through litigation is often considerably cheaper than that of contracts or torts. It may be efficient, for example, for the government to specify appropriate safety standards but to let workers or consumers sue when they feel the standards are violated. Likewise, the government can mandate specific disclosures by a company issuing shares, but then leave litigation to investors. It may be cheaper for investors to establish in a trial that the company has failed to reveal specific information whose disclosure was mandated by law, than to prove the issuer’s negligence in the absence of a statute. Private enforcement of public statutes addresses a number of problems of disorder inherent in pure litigation. First, the burdens on the courts and the plaintiffs of proving liability (or the lack thereof) fall considerably when the statutes describe precisely what facts need to be established to do so. Second, subversion of judges becomes more difficult when they lose discretion. It may be relatively easy to convince a judge – by persuasion or bribery – that a security issuer who concealed important information from investors is not liable when there are no specific rules as to what needs to be disclosed. It is much harder to convince the same judge when the law states specifically what must be disclosed. Perhaps for these reasons, private enforcement of public rules is a highly efficient strategy of enforcing good conduct in some situations (Black and Kraakman 1996, Hay and Shleifer 1996, Hay, Shleifer and Vishny 1998).5 17 At the same time, the creation of public rules – even rules that are enforced privately – raises the risks of dictatorship. Such rules can be used to expropriate politically weak and to favor the politically strong. Mandatory safety precautions in factories, mines, and meat-packing plants during the progressive era in the U.S. at the beginning of the 20th century are sometimes interpreted as an attempt by large established firms to restrict entry by smaller rivals by raising these rivals’ costs of regulatory compliance (Libecap 1992, Coppin and High 1999). Compared to the enforcement strategies described above, public regulation has a number of advantages in controlling disorder. First, unlike judges, public regulators can be expert and motivated to pursue social objectives in specific areas. This, indeed, has been the principal argument for public regulation of securities markets (Landis 1938, Johnson, Glaeser, and Shleifer 2001). A regulator can become expert, for example, in what constitutes material omission from a prospectus, present market participants with specific rules, and then enforce them by imposing its own sanctions or by convincing courts to adopt its rules. Second, to the extent that regulators are empowered to complete the law, they may have an enforcement advantage relative to judges by acting preemptively (Pistor and Xu 2002). Third, because regulators can be incentivized by the sovereign to enforce social policy, they can in principle be much more difficult to subvert than the disinterested judges with either persuasion or bribes (Glaeser and Shleifer 2003). These differences render public enforcement more efficient than private enforcement in some cases. Alas, public regulation obviously has problems, the key one being public abuse of market participants by the officials who are either pursuing their own political interests or are captured by a particular group, including the regulated industry itself (Stigler 1971). Although motivated regulators might be more difficult to subvert than judges, regulated industries have developed a 18 range of techniques to turn regulation into a mechanism of protecting industry rents rather than public welfare. The risks from dictatorship clearly rise as those from disorder decline. The basic implication of our theory is that regulation is only necessary when the level of disorder is too high for private orderings and even courts to deal with successfully. The case for regulation is most compelling when the problem of inequality of weapons between private parties involved in a transaction is severe. Securities issuance is one instance; workplace safety is perhaps another. In contrast, entry of new firms into competitive markets is unlikely to require regulatory control (De Soto 1989, Djankov et al. 2002). Most regulations of competitive labor markets are likewise difficult to justify on efficiency grounds (Botero et al. 2003). Finally, in some situations, nothing short of government ownership can eliminate disorder. If monopolies cannot be tamed by competition or regulation, if quality is essential but cannot be assured except with full state control, if public safety is jeopardized – then there can be a plausible case for state ownership. Hart, Shleifer and Vishny (1997) argue that prisons might be properly publicly- rather than privately-owned because the risk that private jailers mistreat inmates is too high. This is so because inmates have few legal rights and cannot count on the market, the courts, or even the regulators to protect them. Likewise, the military and the police tend to be statecontrolled because the likelihood of disorder from private control is too high. The case for the “state monopoly on arms” is just a reflection of a particular area where the tension between dictatorship and disorder is resolved by going to an extreme. Although in some instances the case for government ownership as a means of dealing with disorder is compelling, state ownership has the obvious problems of dictatorship, illustrated by the miserable performance record of public enterprises and the benefits of privatization (Lopez-de- 19 Silanes 1997, La Porta and Lopez-de-Silanes 1999, Megginson and Netter 2001, Djankov and Murrell 2002). The failure of state socialism as an economic system reveals most dramatically the consequences of dictatorship taken to an extreme, in which all economic problems are solved to maintain political control by the communist party (Kornai 1992). In summary, the framework presented in Figure 1 enables us to discuss systematically the alternative forms of social control of business. As a consequence, it may provide some useful input into thinking about efficient institutional choices. In the next section, we put the model to work and examine three episodes of institutional design. In the following two sections, we move away the assumption of efficiency, and examine alternative views of institutional choice. III. Applications. We examine three applications of the basic framework: the divergence between France and England in the 12th and 13th centuries in their choices of legal systems, the rise of the regulatory state in the progressive era U.S., and post-communist transition. In all three instances, we focus on efficient institutional choice given the institutional possibilities of a country. Legal Origins In the 12th and 13th centuries, England and France established the foundations of their modern legal systems. These systems took rather different forms. England developed a system of common law, characterized by fact finding by juries, relatively independent judges, infrequent appeals, and the reliance on precedents and judge-made law rather than strict codes in reaching judicial decisions. France, following the Roman tradition, adopted a system of civil law, characterized by fact finding 20 by state-employed judges, automatic superior review of decisions, and later the reliance on procedural and substantive codes rather than judicial discretion. Over the centuries, some of these differences have been emasculated, others reinforced, but the basic differences among the two legal systems survive to this day. Why did the two systems diverge? Glaeser and Shleifer (2002) argue that a crucial feature of any legal system is its vulnerability to subversion by the powerful (disorder). The greater are the pressures on courts to rule for the strong rather than the just, the more centralization (dictatorship) is needed to counter these pressures. On the other hand, greater centralization raises the cost of dictatorship: sovereigns rule according to their personal preferences and biases rather than the community standards or laws that the jurors follow. In the 12th and 13th centuries, France was relatively decentralized and disorderly, with local notables successfully able to subvert all local institutions to their own advantage. England, in contrast, was relatively peaceful and the king maintained control over the entire country. To counter disorder, it was efficient for France to adopt a legal system with higher dictatorship than England, even at the cost of greater scope for sovereign abuse of the law. Figure 2 illustrates the relevant choices, where we think of the French IPF as a shift out of the English IPF, but also as a tilt that made marginal increases in dictatorship highly productive in reducing disorder. The efficient choice for France was a legal system with a higher level of dictatorship. In equilibrium, France had a higher level of disorder as well, consistent with the notion that its institutional possibilities were less attractive. Glaeser and Shleifer (2002) further show that state employed judges, reluctance to rely on independent jurors, codification of rules and procedures, and nearly-automatic appeals are all complementary aspects of this choice of greater centralization 21 France Dictatorship Disorder Figure 2: Legal origins England Common law Civil law and dictatorship to counter the problems of disorder. The Rise of the Regulatory State Before 1900, significant commercial disputes in the United States were generally resolved through private litigation over contracts and torts. Courts ruled on corporate liability in industrial accidents, on anti-competitive practices such as railroad rebates, on safety of foods and medicines, and even on the constitutionality of the income tax. Between 1887, when Congress passed the Interstate Commerce Act, and 1917, when participation in the war put an end to the progressive movement, this situation changed radically. Over thirty years, reformers eroded 19th century belief that private litigation was the sole appropriate response to social wrongs. During the progressive era, regulatory agencies at both the state and the federal level took over the social control of 22 competition, anti-trust policy, railroad pricing, food and drug safety, workplace safety, and many other areas. The regulatory state was born in the United States. Glaeser and Shleifer (2003) interpret this experience using a model intimately related to the framework of this paper. They argue that the pre-Civil-War United States was a relatively stable country, without great inequality among the potential litigants, so private litigation was an efficient strategy of social control of business (see Figure 3). But massive industrialization and commercialization of the American economy after the Civil War severely undermined courts as the sole institution securing property rights. The rise of railroads and large firms greatly increased disorder: these firms maimed passengers and workers, destroyed their competitors through aggressive and possibly wasteful tactics, occasionally poisoned and deceived customers, and so on. The growth in disorder resulting from the greater scale of enterprise and rising inequality of wealth shifted the IPF of the economy out, rendering the existing system of private dispute resolution inefficient. The robber barons commanded economic and political resources that overwhelmed consumers, workers, or competitors who complained in court. They did so both legally – by hiring superior lawyers, and illegally – by bribing judges and legislators. Figure 3 illustrates the changes in the U.S. economy during the post-civil-war era, as well as the regulatory reform. Technological change shifted the IPF out, but also twisted it so that the reduction in disorder from a marginal rise in dictatorship increased. When courts were subverted but regulations were new and relatively benign, marginal regulation was highly productive. As Figure 3 shows, the efficient institutional choice in the American economy in response to this shift in the IPF called for more regulation to provide the countervailing power to big business. 23 Post-gilded age Dictatorship Disorder Figure 3: Progressive reforms Pre-gilded age Glaeser and Shleifer (2003) interpret the rise of regulation, of litigation over statutes rather than contracts and torts, as well as many other reforms in the U.S. economy during the progressive era, as precisely such a shift toward the newly-efficient system of social control of business. In equilibrium, disorder may well have increased relative to the mid-19th century – this by itself does not imply that the progressive reforms were ineffective. Dictatorship, as noted by Stigler (1971) and his followers, increased as well. The greater power of the state has led to greater benefits of state capture, and more such capture in equilibrium. Still, the reforms have moved the U.S. economy to a better point than it was at in 1880, as much of the evidence on economic and social progress during the reform period indicates. 24 Institutions in Transition At least some aspects of the transition of Eastern Europe and the Former Soviet Union from socialism to capitalism can be interpreted using our framework. The crucial – though often insufficiently emphasized – aspect of the transition is the collapse of communism, and therefore a sharp decline in dictatorship through most of the region. Figure 4 presents our view of Russia, East European countries, and the FSU non-reformers (with Belarus taken as an example), in the first few years of transition. Two crucial factors separate Russia from FSU non-reformers on the one hand, and from East European reformers on the other. First, Russia has experienced a much more dramatic decline in dictatorship and, consistent with our model, a rise in disorder, than countries like Belarus and Uzbekistan. Second, Russia’s IPF is probably less attractive than that of the East European countries and, at the same time, its shift along the IPF was probably greater. According to Kornai (1992), the communist party is the critical mechanism of keeping economic, political, and social order in socialist countries. Such police order comes at the expense of personal, economic, and political freedom, and as such has enormous human costs, but it is order nonetheless. As the communist control of the Russian economy and society was eliminated by Gorbachev and Yeltsin, so were the mechanisms of keeping order. The immediate benefits of disorder – the free press, the growth of entrepreneurship, the tremendous rise of ties to the rest of the world – were apparent, but so were the costs. The initial impact of the fall of economic dictatorship was extreme economic disorganization (Murrell 1992, Murphy, Shleifer and Vishny 1992, Blanchard and Kremer 1997), but even as markets began to work, the lack of law and order was manifest. The unofficial economy reached perhaps as much as 40% of the total (Johnson, Kaufmann, and Shleifer 1997). The nascent public institutions were subverted by the powerful 25 Russia, 1990 Belarus, 1990-95 Russia, 1995 Dictatorship Disorder Figure 4: Transition Eastern Europe, 1995 Eastern Europe, 1990 through crime, corruption and political influence (Shleifer 1997, Hellman, Jones, and Kaufmann 2003, Sonin 2003). Most importantly, the U.S.S.R. disintegrated and Russia itself – like the U.S. prior to the adoption of its constitution and Argentina today – experienced severe problems of localism, with regional governments undermining both national finances and law and order (Shleifer and Treisman 2000, Zhuravskaya 2000). Russia’s transition can be plausibly described as a sharp move up along its IPF. This experience can be usefully contrasted with that of non-reforming states, such as Belarus and Uzbekistan, where communist dictatorship was replaced by personal dictatorship, but many of the controlling institutions of the communist state remained intact. These countries still utterly lack freedom, but they also did not suffer as sharp a rise in disorder. Tellingly, Belarus and Uzbekistan retained extremely small unofficial economies, a clear indicator of pervasive state control (Johnson, 26 Kaufmann, and Shleifer 1997). They also did not experience as large a decline in officially measured output, again consistent with preserving order. On the other hand, these countries failed to reap the benefits of economic transition from communism, and over the decade performed very poorly. Compared to Russia, Eastern Europe had a more attractive IPF. East European countries lived through a shorter spell of communism, and have, in the cases of the Czech Republic, Hungary, and Poland, rebelled against their Soviet occupiers and local dictators. They had many more independent organizations and, in the case of Poland, an independent Church. They were also more integrated into Western Europe, and from the beginning anxious to join the European community, which imposed rules of cooperation restricting dictatorship and disorder. Compared to Russia (or Yugoslavia), East European countries also did not move as far up along their IPFs, perhaps because they did not face as severe problems of decentralization and localism. A plausible case can be made that Russia in the 1990s stayed on its institutional possibility frontier. It has moved away from the communist dictatorship, but arguably overshot initially in its institutional change toward too much disorder. Under the Putin presidency, Russia has moved down along its IPF toward reducing disorder, even at the cost of some growth in dictatorship. The more general point suggested by this reasoning is that economic and social change in each country should be considered in light of its own institutional possibilities, rather than some idealized view of perfect law and order. IV. Transplantation. Although one can explain at least some institutional diversity by focusing on efficient 27 choices, these are clearly not the whole story. One dramatic deviation from the assumption of indigenous formation of institutions is transplantation. As European powers conquered much of the world in the 19th century, they brought with them their institutions, including their laws. A significant portion of institutional variation among countries, especially in legal systems, is accounted for by transplantation (Watson 1974, La Porta et al 1997, 1998, Berkowitz et al 2002). When the English, the French, the Spaniards, the Dutch, the Germans, and the Portuguese colonized the world, they brought with them many institutions, including language and sports. As we discussed above, there is systematic variation among these institutions in origin countries, shaped by their political economy over the last millennium. England developed a common law tradition, characterized by the independent judges and juries, relatively weaker reliance on statutes, and the preference for contracts and private litigation as a means of dealing with social harms. France, in contrast, developed a civil law tradition, characterized by state-employed judges, great reliance on legal and procedural codes, and a preference for state regulation over private litigation. Germany developed its own civil law tradition, also based in Roman law. There is also a distinctive legal tradition of Scandinavian countries. Finally, and crucially for the 20th century, the U.S.S.R. developed a system of socialist law. Napoleon exported the French legal system during his conquests to Spain, Portugal, and Holland, and through his and their colonial conquests, it was transplanted to all of Latin America, large parts of Europe, North and West Africa, parts of the Carribean, and parts of Asia. The common law tradition was transplanted by England to the U.S., Canada, Australia, New Zealand, East Africa, large parts of Asia (including India), and parts of the Carribean. The German legal system was voluntarily adopted by Switzerland, Austria-Hungary, and later Japan, and through 28 Japan it influenced legal systems of Korea, Taiwan, and China. The U.S.S.R. transplanted its legal system to socialist countries. These channels of both voluntary and colonial transplantation suggest that there might be systematic variation in legal systems among countries. Some of the initial evidence on transplantation of legal systems, assembled by La Porta et al. (1997, 1998), strongly supports this hypothesis. The authors look at the laws governing the protection of investors – shareholders and creditors – from expropriation by corporate insiders using a sample of 49 countries around the world. They find that, generally speaking, common law countries protect shareholders better than do civil law countries, and especially French civil law countries. They also find that common and German civil law countries protect creditors better than do the French civil law countries. La Porta et al. (1997, 1998) and subsequent studies also show that both legal origin and investor protection are strongly correlated with various aspects of financial development (Demirguc-Kunt and Levine 2001, Johnson et al. 2000, Wurgler 2000, Beck et al. 2003a,b, Friedman, Johnson, and Mitton 2003, La Porta et al. 2000, 2002). Djankov, La Porta, Lopez-de-Silanes, and Shleifer (2002, 2003) and Botero et al. (2003) take these cross-country comparisons further by examining various domains of government regulation, including entry by new firms, judicial procedures in courts, and labor markets. Although the data for each of these studies were collected using different procedures and they cover somewhat different samples of countries, some systematic patterns emerge. Countries appear to have pronounced styles of social control of business intimately related to the legal origin of their laws. In all three areas – entry, courts, and labor – socialist and French legal origin countries regulate much more heavily than do the common law countries. On average, the very same countries that regulate entry, also regulate courts and labor markets, and that these correlations are at least in part 29 driven by legal origin (see Table 1). Table I Correlations between regulation measures The table shows the pairwise correlations of regulation indices across 63 countries. No former or current communist countries are included to avoid spurious correlation. The measures of regulation are: (1) "Employment laws index" which measures the level of protection of labor and employment laws and is formed by the normalized sum of: (i) subindex: alternative employment contracts. (ii) subindex: conditions of employment. (iii) subindex: job security; (2) "Number of procedures to start a business;" which is defined as the number of different procedures that a start-up has to comply with in order to obtain a legal status, i.e., to start operating as a legal entity; and (3) "Court formalism index for the collection of a bounced check" which measures substantive and procedural statutory intervention in judicial cases at lower-level civil trial courts, and is formed by various characteristics in the process such as the need for lawyers and legal justifications in the process, written vs. oral elements, statutory regulation of evidence, control of superior review, engagement formalities for the parties, and the number of independent procedural actions. Employment laws index Number of procedures to start a business Number of procedures to start a business 0.6168a 1 Court formalism index for the collection of bounced check 0.6083a 0.5944a a=significant at 1 percent level; b=significant at 5 percent level; c=significant at 10 percent level. Sources: “The Regulation of Labor,” Harvard University manuscript, November 2002 (Botero, Djankov, La Porta, Lopez-de-Silanes and Shleifer); “The Regulation of Entry,” Quarterly Journal of Economics CXVII, February 2002, 1-37 (Djankov, La Porta, Lopez-de-Silanes and Shleifer); “Courts,” Quarterly Journal of Economics, forthcoming 2003, (Djankov, La Porta, Lopez-de-Silanes and Shleifer). The same rankings appear when we look at state ownership. Socialist and French legal origin countries have more government ownership of banks (La Porta, Lopez-de-Silanes, and Shleifer 2002a) and a greater role of state-owned enterprises in the economy (La Porta et al. 1999) than do common law countries. This evidence suggests that transplantation, rather than local conditions, exerts a profound effect on national modes of social control of business, including both state ownership and regulation. This evidence poses a sharp challenge to standard theories of regulation, which emphasize local industry conditions and the power of interest groups, as opposed 30 to broad national tendencies, to explain regulatory practice. The fact that colonial transplantation is such a significant determinant of institutional design suggests that the observed institutional choices may well be inefficient. A legal and regulatory system perfectly suitable for France might yield inefficiently high levels of regulation and state ownership when transplanted to countries with lower civic capital. Likewise, a system of independent courts that works in Australia or the U.S. might fail in Malaysia or Zimbabwe. Can we understand how colonial transplantation can go awry? Figure 5 applies our theoretical framework to transplantation. We think of a legal system as (roughly) limiting dictatorship and disorder in fixed proportions. For a given country, its legal system is the intersection of the ray emanating from the origin defining these proportions with its own IPF. The common law ray has a higher proportion of disorder to dictatorship than the civil law ray. Finally, as the IPFs shift out from developed to developing countries, the marginal amount of dictatorship required to reduce disorder rises: less developed countries (unlike the U.S. in 1900) cannot buy much order with regulation. As a consequence, less developed countries need relatively less dictatorship in equilibrium, i.e., less regulation. Figure 5 describes the transplantation of common and civil law into a country with an IPF further from the origin. Whichever system is transplanted, the equilibrium amounts of dictatorship and disorder rise simply because the institutional possibilities are less attractive. Also, relative to efficiency, the transplantation of both civil and common law will lead to excessive intervention and regulation in lower civic capital countries. Finally, transplantation leads to higher levels of overregulation in civil than in common law countries. Because of the enormous risks of public abuse of business, countries at lower levels of civic capital efficiently need less regulation. But when legal 31 Common law Dictatorship Disorder Figure 5: Transplantation of laws Developed country Less developed country Civil law systems are transplanted, they end with as much relatively speaking, and more in absolute terms, than do the origin countries. This problem is most severe in civil law countries, which end up being especially over-regulated relative to efficiency. Figure 5 is consistent with the findings of Djankov et al. (2002, 2003) and Botero et al. (2003) that levels of regulation are higher in poorer than in richer countries, and in civil law than in common law countries. Moreover, consistent with Figure 5, the evidence shows that, in poor countries, the levels of regulation are often excessive. Higher levels of regulation of entry are associated with larger unofficial economies and no measurable benefits for product quality (Djankov et al. 2002). Higher level of regulation of judicial procedure are associated with higher costs and longer delays, without offsetting benefits in terms of perceived justice (Djankov et al. 2003). Heavier regulation of labor is associated with larger unofficial economies, higher unemployment, 32 and lower labor force participation (Botero et al. 2003). In all these cases, the transplanted regulatory systems appear excessive for the countries that use them. The evidence on the role of legal origin points to some tangible ways in which the existing institutions fall short of their potential, as well as to some possible directions for reform. The evidence suggests that deregulation – particularly in areas such as entry and labor markets, where the forces of competition are potentially so effective – is a high level priority for poor countries. In these countries, regulation is nearly universally associated with poor outcomes because public officials abuse their powers over private agents. Deregulation is likely to diminish the problems of dictatorship without a significant increase in disorder. But the evidence also identifies some pitfalls of reform. One cannot assume, for example, that highly formalized general jurisdiction courts could efficiently resolve disputes in developing civil law countries. The most attractive areas for deregulation in developing economies are those where one can count on competition and market discipline, rather than on courts, to control disorder. In contrast, in the developed countries, courts – especially specialized courts – are becoming an increasingly attractive alternative to regulation. So far, we have focused on the content of transplantation, and more specifically on the transplantation of the legal/regulatory regimes. But obviously, other institutions are transplanted as well. Moreover, how they are transplanted may also matter. Berkowitz, Pistor, and Richard (2003) and Pistor et al. (2002, 2003) stress the importance of “receptivity” for the success of transplantation of legal systems. Acemoglu, Johnson, and Robinson (2001) argue that the quality of institutional transplantation depends on whether the colonizers themselves settled in the occupied land, as they did in the U.S. or New Zealand, or just set up trading posts and exploited the colonies, 33 as they did in most African states. All the evidence points to extraordinary importance of institutional transplantation. Legal origin as we discussed it proxies for the transplantation of institutions of social control of business: it captures the location of the transplanted institutions on the IPF, rather than the location of the IPF itself. As such, we expect legal origin to only affect growth indirectly, in so far as the specific mechanisms of social control affect markets for capital, labor, and entrepreneurs, thereby influence factor accumulation and productivity growth. Other aspects of colonization and transplantation may influence growth directly, by influencing the civic capital of the society, i.e., the location of the IPF. Of these, human capital of the people operating the institutions is probably most important. Understanding the diverse consequence of transplantation is a crucial challenge for the new comparative economics. V. Politics. Politics has a bad name in economics. From Marx (1872), to the Austrians (von Mises 1949, Hayek 1960), to institutional economists (Olson 1965, 1982, North 1990), to public choice scholars (Buchanan and Tullock 1962), to regulation economists (Stigler 1971), to political historians (Finer 1997), writers on institutions have maintained that political choice is often responsible for institutional inefficiency. Generals, dictators, ascendant social classes, democratic majorities, and favored interest groups all choose institutions that entrench them in power, so that they can collect political and economic rents. Constitutions, voting rules, federalist arrangements, organization of army and police are selected by incumbents to keep themselves in power. Olson (1993), Rajan and Zingales (2000), Acemoglu and Robinson (2000, 2002), Aghion, Alesina, and Trebbi (2002), 6 We do not discuss here a large literature on comparative political institutions. See, for example, Laffont (2000) and Persson and Tabellini (2000). 34 Glaeser and Shleifer (2002b), are some of the recent models examining these phenomena. In these models, there is no reason that the equilibrium institutional choice would end up on the IPF, let alone minimize total social losses, because Coasian negotiations between various players fail (Olson 2000, Acemoglu 2003).6 The political perspective predicts that equilibrium institutions can be either excessively disorderly or excessively dictatorial. Recent research on Yelstin’s Russia argues, for example, that the oligarchs who came to strongly influence the government (like the robber barons in the preprogressive-era U.S.) preferred institutional disorder as a strategy of maximizing their rents (Black and Tarassova 2003, Sonin 2003). More typically, as those in power attempt to stay in power, they help themselves and their supporters through excessive dictatorship. State ownership becomes a mechanism of dispensing patronage and maintaining political support for the incumbent politicians (Shleifer and Vishny 1994). Regulations with ostensibly benign goals end up protecting incumbent firms from competition, and offering extensive corruption and political support opportunities to their enforcers (Stigler 1971, De Soto 1989, Djankov et al. 2002, Bertrand and Kramarz 2002, Besley and Burgess 2002, Rajan and Zingales 2003). At the most basic level, the political perspective explains socialism itself – the system that concentrates all political power and economic decision making in the hands of a small elite, thereby providing this elite with the most powerful lever for perpetuating itself – making the whole population of a country dependent on it economically. Despite the value of the political angle, it is premature to blame all institutional failure on politics. Politics is often a stronger force toward institutional efficiency than away from it. After 35 all, even communist dictatorships collapsed and turned into capitalist democracies. There are at least four significant ways in which the political process moves institutions toward efficiency. First, some institutions evolve over time toward more efficient forms as they confront new circumstances. This, fundamentally, is Hayek’s (1960) view of the evolution of common law. Second, the reigns of government are often captured by interest groups favoring efficiency. De Long and Shleifer (1993), for example, show that, over the last millennium, the regions of Europe that were governed by merchants established good institutions and experienced rapid urban growth, whereas regions governed by absolutist princes did not do as well. Third, even when some interest groups oppose change, Coasian bargaining often does lead to efficient institutional choice (Becker 1983). The adoption of Magna Carta in England, and of the U.S. Constitution, are only the most famous examples of such bargaining. Last but not least, despite well-recognized problems with democratic voting (Buchanan and Tullock 1962), it is often a powerful force toward more efficient institutions (Wittman 1989). We have already discussed progressive reforms. On a broader scale, the enormous rise in the world’s wealth during the 20th – democratic – century is the best evidence of the virtues of democratic politics. The tendency of democratic states to seek efficiency is well illustrated by the adoption of that great American invention for balancing dictatorship and disorder – a rigid constitution with checks and balances (Brennan and Buchanan 1980). The clearest embodiment of such rigidity is the idea that courts themselves have the power to check laws passed by the legislature against the constitution. The American idea of constitutional review has spread to countries influenced by the U.S. Constitution, especially those in Latin America, but after World War II to many other parts of the world, including Continental Europe, as constitutional courts became common. La Porta, Lopez- 7 In transition economies and in West European constitutional deliberations, the creation of checks and balances has become a central issue of institutional design (Berglof et al. 2003). 36 de-Silanes, Pop-Eleches, and Shleifer (2002) examine recent constitutions of 71 countries, identify the countries that have adopted judicial review in their constitutions, and find that constitutional review is indeed associated with greater political freedom.7 Not all efficiency-enhancing bargains have the sweet smell of the Magna Carta or the U.S. Constitution, but they may still improve welfare. During the 1990s, Russia’s President Yeltsin fought the possible return of the communist dictatorship and extreme disorder arising from decentralized federalism and economic disorganization. Yelstin’s political, economic, and legal reforms generally required Coasian bargains with significant stakeholders, bargains that to uninformed observers appeared as failures of democratic rule. Privatization, for example, required deep concessions to enterprise insiders, the defeat of communists at the polls in 1996 called for a purchase of political support from the oligarchs, and the preservation of the federation demanded massive giveaways to independent regions (Boycko, Shleifer, and Vishny 1995, Shleifer and Treisman 2000). By the end of the decade, Russia emerged as a democratic, capitalist economy, with much stronger institutions than it had at the beginning of the decade, enabling Yelstin’s successor to continue his program and reap considerable benefits of stability and economic growth. In retrospect, many of Yelstin’s policies and institutional reforms look like welfare-improving strategies of combating dictatorship and disorder. To take this point still further, consider the even more extreme example of Peru’s President Fujimori and his director of Central Intelligence Agency, Vladimiro Montesinos. Montesinos was Fujimori’s right hand man, managing the President’s relations with other politicians, judges, 37 business people, the media, foreign governments, and civil society. A principal tool of Montesinos’ management was corruption: he routinely exchanged favors and bribes with key members of the elite. Unfortunately for Montesinos, he taped his conversations and exchanges. When Fujimori eventually escaped the country, the tapes became public. Ocampo (2003) reviews some of the tapes and documents the corrupt deals between Montesinos and over 100 members of the Peruvian elite. What do we make of this horrifying evidence? On the one hand, this is a story of obscene dictatorship and corruption. But on the other, these deals were part of restoring order in Peru. Before Fujimori was elected, Peru was a country in the state of anarchy, with negative growth rate, continuous – and murderous – threat from the leftwing guerillas, the Shining Path, and incessant political battles among the elite. Fujimori restored order, destroyed the Shining Path, and attained significant economic growth during his rule. He did so, in part, by reducing conflict among the elites and entrenching himself in power through Montensinos’ corrupt deals. From the perspective of Peruvian institutions, these deals reflect a move to eliminate disorder and increase dictatorship, which was probably efficient. The move is not attractive, but neither is the institutional possibility frontier that Peru faced. Eventually, as Fujimori attempted to consolidate his dictatorship, the political process worked and he was driven out of the country. Not surprisingly, disorder increased as well. The often benign influence of politics is subject to an important caveat. Politics is fundamentally a negotiation between different interests, and the success of political negotiation itself relies crucially on the civic capital in the society – the ability to cooperate. Countries with higher civic capital, and the more attractive IPFs, are more likely to have successful political negotiations and to choose an efficient point on the IPF. In this very important way, the location of the IPF, and 38 the political choice of a point on it, are not independent. Even so, it is incorrect to blame all poor institutional outcomes on politics, since the failures of political negotiation are rooted in many of the same factors that undermine institutional opportunities in the first place. VI. Conclusion: Appropriate Institutions. At least since the 18th century, economists have recognized that good institutions – those that secure property rights – are conducive to good economic performance. The appreciation of the benefits of good institutions has grown recently, in light of both the challenges of transition and development and the significant growth of empirical knowledge. But now economics can move further, and recognize that different institutions are appropriate in different circumstances. This, we believe, is the goal of the new comparative economics. In this paper, we tried to put some flesh on this bare bones agenda. We argued that institutional diversity can in part be understood in terms of the fundamental tradeoff between controlling dictatorship and disorder. Many features of successful and unsuccessful institutions can be understood from this perspective. Moreover, this perspective sheds light on a range of historical experiences, including colonial transplantation, the rise of the regulatory state, and the transition from socialism. This perspective also maintains most forcefully that reforms in any country should be evaluated relative to its own institutional opportunities, rather than some idealized benchmark free of dictatorship and disorder. The field of comparative economics has entered a fascinating new stage. The extraordinary turbulence in the world during the last decade – from post-communist transition, to Asian and Latin American financial crises, to economic and social devastation of Africa – has flagged the centrality 39 of institutional reforms, but also the many pitfalls along the way. We are all humbler and wiser now. But we are also keenly aware that the comparative perspective, which identifies both the possibilities and the limitations of individual societies, can serve as a useful framework for future progress. 40 References Acemoglu, Daron, “Why Not a Political Coase Theorem: Social Conflict, Commitment, and Politics,” J. Comp. Econ. 31, forthcoming, 2003. Acemoglu, Daron, Simon Johnson, and James Robinson, “The Colonial Origins of Comparative Development: An Empirical Investigation,” Am. Econ. Rev. 91, 5:1369-1401, Dec. 2001. Acemoglu, Daron and James Robinson, “ Political Losers as a Barrier to Economic Development,” Am. Econ. Rev. 90, 2:126-30, May 2000. Acemoglu, Daron, and James Robinson, “Economic Backwardness in Political Perspective,” NBER Working Paper #8831, 2002. Ades, Alberto, and Rafael Di Tella, “National Champions and Corruption: Some Unpleasant Interventionist Arithmetic,” Econ. J. 107, 443:1023-1042, July 1997. Aghion, Philippe, Alberto Alesina, and Francisco Trebbi, “Endogenous Political Institutions,” Harvard University mimeo, 2002. Alesina, Alberto, Reza Baqir, and William Easterly, “Public Goods and Ethnic Divisions,” Q. J. Econ. 114, 4:1243-1284, Nov. 1999. Allen, Franklin, Jun Qian, and Meijun Qian, “Law, Finance, and Economic Growth in China,” University of Pennsylvania mimeo, 2002. Barth, James, Gerard Caprio, and Ross Levine, “Bank Supervision and Regulation: What Works Best?,” J. Fin. Intermediation, forthcoming, 2003. Becker, Gary, “A Theory of Competition Among Pressure Groups for Political Influence,” Q. J. Econ. 98, 3:371-400, Aug. 1983. Beck, Thorsten, Asli Demirguc-Kunt, and Ross Levine, “Law and Finance: Does Legal Origin Matter?,” J. Comp. Econ. 31, forthcoming, 2003a. Beck, Thorsten, Asli Demirguc-Kunt, and Ross Levine, “Law, Endowments, and Finance,” J. Fin. Econ., forthcoming, 2003b. Berkowitz, Daniel, Katharina Pistor, and Jean-Francois Richard, “Economic Development, Legality, and the Transplant Effect,” Eur. Econ. Rev. 47, 1:165-195, Feb. 2003. Bernstein, Lisa, “Opting out of the Legal System: Extralegal Contractual Relations in the Diamond Industry,” J. Legal Studies 21, 1:115-157, Jan. 1992. 41 Berglof, Erik, Barry Eichengreen, Gerard Rolland, Guido Tabellini, and Charles Wyplocz, Built to Last: A Political Architecture for Europe, London, UK: CEPR, 2003. Bertrand, Marianne, and Francis Kramarz, “Does Entry Regulation Hinder Job Creation? Evidence from the French Retail Industry,” Q. J. Econ. 117, 4:1369-1414, Nov. 2002. Bertrand, Marianne, Paras Mehta, and Sendhil Mullainathan, “Ferreting out Tunneling: An Application to Indian Business Groups,” Q. J. Econ. 117, 1:121-148, Feb. 2002. Besley, Timothy, “Property Rights and Investment Incentives: Theory and Evidence from Ghana,” J. Pol. Econ. 103, 5:903-937, Oct. 1995. Besley, Timothy, and Robin Burgess, “Can Labor Regulation Hinder Economic Performance? Evidence from India,” London School of Economics mimeo, 2002. Black, Bernard, and Reinier Kraakman, “A Self-Enforcing Model of Corporate Law,” Harvard Law Review 109, 8:1911-1981, June 1996. Black, Bernard, and Anna Tarassova, “Institutional Reform in Transition: A Case Study of Russia,” Stanford Law School mimeo, 2003. Blanchard, Olivier, and Michael Kremer, “Disorganization,” Q. J. Econ. 112,4:1091-1126, Nov. 1997. Blanchard, Olivier, and Andrei Shleifer, “Federalism with and without Political Centralization: China versus Russia,” IMF Staff Papers 48, 0:171-179, Special Issue, 2001. Botero, Juan, Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer, “The Regulation of Labor,” Harvard University mimeo, 2002. Boycko, Maxim, Andrei Shleifer, and Robert Vishny, Privatizing Russia. Cambridge, MA: MIT Press, 1995. Brennan, Geoffrey and James Buchanan, The Power to Tax: Analytical Foundations of the Fiscal Constitution. Cambridge, U.K.: Cambridge University Press, 1980. Buchanan, James and Tullock, Gordon, The Calculus of Consent, Logical Foundations of Constitutional Democracy, Ann Arbor, MI: University of Michigan Press, 1962. Coase, Ronald, “The Problem of Social Cost,” J. Law Econ. 3, 1:1-44, April 1960. Coffee, John C., “The Future as History: The Prospects for Global Convergence in Corporate Governance and its Implications,” Northwestern U. Law Rev. 93, 3:631-707, Spring 1999. 42 Coppin, Clayton and Jack High, The Politics of Purity, Ann Arbor, MI: The University of Michigan Press, 1999. Dal Bó, Ernesto, Pedro Dal Bó, and Rafael Di Tella, “Plata o Plomo?: Bribe and Punishment in a Theory of Political Influence,” Harvard Business School Working Paper Series, No. 03-060, 2002. De Long, J. Bradford and Andrei Shleifer,“Princes and Merchants: European City Growth Before the Industrial Revolution,” J. Law Econ. 36, 2:671-702, Oct. 1993. Demirgüç-Kunt, Asli, and Ross Levine, eds., Financial Structure and Economic Growth, Cambridge, MA: MIT Press, 2001. Demsetz, Harold, “Towards a Theory of Property Rights,” Am. Econ. Rev. 57, 2:347-359, May 1967. De Soto, Hernando, The Other Path. New York, NY: Harper and Row, 1989. Diamond, Jared, Guns, Germs, and Steel: The Fates of Human Societies, New York, NY: W. W. Norton, 1997. Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer, “The Regulation of Entry,” Q. J. Econ. 117, 1:1-37, Feb. 2002. Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer, “Courts,” Q. J. Econ. 118, forthcoming, May 2003. Djankov, Simeon, and Peter Murrell, “Enterprise Restructuring in Transition: A Quantitative Survey,” J. Econ. Lit. 40, 3:739-792, Sept. 2002. Easterly, William, and Ross Levine, “Africa’s Growth Tragedy: Policies and Ethnic Divisions,” Q. J. Econ. 112, 4:1203-1250, Nov. 1997. Easterly, William, and Ross Levine, “Tropics, Germs, and Crops: How Endowments Influence Economic Development,” J. Mon. Econ. 50, 1:3-39, Jan. 2003. Ellickson, Robert, Order without Law: How Neighbors Settle Disputes, Cambridge, MA: Harvard University Press, 1991. Engerman, Stanley, and Ken Sokoloff, “Factor Endowments, Institutions, and Differential Paths of Growth among New World Economies,” in S. Haber, ed., How Latin America Fell Behind, Stanford, CA: Stanford University Press, 1997. Engerman, Stanley, and Ken Sokoloff, “Factor Endowments, Inequality, and Paths of Development Among New World Economies,” NBER Working Paper #9259, 2002. 43 Finer, Samuel, The History of Government, Vol. I-III, Cambridge, U.K.: Cambridge University Press, 1997. Fishback, Price and Shawn Kantor, A Prelude to the Welfare State: The Origins of Workers’ Compensation, Chicago, IL: The University of Chicago Press, 2000. Friedman, Eric, Simon Johnson, and Todd Mitton, “Propping and Tunneling,” J. Comp. Econ. 31, forthcoming, 2003. Frye, Timothy, and Andrei Shleifer, “The Invisible Hand and the Grabbing Hand,” Am. Econ. Rev. Papers and Proceedings 87, 2:354-358, May, 1997. Galanter, Marc, 1974, “Why the Haves Come Out Ahead: Speculations on the Limits of Legal Change,” Law Soc. Rev. 9, 95-169, 1974. Glaeser, Edward, “Public Ownership in the American City,” in Urban Issues and Public Finance: Essays in Honor of Dick Netzer (Amy Ellen Schwartz, ed.), Northampton, MA: Edward Elgar Publishing Inc., forthcoming, 2003. Glaeser, Edward, and Andrei Shleifer, “Legal Origins,” Q. J. Econ. 117, 4:1193-1229, Nov. 2002a. Glaeser, Edward, and Andrei Shleifer, “The Curley Effect,” NBER Working Paper #8942, 2002b. Glaeser, Edward, and Andrei Shleifer, “The Rise of the Regulatory State,” J. Econ. Lit. 41, 2:401- 425, June 2003. Glaeser, Edward, Simon Johnson, and Andrei Shleifer, “Coase versus the Coasians,” Q. J. Econ. 116, 3:853-899, Nov. 2001. Glaeser, Edward, Jose Scheinkman and Andrei Shleifer, “The Injustice of Inequality,” J. Mon. Econ. 50, 1:199-222, Jan. 2003. Greif, Avner, 1989, “Reputation and Coalitions in Medieval Trade Evidence on the Maghribi Traders,” J. Econ. Hist. 49, 4:857-882, Dec. 1989. Greif, Avner, “Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders’ Coalition,” Am. Econ. Rev. 83, 3:525-548, June 1993. Greif, Avner, “Cultural Beliefs and the Organization of Society: A Theoretical Reflection on Collectivist and Individual Societies,” J. Pol. Econ. 102, 5:912-950, Dec. 1994. Hall, Robert, and Charles Jones, “Why Do Some Countries Produce so Much More Output per Worker than Others?,” Q. J. Econ. 114, 1:83-116, Feb. 1999. 44 Hamilton, James, James Madison, and John Jay, Federalist Papers, Reprinted in C. Rossiter, ed., Federalist Papers. New York, NY: New American Library, 1788 [1961]. Hart, Oliver, Andrei Shleifer and Robert Vishny, “The Proper Scope of Government.” Q. J. Econ. 112, 4:127-1161, Nov. 1997. Hay, Jonathan and Andrei Shleifer, “Private Enforcement of Public Laws: A Theory of Legal Reform,” Am. Econ. Rev. Papers and Proceedings 88, 2:398-403, May 1998. Hay, Jonathan, Andrei Shleifer and Robert Vishny, “Toward a Theory of Legal Reform,” Eur. Econ. Rev. 40, 3-5:559-567, April 1996. Hayek, Friederich, A., The Constitution of Liberty. South Bend, IN: Gateway Editions Ltd., 1960. Hellman, Joel, , “Winners Take All: The Politics of Partial Reform,” World Politics 50, 2:203-234, 1998. Hellman, Joel, Geraint Jones, and Daniel Kaufmann, “Seize the State, Seize the Day: State Capture, Corruption and Influence in Transition,” J. Comp. Econ. 31, forthcoming, 2003. Hobbes, Thomas, Leviathan, reprinted in R. Tuck, ed., Leviathan. New York, NY: Cambridge University Press, 1651,[1991]. Jin, Hehui and Yingyi Qian, “Public versus Private Ownership of Firms: Evidence from Rural China,” Q. J. Econ. 113, 3:773-808, Nov. 1998. Johnson, Simon, Peter Boone, Alasdair Breach, and Eric Friedman, “Corporate Governance in the Asian Financial Crisis,” J. Fin. Econ. 58, 1-2:141-186, Oct. 2000. Johnson, Simon, Daniel Kaufmann, and Andrei Shleifer, “The Unofficial Economy in Transition,” Brook. Pap. Econ. Act. 2, 1:159-221, Spring 1997. Johnson, Simon, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer, “Tunneling,” Am. Econ. Rev. Papers and Proceedings 90, 2:22-27, May 2000. Johnson, Simon, John McMillan, and Christopher Woodruff, “Courts and Relational Contracts,” J. Law, Econ. Org. 18, 1:221-277. April 2002. Kaufmann, Daniel, Aart Kraay, and Pablo Zoido-Lobaton, “Governance Matters II: Updated Indicators for 2000-2001,” World Bank Policy Research Working paper 2772, 2002. Knack, Stephen and Philip Keefer, “Institutions and Economic Performance: Cross Country Tests Using Alternative Institutional Measures,” Econ. and Pol. 7, 3:207-227, Sept. 1995. 45 Kornai, Janos, The Socialist System: The Political Economy of Communism, Princeton, NJ: Princeton University Press, 1992. Laffont, Jean-Jacques, Incentives and Political Economy, New York, NY: Oxford University Press, 2000. Landes, David, The Wealth and Poverty of Nations, New York, NY: W. W. Norton, 1998. Landis, James, The Administrative Process, New Haven, CT: Yale University Press, 1938. La Porta, Rafael, Florencio Lopez-de-Silanes, Cristian Pop-Eleches, and Andrei Shleifer, “The Guarantees of Freedom,” NBER Working Paper #8759, 2002. La Porta, Rafael, Florencio Lopez-de-Silanes, and Andrei Shleifer, “Government Ownership of Banks,” J. Fin. 57, 1:265-301, Feb. 2002a. La Porta, Rafael, Florencio Lopez-de-Silanes, and Andrei Shleifer,”What Works in Securities Laws,” Harvard University Mimeo, 2002b. La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer and Robert Vishny, “Legal Determinants of External Finance,” J. Fin. 52,3:1131-1150, June 1997. La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer and Robert Vishny, “Law and Finance,” J. Pol. Econ. 106, 6:1113-1155, Dec. 1998. La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer and Robert Vishny, “The Quality of Government,” J. Law, Econ. Org. 15, 1:222-279, Oct. 1999. La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert Vishny, “Investor Protection and Corporate Governance,” J. Fin. Econ. 58,1-2:3-27, Oct. 2000. La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer and Robert Vishny, “Investor Protection and Corporate Valuation,” J. Fin. 57, 3:1147-1170, June 2002. Lopez-de-Silanes, Florencio, “Determinants of Privatization Prices,” Q. J. Econ. 112, 4:965-1026, Nov. 1997. Marx, Karl, Das Kapital, London: Lawrence and Wishart, 1872 [1974]. McKinsey Global Institute, Russia’s Economic Performance, Moscow: McKinsey, 1999. McMillan, John and Christopher Woodruff, “Interfirm Relationships and Informal Credit in Vietnam,” Q. J. Econ. 114, 4:1285-1320, Nov. 1999. 46 Megginson, William, and Jeffry Netter, “From State to Market: A Survey of Empirical Studies on Privatization,” J. Econ. Lit. 39, 2:321-389, June, 2001. Mises, Ludwig von, Human Action, New Haven, CT: Yale University Press, 1949. Montesquieu, M. de Secondat, The Spirit of Laws. Special Edition, The Legal Classics Library. Birmingham, AL: Lucas Printing Company, 1748 [1984]. Murphy, Kevin, Andrei Shleifer, and Robert Vishny, “The Transition to a Market Economy: Pitfalls of Partial Reform,” Q. J. Econ. 107, 3:899-906, Aug. 1991. Murrell, Peter, “Evolution in Economics and in the Economic Reform of the Centrally Planned Economies,” in Emergence of Market Economies in Eastern Europe, Christopher Clague and Gordon Rausser, eds., London and New York: Blackwell, 1992. Murrell, Peter, “Transition According to Cambridge, Mass.,” J. Econ. Lit. 33,1:164-178, March 1995. Murrell, Peter, “Firms Facing New Institutions: Transactional Governance in Romania,” J. Comp. Econ. 31, forthcoming, 2003. North, Douglass, Structure and Change in Economic History. Cambridge, MA: W. W. Norton and Company, 1981. North, Douglass, Institutions, Institutional Change, and Economic Performance. Cambridge and London: Cambridge University Press, 1990. North, Douglass and Barry Weingast, “Constitutions and Commitment: The Evolution of Institutions Governing Public Choice in Seventeenth-Century England,” J. Econ. Hist. 49, 4:803-832, Winter, 1989. Nozick, Robert, Anarchy, State and Utopia, New York, NY: Basic Books , 1974. Ocampo, Luis Moreno, “The Montesino’s Case,” Harvard Law School mimeo, 2003. Olson, Mancur, The Logic of Collective Action, Cambridge, MA: Harvard University Press, 1965. Olson, Mancur, The Rise and Decline of Nations, New Haven, CT: Yale University Press, 1982. Olson, Mancur, “Dictatorship, Democracy, and Development,” Am. Pol. Sc. Rev. 87, 3:567-575, Sept. 1993. Olson, Mancur, Power and Prosperity: Outgrowing Communist and Capitalist Dictatorships, New York, NY: Basic Books, 2000. 47 Persson, Torsten, and Guido Tabellini, Political Economics: Explaining Economic Policy, Cambridge, MA: MIT Press, 2000. Pistor, Katharina, Yoram Keinan, Jan Kleinheisterkamp, and Mark West, “ The Evolution of Corporate Law,” U. Penn. J. Int. Econ. Law 23, 4:791-871, Winter 2002. Pistor, Katharina, Yoram Keinan, Jan Kleinheisterkamp, and Mark West, “Innovation in Corporate Law,” J. Comp. Econ. 31, forthcoming, 2003. Pistor, Katharina and Chenggang Xu, “Law Enforcement under Incomplete Law: Theory and Evidence from Financial Market Regulation,” Columbia Law School mimeo, 2002. Posner, Richard, “Theories of Economic Regulation,” Bell J. Econ. 5, 2:335-358, Autumn 1974. Posner, Richard, “What do Judges Maximize?,” chapter 3 in Overcoming Law, Cambridge, MA: Harvard University Press, 1995. Putnam, Robert, Making Democracy Work: Civic Traditions in Modern Italy, Princeton, NJ: Princeton University Press, 1993. Rajan, Raghuram and Luigi Zingales, “The Tyranny of Inequality,” J. Pub. Econ. 76, 3:521-558, June 2000. Rajan, Raghuram and Luigi Zingales, “The Great Reversals: The Politics of Financial Development in the 20th Century,” J. Fin. Econ. 69, 1:xx-xx, July, 2003. Rodrik, Dani, Arvind Subramanian, and Francisco Trebbi, “Institutions Rule: The Primacy of Institutions over Geography and Integration in Economic Development,” Harvard University mimeo, 2002. Roland, Gerard, Transition and Economics: Politics, Markets, and Firms. Cambridge, MA: MIT Press, 2000. Shleifer, Andrei, “Government in Transition,” Eur. Econ. Rev. 41, 3-5:385-410, April, 1997. Shleifer, Andrei, “State versus Private Ownership,” J. Econ. Persp. 12, 4:133-150, Fall, 1998. Shleifer, Andrei, and Daniel Treisman, Without a Map: Political Tactics and Economic Reform in Russia, Cambridge, MA: MIT Press, 2000. Shleifer, Andrei and Robert Vishny, “Corruption,” Q. J. Econ. 108, 3:599-618, Nov. 1993. Shleifer, Andrei, and Robert Vishny, “Politicians and Firms,” Q. J. Econ. 109, 3:995-1025, Nov. 1994. 48 Shleifer, Andrei and Robert Vishny, The Grabbing Hand: Government Pathologies and their Cures. Cambridge, MA: Harvard University Press, 1998. Smith, Adam, An Inquiry into the Nature and Causes of the Wealth of Nations. Chicago, IL: University of Chicago Press, 1776 [1976]. Sonin, Konstantin, “Why the Rich Favor Poor Protection of Property Rights?” J. Comp. Econ. forthcoming., 2003 Stigler, George, “The Theory of Economic Regulation,” Bell J. Econ. and Man. Sci. 2, 1:3-21, Spring, 1971. Tullock, Gordon, “The Welfare Cost of Tariffs, Monopoly, and Theft,” Western Econ. J. 5, 0:224- 232, 1967. Watson, Alan, Legal Transplants: An Approach to Comparative Law. Charlottesville, VA: University of Virginia Press, 1974. Wittman, Donald, “Why Democracies Produce Efficient Results,” J. Pol. Econ. 97, 6:1395-1424, Dec. 1989. Wurgler, Jeffrey, “Financial Markets and the Allocation of Capital,” J. Fin. Econ. 58, 1-2:1187-214, Oct. 2000. Zhuravaskaya, Ekatherina, “Incentives to Provide Local Public Goods: Fiscal Federalism – Russian style,” J. Pub. Econ. 76, 3:337-368, June 2000.

The Economics of Resources or the Resources of Economics Author(s): Robert M. Solow Source: The American Economic Review , May, 1974, Vol. 64, No. 2, Papers and Proceedings of the Eighty-sixth Annual Meeting of the American Economic Association (May, 1974), pp. 1-14 Published by: American Economic Association Stable URL: https://www.jstor.org/stable/1816009 JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org. Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms is collaborating with JSTOR to digitize, preserve and extend access to The American Economic Review This content downloaded from 132.198.22.255 on Thu, 26 Jan 2023 20:20:04 UTC All use subject to http RICHARD T. ELY LECTURE The Economics of Resources or the R esources of Economics By ROBERT M. SOLOW\* It is easy to choose a subject for a dis- tinguished lecture like this, before a large and critical audience with a wide range of interests. You need a topic that is abso- lutely contemporary, but somehow peren- nial. It should survey a broad field, with- out being superficial or vague. It should probably bear some relation to economic policy, but of course it must have some serious analytical foundations. It is nice if the topic has an important literature in the past of our subject-a literature which you can summarize brilliantly in about eleven minutes-but it better be some- thing in which economists are interested today, and it should appropriately be a subject you have worked on yourself. The lecture should have some technical inter- est, because you can't waffle for a whole hour to a room full of professionals, but it is hardly the occasion to use a blackboard. I said that it is easy to choose a subject for the Ely Lecture. It has to be, because twelve people, counting me, have done it. I am going to begin with a quotation that could have come from yesterday's newspaper, or the most recent issue of the American Economic Review. Contemplation of the world's disap- pearing supplies of minerals, forests, and other exhaustible assets has led to de- mands for regulation of their exploita- tion. The feeling that these products are now too cheap for the good of future generations, that they are being selfishly exploited at too rapid a rate, and that in consequence of their excessive cheapness they are being produced and consumed wastefully has given rise to the conser- vation movement. The author of those sentences is not Dennis Meadows and associates, not Ralph Nader and associates, not the President of the Sierra Club; it is a very eminent economic theorist, a Distin- guished Fellow of this Association, Harold Hotelling, who died at the age of seventy- eight, just a few days ago. Like all eco- nomic theorists, I am much in his debt, and I would be happy to have this lecture stand as a tribute to him. These sentences appeared at the beginning of his article "The Economics of Exhaustible Re- sources," not in the most recent Review, but in the Journal of Political Economy for April 1931. So I think I have found some- thing that is both contemporary and perennial. The world has been exhausting its exhaustible resources since the first cave-man chipped a flint, and I imagine the process will go on for a long, long time. Mr. Dooley noticed that "th' Supreme Coort follows the iliction returns." He would be glad to know that economic theorists read the newspapers. About a year ago, having seen several of those re- spectable committee reports on the ad- vancing scarcity of materials in the United States and the world, and having, like everyone else, been suckered into reading the Limits to Growth, I decided I ought to find out what economic theory has to say \* Professor of economics, Massachusetts Institute of Technology. I This content downloaded from 132.198.22.255 on Thu, 26 Jan 2023 20:20:04 UTC All use subject to http 2 AMERICAN ECONOMIC ASSOCIATION MAY 1974 about the problems connected with ex- haustible resources. I read some of the literature, including Hotelling's classic article-the theoretical literature on ex- haustible resources is, fortunately, not very large-and began doing some work of my own on the problem of optimal social management of a stock of a nonrenewable but essential resource. I will be mention- ing some of the results later. About the time I finished a first draft of my own paper and was patting myself on the back for having been clever enough to realize that there was in fact something still to be said on this important, contemporary but somehow perennial topic just about then it seemed that every time the mail came it contained another paper by another eco- nomic theorist on the economics of ex- haustible resources.1 It was a little like trotting down to the sea, minding your own business like any nice independent rat, and then looking around and suddenly discovering that you're a lemming. Any- how, I now have a nice collection of papers on the theory of exhaustible resources; and most of them are still unpublished, which is just the advantage I need over the rest of you. A pool of oil or vein of iron or deposit of copper in the ground is a capital asset to society and to its owner (in the kind of society in which such things have private owners) much like a printing press or a building or any other reproducible capital asset. The only difference is that the nat- ural resource is not reproducible, so the size of the existing stock can never in- crease through time. It can only decrease (or, if none is mined for a while, stay the same). This is true even of recyclable materials; the laws of thermodynamics and life guarantee that we will never re- cover a whole pound of secondary copper from a pound of primary copper in use, or a whole pound of tertiary copper from a pound of secondary copper in use. There is leakage at every round; and a formula just like the ordinary multiplier formula tells us how much copper use can be built on the world's initial endowment of cop- per, in terms of the recycling or recovery ratio. There is always less ultimate copper use left than there was last year, less by the amount dissipated beyond recovery during the year. So copper remains an exhaustible resource, despite the possi- bility of partial recycling. A resource deposit draws its market value, ultimately, from the prospect of extraction and sale. In the meanwhile, its owner, like the owner of every capital asset, is asking: What have you done for me lately? The only way that a resource deposit in the ground and left in the ground can produce a current return for its owner is by appreciating in value. Asset markets can be in equilibrium only when all assets in a given risk class earn the same rate of return, partly as current dividend and partly as capital gain. The common rate of return is the interest rate for that risk class. Since resource deposits have the peculiar property that they yield no divi- dend so long as they stay in the ground, in equilibrium the value of a resource deposit must be growing at a rate equal to the rate of interest. Since the value of a de- posit is also the present value of future sales from it, after deduction of extraction costs, resource owners must expect the net price of the ore to be increasing exponen- tially at a rate equal to the rate of interest. If the mining industry is competitive, net price stands for market price minus margi- 1 The Review of Economic Stutdies will publish a group of them in the summer of 1974, including my own paper and others by Partha Dasgupta and Geoffrey Heal, Michael Weinstein and Richard Zeckhauser, and Joseph Stiglitz, from all of which I have learned a lot about this subject. I would especially like to thank Zeckhauser for conversation and correspondence, and for the kind of reading of the first draft of this Lecture that one only dares to hope to get because it is so close to Christmas. The final version reflects his comments. This content downloaded from 132.198.22.255 on Thu, 26 Jan 2023 20:20:04 UTC All use subject to http VOL. 64 NO. 2 RICHARD T. ELY LECTURE 3 nal extraction cost for a ton of ore. If the industry operates under constant costs, that is just market price net of unit extrac- tion costs, or the profit margin. If the industry is more or less monopolistic, as is frequently the case in extractive industry, it is the marginal profit-marginal revenue less marginal cost-that has to be growing, and expected to grow, proportionally like the rate of interest. This is the fundamental principle of the economics of exhaustible resources. It was the basis of Hotelling's classic article. I have deduced it as a condition of stock equilibrium in the asset market. Hotelling thought of it mainly as a condition of flow equilibrium in the market for ore: if net price is increasing like compound interest, owners of operating mines will be indiffer- ent at the margin between extracting and holding at every instant of time. So one can imagine production just equal to demand at the current price, and the ore market clears. No other time profile for prices can elicit positive production in every period of time. It is hard to overemphasize the impor- tance of this tilt in the time profile for net price. If the net price were to rise too slowly, production would be pushed nearer in time and the resource would be ex- hausted quickly, precisely because no one would wish to hold resources in the ground and earn less than the going rate of return. If the net price were to rise too fast, re- source deposits would be an excellent way to hold wealth, and owners would delay production while they enjoyed supernor- mal capital gains. According to the fundamental principle, if we observe the market for an exhaustible resource near equilibrium, we should see the net price-or marginal profit-rising exponentially. That is not quite the same thing as seeing the market price to users of the resource rising exponentially. The price to consumers is the net price plus extraction costs, or the obvious analogy for monopoly. The market price can fall or stay constant while the net price is rising if extraction costs are falling through time, and if the net price or scarcity rent is not too large a proportion of tkle market price. That is presumably what has been hap- pening in the market for most exhaustible resources in the past. (It is odd that there are not some econometric studies designed to find out just this. Maybe econometri- cians don't follow the iliction returns.) Eventually, as the extraction cost falls and the net price rises, the scarcity rent must come to dominate the movement of market price, so the market price will eventually rise, although that may take a very long time to happen. Whatever the pattern, the market price and the rate of extraction are connected by the demand curve for the resource. So, ultimately, when the market price rises, the current rate of production must fall along the demand curve. Sooner or later, the market price will get high enough to choke off the demand entirely. At that moment produc- tion falls to zero. If flows and stocks have been beautifully coordinated through the operations of futures markets or a plan- ning board, the last ton produced will also be the last ton in the ground. The resource will be exhausted at the instant that it has priced itself out of the market. The Age of Oil or Zinc or Whatever It Is will have come to an end. (There is a limiting case, of course, in which demand goes asymp- totically to zero as the price rises to in- finity, and the resource is exhausted only asymptotically. But it is neither believ- able nor important.) Now let us do an exercise with this appa- ratus. Suppose there are two sources of the same ore, one high-cost and the other low- cost. The cost difference may reflect geo- graphical accessibility and transportation costs, or some geological or chemical dif- ference that makes extraction cheap at one This content downloaded from 132.198.22.255 on Thu, 26 Jan 2023 20:20:04 UTC All use subject to http 4 AMERICAN ECONOMIC ASSOCIATION MAY 1974 site and dear at the other. The important thing is that there are cost differences, though the final mineral product is identi- cal from both sources. It is easy to see that production from both sources cannot coexist in the market for any interval of time. For both sources to produce, net price for each of them must be growing like compound interest at the market rate. But they must market their ore at the same price, because the product is identical. That is arithme- tically impossible, if their extraction costs differ. So the story has to go like this. First one source operates and supplies the whole market. Its net price rises exponentially, and the market price moves correspond- ingly. At a certain moment, the first source is exhausted. At just that moment and not before, it must become economical for the second source to come into production. From then on, the world is in the single- source situation: the net price calculated with current extraction costs must rise. exponentially until all production is choked. off and the second source is exhausted. (If there are many sources, you can see how it will work.) Which source will be used first? Your instinct tells you that the low-cost deposit will be the first one worked, and your in- stinct is right. You can see why, in terms of the fundamental principle. At the be- ginning, if the high-cost producer is serv- ing the market, the market price must cover high extraction costs plus a scarcity rent that is growing exponentially. The low-cost producer would refrain from undercutting the price and entering the market only if his capital gains justify holding off and entering the market later. But just the reverse will be true. Any price high enough to keep the high-cost pro- ducer in business will tempt the low-cost producer to sell ore while the selling is good and invest the proceed< in any asset paying the market rate of interest. So it must be that the low-cost producer is the first to enter. Price rises and output falls. Eventually, at precisely the moment when the low-cost supply is exhausted, the price has reached a level at which it pays the high-cost producer to enter. From then on, his net price rises exponentially and pro- duction continues to fall. When cumula- tive production has exhausted the high- cost deposit, the market price must be such as to choke the demand off to zero- or else just high enough to tempt a still higher-cost source into production. And so it goes. Apart from market processes, it is actually socially rational to use the lower-cost deposits before the higher-cost ones. You can take this story even further, as William Nordhaus has done in connection with the energy industry. Suppose that, somewhere in the background, there is a technology capable of producing or substi- tuting for a mineral resource at relatively high cost, but on an effectively inexhaust- ible resource base. Nordhaus calls this a "backstop technology." (The nearest we now have to such a thing is the breeder reactor using U238 as fuel. World reserves of U238 are thought to be enough to provide energy for over a million years at current rates of consumption. If that is not a back- stop technology, it is at least a catcher who will not allow a lot of passed balls. For a better approximation, we must wait for controlled nuclear fusion or direct use of solar energy. The sun will not last for- ever, but it will last at least as long as we do, more or less by definition.) Since there is no scarcity rent to grow exponentially, the backstop technology can operate as soon as the market price rises enough to cover its extraction costs (including, of course, profit on the capital equipment in- volved in production). And as soon as that happens, the market price of the ore or its substitute stops rising. The "backstop technology" provides a ceiling for the market price of the natural resource. This content downloaded from 132.198.22.255 on Thu, 26 Jan 2023 20:20:04 UTC All use subject to http VOL. 64 NO. 2 RICHARD T. ELY LECTURE 5 The story in the early stages is as I have told it. In the beginning, the successive grades of the resource are mined. The last and highest-cost source gives out just when the market price has risen to the point where the backstop technology be- comes competitive. During the earlier phases, one imagines that resource com- panies keep a careful eye on the prospec- tive costs associated with the backstop technology. Any laboratory success or failure that changes those prospective costs has instantaneous effects on the capital value of existing resource deposits, and on the most profitable rate of current production. In actual fact, those future costs have to be regarded as uncertain. A correct theory of market behavior and a correct theory of optimal social policy will have to take account of technological un- certainty (and perhaps also uncertainty about the true size of mineral reserves). Here is a mildly concrete illustration of these principles. There is now a workable technology for liquefying coal-that is, for producing synthetic crude oil from coal.2 Nordhaus puts the extraction-and- preparation cost at the equivalent of seven or eight 1970 dollars per barrel of crude oil, including amortization and interest at 10 percent on the plant; I have heard higher and lower figures quoted. If coal were available in unlimited amounts, that would be all. But, of course, coal is a scarce resource, though more abundant than drillable petroleum, so a scarcity rent has to be added to that figure, and the rent has to be increasing like the rate of interest during the period when coal is being used for this purpose. In the meanwhile, the extraction and production cost for this technology is large compared with the scarcity rent on the coal input, so the market price at which the liquefied-coal-synthetic-crude activity would now be economic is rising more slowly than the rate of interest. It may even fall if there are cost-reducing technological improvements; and that is not unlikely, given that research on coal has not been splashed as liberally with funds as research on nuclear energy. In any case, political shenanigans and mo- nopoly profits aside, scarcity rents on oil form a larger fraction of the market price of oil, precisely because it is a lower cost fuel. The price of a barrel of oil should therefore be rising faster than the implicit price at which synthetic crude from coal could compete. One day those curves will intersect, and that day the synthetic- crude technology will replace the drilled- petroleum technology. Even before that day, the possibility of coal liquefaction provides a kind of ceiling for the price of oil. I say "kind of" to remind you that coal-mining and moving capacity and synthetic-crude plant cannot be created overnight. One might hope that the ceiling might also limit the con- suming world's vulnerability to political shenanigans and monopoly profits. I sup- pose it does in some ultimate sense, but one must not slide over the difficulties: for example, who would want to make a large investment in coal liquefaction or coal gasification in the knowledge that the current price of oil contains a large mo- nopoly element that could be cut, at least temporarily, if something like a price war should develop? The fundamental principle of the eco- nomics of exhaustible resources is, as I have said, simultaneously a condition of flow equilibrium in the market for the ore and of asset equilibrium in the market for deposits. When it holds, it says quite a lot about the probable pattern of exploita- tion of a resource. But there are more than the usual reasons for wondering whether 2 As best one can tell at the moment, shale oil is a more likely successor to oil and natural gas than either gasified or liquefied coal. The relevant costs are bound to be uncertain until more research and development has been done. I tell the story in terms of liquefied coal only because it is more picturesque that way. This content downloaded from 132.198.22.255 on Thu, 26 Jan 2023 20:20:04 UTC All use subject to http 6 AMERICAN ECONOMIC ASSOCIATION MAY 1974 the equilibrium conditions have any ex- planatory value. For instance, the flow mar- ket that has to be cleared is not just one market; it is the sequence of markets for resource products from now until the date of exhaustion. It is, in other words, a se- quence of futures markets, perhaps a long sequence. If the futures markets actually existed, we could perhaps accept the no- tion that their equilibrium configuration is stable; that might not be true, but it is at least the sort of working hypothesis we frequently accept as a way of getting on with business. But there clearly is not a full set of futures markets; natural- resource markets work with a combination of myopic flow transactions and rather more farsighted asset transactions. It is legitimate to ask whether observed re- source prices are to be interpreted as ap- proximations to equilibrium prices, or whether the equilibrium is so unstable that momentary prices are not only a bad indi- cator of equilibrium relationships, but also a bad guide to resource allocation. That turns out not to be an easy ques- tion to answer. Flow considerations and stock considerations work in opposite directions. The flow markets by themselves could easily be unstable; but the asset markets provide a corrective force. Let me try to explain why. The flow equilibrium condition is that the net price grow like compound interest at the prevailing rate. Suppose net prices are expected by producers to be rising too slowly. Then resource deposits are a bad way to hold wealth. Mine owners will try to pull out; and if they think only in flow terms, the way to get out of the resource business is to increase current production and convert ore into money. If current production increases, for this or any other reason, the current price must move down along the demand curve. So initially pessi- mistic price expectations on the part of producers have led to more pressure on the current price. If expectations about future price changes are responsive to current events, the consequence can only be that pessimism is reinforced and deepened. The initial disequilibrium is worsened, not eliminated, by this chain of events. In other words, the market mechanism I have just described is unstable. Symmetrical reasoning leads to the conclusion that if prices are initially expected to be rising too fast, the withholding of supplies will lead to a speculative run-up of prices which is self-reinforcing. Depending on which way we start, initial disequilibrium is magnified, and production is tilted either toward excessive current dumping or toward speculative withholding of supply. (Still other assumptions are possi- ble and may lead to qualitatively different results. For instance, one could imagine that expectations focus on the price level rather than its rate of change. There is much more work to be done on this question.) Such things have happened in resource markets; but they do not seem always to be happening. I think that this story of instability in spot markets needs amend- ment; it is implausible because it leaves the asset market entirely out of account. The longer run prospect is not allowed to have any influence on current happenings. Suppose that producers do have some no- tion that the resource they own has a value anchored somewhere in the future, a value determined by technological and demand considerations, not by pure and simple speculation. Then if prices are now rising toward that rendezvous at too slow a rate, that is indeed evidence that owning resource deposits is bad business. But that will lead not to wholesale dumping of cur- rent production, but to capital losses on existing stocks. When existing stocks have been written down in value, the net price can rise toward its future rendezvous at more or less the right rate. As well as beThis content downloaded from 132.198.22.255 on Thu, 26 Jan 2023 20:20:04 UTC All use subject to http VOL. 64 NO. 2 RICHARD T. ELY LECTURE 7 ing destabilized by flow reactions, the market can be stabilized by capitalization reactions. In fact the two stories can be made to merge: the reduction in flow price coming from increased current production can be read as a signal and capitalized into losses on asset values, after which near- equilibrium is reestablished. I think the correct conclusion to be drawn from this discussion is not that either of the stories is more likely to be true. It is more complex: that in tranquil conditions, resource markets are likely to track their equilibrium paths moderately well, or at least not likely to rush away from them. But resource markets may be rather vulnerable to surprises. They may respond to shocks about the volume of reserves, or about competition from new materials, or about the costs of competing technologies, or even about near-term political events, by drastic movements of current price and production. It may be quite a while before the transvaluation of values-I never thought I could quote Nietzsche in an economics paper-settles down under the control of sober future prospects. In between, it may be a cold winter. So far, I have discussed the economic theory of exhaustible resources as a partial- equilibrium market theory. The interest rate that more or less controls the whole process was taken as given to the mining industry by the rest of the economy. So was the demand curve for the resource it- self. And when the market price of the resource has ridden up the demand curve to the point where the quantity demanded falls to zero, the theory says that the resource in question will have been ex- hausted. There is clearly a more cosmic aspect to the question than this; and I do not mean to suggest that it is unimportant, just because it is cosmic. In particular, there remains an important question about the social interest in the pace of exploitation of the world's endowment of exhaustible natural resources. This aspect has been brought to a head recently, as everyone knows, by the various Doomsday forecasts that combine a positive finding that the world is already close to irreversible col- lapse from shortage of natural resources and other causes with the normative judg- ment that civilization is much too young to die. I do not intend to discuss those forecasts and judgments now-this con- vention already has one session devoted to just that-but I do want to talk about the economic issues of principle involved. First, there is a proposition that will be second nature to everyone in this room. What I have called the fundamental prin- ciple of the economics of exhaustible re- sources is, among other things, a condi- tion of competitive equilibrium in the sequence of futures markets for deliveries of the natural resource. This sequence extends out to infinity, even if the com- petitive equilibrium calls for the resource to be exhausted in finite time. Beyond the time of exhaustion there is also equilib- rium: supply equals demand equals zero at a price simultaneously so high that demand is choked off and so low that it is worth no one's while to lose interest by holding some of the resource that long. Like any other competitive equilibrium with the right background assumptions, this one has some optimality properties. In particular, as Hotelling pointed out, the competitive equilibrium maximizes the sum of the discounted consumer-plus- producer surpluses from the natural re- source, provided that society wishes to dis- count future consumer surpluses at the same rate that mine owners choose to dis- count their own future profits. Hotelling was not so naive as to leap from this conclusion to the belief that laissez-faire would be an adequate policy for the resource industries. He pointed to This content downloaded from 132.198.22.255 on Thu, 26 Jan 2023 20:20:04 UTC All use subject to http 8 AMERICAN ECONOMIC ASSOCIATION MAY 1974 several ways in which the background assumptions might be expected to fail: the presence of externalities when several owners can exploit the same underground pool of gas or oil; the considerable uncer- tainty surrounding the process of explora- tion with the consequent likelihood of wasteful rushes to stake claims and exploit, and the creation of socially useless wind- fall profits; and, finally, the existence of large monopolistic or oligopolistic firms in the extractive industries. There is an amusing sidelight here. It is not hard to show that, generally speaking, a monopolist will exhaust a mine more slowly than a competitive industry facing the same demand curve would do. (Hotel- ling did not explore this point in detail, though he clearly knew it. He did men- tion the possibility of an extreme case in which competition will exhaust a resource in finite time and a monopolist only asymptotically.) The amusing thing is that if a conservationist is someone who would like to see resources conserved beyond the pace that competition would adopt, then the monopolist is the conserva- tionist's friend. No doubt they would both be surprised to know it. Hotelling mentions, but rather pooh- poohs, the notion that market rates of interest might exceed the rate at which society would wish to discount future utilities or consumer surpluses. I think a modern economist would take that possi- bility more seriously. It is certainly a potentially important question, because the discount rate determines the whole tilt of the equilibrium production schedule. If it is true that the market rate of interest exceeds the social rate of time preference, then scarcity rents and market prices will rise faster than they "ought to" and pro- duction will have to fall correspondingly faster along the demand curve. Thus the resource will be exploited too fast and exhausted too soon. The literature has several reasons for expecting that private discount rates might be systematically higher than the correct social rate of discount. They fall into two classes. The first class takes it more or less for granted that society ought to discount utility and consumption at the same rates as reflective individuals would discount their own future utility and consumption. This line of thought then goes on to suggest that there are reasons why this might not happen. One standard example is the fact that individ- uals can be expected to discount for the riskiness of the future, and some of the risks for which they will discount are not risks to society but merely the danger of transfers within the society. Since there is not a complete enough set of insurance markets to permit all these risks to be spread properly, market interest rates will be too high. Insecurity of tenure, as William Vickrey has pointed out, is a special form of uncertainty with particular relevance to natural resources. A second standard example is the exis- tence of various taxes on income from capital; since individuals care about the after-tax return on investment and society about the before-tax return, if investment is carried to the point where the after-tax yield is properly related to the rate of time preference, the before-tax profitability of investment will be too high. I have nothing to add to this discussion. The other class of reasons for expecting that private discount rates are too high and will thus distort intertemporal deci- sions away from social optimality denies that private time preference is the right basis for intertemporal decisions. Frank Ramsey, for instance, argued that it was ethically indefensible for society to dis- count future utilities. Individuals might do so, either because they lack imagina- tion (Bohm-Bawerk's "defective telescopic faculty") or because they are all too conThis content downloaded from 132.198.22.255 on Thu, 26 Jan 2023 20:20:04 UTC All use subject to http VOL. 64 NO. 2 RICHARD T. ELY LECTURE 9 scious that life is short. In social decision- making, however, there is no excuse for treating generations unequally, and the time-horizon is, or should be, very long. In solemn conclave assembled, so to speak, we ought to act as if the social rate of time preference were zero (though we would simultaneously discount future consump- tion if we expect the future to be richer than the present). I confess I find that reasoning persuasive, and it provides another reason for expecting that the market will exhaust resources too fast. This point need not be divorced so com- pletely from individual time preference. If the whole infinite sequence of futures markets for resource products could ac- tually take place and find equilibrium, I might be inclined to accept the result (though I would like to know who decides the initial endowments within and be- tween generations). But of course they cannot take place. There is no way to col- lect bids and offers from everyone who will ever live. In the markets that actually do take place, future generations are represented only by us, their eventual ancestors. Now generations overlap, so that I worry about my children, and they about theirs, and so on. But it does seem fundamentally implausible that there should be anything ex post right about the weight that is actually given to the wel- fare of those who will not live for another thousand years. We have actually done quite well at the hands of our ancestors. Given how poor they were and how rich we are, they might properly have saved less and consumed more. No doubt they never expected the rise in income per head that has made us so much richer than they ever dreamed was possible. But that only reinforces the point that the future may be too important to be left to the accident of mistaken expectations and the ups and downs of the Protestant ethic. Several writers have studied directly the problem of defining and characterizing a socially-optimal path for the exploitation of a given pool of exhaustible resources. The idea is familiar enough: instead of worrying about market responses, one imagines an idealized planned economy, constrained only by its initial endowment, the size of the labor force, the available technology, and the laws of arithmetic. The planning board then has to find the best feasible development for the economy. To do so, it needs a precise criterion for comparing different paths, and that is where the social rate of time preference plays a role. It turns out that the choice of a rate of time preference is even more critical in this situation than it is in the older literature on optimal capital accumulation without any exhaustible resources. In that theory, the criterion usually adopted is the maxi- mization of a discounted sum of one-period social welfare indicators, depending on consumption per head, and summed over all time from now to the infinite future. The typical result, depending somewhat on the particular assumptions made, is that consumption per head rises through time to a constant plateau defined by the "modified Golden Rule." In that ultimate steady state, consumption per head is lower the higher is the social rate of dis- count; and, correspondingly, the path to the steady state is characterized by less saving and more interim consumption, the higher the social rate of discount. That is as it should be: the main beneficiaries of a high level of ultimate steady-state con- sumption are the inhabitants of the dis- tant future, and so, if the planning board discounts the future very strongly, it will choose a path that favors the near future over the distant future. When one adds exhaustible resources to the picture, the social rate of time prefer- ence can play a similar, but even more critical, role. As a paper by Geoffrey Heal This content downloaded from 132.198.22.255 on Thu, 26 Jan 2023 20:20:04 UTC All use subject to http 10 AMERICAN ECONOMIC ASSOCIATION MAY 1974 and Partha Dasgupta and one of my own show, it is possible that the optimal path with a positive discount rate should lead to consumption per head going asymp- totically to zero, whereas a zero discount rate leads to perpetually rising consump- tion per head. In other words, even when the technology and the resource base could permit a plateau level of consumption per head, or even a rising standard of living, positive social time preference might in effect lead society to prefer eventual ex- tinction, given the drag exercised by ex- haustible resources. Of course, it is part of the point that it is the planning board in the present that plans for future extinc- tion: nobody has asked the about-to- become-defunct last generation whether it approved of weighting its satisfactions less than those of its ancestors. Good theory is usually trying to tell you something, even if it is not the literal truth. In this context, it is not hard to interpret the general tenor of the theoret- ical indications. We know in general that even well-functioning competitive markets may fail to allocate resources properly over time. The reason, I have suggested, is because, in the nature of the case, the future brings no endowment of its own to whatever markets actually exist. The intergenerational distribution of income or welfare depends on the provision that each generation makes for its successors. The choice of a social discount rate is, in effect, a policy decision about that intergenera- tional distribution. What happens in the planning parable depends very much perhaps dramatically-on that choice; and one's evaluation of what happens in the market parable depends very much on whether private choices are made with a discount rate much larger than the one a deliberate policy decision would select. The pure theory of exhaustible resources is trying to tell us that, if exhaustible re- sources really matter, then the balance between present and future is more deli- cate than we are accustomed to think; and then the choice of a discount rate can be pretty important and one ought not to be too casual about it. In my own work on this question, I have sometimes used a rather special cri- terion that embodies sharp assumptions about intergenerational equity: I have im- posed the requirement that consumption per head be constant through time, so that no generation is favored over any other, and asked for the largest steady consumption per head that can be main- tained forever, given all the constraints in- cluding the finiteness of resources. This criterion, like any other, has its pluses and its minuses and I am not committed to it by any means. Like the standard criterion the discounted sum of one- period utilities this one will always pick out an efficient path, so one at least gets the efficiency conditions out of the anal- ysis. The highest-constant-consumption criterion also has the advantage of high- lighting the crucial importance of certain technological assumptions. It is clear without any technical appa- ratus that the seriousness of the resource- exhaustion problem must depend in an important way on two aspects of the tech- nology: first, the likelihood of technical progress, especially natural-resource-sav- ing technical progress, and, second, the ease with which other factors of produc- tion, especially labor and reproducible capital, can be substituted for exhaustible resources in production. My own practice, in working on this problem, has been to treat as the central case (though not the only case) the as- sumption of zero technological progress. This is not because I think resource-saving inventions are unlikely or that their ca- pacity to save resources is fundamentally limited. Quite the contrary- if the future is anything like the past, there will be proThis content downloaded from 132.198.22.255 on Thu, 26 Jan 2023 20:20:04 UTC All use subject to http VOL. 64 NO. 2 RICHARD T. ELY LECTURE 11 longed and substantial reductions in natural-resource requirements per unit of real output. It is true, as pessimists say, that it is just an assumption and one can- not be sure; but to assume the contrary is also an assumption, and a much less plau- sible one. I think there is virtue in analyz- ing the zero-technical-progress case be- cause it is easy to see how technical prog- ress can relieve and perhaps eliminate the drag on economic welfare exercised by natural-resource scarcity. The more im- portant task for theory is to try to under- stand what happens or can happen in the opposite case. As you would expect, the degree of sub- stitutability is also a key factor. If it is very easy to substitute other factors for natural resources, then there is in prin- ciple no "problem." The world can, in effect, get along without natural resources, so exhaustion is just an event, not a catastrophe. Nordhaus's notion of a "back- stop technology" is just a dramatic way of putting this case; at some finite cost, pro- duction can be freed of dependence on exhaustible resources altogether. If, on the other hand, real output per unit of resources is effectively bounded- cannot exceed some upper limit of produc- tivity which is in turn not too far from where we are now-then catastrophe is unavoidable. In-between there is a wide range of cases in which the problem is real, interesting, and not foreclosed. For- tunately, what little evidence there is sug- gests that there is quite a lot of substituta- bility between exhaustible resources and renewable or reproducible resources, though this is an empirical question that could absorb a lot more work than it has had so far. Perhaps the most dramatic way to illus- trate the importance of substitutability, and its connection with Doomsday, is in terms of the permanent sustainability of a constant level of consumption. In the simplest, most aggregative, model of a resource-using economy one can prove something like the following: if the elas- ticity of substitution between exhaustible resources and other inputs is unity or big- ger, and if the elasticity of output with respect to reproducible capital exceeds the elasticity of output with respect to natural resources, then a constant population can maintain a positive constant level of con- sumption per head forever. This perma- nently maintainable standard of living is an increasing, concave, and unbounded function of the initial stock of capital. So the drag of a given resource pool can be overcome to any extent if only the initial stock of capital is large enough. On the other hand, if the elasticity of substitution between natural resources and other inputs is less than one, or if the elasticity of out- put with respect to resources exceeds the elasticity of output with respect to repro- ducible capital, then the largest constant level of consumption sustainable forever with constant population is-zero. We know much too little about which side of that boundary the world is on-techno- logical progress aside-but at least the few entrails that have been read seem favorable.3 Perhaps I should mention that when I say "forever" in this connection, I mean "for a very long time." The mathematical reasoning does deal with infinite histories, but actually life in the solar system will only last for a finite time, though a very long finite time, much longer than this lecture, for instance. That is why I think it takes economics as well as the entropy law to answer our question. I began this lecture by talking of the conditions for competitive equilibrium in the market for natural resources. Now I have been talking of centralized planning I See pp. 60-70 in William D. Nordhaus and James Tobin. This content downloaded from 132.198.22.255 on Thu, 26 Jan 2023 20:20:04 UTC All use subject to http 12 AMERICAN ECONOMIC ASSOCIATION MAY 1974 optima. As you would expect, it turns out that under the standard assumptions, the Hotelling rule, the fundamental principle of natural-resource economics, is a neces- sary condition for efficiency and therefore for social optimality. So there is at least a prayer that a market-guided system might manage fairly well. But more than the Hotelling condition is needed. I have already mentioned one of the extra requirements for the intertemporal optimality of market allocations: it is that the market discount future profits at the same rate as the society would wish to dis- count the welfare of future inhabitants of the planet. This condition is often given as an argument for public intervention in re- source allocation because as I have also mentioned there are reasons to expect market interest rates to exceed the social rate of time preference, or at least what philosophers like us think it ought to be. If the analysis is right, then the market will tend to consume exhaustible resources too fast, and corrective p(iblic interven- tion should be aimed at slowing down and stretching out the exploitation of the resource pool. There are several ways that could be done, in principle, through con- servation subsidies or a system of gradu- ated severance taxes, falling through time. Realistically speaking, however, when we say "public intervention" we mean rough and ready political action. An only moderately cynical observer will see a problem here: it is far from clear that the political process can be relied on to be more future-oriented than your average corporation. The conventional pay-out period for business is of the same order of magnitude as the time to the next election, and transferring a given individual from the industrial to the government bureau- cracy does not transform him into a guard- ian of the far future's interests. I have no ready solution to this problem. At a min- imum, it suggests that one ought to be as suspicious of uncritical centralization as of uncritical free-marketeering. Maybe the safest course is to favor specific poli- cies-like graduated severance taxes rather than blanket institutional solutions. There is another, more subtle, extra requirement for the optimality of the com- petitive market solution to the natural- resource problem. Many patterns of ex- ploitation of the exhaustible-resource pool obey Hotelling's fundamental principle myopically, from moment to moment, but are wrong from a very long-run point of view. Such mistaken paths may even stay very near the right path for a long time, but eventually they veer off and become bizarre in one way or another. If a market- guided system is to perform well over the long haul, it must be more than myopic. Someone-it could be the Department of the Interior, or the mining companies, or their major customers, or speculators- must always be taking the long view. They must somehow notice in advance that the resource economy is moving along a path that is bound to end in disequilibrium of some extreme kind. If they do notice it, and take defensive actions, they will help steer the economy from the wrong path toward the right one.4 Usually the "wrong" path is one that leads to exhaustion at a date either too late or too soon; anyone who perceives this will be motivated to arbitrage between present and future in ways that will push the current price toward the "right" path.5 It is interesting that this need for some- 4This sort of process has been studied in a different context by Frank Hahn and by Karl Shell and Joseph Stiglitz. 5 For example, suppose the current price is too low, in the sense that, if it rises according to the current principle, the demand path will be enough to exhaust the resource before the price has risen high enough to choke demand to zero. A clever speculator would see that there will be money to be made just after the date of exhaustion, because anyone with a bit of the resource to sell could make a discrete jump in the price and still find buyers. Such a speculator would wish to buy now This content downloaded from 132.198.22.255 on Thu, 26 Jan 2023 20:20:04 UTC All use subject to http VOL. 64 NO. 2 RICHARD T. ELY LECTURE 13 one to take the long view emerged also when the question at hand was the poten- tial instability of the market for natural resources if it concentrates too heavily on spot or flow decisions, and not enough on future or stock decisions. In that context too, a reasonably accurate view of the long-term prospects turns out to be a use- ful, maybe indispensable, thing for the resource market to have. This lecture has been-as Kenneth Burke once said about the novel-words, all words. Nevertheless, it has been a dis- course on economic theory, not on cur- rent policy. If some of you have been day- dreaming about oil and the coming winter, I assure you that I have been thinking about shadow prices and transversality conditions at infinity. If I turn briefly to policy at the end, it is not with concrete current problems in mind. After all, nothing I have been able to say takes account of the international oil cartel, the political and economic ambitions of the Middle Eastern potentates, the speeds of adjustment to surprises in the supply of oil, or the doings of our own friendly local oligopolists. The only remarks I feel en- titled to make are about the long-run pur- suit of a general policy toward exhaustible resources. Many discussions of economic policy- macroeconomics aside-boil down to a tension between market allocation and public intervention. Marketeers keep thinking about the doughnut of allocative efficiency and informational economy and dirigistes are impressed with the size of the hole containing externalities, imperfec- tions, and distributional issues. So it is with exhaustible resources. One is im- pressed with what a system of ideal mar- kets, including futures markets, can ac- complish in this complicated situation; and one can hardly miss seeing that our actual oligopolistic, politically involved, pollution-producing industry is not ex- actly what the textbook ordered. I have nothing new to add to all that. The un- usual factor that the theory of exhaustible resources brings to the fore is the impor- tance of the long view, and the value of reasonable information about reserves, technology, and demand in the fairly far future. This being so, one is led to wonder whether public policy can contribute to stability and efficiency along those lines. One possibility is the encouragement of organized futures trading in natural re- source products. To be useful, futures con- tracts would have to be much longer-term than is usual in the futures markets that now exist, mostly for agricultural products. I simply do not know enough to have an opinion about the feasibility of large scale futures trading, or about the ultimate con- tribution that such a reform would make to the stability and efficiency of the market for resource products. But in principle it would seem to be a good idea. The same considerations suggest that the market for exhaustible resources might be one of the places in the economy where some sort of organized indicative planning could play a constructive role. This is not an endorsement of centralized decision- making, which is likely to have imperfec- tions and externalities of its own. Indeed it might be enough to have the govern- ment engaged in a continuous program of information-gathering and dissemination covering trends in technology, reserves and demand. One could at least hope to have professional standards govern such an exercise. I take it that the underlying logic of indicative planning is that some comparison and coordination of the main participants in the market, including the and hold for sale then. But that action would tend to raise the current price (and, by the fundamental prin- ciple, the whole price path) and reduce demand, so that the life of the resource would be prolonged. The specula- tion is thus corrective. This content downloaded from 132.198.22.255 on Thu, 26 Jan 2023 20:20:04 UTC All use subject to http 14 AMERICAN ECONOMIC ASSOCIATION MAY 1974 government, could eliminate major errors and resolve much uncertainty. In the case of exhaustible resources, it could have the additional purpose of generating a set of consistent expectations about the distant future. In this effort, the pooling of infor- mation and intentions from both sides of the market could be useful, with the effect of inducing behavior that will steer the economy away from ultimately inferior paths. It is also likely, as Adam Smith would have warned, that a certain amount of conspiracy against the public interest might occur in such sessions, so perhaps they ought to be recorded and the tapes turned over to Judge Sirica, who will know what to do with them. REFERENCES P. Dasgupta and G. Heal, "The Optimal De- pletion of Exhaustible Resources," Rev. Econ. Stud., forthcoming, 1974. F. H. Hahn, "Equilibrium Dynamics with Heterogeneous Capital Goods," Quart. J. Econ., Nov. 1966, 80, 633-646. H. Hotelling, "The Economics of Exhaustible Resources," J. Polit. Econ., April 1931, 39, 137-175. W. D. Nordhaus, "The Allocation of Energy Resources," Brookings Papers on Econ. Activ., forthcoming. and J. Tobin, "Is Economic Growth Obsolete?" in National Bureau of Economic Research, Economic Growth, 50th Anni- versary Colloq. V, New York 1972. K. Shell and J. E. Stiglitz, "The Allocation of Investment in a Dynamic Economy," Quart. J. Econ., Nov. 1967, 81. R. M. Solow, "Intergenerational Equity and Exhaustible Resources," Rev. Econ. Stud., forthcoming, 1974. J. E. Stiglitz, "Growth with Exhaustible Nat- ural Resources," Rev. Econ. Stud., forth- coming, 1974. M. Weinstein and R. Zeckhauser, "Use Pat- terns for Depletable and Recyclable Re- sources," Rev. Econ. Stud., forthcoming, 1974. This content downloaded from 132.198.22.255 on Thu, 26 Jan 2023 20:20:04 UTC All use subject to http

THE ECONOMICS OF INDUSTRY I I I ALFRED MARSHALL, ii AND MARY PALEY MARSHALL. Hontton : M A C M I L L A N A N D CO. 1881 à [ The Right of Translation is reserved.] Samtatoge : PRINTED BY C. J . CLAY, M.À. & SON, AT THE UNIVERSITY PRES8. 2 7/s 6 PREFACE. T his book was undertaken at the request of a meeting of Cambridge University Extension lecturers, and is designed to meet a want which they have felt. It is an attempt to construct on the lines laid down in Mill’s Political Economy a theory of Value, Wages and Profits, which shall include the chief results of the work of the present generation of Economists. The main outlines of this theory have been, Rested during many years in lectures at Cambridge, and more recently at Bristol. An inquiry into the subjects of Banking, Foreign Trade and Taxation is deferred to a companion volume on the “ Economics of Trade and Finance.” The authors wish to acknowledge their obligations to Mr H. Sidgwick, Mr H. S. Foxwell and the Rev. W. Moore Ede for suggestions and aid in preparing the book for the press. PREFACE TO THE SECOND EDITION. T he present volume contains an outline of the theory of Value, Wages and Profits. This theoiy, as it was left by English Economists of the last generation, made too great pretensions to finality ; and by a natural reaction their work has been severely criticised. But on the whole the progress of inquiry has tended to vindicate it, and to show that while m ost of it is very incomplete, there is but little in the careful exposition of it given by John Stuart Mill which is not, when properly interpreted, true as far as it goes. It seems however necessary to go a good way apart from Mill with regard to one important question. He never worked out fully the applications of his own principles to the problem of Distribution: his last utterance on the question in his review of Thornton, left part of it avowedly in an unsatisfactory state; he gave indeed some hints as to the direction in which a solution was to be looked for, but he did not pretend to work it out himself. In this volume an attempt is made to supply the solution, and to show that there is a unity underlying all the different parts of the theory of prices, wages and profits. The remuneration of every kind of work, the interest on capital, and the prices of commodities, are determined in the long run by competition according to what is fundamentally the same law. This law of Normal Value has many varieties of detail, and takes many different forms. But in every form it exhibits value as determined by certain relations of demand and supply; and Cost of Production as taking the chief place among the causes that determine supply. The Second Book begins with a general statement of this law as applied to commodities. This is followed by a discussion of some peculiarities that are found in thç laws of supply of unskilled VI PREFACE. labour, of skilled labour and of business power. The is thus cleared for the inquiry how Normal demand and supply determine first the share of the joint produce of capital and industry which goes as interest to capital, and secondly how the share that goes to industry is divided among its different ranks. The volume ends with a discussion of the way in which the Normal action of economic forces is hindered, or even overridden, but never destroyed by friction, by combination or by those passing events which exercise a restless influence on Market values. By an oversight no formal definition of the term free competition has been given in the text; and as, the book™ being stereotyped, there is some difficulty in introducing it at the proper place, it may be given here. A man competes freely when he is pursuing a course, which without entering into any combination with others, he has deliberately selected as that which is likely to be of the greatest material advantage to himself and his family. He ’ is not supposed to be selfish: in fact the Normal supply o f ! all grades of industry, except perhaps the lowest, depends on the unselfish sacrifice by parents of their own pleasures | for the benefit of their children. But he is supposed to be i consulting his own material advantage and that of his family to the comparative neglect of the welfare of others. If every one always found his greatest happiness in trying to do that which was best for others, the world would have no theory of Normal values as it is described in this volume : some such Communism as that which prevailed among the early Christians would be the basis of economic theory. But in this world, as it is, the chief active principle in business is the desire of each man to promote the material interests of himself and his family. Normal results in Economics are therefore those which Would be brought about in the long run by this active principle, if it had time to overcome— as it necessarily would in sufficient time — custom, inertness, ignorance, and all the other passive elements which make up economic friction. It ha \* ' urged that as custom is often more powerful than cc tion it ought not to be spoken of slightingly as i friction. But this is entirely to misapprehend the m of the term friction. A friction is not necessarily a i PREFACE. vu P hing, but it is a passive resistance ; and an active force, i owe ver small it is, acting on a material that is not perfectly igid, will in the long run overcome any amount of friction, lum an nature is never absolutely rigid ; and custom never îolds its own in opposition to a strong active economic orce working for many generations persistently in the same direction. Normal results are those which competition would bring\* ibout in the long run. The periods to which they relate nust be sufficiently long to give time for the active forces of i :ompetition to overcome the passive resistances of ignorance, prejudice, custom, etc. They must be sufficiently long to i mable us to neglect temporary fluctuations of supply and demand, the influence of good or bad harvests, etc., and to regard these alternating changes in opposite directions as neutralizing one another. We must be able to see standing lip in clear outline the broad effects of the constant action of the forces of competition. It must however be admitted that there are several difficulties in the way of a precise definition of the period of time to which normal results apply, a great deal must be left to the judgment in each case. In particular it is impossible to lay down general rules as to the time that must be allowed for the spreading of knowledge with regard to new inventions, and to changes in the markets for goods and for labour. Every one has a tendency to seek the most advantageous occupation for himself and for his children, and this active tendency will in the long run overcome the passive resistance of ignorance. But ignorance may act as a drag for a long time, and the advantages of different occupations may vary rapidly. One of the most important of the unwritten chapters of Economics is that on the time that elapses between economic causes and their effects in consequence of the slowness with which knowledge diffuses itself. The first chapters were printed at a time when it was proposed to give the volume a more elementary character than was ultimately found advisable; and the difficulties which surround the definition of economic terms were ignored as far as possible ; but in this edition some l further discussion is given of the definitions of wealth and VIII PREFACE. capital. There is however no really satisfactory meaning to be given to these terms without departing too much from precedent f In particular the account which ordinary English usage compels us to give of capital when looked at from the point of view of the individual is not well suited for scientific purposes. Not only is it indefinite in many ways; but it compels us to say that if a boat builder has rented his carriage from a carriage builder who meanwhile has rented a boat from him, the capital of each would be diminished if each were to buy the things that he had been hiring. i We are thus brought to a fringe of difficulties which it would be out of place to discuss in a treatise such as the present. They are not merely verbal; but yet they ard of such a nature that we can pass by them. For it wil| be found that none of our reasonings depend on our bein§jt able to give an exact list of the things which we regard as capital. The notion of providing enjoyment for the future rather than for the present is that which underlies all the uses of the term Capital; and the property of providing for the future admits of all degrees and suggests no precise dividing line. The direction in which we ought to look for a solution of the difficulty seems to be a classification of all kinds of wealth according to the extent to which they possess this property. French writers have always used the word Capital in a broad sense, sometimes as practically synonymous with wealth. Some German writers have also developed a hint of M. Say’s as to the division of capital into productive capital and capital for consumption. But their suggestions, though not without value, do not seem to go to the root of the matter. It may be mentioned that the phrase Net Income is used in several senses in different parts of the book. “ Net” in each case means that which remains after making the deductions described in the context \ August, 1881. CONTENTS, i f BOOK I. L A N D , LABOUR AND CAPITAL. CHAPTER I. INTRODUCTORY. § i The work of Political Economy or Economics. [§§ a, 3 The relation of science to practice. The methods of moral or social sciences.] § 4 Definition of the terms Economics, § 5 Wealthy Material Wealthy Personal or Non-material Wealthy § 6 Productive [§ 7 Wealth, farther considered] .... pp. 1— 7 CHAPTER II. AGENTS OF PRODUCTION. I 1 Man’s worl^ § a Is aided by machinery, § 3 And controls Nature’s forces. § 4 Man’s efficiency depends on his physical vigour ; § 5 On his knowledge and mental ability, § 6 And hfe moral c h a r a c t e r ..........................................................pp. 8— 12 CHAPTER III. CAPITAL. 1 1 Capital, § a Though saved it is spent. § 3 "What is seen and what is not seen.” Labour requires support and aid from capital. § 4 Trade cannot be benefited by destruction of goods ; § 5 Or hindrances to free production. § 6 Remuneratory or Wage Capital. Auxiliary Capital. § 7 Circulating and Fixed Capital. Specialised i and Non-specialised Capital. § 8 Personal Capital. pp. 13— 20 X CONTENTS. § i An illustration. § 2 The Return due to a Dose of Capiti 1 § 3 Law of Diminishing Return. § 4 The influence of improv » ments in production. [§ 5 Agriculture in a new country.] § 6 Pi » duce of m i n e s ..................................... . . pp. 21— CHAPTER V. GROWTH OF POPULATION. MALTHUS , POOR LAWS. I § i The multiplication table. § 2 Recklessness of backward race. Standard of Comfort. § 3 Law of population, § 4 Malthus' Essaj. § 5 Qualifications of his doctrine. §§ 6, 7 State aid and voluntary aid in P o o r -r e lie f..................................................pp. «7—$ | CHAPTER V I.C GROWTH OF CAPITAL.'\* § i The power to save. § 2 The will to save. § 3 Savings from re\*, wages and profits. § 4 The growth of capital depends -partly on tk rate of profits, § 5 And must ultimately be slow. . pp. 36—41 CHAPTER VII. ORGANIZATION OF INDUSTRY. § 1 Uncivilized tribes. The Agricultural state. § 2 Greeks, Romas, and German races. § 3 Mediaeval Gilds. § 4 Transition lo modern times. § 5 Circulation of capital . . pp. 43—-f \ ^ CHAPTER VIII. DIVISION OF LABOUR. § 1 Division of labour increases skill, § 2 Economizes skill, § 3 A of machinery. § 4 Production on a large scale. § 5 Subsidiajj 1 industries. § 6 Localisation of industry. § 7 Advantages ofl ' large factory. § 8 Modem changes in manufacture. § 9 Indire\* results of division of labour. § 10 The law of Division of Labot or Increasing Return. § 11 Application to Agriculture. § 0 The Peasant P ro p rie to r..........................................pp. 49— ÿ CHAPTER IX. TENURE OF LAND. § 1 Early tenures. §§ 2, 3 Modem tenures . . . p f CHAPTER IV. LAW OF DIMINISHING RETURN. l! J t XI I ^CONTENTS. / V BOOK IL ) N O R Ï T ^ - V A t U E . CHAPTER I. DEFINITIONS. LAW OF DEMAND. § i Competition only one of the causes which determine value» but in this second Book all others are to be neglected. [§ ? The term Normal.] § 3 Definition of Market^ § 4 Value in Use9 Value in Exchange, Price, § 5 Buyert Seller, § 6 Relation between utility and value, Final utility, § 7 Law of Demand . pp. 65— 71 CHAPTER II. LAW OF SUPPLY. § 1 Introductory. § 2 Cost of production. Expenses of production, § 3 Relation between Expenses of production and supply, firstly by men working on their own account, § 4 Secondly by men working under an employer. § 5 Law of Normal Supply, § 6 It assumes free competition. § 7 Law of Normal Value, §§ 8, 9 Expenses of production include Insurance against Risk. [§ 10 “ Cost of Reproduction.\*\*].................................pp. 72— 80 CHAPTER III. RENT. X ) § Illustration. § 2 Law of Rent, § 3 Special case in which there is land on the margin of cultivation. § 4 The effects on rent o f, increased demand, of free importation, and of improvements in' production. [§ 5 Rent of mines.] .... pp. 81—87 CHAPTER IV. RENT IN RELATION TO VALUE. §/i Meaning of the phrase: “ Rent does not enter into Expenses of I production.\*\* Law of Normal value of a commodity, different I portions of which have different Expenses of production. [§ 2 Ground-rent of f a c t o r ie s .]....................................PP\* 88— 90 I xii CONTENTS. § i A change which affects the Expenses of production of all comme £ ties in like proportion, does not alter values, § 2 But an increased demand alters values. § 3 Consequent addition to Law of Nona! v a l u e ..................................................................pp. 91—^3 CHAPTER V. INFLUENCE OF DEMAND ON VALUE , CHAPTER VI. DISTRIBUTION. § i Expenses of production can be divided generally into Wages ai Profits. Relation between Cost and Expenses of producti®. § 2 The Wages-and-profits-Fund. § 3 Earnings of Managemai to be classed with other earnings. The Eamings-and-interest FtaL [§ 4 Ambiguous use of the term “ Cost of production.” § 5 That is no separate class of consumers. § 6 Further examination of the term : “ the real net annual income of the country.”] pp. 94— 100 CHAPTER VII. SUPPLY OF UNSKILLED LABOUR, § j Real wages and Nominal wages. Time wages and Task wages. The family as a unit. § 2 The supply of efficient labour as depend®! on wages. § 3 Net Advantages of an occupation . pp. 101— 10) I CHAPTER VIII. SUPPLY OF SKILLED LABOUR. | § 1 The Law of Normal supply applies' to skilled labour in scie ; respects, § 2 But not in all. § 3 Mill’s four grades of labott I § 4 Net advantages further examined. § 5 Difficulties which k« i down the supply of labour in a trade. § 6 Specialised s^l decreasing in importance relatively to Non-specialised ability. § 7 Apprenticeships. Investment of capital in education by Govetf m e n t .........................................: pp. 104— ^ CHAPTER IX. SUPPLY OF BUSINESS POWER. J ! § i The work of business management is increasing in difficulty. 31 Transference of business power from one occupation to another. \*j{ Supply adjusted to demand in the case of business poi case of skilled l a b o u r .................................. p CONTENTS. § i Interest depends firstly on the amount of the Eamings-and-interestFund ; § 3 Secondly on the demand of industry for the aid of capital. § 3 The connexion between the rate of interest and the amount of capital demanded. § 4 Law of demand for capital. ^ § 5 Supposition that the amount of capital is fixed. § 6 Supposition that the growth of capital is completely controlled by the rate of interest. § 7 Law of Supply of capital. Law of the Normal rate of interest. § 8 The Normal rate in England, § 9 And in new countries. § 10 Tendency of the rate of interest to diminish.................................................................. pp. 119— 137 I CHAPTER X. INTEREST. Pt CHAPTER XI, WAGES, § 1 Competition for the aid of a trade. § 3 Normal wages of unskilled labour. § 3 Law of Normal wages of skilled labour. § 4 A rise - of Time-wages, if Task-wages are unaltered, does not lower profits. [§ 5 Wages equal the Discounted Value or Net Return of labour.] pp. 128— 134 CHAPTER XII. EARNINGS OF MANAGEMENT. 3 § i Trading with borrowed capital. Trade Risks. Personal Risks. § 2 Earnings of Management in‘private partnerships and Joint Stock companies. § 3 Profits on equal capitals in different trades. § 4 Profits on unequal capitals. § 5 Exceptional cases. § 6 Law of Normal Earnings of Management. § 7 They include much “ Rent” of rare natural abilities. § 8 Tendency of profits to diminish...................................................................pp. 135— 145 CHAPTER XIII. RELATION OF NORMAL TO MARKET VALUE. § 1 Summary of the theory of Normal value. [§ 3 In nature changes re-act on one another.] § 3 The theory of Normal value a necessary introduction to that of Market value . . pp\* 146— M9 ! XIV QONTENTS. BOOK III. MARKET VALUE. CHAPTER I. CHANGES IN THE PURCHASING POWER OF MONEY. § i Influence on prices exerted by changes in the supply of the precious metals ; § 2 And by the modem system of credit. § 3 Commerçai fluctuations ; Crises. [§ 4 Commercial depression gives rise to die false notion that there has been general Over-production. § 5 How a fall of prices injures traders, § 6 And thus makes the evils of commercial depression appear greater than they are.] pp. 150— 157 CHAPTER II. MARKET FLUCTUATIONS. | § ! Introductory. § 2 Market fluctuations due to uncertainties of | supply. § 3 The higgling and bargaining of the market. § 4 Flue- 1 tuations due to changes in demand, g 5 Fluctuations are great in trades which use much Fixed capital, § 6 And in trades whkb | make Fixed capital. § 7 Commodities which have Joint expenses ' of production. § 8 Market Fluctuations in prices cause, and sc not caused by, fluctuations in wages, [§9 Although a change of Normal Expenses of production is the cause of a change of Normal va lu e.].................................................................. pp. 158— 167 CHAPTER III. INFLUENCE OF CUSTOM. LOCAL VARIATIONS OF PRICES AND WAGES. § 1 Laws of Local Variation of Normal prices. 4 § 2 The value of land. § 3 Movements of prices in town and country. § 4 Local variations greater in Time-wages than in Task-wages. [§ 5 H ow such variations should be measured. § 6 The influence of die competition of rival producers in neutral marked.1 T VT women. § 8 Local variations of profits, g 9 Re • CONTENTS. xv § i Complete and partial monopolies. How monopoly price is determined. § a A trade combination often aims at a partial monopoly. Difficulty of securing the adhesion of those already in the market. § 3 Difficulty of excluding competition from a distance. § 4 Markets under the control of a few dealers. § 5 Conditions of the success of a combination. § 6 Combinations among employers to lower w a g e s .......................................................... pp. 180—186 CHAPTER IV. MONOPOLIES. COMBINATIONS. CHAPTER V. TRADES UNIONS. § 1 Trades unions an outcome of the Anglo-Saxon spirit. § 1 Town gilds and Craft gilds. § 3 The mechanical inventions of last century brought about the modern wages question. The early history of Trades unions. § 4 Their objects. § 5 Their organization. § 6 Trades union congresses. Trades Councils. § 7 The hold that unions have on the best workmen. § 8 Cost of strikes. § 9 Union rules: Apprenticeships; § 10 A minimum wage; piecework; § 11 The length of the day’s work. § is The indirect influence of unions. § 13 Picketing : Rattening : the use of m a c h in e ry ..........................................................pp. 187— 198 CHAPTER VI. INFLUENCE OF TRADES UNIONS ON WAGES. § 1 The labourer's disadvantage in bargaining for wages. §§ a, 3 It is not impossible for unions to obtain a general risp of wages ; but this will not be permanent if the change diminishes the Wages-andprofits-Fund. [§4 “ The Wages-Fund theory.”] • pp. 199— 205 CHAPTER VII. INFLUENCE OF TRADES UNIONS ON WAGES (CONTINUED). §§1,2 The relations between different groups of workmen associated in the same process of production, compared with their relations to their employers. § 3 Conditions under which a union can obtain a rise of wages in a single trade. § 4 The relation of the moral to the economic elements of the problem . . • pp. 206— 213 xvi CONTENTS. § i Boards of Conciliation and Courts of Arbitration. § 2 The geneal principle on which they should act. § 3 Sliding scales. § 4 Indi\* rect benefits of Conciliation................................. pp. 214— 21J CHAPTER VIII. ARBITRATION AND CONCILIATION. § 1 The ultimate aims of Co-operation. § 2 Industrial Partnership § 3 Co-operative workshops. § 4 In what trades they are most | likely to succeed. 9 5 Their work in taking sub-contracts. |6 . Co-operative credit associations. § 7 Co-operative stores. § 8 The | principles on which they are based. § 9 The advantages which 1 they have in competing with shops. § 10 The probable futwe | of the retail trade. 9 n The Co-operative Wholesale society. 1 The Federation of Co-operators . . . . pp. 218— 218 | In d e x ............................................................................pp. 229— 231 J \* #\* A few discussions are included in square brackets to show that they should be omitted by beginners in tint first time of reading. Words used in a technical sense art printed with capital initial letters. CHAPTER IX. CO-OPERATION. BOOK I. LAND, LABOUR, AND CAPITAL. C H A P T E R I. INTRODUCTORY. 0 § i. “ R o u s s e a u has said, ‘ Much philosophy is wanted for the correct observation of things which are before our eyes.’ And such are the events and customs of every day social life. Habit has so familiarized us with them that we fail to observe them, unless something striking and exceptional forces them on our notice. “ Let us take, by way of illustration, a man in the humble walks of life— a village carpenter, for instance— and observe the various sérvices he renders to society, and receives from it ; we shall not fail to be struck with the enormous disproportion between them. < “ This man employs his day’s labour in planing boards, and making tables and chests of drawers. What does he receive from society in exchange for his work ? u First of all, on getting up in the morning, he dresses himself ; but he has himself made none of his clothing. In order to put at his disposal this clothing, simple as it is, an enormous amount of labour, and many ingenious inventions, must have been employed. Americans must have produced cotton, Indians indigo, Englishmen wool and flax, Brazilians hides ; and all these materials must have been transported to various towns where they have been worked up, spun, woven, dyed, etc. I 2 “ He sends his son to school, and the simple teaching which is given there, is itself due to the work of many thousand mind& “ If he undertakes a journey, he finds that, in order to saw him time and exertion, other men have removed and levelled up the soil, filled up valleys, hewed down mountains, united the banks of rivers, and brought the power of steam into subjection to human wants. i “ It is impossible not to be struck with the measureless disproportion which exists between the enjoyments which this man derives from society, and what he could obtain by his own unassisted exertions. The social mechanism then must be very ingenious and very powerful, since it leads to this singular result, that each man, even he whose lot is cast in the humblest condition, obtains things every day which he could not himself produce in many ages. “ The study of that mechanism is the business of Political Economy1.” In other words,'we may say that Political Economy examines the Production, the Distribution and the Consumption of weaUb. It seeks for the causes which determine wages, profits, and rent ; it inquires how far these causes are fixed by unchange- ; able natural laws, and how far they can be modified by human effort. Last, but not least, it traces the connexion that there • is between the character of the workman and the characterof I his work. “ As a man thinketh, so is he” ; as the work is, so is the worker ; as the worker is, so is the work. The nation used to be called “ the Body Politic.” So long as this phrase was in common use, men thought of the interests of the whole nation when they used the word “ Political” ; aid then “ Political Economy” served well enough as a name for.tjie I science. But now “ political interests” generally mean the fc- I terests of only some part or parts of the nation; so that it | seems best to drop the name “ Political Economy,” and to speak , simply of Economic Science, or more shortly, Economics. The present volume is called The Economics of Industry | because it treats of the affairs of producers, both employers aid workmen. The discussion of banking, foreign trade and finance is deferred to the companion volume. |i § 2. Economics is a science because it collects, arranges, w and reasons about one particular class of facts. A science brings I together a great number of similar facts and finds that they are ] special cases of some great Uniformity which exists in nature, l i t describes this Uniformity in a simple and definite statement, j I or Law. I BOOK I. CHAP. I. § 2. 1 Bastiat’s Harmonics of Political Economy. j A Law of Science states that a certain result will be produced whenever a certain set of causes are present\*. P Science traces the connexion between different Laws, often shewing that some of them are explained by, or contained in others. It reasons from these Laws, applying them to new cases of gradually increasing difficulty, and finds out the conclusions to which they pomt. It then inquires how far these conclusions are consistent with ooservation, so as to verify its work. If necessary, it goes back to its original Laws, and corrects, or modifies or adds to them, so as to make them represent Nature more truly. Thus gradually science becomes able to predict future events with increasing confidence and accuracy3. But this is all that a science can do ; it cannot claim to be a guide in life, or to lay down rules for the practical conduct of affairs. That is the task of what in old times used to be called an Art. An Art considers some important practical end, and directs men in their efforts to obtain it. First it inquires generally into the various conditions of the case. Then taking one of them at a time it seeks out the science whose special business it is to answer questions relating to this particular class of conditions, and demands of this science an answer to a question which bears directly on the end in view. Having collected such answers from many sciences, Art puts them together ; and says, Since we are told by the sciences that such and' such effects will follow from such and such causes, therefore it is best to pursue such and such a course: this course will, all things being considered, lead us up to, or near to, our desired end, so as to cause as much good and as little evil as possible. 0 Thus the railway engineer is a man who devotes himself to the Art of making railways : and when it is decided that a railway is to be made from one town to another, he consults Geology and other sciences and obtains their answers to certain definite questions before he decides which route to adopt. 1 1 It is unfortunate that the word Law is used in this sense, and also in the sense of a command by authority. The law which speaks in the Indicative mood, and says, A is a cause of Z?, is not of the same character as a Law which speaks in the Imperative mood, and says “ Do this, Avoid that.” \* Beginners should omit all passages contained in square brackets. 8 Science, when obtaining new Laws, is said to be Inductive ; when reasoning from them and finding how they are connected with one another, it is said to be Deductive ; its third task, that of Verification, has j ust been described. There has been a controversy as to whether Economics is an Inductive or a Deductive Science. It is both : its Inductions continually suggest new Deductions ; its Deductions continually suggest new Inductions. INTRODUCTORY. 3 I - 2 '4 BOOK I. CHAP. I. § 3— 5. But the statesman or the financier who decides that these two towns are to be connected by a railway is a man of Art in a , yet broader sense of the term. For he has to consider not only I what it will cost to make the railway but also what nett-profit it , will bring in, and perhaps what will be its indirect pdlitical, social and moral effects. In doing this he has to make many inquiries of Economics ; for this science examines the laws that I determine the growth of trade in particular channels, and those that determine the cost of making and working the railway. Economics then cannot by itself be a guide in the practical affairs of life ; but it answers a number of difficult questions which must be asked of it by the statesman, the man of business, I and the philanthropist. Economics is to be classed with the IMoral or Social Sciences ; because it deals only incidentally with Iinanimate things. Its main purpose is to seek for the moral and 1 social Laws by which men’s conduct is determined in the every | day work of their lives : the motives which cause them to seek | one trade and occupation rather than another, and which govern | their behaviour to others with whom their trade brings them 1 into contact. Economics investigates the causes which detetmine the work of a man’s daily life, the manner in which he spends his income, and the influence which his work exerts on I his character. J§ 3. Social sciences have made slower progress than I physical sciences. One reason of this is that men have only recently begun to apply to social sciences those methods of classification, and that systematic study of each class of truths, which have been so successful in the physical sciences. But now that men have set themselves to study each separate group of social facts by itself, these sciences too are beginning to advance steadily. In any history of the physical sciences we may read how j men failed to make rapid progress so long as they persisted in the vain attempt to discover some simple explanation of all ^ the various natural phenomena. The ancients used to be con- | tinually starting new theories for the explanation of the universe, : which succeeding ages had to cast away. As time went on, men learnt that they must separate the study of inorganic life '■from that of organic, the study of chemistry from that of mechanics, and so on. And when men had thus begun to concentrate their attention on one particular class of natural I phenomena at a time, to trace by careful and steady work their | Laws, they made solid progress. Of course they seldom 4i obtained results that were completely true. But the new I results were always nearer the truth than those which they displaced, so that each generation started from a more advantageous position than its predecessors ; and thus by gradual steps m an h a s obtained a command over nature similar to that which fancy used to attribute to the fairy or the magician. A t the same time it is true that in the Moral Sciences, even more than the physical, a man who confines himself entirely to one narrow branch of inquiry is not likely to make good progress in it. T he economist should know something of the history of manners and customs and laws, and of the principles of mental, m oral, legal and political science. He must avoid the error of regarding “ the present experience of mankind as of universal validity, mistaking temporary or local phases of human character for human nature itself, having no faith in the wonderful pliability of «the human mind; deeming it impossible in spite of the strongest evidence that the earth can produce human beings o f a different type from that which is familiar to him in their ow n age and even perhaps in his own country. The only security against this narrowness is a liberal mental cultivation... A person is not likely to be a good economist who is nothing else. Social phenomena acting and reacting on one another, they cannot rightly be understood apart ; but this by no means proves that the material and industrial phenomena of society are not themselves susceptible of useful generalisations, but only that these generalisations must necessarily be relative to a given form of civilisation and a given stage of social advancement1. ” Thus the economist must pause sometimes to consider the connexion between that element of well-being with which he is chiefly concerned, and the other elements; for it is only by this means that he can ascertain the real significance of his own results, and can learn in what direction it is most important to extend his inquiries.] § 4. This account of Economic science may be summed up in the following definitions: Those portions of human conduct which are directed towards the acquirement of material wealth, and those conditions of human well-being which directly depend on material wealth, are called Economic. The Science of Economics collects, examines, arranges and reasons about the facts which aie connected with the economic habits and conditions of well-being in various countries at various times. § 5. The subject-matter of Economics is Wealth. But there is some difficulty in ascertaining the meaning of this word. Wealth must be distinguished from well-being. All things which are useful or pleasurable are elements of Well-being, whether they are material things, or human faculties and capacities for enjoyment. INTRODUCTORY. ç Mill, On Comte, pp. 81—83. 6 But though some of these human qualities should not be\* included under the term wealth ; yet this term should not be confined to those things which can be bought and sold in the market. For in estimating the wealth of a country every one would include the market-value of the power of work possessed by a cart-horse or a slave. And it seems unreasonable to exclude that of a free man simply because he cannot be sold, and so has no market-price. But we want some term which will fitly describe such things as are capable of being exchanged and of having their value definitely measured; and we find such a term in “ Material Wealth.” Wealth then may be said to consist of Material wealth and Personal or non-material wealth. Material Wealth consists of the material sources of enjoyment which are capable of being appropriated and therefore of being exchanged. Thus it includes not only commodities (or things the possession of which can give enjoyment directly), but also machinery and otjier things which are made or appropriated in order to aid man in producing commodities. “ Personal” or “ non-material wealth” consists of those human energies faculties and habits, physical mental and moral, which directly contribute to making men industrially efficient, and which therefore increase their power of producing material wealth. Thus manual skill, intelligence, and honesty may be included in the personal wealth of a country. All other human faculties and qualities which it is an advantage to have, and all other sources of enjoyment, are elements in the well-being of a man, but are not included under the term wealth. Thus the power of appreciating and deriving pleasure from music is an element of well-being, but it is not called wealth ; because generally speaking it does not make men’s work efficient in the production of material wealth. § 6. The term Productive has been used by Economists in many senses, and has caused much misunderstanding. It seems best that Productive when used simply as a technical term should mean productive of wealth. Labour is Productive when it produces wealth, whether Personal or Material. But whenever there is any room for doubt, mention will be made of the particular kind of thing which is produced. Thus it may be said that labour, is “ productive of wealth,” or to anticipate the use of terms which will soon be defined, “ productive of capital,” or “ productive of wage-capital,” etc. BOOK I. CHAP; I. § 6. INTRODUCTORY. Productive labour cannot generally be divided off by a clearly defined line from Unproductive. A minister of religion is often classed as an Unproductive labourer, but if by exerting moral influence he makes labourers more sober, honest and efficient, he is so- far productive of Personal wealth. Again, since some recreation is necessary for the highest efficiency of labour, it is quite possible that a musician may indirectly increase the wealth of a nation, and be indirectly Productive. The word Productive is often applied not only to labour, but also to consumption. This is indeed somewhat inaccurate, for it is labour that produces ; and consumption can have no further claim to being productive than that it supports the labour which is productive. ► I§ 7. Wealth may be looked at from two different points of view— that of the individual and that of the nation. For the individual, material wealth includes not only all the material things which are actually in his possession, but all rights to material things which are recognised by Government or custom. Thus it includes such things as rights of way, patent rights, mortgages, Government securities, shares in public and private trading companies. On the other hand all debts that he owes have to be subtracted from the sum total of his possessions, so that his wealth may be a negative quantity. His Personal wealth includes his faculties and habits of industry, his business reputation, his knowledge of others, and generally everything that enables him to do a good trade or to obtain material advantages in any other way. National wealth includes the wealth of the individual members of the nation. But in estimating it, any debts due from one member of the nation to another may be omitted altogether. On the other hand account must be taken of the internal an’d external political organization of the nation, in so far as this affects the freedom and security of its industry. And, though it must be confessed there is some doubt on the subject, it seems best on the whole to count in all the natural resources of the country, such as the highways afforded by its rivers and seas, and any advantage in production that is derived from a favourable climate, provided it has not already been reckoned for in the value of the land. All wealth is useful, that is, it has the power of satisfying some desire. But things, however useful they may be, have no value in exchange if they can be obtained without effort in as large quantities as they are desired. There has been much discussion as to whether they must not be excluded on this ground from the list of things which constitute wealth ; but the question seems to be of no moment for the purposes either of science or practice.] 7 C H A PT E R IL AGENTS OF PRODUCTION. "j § i. M a n when producing wealth acts upon the things which N ature supplies. The gifts of Nature to man are firstly materials such as iron, stone, wood, etc., and secondly, forces such as the power of the wind, and the heat of the sun, the source whence I all other powers are derived. IAll that man’s hands can do is to move things. A carpenter I for instance takes some planks, cuts them up and fits them 'i together in the form of a box ; he does not in the strict sense of I the word make, or create; he only arranges. “ If we examine any case of what is called the action of man or nature we shall find” that it consists merely in uputting things into fit places for being acted upon by their internal forces and by those 1 residing in other natural objects. .. man only moves one , thing to or from another. He moves a seed into the ground; ---- a«d-the natural forces of vegetation produce in succession a root, a stem, leaves, flowers and fruit. He moves a spark to fuel and it ignites, and by the force generated in combustion i t . J cooks the food, melts or softens the iron, converts into beer or ' 1 sugar the malt or cane juice, which he has previously moved to !the spot. He has no other means of acting upon matter than jby moving it. Motion or resistance to motion are the only jthings which his muscles are constructed for1.” j Since then production does not mean creation but only re- jarrangement, it is a mistake to suppose, as some have done, that ^/ the work of those who carry or sell goods cannot be Productive. J( The carpenter who makes a box, takes certain boards from V an arrangement in which they were of little use, and puts them j together in an arrangement in which they are of greater use. à1 Mill, Principles of Political Economy, Book I. chap, i, § 2. AGENTS OF PRODUCTION. So the carrier or the trader takes the box from the manufactory where it is made, and where it is of little use, and delivers it to the purchaser. It is true that the carrier and the trader make no permanent change in the form of the box ; but that does not prevent them from being Productive, for they as well as the carpenter contribute to render the material which is provided by nature more useful to man. § 2. As civilisation advances the relative importance of mental to manual labour changes. Every year mental labour becomes more important, and manual labour less important. W ith every fresh invention of machinery work is transferred from the muscles, or vital force, to natural force. Even with the imperfect machinery we now have, one pound of coal will raise a hundred pounds twelve thousand feet high ; and the daily work o f a man cannot exceed this even if we make him a mere working machine, and obtain instead of a man’s life so much pulling and pushing, and hewing and hammering. With an ordinary tide, the water rushing in and out of a reservoir of a mile in area, even if three-fourths of its force were wasted through imperfections of machinery, would do as much work in a day as the muscles of a hundred thousand men. § 3. The agents of production are then Nature’s forces, and Man’s force; man’s force being generally most efficientL when it is so applied as to control and direct nature’s forcesJ rather than to counteract them. And the wealth of a country/ depends upon the manner in which nature’s forces and man’s force work together in the production of wealth. Let us first regard the wealth of a country in so far as it depends on the liberality of nature. This consists not only in her gifts of fertile land and rich mines, but also in a convenient arrangement of her gifts. Before the invention of railways, a district could not have a prosperous trade unless it had easy means of communication by river or by sea. Iron mines are of comparatively little value if there is no coal near them. England’s present position in the world is in a great measure due to the fact that she not only has coal mines and iron mines, but also her coal and iron mines are near together. In the course of generations man works on the face of nature, and improves her gifts, or wastes them. The patient industry of the Dutch has turned their barren sands into fertile meadows ; while the wasteful carelessness of the slave-owning cotton-planters of Southern America has turned some of the richest districts of the world into a wilderness. § 4. Turning next to the efficiency of man’s labour in production, we may classify the conditions on which it depends as (i) his physical strength and energy, (ii) his knowledge anf‘ mental ability, and (iii) his moral character. 9 IO Firstly, with regard to man’s physical strength and eneigy. ! While man is altering the face of nature, nature is ever changing the quality of man. A healthy invigorating climate is one ' of the most important of the gifts that nature bestows. Extreme heat enervates man; and in tropical countries makes him lend a ready ear to the suggestion that he should live in idleness on the fruits that nature showers down upon him. ' In many places even in the temperate zones work is almost j entirely suspended during the extreme heats of summer. And | in some parts of America men are not only prone to take a j holiday during “ the heated term,” but are also prevented from doing such work as that of carpenters in the open air during the extreme severity of winter. England is fortunate in having I a climate in which men can work with vigour out of doors ! almost all the year round ; and which, by thus fostering ener- ] getic and steady habits of labour, contributes much to her 1 greatness. | The physical power and energy of man is of course partly dependent upon inherited race-qualities. But modern science i shews that the character of a race may be greatly modified by i changes in its habits of living, in diet, in cleanliness, in houseroom, etc. Thus the physical vigour of a race depends partly on its wealth. To him that hath, to him is given. It must be remembered that the average efficiency of labour in a country depends not only on the efficiency of the labourer in the prime of life, but also on the number of years during which he is really efficient. It is better even as regards material wealth that a man should work with moderate energy up to die age of sixty, than that he should overwork himself so as to become an old man at forty. § 5. Next with regard to knowledge and mental ability. It has already been remarked that the skill and intelligence which are required fox commanding and directing the forces of nature are growing in importance relatively to the brute strength by which uncivilised races struggle with nature. Indeed a thorough general education, together with a special training for some particular employment\* is becoming more necessary to the working man every year. There is scarcely any work which does not need some mental effort. Even in agriculture machinery is being introduced, the management of which requires much skill and intelligence. A man does work that does not need skill all the better if he knows more than is actually required for him. Education makes him quick to understand whatever directions may be given to him : if his machinery gets out of order, or the plan of his work miscarries in any other way, he can set things to right at once and thus prevent much loss. In this and BOOK 1. CHAP. II. §§ 5, 6. I other ways, every increase in the intelligence of the workm an diminishes the amount of supervision required from the em p loyer and his foremen. And as civilisation advances, further progress becomes more and more dependent upon th e diffusion of education among the working classes. T h is education may be classed as general and technical\ A s M ill says, “ the aim of all intellectual training for the m ass o f the people, should be to cultivate common sense, to q u alify them for forming a sound practical judgment of the circum stances by which they are surrounded. Whatever in the intellectual department can be superadded to this is chiefly ornam ental ; while this is the indispensable groundwork on w h ich education must rest. An education directed to diffuse go od sense among the people, with such knowledge as would qualify them to judge of the tendencies of their actions, would b e certain, even without any direct inculcation, to raise up a public opinion by which intemperance and improvidence of every kin d would be held discreditable.” General education should then aim at causing a man to form an intelligent opinion with regard to the ordinary matters of life, and to be full of resources for meeting new emergencies. Technical education should aim at enabling him to understand the processes and the machinery of the special work in which he is engaged. It should help him to understand the reason of everything that goes on in his trade, and thus enable him to accommodate himself to new machinery or new modes of production. And it should train him in the use of his fingers. T h is technical education should be begun at school, but a great deal of the education that is wanted in many trades can only be got in workshops. § 6. Next with regard to riiotal character. Uprightness and mutual confidence are necessary conditions for the growth of wealth. “ Wherever there is a great store of wealth, there must be a people living to a considerable degree under moral restraint, and possessed of a more or less accurate code of duty ; and a land dotted with bursting stack-yards, mapped out into well-tilled fields, and noisy with the hum of looms and the clang of hammers, is evidence that there is at hand no small portion of the stuff out of which martyrs and heroes are formed. Though fine names may not be given to the qualifications of a busy people, skilled in many crafts and trades, producing articles cheaply and well ; it is patience and sobriety, faithfulness and honesty that have gained for them eminence1.” AGENTS OF PRODUCTION. 11 1 Macdonell, Survey of Political Economy\* \± BOOK I. CHAP. II. §6. The character of a nation depends chiefly on that of the mothers of the nation— on their firmness and gentleness and sincerity. It is in childhood, and at home, that the workman must learn to be truthful and trusty, cleanly and careful, energetic and thorough, to reverence others and to respect himself. Finally, industry cannot attain full freedom and efficiency unless, as Mill says, it is protected by the Government and from the Government. We have now seen how land and labour are two of the requisites of production, the third requisite is capital. \ C H A PT E R i Ï Ï .l CAPITAL. § i. IN a savage state man thinks only of satisfying his immediate wants ; in a civilized state he devotes much labour to preparing the roads, buildings, tools, materials, etc. which may be of service to him in the future. He abstains from seeking immediate enjoyment from the whole produce of his labour, and devotes some part of it to producing things which will assist him in his future work. These requisites of production are called Capital. Capital is that part of wealth which is devoted to obtaining further wealth. Regarded firstly from the point of view of the individual it is that part of his wealth which is destined to be used as a source of income. W e say "destined” or devoted, because it is not always possible to tell whether a thing is capital or not, without knowing how the owner intends to use it. Thus oats are capital if they are to be given to a cart-horse, but not if they are to be given to a racehorse. Again, many things are used sometimes for business, therefore as capital ; sometimes for pleasure, and therefore not as capital. A French peasants cart is used as capital in the field, but not as capital when it carries him and his family for a jaunt on a holiday. Again, it is not always clear to what extent a doctor’s house and his carriage should be regarded as capital required for his business. "T o familiarise ourselves with the conception,let us consider what is done with the capital invested in any of the branches of business which compose the productive industpr of a country. A manufacturer, for example, has one part of his capital in the form of buildings, fitted and destined for carrying on his branch of manufacture. Another part he has in the form of machiner\* 14 BOOK I. CHAP. III. §§ 2, 5. A third consists, if he be a spinner, of raw cotton, flax, or wool ; if a weaver, of flaxen, woollen, silk or cotton, thread ; and the 1 like, according to the nature of the manufacture. Food and I clothing for his operatives, it is not the custom of the present age that he should directly provide ; and few capitalists, except the producers of food or clothing, have any portion worth ■mentioning of their capital in this shape. Instead of this each | capitalist has money, which he pays to his workpeople, and so enables them to supply themselves : he has also finished goods | in his warehouses, by the sale of which he obtains more money, j to employ in the same manner, as well as to replenish his stock of materials, to keep his buildings and machinery in repair, \* and to replace them when worn out. His money and finished 1 goods, however, are not wholly capital, for he does not wholly I devote them to these purposes : he employs a part of the one, and of the proceeds of the other, in supplying his personal consumption and that of his family, or in hiring grooms and valets, or maintaining hunters and hounds, or in educating his children, or in paying taxes, or in charity. What then is his capital ? Precisely that part of his possessions, whatever it be, which is to constitute his fund for carrying on fresh productions1. Luxuries and other things which are not intended to be used productively, are thus regarded as private capital if they are in the hands of a trader who derives an income from selling them j or hiring them out, but not if they are in the hands of a private consumer. I But secondly, capital may be regarded from a national point of view as that part of a nation’s wealth which is destined to be employed productively. In estimating it we have not to inquire whether goods are in the hands of a trader or consumer, we 1 have not to inquire how much command over capital each per- I son has at his disposal in the shape of money or of any of the , substitutes for money; (except indeed when it becomes necessary | to take account of the commercial relations with other countries). W e have simply to consider what part of the nation’s wealth, material and personal, is destined to be used in producing further wealth: the sum total of this wealth constitutes the national capital. According to ordinary though not universal English usage the lands and water-power of a country are not counted in the capital of the country, because they have not been made by man ; it is however doubtful whether it is best to continue to exclude them. § 2. Capital is for the most part the result of labour and ' abstinence ; it is saved. But it is also used. \* 1 Mill, Principles, Bk. 1. Ch. iv. § 1. CAPITAL. It is true that some wealth is hoarded; and that while being hoarded it is not being used ; but hoarding has gone out of fashion in civilized countries. An Englishman, when he saves capital, intends either to use it himself, or to lend it out to be used by others ; and capital when it is used is almost always spent : but it is so spent as to be reproduced : it is spent Productively. W e may find a good instance of that Productive expenditure which is realiy saving in the career of the great Duke of Bridgewater. In his time English industries were sadly hampered by the enormous cost of carrying goods from one part of the country to another. None of the roads leading into Manchester were passable by carts. Coals, corn, cloth and other things were carried on horses\* backs in summer. But in winter, when the roads were bad, Manchester was like a beleaguered town. The duke conceived the daring project of making canals to connect the manufactures of Manchester with the coal districts on the one hand, and with the sea at Liverpool on the other. He devoted all his wealth and energies to the work. He lived in the plainest manner, had long consultations with Brindley the engineer in rough log huts ; and derived from the excitement of his enterprise a keener pleasure than he would have Obtained from spending his wealth in luxury. He bequeathed vast wealth to his descendants, but in the act of saving it he gave employment to vast numbers of working men. His canals are a source of the prosperity of his country, and afford permanent employment to thousands. Capital devoted to making things that are to be consumed unproductively does indeed sustain labourers during the process, but it does not produce a fresh supply of capital which will in its turn support and assist labour. A man who employs people in making lace or in laying out ornamental grounds, sustains them while they are at work, and then this means of support comes to an end. On the other hand a man who invests capital in a coal mine not only benefits labourers during the time, but prepares a store of coal which may again be used as capital : his expenditure has the after effect of increasing the amount of those things which either support or assist labour. § 3. And now we may pause to consider Bastiat\*s distinction between those effects of an action which are seen and those which are not seen. When a spendthrift is wasting his substance it is easy enough for his neighbours to see how he gives employment to cooks and valets and horse-jockeys. But when a man saves his money and invests it, say in a new canal or railway, it is not so easy to see that his wealth employs navvies ai^t\*other workmen in the present, and will go on affording employment to other labourers in other generations. Though his wealth is spent, it is not spent by himself : so that 15 i6 BO O K I. CHAP. III. §§ 3, 4. his neighbours do not see that it is spent at all. And though the future effect of his spending will be to increase the means of employing labourers, these effects are not seen ; but they may | be foreseen. “ In the department of economy, an act, a habit, an institution, a law, gives birth not only to an effect but to a I series of effects. Of these effects the first only is immediate ; it I manifests itself simultaneously with its cause ; it is seen. The I others unfold in succession— they are not seen; it is well for us ! if they are foreseen. Between a good and a bad economist this | constitutes the whole difference : the one takes account only of | the visible effect : the other takes account of the effects which are seen and also of those which it is necessary to foresee. Now | this difference is enormous, for it frequently happens that when 1 the immediate effect is favourable, the ultimate consequences are fatal and conversely. Ignorance surrounds mankind in its 1 cradle; then its actions are determined by their first consequences, the only ones which in its first stage it can see. It is only gradually that it learns to take account of the others. It has to learn this lesson from two very different masters— experience and foresight. Experience teaches effectually but brutally. It makes us acquainted with all the effects of an action ' by causing us to feel them, and we cannot fail to finish by 1 knowing that a fire burns when we have burned ourselves. For | this rough teacher I should like if possible to substitute a more gentle one— I mean Foresight. For this purpose I shall examine the effects of various economical phenomena by placing in opposition to each other those which are seen and those which are not seen.” There are, as Bastiat goes on to shew, many important consequences of men’s action which are not seen, but which may be foreseen by the aid off economic principles. \*One of the most important of these principles has been expressed by saying “ Industry is limited by Capital1.” But 1 when so stated it is often misunderstood. It means that :— | Labour requires support and aid from capital. The I - demand for labour in a district cannot in the long run 1 be increased by any device that does not lead to an ' increase of the supply of capital in it. j [ I f the efficiency of labour were to be doubled by a magician’s j wand, while the material capital in the country were unchanged 11 The sentence “ Industry is limited by Capital” is the first of Mill’s four “ fundamental propositions about Capital.” His fourth proposition expresses another side of the same truth ; it is “ Demand for commodi- ' ties ” (payment for them being made after they are finished as is usually done) “ is not demand for labour” (payment for it being made beforehand, as is usually done). Mill’s second proposition is to a similar effect ; it is “ Capital is the result of saving.” This and his third proposition “ Capital is consumed” have already been explained. CAPITAL. there would be a great immediate rise in the real wages of labour, that is in the necessaries, comforts and luxuries which each labourer could purchase by his wages. The stores of capital already in existence would be distributed among the labourers more rapidly than they would otherwise have been; and the increased efficiency of work would speedily replenish the diminished stores. The fact is that increase in the efficiency of labour would really lead to an increase in the supply of capital. T he overstrained interpretation of the proposition— Industry is limited by Capital— has caused many errors1.] § 4. Let us now look at some conclusions which follow from this proposition. Firstly : The destruction of things is not good for trade. For instance, it is not good for trade to have dresses made of material which wears out quickly. For if people did not spend their means on buying new dresses, they would spend them on giving employment to labour in some other way. The dressmaking trade will be as much benefited by their ordering new dresses for the poor as by their ordering new dresses for themselves to the same amount. But it is true that if there were a sudden falling off in the demand for dresses, dressmakers would suffer through being compelled to seek a new occupation. This brings us to the question whether it is right to make work for those who are out of employment. It may be right, if tne suspension of work is temporary, to relieve those who are in distress by making work for them ; because this method of relief does not injure their self-respect. I f however the distress in a trade is not merely temporary, it is not true kindness to dissuade people from leaving it. A great mistake was made by those who tried to bolster up the trade of hanï-weaving, after it had been proved that the handloom must in the long fun give place to the power-loom. Those who doled out poor relief to supplement the miserable wages of the hand-weavers would have done better by helping them to earn their living in another way. But while it is not good for trade to spend money on dress ostentatiously, or to change the fashions rapidly, it is true that those who dress well add, if not to the wealth, yet to the well-being of the nation. A service is done to society by every one who provides the means of refined pleasure without unnecessary expenditure. When a man buys a good picture he devotes wealth to one of its best ends ; and a really artistic dress educates taste just as a picture does. A time may come when it will be thought absurd for a woman to display her wealth by 1 These have chiefly appeared in various forms of what has been called the “ Wages fund theory.” See Book in. chap. vi. \*7 M. a 18 carrying about a great quantity of expensive materials and embroidery; just as a painter who advertised the price he had paid for the paints on his picture would be thought vulgar. It may be noticed that some acquaintance with luxuries seems to be a necessary condition of progress. Many nations have been aroused from a state of sluggish apathy by a desire to obtain foreign luxuries, which has made them exert themselves. But every increase in the consumption of luxuries, unless it leads to increased energy in the production of wealth, diminishes saving and so checks the accumulation of wealth. § 5. The fundamental truth, that the employment of labour cannot in general be increased by any artificial device that does not increase the supply of capital, shews us that :— It is not in the long run good for trade to prevent goods from coming freely into the country so as to cause an artificial demand for some kinds of home productions. Bastiat tçl|5 a story of the candlemakers, who petition the authorities to pass a law ordering the shutting up of all windows and openings, by which the light of the sun can come in. This, they say, will create such a demand for artificial light that their manufactures will be greatly encouraged: being prosperous they will bp able to purchase the goods made by other trades, so that the plan will benefit trade generally. Doubtless government could thus give employment to many industries: the effects which were seen would be beneficial. But the candle- I makers did not foresee that the capital which came into their industry must be subtracted from some other industry, say, that of growing corn, in which it was giving good employment to labour; and that the corn growers would have purchased j from other trades just as much as the candlemakers would. Government would not then increase the employment of labour ; but , would only cause it to be employed in producing candles that were not wanted instead of corn which was wanted. Bastiat wrote this story for the instruction of the French | Government, who often imposed heavy duties on manufactured commodities in order to give a premium to domestic manufacturers, and then plumed themselves on the “ multitude of tall chimneys” they had called into existence. They did not see that the capital invested in such industries was not an addi- ! tion to that which existed in the country ; they did not see that | it had been artificially drawn into these industries at the expense of others. We shall find that many of the most important j practical conclusions of Economics are contained in the state- I ment :— Regulations which prevent men from doing whatever they are best qualified to do cannot benefit the country BOOK I. CHAP. III. § 5—7- J as a whole, whether they are made by Governments, by Professions or by Trades-Unions. § 6. Labour requires support and aid from capital. It is convenient to have separate names for that portion of capital which supports, and that portion which aids labour. Remuneratory Capital or Wage-Capital consists of the food, clothes, shelter, etc. which support labour. Auxiliary Capital is that which aids labour. It consists of tools, machines, factories and other buildings that are used for trade purposes, railways, canals, roads, ships, etc. ; also raw materials. A very rapid conversion of Remuneratory into Auxiliary Capital may temporarily injure labourers. Suppose for instance that a great amount of labour and capital is diverted from agriculture to making railways or steam engines. The corn which would have fed farm labourers while growing com, now feeds some of the same labourers who have left their farms, and gone to work on railways or in machine shops. At the end of the year there will be more railways or steam engines in existence than would otherwise have been, but less com : and in consequence labourers may for the time have a scanty supply of food. But this injury will be only temporary. A t the time of the wRailway mania” capital was sunk in railways on such a gigantic scale that there was not enough left for carrying on the general business of the country, and the Railway panic of 1847 caused much suffering. Yet in the long run railways have conferred great benefits on the working classes of England by increasing that general wealth of the country on which their wages depend. The amount that is now subtracted in a year from the capital of the country for building new railways is slight in comparison to the addition which the railways make to the Remuneratory capital of the country. § 7. This distinction between Remuneratory and Auxiliary Capital is closely connected with one that must be made between Circulating and Fixed Capital. M Capital which fulfils the whole of its office, in the production in which it is engaged, by a single use, is called Circulating Capital” “ Capital which exists in any durable shape and the return to which is spread over a period of corresponding duration, is called Fixed Capital1.” These terms have sometimes been used as though they were convertible with Remuneratory and Auxiliary capital respectively. But the cottage which a farm labourer has rent free is 1 Mill, Book 1. chap. vi. CAPITAL. 19 2— 2 20 BO O K I. CHAP. III. §8. Remuneratory Capital, but not Circulating. Again raw materials are Circulating capital, but not Remur.eratory. We shall presently see the importance of the distinction between Fixed and .Circulating Capital. We shall also see the importance of a third distinction for which these names are sometimes used, but for which other names seem to be required. Whenever capital has been designed for use in one trade there is some difficulty in diverting it to another; if this difficulty is great the capital is Specialised, if not great, the capital is Non-specialised1. No clearly defined line can be drawn between these two kinds of capital. Labourers’ food and clothing, many kinds of tools, and such materials as wood and metals are Non-specialised, for they can be employed in a number of different ways. . Again, many kinds of offices and commercial buildings and workmen’s dwellings in towns are Non-specialised. But agricultural improvements, including farm labourers’ cottages, cannot often be used except for their original purposes. So also railroads, docks, ironworks, and again printing and reaping machines may be considered as Specialised capital. But it is often difficult to decide whether such machinery should be called Specialised, or Non-specialised : a factory fitted up with a steam engine and “ over-head gear,” can scarcely be called Specialised capital, because it may be almost equally adapted to many branches of the textile trades, and of the lighter metal and wood trades. § 8. Almost all Personal Wealth is or may be Personal Capital. As Adam Smith says :— “ The acquisition of all useful abilities, by the maintenance of the acquirer during his education, study or apprenticeship, always costs a real expense, which is a capital fixed and realised, as it were, in his person. Those talents, as they make a part of his fortune, so do they likewise of that of the society to which he belongs. The improved dexterity of a workman may be considered in the same light as a machine or instrument of trade which facilitates and abridges labour, and which, though it costs a certain expense, repays that expense with a profit.” 1 Compare Jevons’ Theory of Political Economy, pp. 134— 234. Ç H A PT E R IV. LAW OF DIMINISHING RETURN. § i. W e have seen that the requisites of production maybe classed as land, labour and capital. We have now to seek for the Law of fertility of land, the Law of the increase of population, and the Law of the growth of capital. The latter two Law s depend on the first, which goes by the name of the Law o f Diminishing Return. We may explain the meaning of this name by an illustration. W e shall presently see that the Law does not apply to all new countries in which, though the land is fertile, the population is very sparse. But let us suppose that in a certain district there are twenty agricultural labourers to the square mile, and that in the seasons in which they grow wheat they raise 2000 quarters, which is at the rate of 100 quarters per man. If now the population increases so that there are thirty agricultural labourers to the square mile, there will of course be an increase in the produce raised, but not a proportional increase. Perhaps the produce m ay now be 2600 quarters, so that the amount due to the labour o f the additional ten labourers is 600 quarters, which is at the rate of 60 quarters per man. Now let there be a further increase in the population, till there are thirty-five agricultural labourers to the square mile : there will again be an increase in the total produce raised, but again not a proportional increase. - Perhaps the produce may now be 2850 quarters, so that the amount due to the labour of the last five labourers is 250 . quarters, which is at the rate of 50 quarters per man. That increase in the amount of corn raised which is due to the labour of each additional labourer may be called the Return due to his labour ; and we may then say, that in this case the greater the number of men employed on the land, the less is the Return which would be due to the labour of an additional man. This result illustrates the meaning of the Law of Diminishing Return. § 2. . But so far no account has been taken of the fact that 22 the cultivation of land requires the farmers capital as well as the labourer’s toil It is indeed best to look at the matter from the point of view of the farmer : for since he generally advances the wages of the labourer, the produce may be regarded as due to his capital— to his Auxiliary capital in the form of implements, seed, cattle, etc. and to the Remuneratory capital with which he pays his labourers’ wages. The previous illustration represented the farmer as sending additional labourers into the farm, and then noticing the increase of produce due to their labour, or the Return to their labour. W e may now suppose that he increases by successive doses1 the capital he applies to the land, and then notices the increase due to each successive dose, or as we may say the Return due to each Dose. Suppose, for instance, that by expending in one way or another four doses of £ 1 each, on an acre of land he can raise from it 20 bushels of wheat, which is at the rate of five bushels for each dose of capital. If he applies a fifth dose of £ 1 to the acre, whether in the form of extra manure or extra ploughing or what not, the produce of course will be increased, but not by as much as 5 bushels. Probably he may now obtain 24 bushels from each acre, the return due to the fifth dose of capital being 4 bushels. If he applies a sixth dose of capital, and with the expenditure of £6 obtains 27 bushels an acre, the return due to the sixth dose will be 3 bushels, and this perhaps would not remunerate him for the outlay. If so, he will not apply more than £5 to the acre. § 3. Thus, as the experience of every English farmer tells us, when land is already well cultivated the application of additional doses of capital will cause a less than proportionate increase in the return ; or, as we may say, they will obtain a Diminishing Return. We may then state the Law of Diminishing Return thus :— An increase in the amount of capital (or, to speak more generally, in the amount of effort) applied in the cultivation of land causes in general a less than proportionate increase in the produce raised3. But this Law is subject to two conditions, one relating to the progress of the arts of agriculture, and the other to the special circumstances of a new country in which but very little capital is applied to land. We may consider these separately. 1 This term was applied by James Mill to denote equal amounts of capital applied to land. ' \* The statement that after a certain number of doses have been applied to any given piece of land, the return due to each additional dose BOOK I. CHAP. IV. § 3, 4. § 4. Firstly, with regard to progress in the arts of agriculture. New methods of cultivation are from time to time discovered, will be less than those due to the preceding doses, may be illustrated by figures. LAW OF DIMINISHING RETURN. 23 Let the line Ox be divided into equal parts, the divisions representing equal doses of capital, and let them be called OA, AB, BC, etc., each successive dose gives a return which may be represented by a rectangle at right angles to Ox, and as thick as the corresponding division of Ox. Then if five such doses, represented by OE, have been applied, the returns are shewn by the figure OPQE. The Law of Diminishing Return states that after a certain number of these doses have been applied, the corresponding rectangles are shorter. But the Law of Diminishing Return does not state that this diminution will in every case commence at once ; thi^^ums to the first few doses may be small, and the returns to the laterabses may be larger, as they are in fig. 2. All that the Law states is that after a great number of doses have been applied, the return must diminish. We may make the divisions in Ox as small as we please, so that the rectangles become what may be called thick straight lines. The tops of these straight lines will lie on curves, the shapes of which are represented in figs. 3 and 4. Fig. 3 represents the increase of produce of land, the return from which always diminishes : and fig. 4 that of land the return from which first increases and then diminishes. new rotations of crops, new methods of draining land. Science ,gives instruction in the feeding and breeding of cattle, and in the'-application of different manures to different kinds of land. Machinery is being invented which lightens the work of agriculture— weeding, mowing and reaping machines ; steam ploughs, steam threshing and chaff-cutting machines. Meanwhile, the education and energy of the farmer and labourer are increasing. The knowledge of new inventions and methods is diffused by agricultural shows and farmers’ newspapers. Again, when a country becomes very thickly peopled it has great facilities for organising a supply of food from abroad, and the emigration of its people to other countries. Thus the progress of civilisation, while it presses on the resources of land, enlarges those resources. [§ 5. Secondly, with regard to the special circum stances o f a new country. T h e Law o f D im inishing Return has been disputed b y M r C arey, who is at the head o f a set o f Econom ists ca llin g them selves the A m erican school. M r C arey’s argum ents a gain st the L aw appear to be founded on a m isconception o f its re a l nature; and the practical conclusions for which he contends have not m uch application to a densely peopled country such a s E n glan d is. But yet the writings o f the A m erican school a re in m any w ays suggestive, and their opinions have attracted m uch attention on the continent o f Europe, so that it m ay be well to say som ething about them here. These economists argue that history shews that the best \* lands are not those which are cultivated first, but that the order of settlement of new countries is the passage from poorer to richer soil. The causes of this are various. Mountain districts have sometimes been selected on account of the means of defence which they offer against enemies ; but more often the steep and self-draining mountain sides were chosen because the low rich lands, until they are drained, are infested by malarious fevers. In fact, if land is very rich, full of luxuriant undergrowth or marshy, it is not possible to cultivate it at all with only a small expenditure of capital and labour. But when the growth of population and the advance of civilisation give the means of bringing such land under the plough, the return which it gives will abundantly repay the pains that have been spent on it. The tasks of draining marshy lands and freeing them from malaria, and of making roads and railways, are not easily performed when the population is thin and scattered. And American writers have done good service in insisting on the fa ct. that up to a certain point, the greater the numbers in a country the greater will be the power of organizing labour and capital, and the greater therefore will be the return from land. It may 24 BOOK I. CHAP. IV. § 5, 6. LAW OF DIMINISHING RETURN. 25 be conceded to them that until this point is reached, land may be said to yield an Increasing, not a Diminishing Return ; and that perhaps more than half of the richest land in the globe is yet uncultivated. But this fact is not inconsistent with the Law of Diminishing Return which merely asserts that the return to capital applied to land diminishes, provided that there is already a dense and rapidly increasing population, and comparatively little improve\* ment in the arts of cultivation. One cause of this difference of opinion between American and English Economists is probably that in America there is, and in England there is not a large amount of rich land waiting to be cultivatedJ § 6. We may next consider the Law of increase of mineral produce. The Law of Diminishing Return is said to apply to mineral as well as to agricultural produce. This is not strictly true: at least there is a fundamental difference between the two cases. The richness of cultivated lands is likely to increase: so that if the demand for produce were to remain stationary, it could be satisfied with continually diminishing effort But every mine is being impoverished by being worked. And when the richest mineral strata have all been discovered, the difficulty of satisfying the demand for mineral produce must increase, even if the demand should remain stationary. Improvements in the arts o f mining may retard the operation of this Law, but cannot entirely prevent it. Another difference between agriculture and mining is this. It is impossible to raise by any amount of labour ten times the ordinary produce from a well-cultivated garden in one year. But the produce raised in a year from a mine might be increased tenfold without in c a sin g more than tenfold the labour employed on it ; provided that there were a sufficient supply of mining plant and skilled miners. But in fact the difficulties and dangers of mining deter those who have not been accustomed to them from childhood. A rapid increase in the supply of coal can only be obtained by paying high wages to new men for unskilful work. Again, the machinery, shafts and galleries that are required for opening up new seams cannot be provided without some delay. It is owing to these causes that a sudden increase in the demand for coal forces its price up very high. These causes would act even if there were an unlimited supply of good mines ready to be opened ; they are independent of the Law of Diminishing Return. The gradual exhaustion of our coal mines has very little to do with the sudden rise in the price of coal in 1873. This rise was caused by a sudden increase of demand ; as is shewn by the facts that it attracted a large amount of capital and labour to tbe coal mining trade, and the consequent increase in the supply of coal forced down I prices below their old level. But though the Law of Diminishing Return does not apply | to mineral produce in the same way as it does to agricultural, yet the exhaustion of the mineral strata of the earths surface may I ultimately exert a most serious influence on the history of the world. When steam power becomes expensive, science will doubt- . less teach us to substitute for it the forces of the air and \* water. In the wind and waves, in tides and waterfalls, we are supplied with forces many thousand times as great as can ever be required by the whole population of the globe. At present coal is preferred to these sources of power, because coal can 1 easily be carried wherever it is wanted and any quantity of | steam power can be obtained whenever it is wanted ; but the forces of Nature cannot. Means will however probably be discovered of transmitting the power that is supplied by Nature from place to place, so that it may be applied wherever it is | wanted, and of storing up this force in reservoirs so that it may be used whenever it is wanted ; and then we shall be able almost to dispense with the use of steam power. But there seems to be no prospect of obtaining any important cheap substitute for. coal as a means of producing warmth. And it is quite possible that in the colder portions of the earths 1 surface the growth of population will ultimately be restrained by the need of warmth, more than by the need of food. 1 126 BOOK I. CHAP. IV. § 7. 14 C H A P T È R V. GROWTH OF POPULATION. MALTHUS. POOR-LAWS. § I. ~ T h e Law of Diminishing Return tells us that when population has reached a certain density, an additional amount o f labour and capital will not raise a proportionately increased supply of food. The operation of this Law is delayed by the progress of the arts of agriculture and manufacture, and by, bringing fresh land under cultivation. It is possible that when the whole world is well cultivated, it may afford support for five or even ten times as many people as there are acres in the earths surface. But a limit to the growth of population must b e reached at last. The surface of the globe, including sea and land, is about 600,000,000,000,000 square yards. If we suppose that each yard allows standing room for four persons, this calculation gives room for 2,400,000,000,000,000 persons. Next, looking at the rate of increase of the population of England and Wales, we find that it doubled between the years 1801 and 1851. At this rate of increase population would in 100 years multiply itself by 4 200 ,, „ ., 16 4°° » y> 99 256 500 „ „ „ (1024, say) 1000 1000 „ „ „ 1,000,000 2000 „ „ „ 1,000,000,000,000 3000 „ „ „ 1,000,000,000,000,000,000. At this rate the descendants of a single pair would in 3,000 years form a solid column covering the surface of the globe more than eight hundred deep. These facts shew that sooner or later the growth of population must receive a check ; but they do not shew that it need be checked at present. The question whether there is such a need is a difficult one, and we shall be occupied during the greater part of the present chapter in considering it. § 2. In savage countries people marry very early, and the growth of population would be very rapid, if it were not re 28 BOOK I. CHAP. V. § 3, 4. strained by infanticide and war, by pestilence and by famine. ! When men have left the savage state and live under settled , government, they very often retain the habits of improvidence | of the savage. In fact they almost always do, unless they are j educated, and have some sort of ambition for themselves and I their children. Not only the Chinese and other Asiatic races, : but the agricultural labourers in Ireland and even in some ) parts of England marry recklessly without any Jthought of the j morrow. For indeed their future is a matter neither of hope j nor of fear to them ; they have little hope of bettering their condition, and there is scarcely a lower level to which they can fall. As a result so many children are born that the Iresources of the family will not suffice to nourish and educate j them : and though in Europe at least infanticide is rare, and 1children do not often die of actual starvation, yet the rate of j mortality among the children of the very poor is terrible. 1Insufficient food and clothing, neglect, dirt, foul air and \* infectious diseases hurry off vast numbers of the children of |the poorer labourers in town and country to an early grave. 1Skilled artizans are in a transitional state. We may hope |that before long they will not marry without that sense of responsibility with regard to the education of their, children which I the middle classes feel now. This change may be described by saying that their Standard of Comfort is rising. For :— When any class of the population has obtained such habits of forethought as to be unwilling to marry with- | out the expectation of being able to enjoy a certain given amount of the necessaries, comforts, and luxuries of life, then this amount is called the Standard of Comfort for that class of the population. These necessaries, comforts, and luxuries are for a man’s children as well as for himself ; indeed the chief of them is a . good physical, mental, and moral education for his children. Economic progress depends much on changés in the Standard of Comfort of the people,^and therefore oh the strength of their family affections. ^ § 3. If there be a rise in the incomes of any class of the j people, we may expect an increase in the number of marriages and births among this class. It is in fact well known that marriages are more frequent among all classes of the people in a period of commercial prosperity than in one of depression ; and the statistics published by the Registrar-General during the last thirty years prove that there have been as a rule more marriages in the years in which bread has been cheap, than in those in which it has been dear. \* It may however happen that a rise in the incomes of any class, is accompanied by a rise in their Standard of Comfort ; G R O W T H O F P O P U L A T I O N . 29 which prevents any increase in the number of births among them. But a rise in the Standard of Comfort brings with it improved means of caring for and supporting infant life, so that out of every hundred children that are born, a larger number grow up to be efficient workers in the next generation if the Standard of Comfort is high than if it is low. We see then that the Law of Population is that :— A rise in the average income of any class of the people causes either a rise in their Standard of Comfort, or an increase in the number of marriages and births in that class. A rise in the Standard of Comfort is almost sure to increase the percentage of children who grow up. Therefore a rise in average income almost always increases, and a fall almost always diminishes the rate of growth of population. As a rapid growth of population has almost always followed on a rise of wages, Economists from the time of Locke have assumed that the wages of unskilled labour have never been much more than sufficient to enable them to continue the species. But Adam Smith’s sagacity remarked that it is quite possible for the labouring classes to change their definition of necessaries. Shoes are necessaries in England, not in Scotland. Wheaten bread was a necessary in England before the reign of Henry VII. ; after that the working classes ate a great deal of rye-bread ; but again at the end of the seventeenth century wheaten bread became a necessary. At the present time the definition varies much from one country to another, and even from one part of the same country to another. § 4. Malthus published his famous Essay on Population in 1798. Perhaps no other book has been so much discussed by persons who have never read it as this Essay has. It may be worth while therefore to say something here about it and the circumstances under which it was written. A t the end of the last and during the early years of the present century, the English poor-law acted in such a way as to foster the growth of population upward indeed in point of quantity, but downward in point of quality. It put a premium on early and reckless marriages, at the same time that it discouraged not only every kind of thrift and forethought, but also all strength and manliness of character. “ Under it the English peasant had a legal right to support in the shape of a poor-rate, doled out to him when his wages were judged insufficient. Children called for no self-denial on his part, the rate was doled out to him in proportion to their number. A labourer could not by deferring marriage earn a little portion to start with, or at least such a course would have required extreme self-denial ; for while unmarried and childless, the payment of his service 30 BOOK I. CHAP. V. § 4 ,5 . | was depressed to the sum just sufficient for his absolute neces- I slty. Like a slave or like a horse, he had his food nearly the same at every time, and work as he might, was unable to reserve I a store for his needs1.” T h e consequences o f these m istaken arrangem ents w ere shew ing them selves in w ide-spread m isery ; and M alth u s, a , kind-hearted clergym an, set h im self to w ork to inquire w hether | after all it w as right thus to prom ote the in crease in the , quantity o f the population without caring for the quality. T h e I conclusion at w hich he arrived is this :— T h e natural propensities o f m an w ill cause population to increase as fast or faster than the m eans o f supporting them ; therefore the efforts o f the I legislator and o f the m oralist should be directed tow ards im - j proving the character o f the people ; and towards d iscou ragin g I rather than encouraging a rapid increase of numbers. j Accordingly he urged “ moral self-restraint” as a “ Preventive Check” to the excessive growth of population. But what he has \* said on this subject has been much misunderstood. It has | been thought that he wished to impose great hardships on the i poor. But in fact “ he again and again explained to inquirers | that he wished to put no more restraint on the marriage of the poor than such as every prudent parent in the middle classes would wish to place on that of his children. He advises delay ! in marriage, and he argues that, without strong and lasting | attachments, the married state is generally more productive of 4 misery than of happiness, and that for the formation of such I attachments, time must be given to find out kindred disposi- ] tions1.” Now it seems easy to object to Malthus, that the world has I gone along very well for thousands of years without men’s troubling themselves about his Preventive Checks to population ; ! and that therefore the world may be expected to go on well for j many more thousand years, even if men pay no attention to ] Malthus. This objection he anticipated and met by one of the i most crushing answers that patient and hard-working science has | ever given to the reckless assertions of its adversaries. He ^ examined one after another the various countries of the world i in ancient, mediaeval and modern times. He found that in ; some places Preventive Checks had been actually in operation for many centuries; that in some parts of Switzerland, for instance, large numbers of the population remain unmarried. i He found that in many other countries the hand of man had I habitually restrained population by infanticide and war; and that in every country which had been inhabited for many centuries, and in which man had not controlled population by < 1 F. W. Newman, Malthusianism true andfalse. M A L T H U S . • 31 these means, the stern hand of nature had kept it down by what he called Positive Checks, that is, by want and famine, by infant mortality, and by terrible disease and pestilence. So that it is not a fact that the world has gone on very well hitherto : and Malthus argued that the faculties of thought and forethought were bestowed upon men to enable them to make the world go on better. [H e used the expression that population tends to increase indefinitely in a geometrical progression, and that under favourable circumstances it was found to double in 25 years, while food only increased in arithmetical proportion, and that therefore population must increase faster than food. The expression “ arithmetical” is slovenly and no one defends it now, but he intended to express what is stated in the Law of Diminishing Return ; viz. that if labour continued to increase at a great rate for a long time, the produce which it could obtain from land would not go on increasing at the same rate.] § 5. Malthus\* statements with regar.d to the misery that has existed in past ages have been confirmed by more recent historians : but the practical conclusions that he deduced from them are more liable to be disputed. For he could not foresee the inventions and discoveries which were just about to be made when he wrote. He could not foresee how the growth of steam traffic would enable England, on the one hand to import food from countries where there was a scanty population; and on the other to send out her surplus population to cultivate new soils, and to spread the energy and genius of the English people over the earth. There can be no doubt that this extension of the English race has been a benefit to the world. A check to the growth of population would do great harm if it affected only the more intelligent races, and particularly the more intelligent classes of these races. There does indeed appear some danger of this eviL For instance, if the lower classes of Englishmen multiply more rapidly than those which are morally and physically superior, not only will the population of England deteriorate, but also that part of the population of America and Australia which descends from Englishmen will be less intelligent than it otherwise would be. Again, if Englishmen multiply less rapidly than the Chinese, this spiritless race will overrun portions of the earth that otherwise would have been peopled by English vigour. It must be remembered that the growth of population depends not on the number of those who are bom, but on the number of those who grow up to maturity : that infant mortality is the natural consequence of improvident marriages; and that the character of the population of a country will be lowered ir 32 BOOK I. CHAP. V. §6.. people marry early on insufficient means, and have families so large that, though they can just rear them, they cannot j properly care for their physical, moral and mental education. I The best practical answer to the question which we have ; set ourselves to discuss in this chapter seems to be :— Just as a man who has borrowed money is bound to pay it back with interest, so a man is bound to give his | children an education better and more thorough than \* he has himself received. I When people are in a position to do this they confer a benefit on the State by marrying. This practical principle measures equitably the same measure of public duty to rich and 1 to poor. Its general adoption would cause the spectre of j Malthusianism, which casts a gloom over economic speculation, I to disappear, or at any rate to be no longer dreaded ; and Iwould rid us of the competition for food which seems to dog !the heels of progress. § 6. It will be convenient here to pause in our discussion |of the Laws of increase of the three requisites of production and i to make a short digression on the question which Malthus raised as to the proper administration of the Poor-law. At present the management of the rates that are levied for the relief of the poor is entrusted to Guardians of the Poor. These Guardians, with the aid of paid officials, decide what aid | from the rates is to be given in each case. This relief is either ^ Indoor, that is within the doors of the Parish workhouse, or j Outdoor. As Bastiat would say, the evils of Indoor relief are ' “ seen” ; those of Outdoor relief are often “ not seen.” Indoor relief is unpopular ; it resembles imprisonment, and seems too hard a fate for those whose poverty is not the result of positive vice. When a man goes into the workhouse, his home is generally broken up, so that he cannot easily leave it and i start in life afresh. j It seems therefore at first sight much better to give Outdoor j relief to those who are not hopelessly infirm. Outdoor relief | may be eked out by any little earnings of the man or his \* family, and perhaps by some assistance from relations or other | friends. He rubs along in his old abode and with his own furniture: the cost of relieving him in his home is thus, even while it lasts, less than the cost of Indoor relief ; and, as soon as he is again able to earn considerable wages, the Outdoor relief can be stopped. Thus in its immediate result Outdoor relief is less burdensome to the rate-payer, as well as more agreeable to those who receive it, than Indoor relief is. But Outdoor relief leads to great evils. If given in aid of < wages, it is likely, as experience has shewn, to cause wages to be so low that labourers cannot live without support from the POOR LAW S. 33 rates. And since 1834 it has been generally agreed that Outdoor relief should not be given to labourers who are in full health unless to enable them to tide over some passing emergency. But even with this restriction, Outdoor relief is often the booty of the idle, the dissipated, the crafty and the hypocritical ; for the Guardians, and Relieving officers employed by them, have not the time required for investigating properly the merits of ) each case ; and they have not as yet entered into systematic alliance with private distributors of charity. Thus while the honest poor suffer great hardships, those who are dishonest often receive Outdoor relief at the same time that their wants are abundantly provided for by private charity. And on the whole, it has been found that wherever Outdoor relief has been given freely, a large part of the population has become idle, thriftless, and base, in short “ pauperised.” Yet Poor-law Guardians are under a great temptation to ■'\* give Outdoor relief, because its evil results are distant, while the expense and the hardships of Indoor relief are present. It has therefore been thought right to offer an inducement to Guardians to apply “ the Workhouse test ” ; that is, to offer only Indoor relief in ordinary cases, so that the inability of the applicants to provide for themselves might be tested by their willingness to submit to the restraints of the Workhouse. An Act was passed on Mr Goschen’s motion in 1870, by which the expense of the v Workhouses throughout the whole of London was borne by a Metropolitan rate. Each body of Guardians now knows that the whole of the Outdoor relief, but only a small portion of the Indoor relief which they grant, must come from the pockets of the rate-payers by whom they are elected. This Act, while it made the rich districts of London to contribute their due share to the support of the poor, diminished the pauperism of London, which till then had continually increased. It is to be hoped that the principle of this Act will be applied f throughout England ; and that the expense of workhouses will be thrown upon counties, or other large areas, while each district bears the expense of its own Outdoor relief. Whether this is done or not, it should be a rule that when a man applies for Outdoor relief, the burden of proving that he deserves it lies with him. It should be assumed that he does not deserve it unless he can shew either, that though he did not marry recklessly, it had never been possible for him h to save ; or that he had made some effort to provide against the day of misfortune by subscribing to a provident society or by saving in some other way1. There are indeed some who think that every change in the poor-law should aim at the ultimate abolition of Outdoor relief. 1 See Mr George Bartlett on Thrift as a test of outdoor relief, M. 3 34 BOOK I. CHAP. V. § 7. But the deserving poor feel, and ought to feci, great anguish when they are forced into the workhouse. When a man has not undertaken the responsibilities of marriage without a fair chance of being able to provide for his children, when he has led a hard-working unselfish life and has saved to the utmost of his | power, but has been weighed down by accumulated misfortunes ; every hardship, that is imposed on him needlessly, is an 1 unjustifiable cruelty. It is true that the abuses of Outdoor relief ( are at present so great that it should be abolished if they could 1 not be diminished. But it has not been proved that it is impossible to separate the deserving from the undeserving poor. § 7. It is indeed scarcely too much to say that the means by which alone the task coiild be accomplished have not yet 1 been fairly tried. It seems clear that paid officials have not the I time required for detecting the falsehoods of the unworthy. Still less have they time, even if they had the other qualities required for getting to the bottom of the sad tale which is \* unwillingly and shrinkingly told by those who have nobly strug- | gled against misfortune. “ If the poor are to be raised to a permanently better condition, they must be dealt with as individuals and by individuals ; for this hundreds of workers are necessary ; arid this multitude of helpers is to be found amongst volunteers— whosfe aid, as we arrange things at present, is to a great extent lost. The problem to be solved, therefore, is how to collect our volunteers into a harmonious whole—the action of each being free, yet systematized ; and how thus to administer relief through the united agency of corporate bodies and private individuals ; how, in fact, to secure all the personal intercourse and friendliness, all the real sympathy, all the graciousness of individual effort, without losing the advantage of having relief voted by a central committee, and according to definite principes1.” !There are in England a large and increasing number of people who are ready to take part in the work ; who have the leisure, the means, the education and the will required for it ; “ on all sides we hear of people willing to give their time, if only they could be sure of doing good. They are dissatisfied, they say, with district visiting, which creates so much discontent and poverty, and does so little lasting good ; they want to know of some way in which their effort may fit in with more organized work....With our volunteers, home claims must and should come first ; and it is precisely those whose home claims are deepest, and whose family life is the noblest, who have the most precious influence in the homes of the poor. But if the work is to be valuable, we must find some way to bind together broken scraps of « time, and thus give it continuity in spite of changes and breaks.” 1 Homes of the London Poor, by Miss Octavia Hill, p. 113. POOR LAWS. It seems then that the case of every applicant for relief should be decided on by a committee, who may be the Poor-law Guardians, or a Volunteer Committee acting in harmony with them ; and who should receive three reports. The first should be from a paid officer, who Should make it his study tb learn all the guiles of the dishonest, and who, before reporting on the case, should have asked as a matter of business and without reserve all questions necessary for his purpose. The second should be a result of the organization of the various charitable societies in the neighbourhood, and should enable the Committee to know for certain what aid the applicant is already receiving. Thé third should come from a volunteer visitor who has gone into the applicant’s history, and has drawn forth by sympathy whatever good there was in him. The Committee might find that the case should be referred to some private charitable society ; or that relief could \* be given in some better form than a money grant. In many places there are so many charities that, if they were properly organized, Outdoor relief would be rendered almost if not quite unnecessary. But if the Committee decided that the case of the applicant was not met by any existing fcharity ; that he had been prudent, self-denying and industrious ; and that he ought to receive Outdoor relief; then such relief might be given without fear that it would tend to pauperise the people. Very little Outdoor relief would be given under such a system ; but deserving people would not be forced into a workhouse, as they often are now, in consequence of their case being misunderstood1. Miss Octavia Hill has done much to shew the need and the feasibility of some suçh plan as this for diminishing the evils of pauperism, and has also thrown much light on the problem of preventing pauperism by improving the homes of the poor. She suggests the following rules for those who would share in this work: “ It is best to enforce fulfilment of all such duties as payment of rent, etc. “ It is far better to give work than either money or goods. “ It is most helpful of all to strengthen by sympathy and counsel the energetic effort which shall bear fruit in time to come. “ It is essential to remember that each man has his own view of his life, and must be free to fulfil it ; that in many ways he is a far better judge of it than we, as he has lived through and felt what we have only seen. Our work is rather to bring him to the point of considering, and to the spirit of judging rightly, than to consider or judge for him. “ The poor of London (as of all large towns) need the development of every power which can open to them noble sources of joy.” 1 Portions of such a system are in work at Boston, and at Elberfeld. 35 3 - 2 C H A PT E R VI. GROWTH OF CAPITAL. § i. We have discussed the Laws of increase of two elements of production, viz. the Law of increase of production from land, and the Law of the increase of labour. The next step is to examine the Law of increase of the third element of production, that which supports and aids labour, viz. capital. The growth of capital depends upon the power and the w ill to save. The power of saving depends on the amount of wealth out of which saving can be made. Some countries, which have a large population and produce a great amount of wealth, have very little power of saving. The whole continent of Asia, for instance, has less power of saving than England has. The total produce indeed of its industry is larger than that of England ; but the number of people among whom this is divided is so great that they are compelled to consume almost the whole of it in supporting life. As Mill says, 4 the business fails and ceases. Everything depends on the correctness of the unseen decisions,' on the secret sagacity of the determining mind.” § 3. Next with regard to the economy of machinery. JuSt as there is waste whenever a skilled man is engaged on work in which his skill is useless, so there is waste whenever a machine lies idle. A blacksmith could sometimes find use for a steam hammer, but there would be so much waste in keeping it idle the greater part of the day, that he cannot afford to have it. Many things continue to be made by hand which would long ago have been made better and more cheaply by machinery, if there had been a large demand for them. For the trouble, expense and risk involved in inventing and patenting a new process, or a new machine, is the same whether the invention is to be applied to the production of many goods or few. But the profit to be derived from the invention depends on the extent of the manufacture to which it can be applied. The substitution of machinery for handwork has sometimes to wait for a brilliant invention by which a great practical difficulty can be overcome : but most new machines are really adaptations of old ones, which are sure to be made as soon as there is a demand for them. In fact, the great advantage that the human hand has over machinery is that the hand can bring tools to work upon the material in almost any position : and many important mechanical inventions are merely new plans for copying the way in which the tools and the material are adjusted to one another by the 5'i 4 - 2 52 BOOK I. CHAP. V III. § 4— 7- hand. This is sometimes done by giving new movements to the tools that are driven by a machine, but more often it consists in making some new fixed or moving socket for holding firmly the material exactly as it is wanted. § 4. We have seen that the advantages of division of labour cannot be obtained in the production of any commodity unless it is one for which the demand is very great, so that it is produced in very large quantities. We have next to inquire how far these advantages are dependent on the size of the factories in which the work is done. This question is a very important one, because the hope of rising in the world is one of the chief inducements to energetic action, and to thrifty habits. The prosperity of the nation depends greatly upon how far the upper classes are recruited by the best strength and ability that is born among the lower classes. And one of the easiest and healthiest routes upwards which a working man can follow is that of first saving a little money while working as foreman or overlooker, and then starting a small establishment in that branch of his business of which he is a master. W e shall find that some of the advantages of division of labour can be obtained only in very large factories, but that many of them, more than at first sight appears, can be secured by small factories and workshops, provided there are a very great number of them in the same trade. The manufacture of a commodity often consists of several distinct stages, to each of which a separate room in the factory is devoted. But if the total amount of the commodity produced is very large, it may be profitable to devote separate small factories to each of these steps. If there are many factories, large or small, all engaged in the same process, Subsidiary Industries will grow up to meet their special wants. § 5. Firstly, there are the industries which make the special tools and machinery required for the process. There are, for instance, large works in which the machinery wanted for the woollen trades, is made by a vast number of complex machines. Secondly, there are the Subsidiary Industries which facilitate communication between various branches of a trade. Their aid is important to all classes of factories, but especially to those which are devoted entirely to one stage of a process of manufacture. Under this head come carriers, railway companies, and all classes of agents and intermediate traders, and again those who collect and spread information by trade newspapers, and in other ways. Bankers also come under this head. The payment for goods is facilitated by them as the transfer of goods is by the railways. But further, they pass command over capital from hand to hand, and the help which they thus give to new men, who have but little capital of their own, is perhaps the most important of the forces which oppose the modern tendency towards the concentration of manufacture in the hands of a few large firms. But small factories, whatever their numbers, will be at a great disadvantage relatively to large unless many of them are collected together in the same district. We may then consider the advantages of the Localization of Industry. § 6. The Localization of Industry promotes the education of skill and taste, and the diffusion of technical knowledge. Where large masses of people are working at the same kind of trade, they educate one another. The skill and the taste required for their work are in the air, and children breathe them as they grow up. This is seen particularly in such manufactures as those of glass and pottery. Again, each man profits by the ideas of his neighbours : he is stimulated by contact with those who are interested in his own pursuit to make new experiments ; and each successful invention, whether it be a new machine, a new process, or a new way of organizing the business, is likely when once started to spread and to be improved upon. In a district in which an industry is localized a skilled workman is sure of finding work to suit him ; a master can easily fill a vacancy among his foremen ; and generally the economy of skill can be carried further than in an isolated factory however large. Thus both large and small factories are benefited by the localization of industry and by the assistance of subsidiary trades. But these benefits are most important to the small factories, and free them from many of the disadvantages under which they would otherwise labour in competition with large factories. § 7. Still a large factory has many special advantages. Firstly, greater economies can be attained by a large than by a small factory in such matters as the arrangement of buildings, steam engines, and other machinery ; and again in such work as that of clerks, doorkeepers, stokers, repairers of machines, &c. One high chimney can make a draught for a large furnace as well as for a small one ; one doorkeeper can admit five hundred men as easily as fifty. Again, a large factory can often afford to have a machine to do work that is done by hand in a small factory. It is true that a small factory devoted to one short stage of the process of manufacture may have the best and most highly specialized machinery. But such a factory would not come into existence until the advantage of having special machinery for this stage had become well established. If there is a large demand for any kind of machine it may be worth a man’s while to spend DIVISION OF LABOUR. 53 54 BOOK I. CHAP. V III. § 8, 9. much money and trouble in trying to make it, with the intention of patenting it, and getting a royalty on each machine that is made. But in fact a very small number of the improvements ^ that are made are patented ; and a large manufacturer has , greater inducements than a small one to make experiments I which are not certain of success. He can spend a large sum on trying a new process or a new kind of machine, without feeling the loss if the experiment fails. If it succeeds, he is j much more likely to get a good return on his outlay before others are able to take advantage of it than a smaller manufacturer would be. But very often the most important advantages to large firms ^ consist in their facilities for buying and selling. The expense jinvolved in buying a large quantity of goods is always smaller 'in proportion than that of buying a small quantity. A large Ifirm gains in the transport of goods ; particularly if it has a Jrailway siding. It gets prompt and trustworthy information from highly paid agents ; and thus learns when and from whom to buy, when and what to manufacture, when and to whom to sell, and lastly whom to trust and whom not to trust. It can ,afford to advertise largely in print and by means of commercial "Ttravellers. Something may also be allowed for the confidence |which consumers feel that a large house will not descend to jpetty tricks and dishonesties ; it has too much to lose. | In some trades a large firm gains much by the variety of i its wares. A builder, who wants a great number of brass fittings, likes to order them of a firm which has a vast number of moulds and can fill up a large and various order from its own resources. A large firm gains often in the matter of superintendence ; but not so much as at first sight appears. It is true that in small works the master or manager looks after 1many things which in large works are left to the care of a 4 foreman or overlooker; but hired overlookers have not the \*same interest and energy in preventing waste through careless- jness or dishonesty that the master himself has. There are Isome trades, particularly those concerned with the more valu- - able metals, in which large firms are on the whole at a disadvantage as compared with small ones in the matter of j superintendence. '§ 8. There are many branches of trade in which the advan- -4 tages of division of labour and specialization of machinery seem to have reached their limit. In cotton spinning, for instance, a large factory contains many rooms which are exactly alike in all respects ; so that it resembles several smaller factories joined together. It has then no great advantage over smaller factories ]in the work, of production, but it has those general advantages in buying and selling and organizing which a large capital must DIVISION OF LABOUR. always have over a small one. When the owner of such a factory wants to put more capital into his business, it may answer his purpose to put up looms for making his yam into calico. He then has an advantage over those who only spin, because he has no trouble about finding a market for his yarns ; and he has an advantage over those who merely weave, because he has no trouble about bargaining for and testing the quality of the yarn that he uses. What used to be the work of several distinct trades is all done now under one roof. • This change is partly due to the modem facilities for the growth of firms which manage vast sums of capital. The number of very wealthy men has increased rapidly of late ; and the Limited Liabilities A cts1 have enabled those who are not rich, and who perhaps have no time for business, to take shares in large trading concerns, without imperilling their whole fortunes. This has led to the formation of vast companies in almost every branch of business. Some of them undertake enterprises for which no private capital could suffice. The London and North Western Railway Company wields a capital greater than the whole accumulated wealth of many of the states whose names occupy a large place in history. There does not seem to be any limit to the amount of wealth which a single trading company can profitably manage in a business which can be managed by routine, and does not require the bold enterprise and prompt decision of a single mind. But this great tendency to the concentration of capitals is opposed by the increasing variety in the number of things that hâve to be done, and in the modes of doing them. Industries subsidiary to the old established industries are springing up continually, and when they become well established, other industries subsidiary to these make their appearance. The growing variety of the wants of man, the growing resoiirces of invention by which they are met, continually make openings by which new men edge their way into business. A glance at the Trade Directory of London or almost any other large manufacturing town will discover an astonishing variety of trades which are almost entirely in the hands of small masters. « § 9. Whatever may be the result of the contest between large and small factories, it appears certain that the division of labour will continually increase. This increase is one of vital importance. It adds to men’s power over nature, and furthers social progress by increasing wealth. Its effects are for the 1 These Acts, passed in 1855— 1862, enable a Company by writing “ Limited” after its title to secure its shareholders against losing more than the amount of their shares by its failure. 55 56 \* BOOK I. CHAP. V III. §9 — i i . most part good, but in some part evil. Division of labour is said to increase the uncertainty of industry. A man whose skill can be turned to account in only one trade is likely to suffer much when that trade is depressed, or his skill is displaced by machinery. On the other hand, division of labour accustoms producers to sell in widely distant markets, and it is not very likely that all of them will be depressed at the same time. Again the division of labour sometimes enables a man to pass easily between trades which used to be totally distinct. A country watchmaker could not easily become a gunmaker or vice versa : but many of the men in a large watch factory could easily find employment in a large rifle factory, and vice versa. When the late American war came to a close, a famous rifle factory devoted itself to making sewing machines. As we proceed further, we shall find other reasons for thinking that whatever tendency there is to a growing uncertainty of industry's to be ascribed to other causes than the division oTlabour. But division of labour does unquestionably cause some monotony. This is a very great evil in the case of work which involves continued muscular strain or long hours of work in a bad atmosphere. But when the work is light, and the hours of work not excessive, monotony is not very injurious. A s Mr Nasmyth says1, “ If you call for the brute force of a man you will degrade the man. He goes to his house so physically exhausted that it is an utter absurdity to say to that man \* Read and improve yourself/ He would fall asleep immediately : he must go and take some excitement. But if you take the man who has been superintending some piece of machinery all day, in which there is very little or only a minimum of call for his brute force, you will find that that man’s intellectual power, if he has any at all, will come forth, and he will be a reader and a self cultivated man. I have found that again and again. I think this is the result of machinery, that it takes away the necessity for brute labour, and very much elevates the intellectual and moral position of the working classes. The management of machinery requires much judgment and resource.” The experience of Englishmen in backward countries shews that costly and delicate machinery cannot be worked profitably by a dull and ignorant people, however low may be the rate of wages which they are willing to take. The work of the mechanic, even where it is apparently monotonous, makes him .shrewd, cautious and prompt. Again even when division of ' labour makes the work of an individual monotonous and uniform^ it makes the work of the country at large changeful and various. The worker in a town 1 Tenth Report of Trades Union Commissioners, 1868, p. 65. DIVISION OF LABOUR. whose mental and physical energies are not strained by his work can hardly fail to be educated by the variety and excitement of the various work that is going on around him. His neighbours have interests in life sufficiently like his, and yet sufficiently different, to enable him to learn from them new ways of looking at things. • There is more division of labour in the town than in the country ; but the agricultural labourer lags behind the town workman in intelligence. The advances that have recently been made in agricultural science are chiefly due to the mental activity of towns-people. § io. It will be useful to refer to the Law of Division of Labour, which may be stated thus :— When the demand for a commodity becomes very large, the process of making it is generally divided among several distinct classes of workers, each with its proper appliances, and each aided by Subsidiary industries. It leads the way to the Law of Increasing Return, which is :— The Division of Labour tends to diminish the difficulty of making a commodity, and therefore to increase the return obtained by a given amount of effort. The total difficulty of producing a commodity of any kind is generally governed by the concurrent action of the Laws of Diminishing and of Increasing Return. Of these two conflicting influences sometimes the one preponderates, sometimes the other. The former preponderates in such a case as that of blankets, in which the cost of the raw material is great and there is little room for further economies in manufacture ; while watchmaking and other light metal trades afford good instances of the preponderance of the Law of Increasing Return. § ii. The largest industry is that of agriculture ; but there is scarcely any other industry which is ableto make "Éo little use of the advantages of division of labour and of production on a large scale. For agricultural labourers cannot be grouped together in large masses ; they must be scattered over the country. And each season of the year has its special work: a man cannot spend his life in reaping. So that the work of agriculture cannot be broken up into a vast number of parts each of which is performed by a band of labourers who devote their lives to acquiring a special skill in this class of work. Agriculture, however, seems to be following in the steps of manufacture. Field steam-engines are becoming common, and new machines to be worked by them or by horse power are appearing in rapid succession. The fields demand every day a smaller number of dull labourers and a greater number of intelligent mechanics. 57 58 BOOK I. CHAP. V III. § 12. This change is exercising an important influence in the competition between small and large farms. The small farmer cannot always afford to have a field steam-engine; he cannot afford to have a great number of machines for occasional use. Thus every year puts him at a greater disadvantage relatively to the large farmer. This disadvantage is diminished but not removed by the rapid growth of a subsidiary industry, which undertakes steam ploughing threshing, &c. for farmers. The growth of this industry is the most important step towards obtaining the advantages of division of labour that has ever been made by agriculture. In comparison with a small farmer a large farmer gains something in economy of buildings, and in economy of materials. He is able tlTfiave a better rutation of crops J"1iu ran suid-a great many labourers into a field in which there is anything to be done quickly. He can, as a rule, borrow capital from the banks more easily than a small farmer can. Lastly, the large farmer is likely to have more knowledge and greater skill and enterprise than the small farmer. He probably received a better education at starting; and he can afford to leave to subordinates much work that the small farmer does himself, so that he has more time and opportunity for increasing his knowledge. And as farms change hands from time to time, the ablest farmers are likely to find their way to the largest farms. Thus the economy of skill is carried further under a system of large, than under one of small, farms. On the other hand the large farmer loses in the matter of superintendence. The small farmer works hard himself : he watches for every trifling gain and every small saving : and those who work under him have little opportunity of being idle or dishonest. § 12. These advantages enable spade husbandry to be as successful on small holdings as on large. A large market garden has few advantages over a small one except in buying and selling. An intelligent market gardener who cultivates a few acres by the labour of himself, his family, and perhaps one or two hired labourers, can pay a high rent for his garden, and yet earn very good wages for himself and very good profits on his capital. Similarly a well managed small vineyard in Southern Europe seems to be able to hold its own against the competition of larger vineyards. Each vine has its own history, the soil is seldom exactly of the same character for many consecutive yards; and the cultivator of a large' vineyard could not carry in his head many of the trifling details that guide the action of him who devotes all his life to one little plot of ground. But the greater part of the small holdings on the continent have generally the advantage of being owned by those who DIVISION OF LABOUR. cultivate them. It has already been remarked1 that the peasant proprietor has in land a constant source of pleasure and excitement and the safest and most convenient of savings-banks. He invests his capital and labour in his land, without requiring as high a profit on his capital as the wealthy farmer would, and without expecting as high wages for his toil as would be demanded by the hired labourer. He may through want of machinery, or of knowledge retain methods of cultivation that have been discarded as wasteful by the English farmer. But still he is content and happy. His produce is generally less in proportion to the amount of labour spent in raising it than is the case with an English farm. But his produce per acre is often large, and on the whole he contributes his full share to the agricultural wealth of the country. 59 1 Chap. vi. § 3. C H A P T E R IX. TENURE OF LAND. § i. A GREAT deal has already been said incidentally about the tenure of land. In tracing the gradual organization of industry, we noticed how in very early times land was no man’s property how tribes of savages wandered over it and supported themselves by hunting ; how, afterwards adopting a pastoral life, they drove half domesticated herds slowly from one pasture to another. We saw how, when agriculture appeared, land became the property of village communities. Recent historical research has shewn that in almost every part of Europe, in many parts of Asia, and probably in some other places, the land of each such village community was generally divided into three parts, or to use the Teutonic name, into three Marks. The Town Mark contained the houses, which were the private property of the several families who lived in them. The Arable Mark was divided into three fields, one of which was left fallow each year, the other two being cultivated. Each family had for its use a lot in each field : so as always to have its proper share of land in cultivation. In most countries the lots were periodically redistributed. The remainder of the land was the Common Mark. This was not cultivated, and each family had equal rights of pasture and of cutting wood in it. The system of village communities somewhat modified prevails in Russia and India in the present day. It has some advantages; extreme misery is seldom found in it; men lead peaceful and contented, but monotonous lives. The community watches jealously to prevent any one from adopting methods of cultivation that are opposed to its interests, or even to its habits TENURE OF LAND. 61 and prejudices. Thus grows up in the course of time a network of customary rules, which hampers the freedom and enterprise of individuals, and hitherto has been found to check and hinder agricultural improvement of every kind. In Western Europe this system was transformed into the military system of Feudalism by the wars and conquest of the Middle Ages. In Feudalism the notions of ownership and government were so blended that the sovereign was regarded as having a kind of ownership of the land. His subjects held it from him on condition of rendering him military service when required. Gradually the rights of the Sovereign to land have fallen into abeyance, and private persons have now practically undisputed possession of it, but even to the present day bargains about land are not determined by free competition in quite the same way as bargains about other things. Each nation has special laws, customs, and sentiments with regard to the transfer and tenure of land. § 2. A large part of the Continent is owned by Peasant Proprietors. We have seen how the peasant loves his land as his friend, how gladly he invests his earnings in it ; how he gets to know the history of every square yard of it. He may not know of, or may be unable to afford the advanced methods of the rich English farmer. But in some kinds of cultivation he excels ; and even though he does not generally turn his labour to the best account, his untiring zeal often raises a large gross produce. In some portions of Southern Europe the Metayer tenure prevails. The metayer has an hereditary right to cultivate a small piece of land on condition of giving a certain portion of the produce, generally one half, to the landlord. The landlord supplies the whole or a part, according to local custom, of the capital required for working the land1. The metayer resembles the peasant proprietor in having fixity of tenure. But in some other respects he resembles the.members of a village community. For he is harassed and hampered by the rules which have grown up to secure the landlord’s share of the produce ; and since he retains only a fixed portion of the fruits of his labour, he has not as strong an incentive to exertion as the peasant proprietor has. • The land of America is cultivated by those who own it2. 1 With regard to Metayers, Irish Cottiers and Peasant proprietors, the reader is referred to Mill, Bk. n., and Cliffe Leslie’s System of Land Tenure. \* “ The American farmer is at least in nine hundred and ninety-five cases out of a thousand the owner of the land he cultivates.” Mr Ruggles\* report published by the New York Chamber of Commerce. 1874. Ô2 BOOK I. CHAP. IX. § 2. The ease with which men get land prevents the growth of any I considerable class of agricultural labourers at present, so that the tenure of land in America resembles in many ways that of peasant proprietors. But the farms are not small, and the ^farmers are generally educated men, full of restless energy. | They frequently sell their farms and move westwards to larger ' farms or richer soil; they are always on the look out for improved machinery and improved methods of cultivation. And ' in many other ways they present a striking contrast to the rpatient and unenterprising peasant proprietors of Europe, among I whom the land and the method of cultivating it descend with little change from father to son. JThe share of the produce that the metayer pays to the v N landlord is sometimes called “ rent.” But in this book the word will always mean that payment which the owner of land can obtain by free competition for lending out the use of it to others. Rich land affords a larger return to the capital employed lon it than could be obtained by applying the same amount of capital to the cultivation of poor land. In a populous country in which there is a great demand for food, and in which there- jfore some food has to be raised at great expense from poor ^land, the value of the produce raised from the rich land will be w more than sufficient to pay the expenses of raising it This ^ surplus value the owner retains if he cultivates the land himself. iBut in England and some other countries there are always capitalists willing to cultivate the land with their own capital, i>iand.to pay this surplus to the landlord in the form of rent. This system has the great advantage of giving the management of the land to those who have capital, agricultural skill and liking for the work. We have already seen that the advantages of division of labour and of productions on a large scale are of less importance in agriculture than in manufacture: but their ^importance is increasing. The progress of the arts of agri- jculture brings with it a continually increasing demand for capital and for highly trained agricultural skill. And the pro- gress of the nation in wealth and intelligence increases the number of able farmers who have a considerable command over capital. The combined action of all these causes is increasing the average size of farms, and raising the status of the farmer. In Scotland and some parts of England there prevails a system of long leases which secures to the farmer nearly the whole benefit of his skill and energy. If his lease is short he receives but scanty protection from the law. But custom so far shields him that he is seldom in great danger of losing all the benefits of the improvements he has made in his farm, through having his rent raised, or being ejected without compensation, by an unscrupulous landlord. TENURE OF LAND. The Irish Cottier pays a rent for the use of his land and cultivates it at his own risk. But here his resemblance to the English farmer ends. The Irish Cottier is a poor and uneducated peasant who rents a small plot of land either directly from its owner or from a middleman who makes a living by subletting land. The ignorance and recklessness of the Irish peasant and his inherited thirst for land often induce him under the stress of competition to undertake to pay a rent higher than he can pay. Some of the smaller landowners and many of the middlemen grasp at such promises ; and then the cottier finds idleness his best policy ; thrift his worst. He has no Standard of Comfort, no inducement to prudence in marriage ; and population is restrained chiefly by poverty, disease and famine. The misery of the cottiers cannot be removed without first removing the causes of their recklessness. This recklessness is to some extent— it is a matter of controversy to what extent — due to the bad legislation of our forefathers. There is good reason to believe that it is being diminished under the wise legislation of recent years. The discussion of the question as to which is the best system of land tenure has been complicated by some uncertainty as to the meaning of the term “ the best system.” By the best system some mean that which gives the greatest gross produce, others mean that which gives the greatest net or surplus produce after deducting the necessaries of life for the labourers, while others again mean that which contributes most to man’s general wellbeing. We have seen reason for thinking that the greatest net produce is on the average obtained under the system of large farms ; and that the largest gross produce is obtained in some of those districts in which there is an intelligent and energetic race of peasant proprietors ; for their untiring zeal keeps on applying more labour to the land long after the return from it has diminished so far that a capitalist farmer would have ceased from further cultivation. Economists are not agreed as to what system best promotes general well-being. If a vote could be taken from all Economists throughout the world, it would probably be given in favour of the system under which the land is owned by its cultivator, whether in large farms as in new countries, or in small plots as in old, and this view is adopted now by many Englishmen. But nearly all English Economists of the past generation had a strong preference for the system of large farms. This is partly due to the same causes that have promoted the employment of large capitals in English manufactures1, partly to the fact that 63 1 See Book 1. ch. viL § 4. 64 BOOK I. CHAP. IX. §3. England’s position in the Great French war was largely owing to the vast net produce of her manufactures and agriculture1. § 3. Enough has been said now to shew that when land is let on the English system its rent is the surplus return which it gives to the farmer’s capital after deducting what is necessaiy \ to replace his capital with profits. It is easy to see that this rent will be increased by anything that enables the farmer to raise a greater produce with a given outlay, or by anything that gives him a better market for his produce. But it is more difficult to analyse the various parts of his outlay ; to measure the total value of his produce ; and finally, by subtracting from this value enough to return him his outlay with profits, to determine the amount of his rent. We shall be in a better position for doing this when we have examined the first principles of the theory of value. 1 For instance Ricardo sàys, “ Adam Smith constantly magnifies the advantages which a country derives from a large gross, rather than a large net income....Provided the net real income of a nation, its rent and profits be the same, it is of no importance whether it consists of ten or of twelve millions of inhabitants. Its power of supporting fleets and armies, and all species of unproductive labour, must be in proportion to its net, and not in proportion to its gross income.” BOOK II N O RM AL V A L U E . C H A PT E R I. DEFINITIONS. LAW OF DEMAND. § I. In the present Book we are to inquire into the influence that competition exerts upon wages, profits, and prices. O f course competition is only one out of many causes by which they are determined. In backward countries competition exerts'but little influence men do not forecast the future, and deliberately shape their course by a calculation of distant advantages.' They rather drift along under the influence of custom, doing the same work in the same way, and for the same remuneration as their fathers did it. But in advanced countries, particularly in Western Europe, Northern America and Australia, competition is far the most important of all the influences that affect wages profits and prices. And it will therefore be our best plan to begin our examination of wages profits and prices by inquiring how they would be determined if free competition were the only influence acting upon them. In Book III., when we are discussing Market values, our attention will be directed chiefly to other influences. But in this Second Book we have to shut our eyes to those other influences as far as we can, and think only of the effects which would be produced in the long run by free competition, if every one were quick to seek out and follov M. 5 66 his own economic interests. W e have in this Book to examine Normal Values. For:— That condition of a thing which would be brought about by the-undisturbed action of free competition is called its Normal Condition. § 2. [That which is according to the laws or precepts made by authority is legal: That which is according to the Laws of nature in their ordinary operation is normal. Of Course everything in nature happens, in one sense, according to the Laws of nature. But very often we have in view some particular set of Laws ; and when it is understood that special reference is made to these Laws, we may say that that condition of a thing which is brought about by their undisturbed action, is its normal condition. Every tree grows according to the Laws of nature, but if a tree is planted in such a position that it cannot grow “ naturally,” that is according to the laws of its own nature, its shape is said to be abnormal. So it is in one sense according to the Laws of nature that the branches of a tree are swayed by the wind. Vet we say that they are in their normal position only when the wind is still. This is the position which they adopt, and in which they rest when not disturbed. When we speak of Normal values, or Normal prices, or Normal wages, or Normal profits, the particular set of Laws which we have in view are those Laws of human nature and human conduct which are brought into play when competition is perfectly free. It is true that when man is influenced by custom or prejudice, or when he is prevented by ignorance or apathy from competing freely, his action is according to the Laws of his nature, and is in one sense natural. But it is not according to those Laws of his nature which we have specially in view when discussing the economic condition of highly civilised countries. We therefore do not call his action Normal in the special or technical sense in which the word is used here. If the wind blew equally from all quarters the mean or average position of a branch would be the same as its normal position. But if the wind blew more towards the east than towards the west, the mean position of the branch would be to the east of its normal position. And in the same way with regard to value : if competition were perfectly free, the average value of a thing would be the same as its Normal value. But since in fact competition is not perfectly free, the average value of a thing may differ from its Normal value ; although the two seldom differ much. A IvOWu\*. Adam Smith and the older Economists spoke of the “ Natural” rates of wages profits and prices. They used the word Natural to mean that which is according to man’s nature when competition is free,. But it has been found best to use BOOK II. CHAP; I. §2,3. D E FIN ITIO N S. 67 "N orm al” for this purpose, because the word " Natural” has been used loosely : men often call an arrangement “ Natural ” merely because they approve it, and without taking the trouble to examine whether the Laws of Nature actually tend to bring it about.] § 3. Before entering upon the theory of Normal value it will be well to define some words which will be used in it. The growth of the spirit of competition causes men to look about them to see where they can buy cheapest, and where they can sell dearest. A man may not trouble himself much about small retail purchases : he may give half a crown for a packet of paper in one shop which he could have got for two shillings in another. But it is otherwise with wholesale prices. A manufacturer cannot sell a ream of paper for six shillings while his neighbour is selling it at five» For those whose business it is to deal in paper know almost exactly the lowest price at which it can be bought, and will not pay more than this. The manufacturer has to sell at about the market price, that is at about the price at which other manufacturers are selling at the same time. When the competition among dealers is perfect, there can be but one price in the same market : so that we may say A Market for a ware is a place where there is such competition among buyers, and also among sellers, that the ware cannot have two different prices at the same time. "Originally,” says Mr Jevons, " a market was a public place in a town where provisions and other objects were exposed for sale ; but the word has been generalized, so as to mean any body of persons who are in intimate business relations and carry on extensive transactions in any commodity. A great city may contain as many markets as there are important branches of trade, and these markets may or may not be localized. The central point of a market is the public exchange, mart or auction rooms, where the traders agree to meet and transact business. In London, the Stock Market, the Corn Market, the Coal Market, the Sugar Market, and many others are distinctly localized ; in Manchester the Cotton Market, the Cotton Waste Market, and others. But this distinction of locality is not necessary. The traders may be spread over a whole town, or region of country, and yet make a market, if they are, by means of fairs, meetings, published price lists, the post office or otherwise, in close communication with each other.” That is, they will all pay the same price for the same thing in the same place. When it is said that the price is the same for all the dealers in a market that extends over a large district, each buyer is supposed to pay extra the expense of its being delivered to him. 5 - 2 68 \* BOOK II. CHAP. I. § 4 - 6 . § 4. There are two ways in which a person may regard anything that he has. He may consider its use to himself, or he may consider what he can obtain in exchange for it. From the one point of view he looks at what Adam Smith calls its value in use, from the other at its value in exchange. The value in use or utility of a thing to a person is the amount of pleasure or satisfaction which he derives from possessing it. The value in exchange or exchange value of a thing is the power of purchasing other goods which its possession conveys. “ Value” standing alone always means value in exchange, and never value in use. Value implies a relation between commodities. F or instance, the value of a pound of beef may be its power of purchasing three pounds of sugar or five grains of gold. If sugar were to become scarce or gold were to become more plentiful, the value of a pound of beef might become the power of purchasing two pounds of sugar or six grains of gold. Such a change would not justify us in saying that the value of beef had fallen or risen, but only that it had fallen relatively to sugar, and risen relatively to gold. But if we found that a pound of beef would exchange for a greater quantity of almost every other commodity than before, we might then say that its value had risen. Its value would have risen relatively to commodities in general ; it would have obtained a greater purchasing power over commodities in general. In every civilised country some commodity is chosen as a medium of exchange and a measure of the value, or general purchasing power, of other things. This commodity is generally one of the precious metals, gold and silver ; coins of fixed weight are made of these metals and stamped by Government, and are the money of the country. The number of them for which a pound of beef can be exchanged represents its value and is called its price. If beef becomes scarce, and rises in price, while the prices of all other things remain unchanged, there is a rise in the general purchasing power or value of beef. Of course the value of money itself may have fallen ; that is, each coin may purchase less of all other commodities than before. In that case, though the price of beef has risen, its value or general purchasing power may have remained stationary ; the greater price for which it is sold may only give the power of purchasing the same amount of other commodities as before. In a later chapter1 something will be said of the changes that may occur in the purchasing power of money. But while 1 Book ill. chap. 1. D E FIN IT IO N S. LAW OF DEM AN D. 69 examining the theory of Normal value we shall, for convenience, assume that the purchasing power of money remains unchanged. So that a rise or fall in the price of a thing will always mean a rise or fall in its general purchasing power or exchange value. § 5. If things were exchanged for one another without the use of money there would be no distinction between buyers and sellers; but this distinction makes its appearance as soon as money comes into use to represent general purchasing power. A buyer is one who wishes to obtain a particular commodity and offers in exchange for it a certain amount of money, that is, of command over commodities in general. A seller is one who is willing to part with a particular commodity that he has in his possession, and demands in exchange for it a certain amount of money, that is, of command over commodities in general. The price of any commodity in a market is determined by the eagerness of buyers on the one hand, and the eagerness of sellers on the other. The remainder of the chapter will be occupied in explaining the Law of the eagerness of buyers, or the Law of Demand. § 6. It is a matter of common experience that the larger the stock which sellers determine to sell, the lower will be the price at which it can be got rid of. Vice versâ, the lower the price at which anything is offered for sale, the greater is the amount of it which can be sold off. Examples occur to us every day. In a good apple year, the price of apples is low ; in a bad year it is high. At the end of the season a fashionable shop sells off at a great reduction, and so gets many customers. These facts shew how the Utility of anything to a man, its power of satisfying his wants, depends partly upon the quantity of things of the same kind that he has already. The more he has of it the less will be the utility of more of it to him. Suppose he wants to buy some flannel. If he could not get it for less than five shillings a yard he might be willing to buy a single yard at this price: that is, the Value in Use or utility of a yard to him may be greater than the satisfaction that he could obtain by spending the five shillings in any other way. But we may suppose that he is able to get flannel at one shilling a yard, and that at this price he buys twenty yards. This shews that the utility to him of the twentieth yard is not less than the satisfaction he could get by spending the shilling in other ways ; but that the utility of a twenty-first yard would be less than this satisfaction. In other words a shilling just measures the utility of the twentieth yard, the final yard which he buys. To 70 BOOK II. CHAP. I. § 7. use Mr Jevons’ happy phrase, the Final Utility of a yard of flannel to him is measured by one shilling. In speaking of the utility of a commodity we must always have in our minds some particular amount of that commodity and the particular person to whom it is useful. The utility of a commodity to any one depends on the amount of it he has at the time, and the opportunity he has or expects to have of getting it, or other things that will serve as substitutes for it. But further, the price which he is willing to pay for a thing depends not only on its utility to him but also on his means; that is, the amount of money or general purchasing power at his disposal. A greater utility will be required to induce him to buy it if he is poor than if he is rich. A shilling is the measure of less pleasure to a rich man, than to a poor one. A rich man in doubt whether to spend a shilling on a single cigar, is weighing against one another smaller pleasures than a poor man, who is doubting whether to spend a shilling on a supply of tobacco that will last him for a month. The clerk with ^100 a year will walk into business in a much heavier rain than the clerk with £300 a year; for a sixpenny omnibus fare measures a greater utility to the poorer man than to the richer. If the poorer man spends the money, he will suffer more from the want of it afterwards than the richer would. The utility, or satisfaction, or value in use that is measured in the poorer man’s mind by sixpence is. greater than that measured by it in the richer man’s mind. If the richer man rides a hundred times in the year and the poorer man twenty times, then the utility of the hundredth ride which the richer man is only just induced to take is méasured to him by sixpence ; and the utility of the twentieth ride which the poorer man is only just induced to take is measured to him by sixpence. For each of them the Final Utility is measured by sixpence ; but this Final Utility is greater in the case of the poorer man than in that of the richer. § 7. The lower the price that a man has to pay for a thing, the more of it is he likely to buy. A fall in price will not indeed make every purchaser increase his purchases. It might in the case of sugar ; it would not in the case of carpets. But a fall in the price of carpets would induce some of the householders in a large market to buy new carpets ; just as an unhealthy autumn increases the mortality of a large town, though many persons are uninjured by it. For in a large market there must be some who are doubting whether to replace an old carpet by a new one ; and their decision will be affected by a fall in the price of carpets. There will not be any exact relation between the fall in price and the increase of demand. A fall of one-tenth in the price may increase the sales LAW OF DEMAND. 7i by a twentieth or by a quarter, or it may double them. But in a large market every fall in price will cause an increase of demand. The Law of Demand then is The amount of a commodity which finds purchasers in a market in a given time depends on the price at which it is offered for sale ; and varies so that the amount demanded is increased by a fall in price and diminished by a rise in price. Its price measures its Final Utility to each purchaser, that is, the value in use to him of that portion of it which it is only just worth his while to buy. C H A P T É R IL LAW OF SUPPLY. § i. It was said in the last chapter that the exchange value or the price of a commodity is determined by the eagerness of buyers on the one hand, and the eagerness of sellers on the other. The Law of Demand is a general statement of the action of buyers. The Law o f Supply, to which we now turn has to do with the action of sellers. The Laws concerning the action of sellers fall into two distinct classes, according as the commodities which are sold can be produced freely by anyone, or are monopolized. Deferring the consideration of the latter, we may at present confine our attention to things the production of which is free to all, those which anyone may produce. We have to inquire how a man calculates the price which will remunerate him for producing a commodity. The inquiry is difficult ; and therefore it will be best to consider first the simple case of a producer who makes things with his own hands. Let us suppose then that a carpenter is doubting whether it is worth his while to make boxes for sale. In calculating the price which will remunerate him, he must allow for wear and tear of tools, for the price of material, for interest on the capital he has invested in tools and materials, and for rent of his workshop. Next he must allow for his own wages. Of course he does not pay himself wages. But he calculates the wages he could earn by other work that is as much to his taste and not more fatiguing. It will not be worth his while to produce the box unless its price will repay him the first mentioned expenses, and afford him such remuneration for his labour as shall not be less than these wages. This fact may be expressed by saying that it will not be worth his while to produce it, if its price is less than its Expenses of Production. We must carefully examine the meaning of this term. LAW OF SUPPLY. § 2. The production of a commodity requires the use of tools, machinery, workshops, etc.; it also consumes raw material and labour. Thus it requires capital and labour ; and capital, as we have seen, is the result of labour and abstinence. Now there are two ways of estimating these various elements. Firstly, we may regard them as so much labour and abstinence. The carpenter’s work in making the box involved muscular exertion and fatigue; and abstinence from immediate consumption was necessary in order to provide the tools with which he worked. Next with regard to the wood of which the box was made. It may have been grown on land that paid rent ; but in order to avoid considering rent at present, we may suppose that it was got from some of the wild forests of South America. In this case there was clearly no rent paid for it; but we must reckon the labour of those who cut the tree down, and of those who brought it to the carpenter. And so on. Then , again there is another set of efforts and abstinences of which account must be taken. The efficiency of the carpenter’s work in making the box depends not only on the exertion which he undergoes at the time of making it, but also upon his skill in carpentering. And to acquire this skill, he and his teachers had to exert themselves ; and his parents had to use self-denial to pay the expenses of his education. In fact his skill is Personal Capital, which owes its existence to labour and abstinence. Thus then the production of the box is the result of a great number of efforts and abstinences undergone by different people at different times. It will be convenient to have a collective name for all these efforts and abstinences reckoned together. This is found in the term “ Cost of Production.” The Cost of Production of a thing then consists of the efforts and abstinences required for producing it. The carpenter, however, in deciding whether to make the box or not, would not care to examine all these efforts and sacrifices; he would decide in a much easier way. He would calculate what it is proposed to call its Expenses of production. He would, as we hg.ye seen, want to know what prices the various efforts and sacrifices in question would command in the open market ; he would want to know what price he would have to pay for his material, what wages he could obtain for his own labour, what was the rate of interest at which he could borrow such capital as he wanted, and so on. These various sums of money when taken together may be called the Expenses of Production of the box. It has already been seen that the work of man in production consists really in moving things and rearranging them ; so that the work of taking a box to market is as much productive as that of making it The Expenses of pro73 74 BOOK II. CHAP. II. §3,4. duction o f a box when offered for sale in any market include therefore the expense of carrying it to that market. W e may then say:— The Expenses of Production of a thing in a market are the sum of the prices of the effort^ and sacrifices which are required for its production there : or, in other words, the sum of the expenses which would have to be incurred by a person who should purchase them at their market prices. § 3. The carpenter, would decide to make boxes if the price which he could obtain for them covered their Expenses of production ; and other carpenters in similar circumstances would do the same. Thus the supply of boxes would be increased ; and the increased competition of sellers would lower the price of boxes towards their Expenses of production. If many carpenters thought that the demand for boxes would continue to be large relatively to the supply, and that therefore the price of a box would long remain greater than its Expenses of production ; then they would not only keep to the bpx-making trade themselves, but also bring up their sons to the work. By so doing they would increase the supply of boxes, and then perhaps the increase might be so great as to force down the price below the Expenses of production. If the price falls below the Expenses of production, carpenters will seek every convenient opportunity of turning to some other occupation. Those who were specially skilful in boxmaking would indeed lose the benefits of this skill if they sought another employment. But some would gradually leave the trade, and carpenters would not bring up their sons to it. So the supply of boxes would be lessened ; the competition of sellers would diminish and that of buyers would increase until the price was again brought up to the Expenses of production. § 4. The carpenter then working with his own tools, raw material, etc., calculates the Expenses of production of a box, including his own remuneration among them, that is, including what he could earn in other ways by labour that was equally difficult and equally attractive to him. If now we look at an employer who has men to make boxes\* we find that he will calculate the Expenses of producing a box very nearly as the carpenter would. He would have to arrange for making payments not only for his raw material, etc., but also for the wages of those whom he employed. And he would have to allow for the remuneration of his own labour in managing the business — his own Earnings of Management1, as we may say—just as the carpenter had to allow for the remuneration of his own work. 1 These are sometimes called “ Wages of Superintendence.” LAW OF SUPPLY. At a later stage we shall see that there is a market for business ability very much as there is a market for carpenters’ work ; that an employer gets to know the value of his own time, the Earnings of Management which he may fairly expect to get, almost as exactly as a carpenter knows what wages he may expect. He generally i reckons these Earnings together with the interest on his capital ; ! he calls the two togetherprofits ; ” and expects to make a certain rate of profits in his business. He expects to have returned to ? him with profits all his Circulating capital, such as raw material, which is consumed in a single use, and the wages which he pays | away. And he expects to receive profits, together with an allow- ! ance for Depreciation, on his Fixed capital1. Depreciation in- \* eludes the wear and tear of his buildings, machinery, etc., and | their tendency to become obsolete through the progress of invention, and the changes of trade. The interest which he expects to receive, depends not only on I the amount of capital invested, and the annual rate of interest ; but also on the time which elapses between his making each outlay on the production of goods and his receiving the price of these goods. Let the rate of interest be five per cent, a year, w If then he spends ^ioo in the production of a thing one year \* before it is ready for sale, there will be a corresponding sum of £5 to be reckoned under the head of interest in its Expenses of production. But if he has to spend the ^100 two years before the thing is ready for sale, the corresponding sum to be reckoned under interest would be ^10; or, at compound interest, something over ^10. In deciding for instance whether to make some boxes, the manufacturer calculates the price that he will get for them, and he calculates their Expenses of production. He makes a definite allowance for the remuneration of his own labour just as the j carpenter does for his. He reckons in these expenses not only the outlay of money that he will make, but interest on this outlay together with his own Earnings of Management, or more strictly, profits. If he sees his way to getting a price that will , cover these Expenses of production, and therefore give him k adequate profits, he is content and continues his production. If he sees his way to getting a still higher price, one which will afford him exceptionally high profits, he strains every nerve and borrows more capital in order to extend his business and increase ^ his supply. But if he expects that the price will be less than these Expenses of production and therefore not afford him adequate profits he checks his supply, and perhaps begins to think about directing his capital and his energies to some other branch of production. 75 1 See Book 1. ch. 111. § 7. 76 B O O K IL CHAP. IL § 5 - 7 . § 5. Thus we see that the interest of every producer of a commodity is always to calculate the amount of it that is being produced for market If this amount seems likely to be small so that its price will rise above its Expenses of production, he will produce as much as he can, so as to derive as much benefit as possible from the high price which he anticipates. If, on the other hand, the amount brought to market seems likely to be so great that its price will fall below its Expenses of production, then he will check his own production so far as he conveniently can. So that if the price of a commodity is likely to be higher than its Expenses of production, it is the interest of each producer to do what he can to increase the supply; and the effect of this is to lower the price towards its Expenses of production. And if its price is likely to be lower than its Expenses of production, it is the interest of each producer to do what he can to check the supply ; and the effect of this is to raise the price towards the Expenses of production. That is to say :— Every producer of a commodity calculates the j>rice at which he will be able to sell it, and the Expenses of producing it. He thus determines whether to increas.e or diminish his production. If there is free competition, that is, if he is not acting in combination with other producers, he increases or diminishes his supply according as the price of the commodity seems likely to be greater or less than its Expenses of production. Thus he is led by his own interests to act in the same way as he would if his only object were to regulate the amount produced so that it could just be sold off at a price equal to its Expenses of production. This Law of Supply may be called the Law of Normal Supply, because it refers to the results that are in the long run brought about by free competition. Of course all the Expenses of production of a thing are themselves liable to variation. Wages may rise and fall, the rate of interest may rise and fall ; and so on. But we shall presently find that with some exceptions these Expenses are themselves governed in the long run by Economic Laws ; and we shall presently get to understand what is meant by the Normal wages, the Normal rate of interest and so on. The Expenses of production with which we are here concerned are these Normal Expenses. § 6. In the above illustration it was assumed that if the trade of box-making became exceptionally profitable, a great many people would begin to make boxes, or would make more boxes than before : that is, it was supposed that the box-making trade is open to free competition. We might have supposed that the boxes were of a peculiar LAW OF SUPPLY. kind and patented, so that only one person had the right of making them: or that all the box-makers combined together, and not only kept new comers out of the trade, but also agreed among themselves as to how many boxes they would make. In these cases the supply would not be governed by the Law of Normal Supply. The supply might be small, and the price very much above the Expenses of production, and yet producers might not increase, but rather diminish the supply so as to raise the price still further. This could not happen if competition were free, because then any one who checked his production when prices were exceptionally high, would lose by so doing. For others would go on producing fast, and before long the price would fall, and his opportunity of making boxes and selling them at a high price would have passed away. Every one must make hay while the sun shines, because he cannot make it shine when he wants to. Under free competition a single producer cannot control prices; and therefore each one tries to get as much profit as he can out of the time of high prices ; he makes and sends to market as many goods as he can whenever their price is above their Expenses of production. § 7. The Laws of Demand and Supply tell us that a rise in price lessens the amount demanded and increases that supplied, and that a fall in price increases the amount demanded and lessens that supplied. This competition tends to make the exchange value such as just to “ equate supply and demand,” i.e. such that the amount that people are willing to sell at that value is equal to the amount which can find purchasers at that value. This tendency of competition to equate Supply and Demand is the central fact of the Theory of Exchange, whether applied to Normal values or to Market fluctuations of value. But in considering Normal values we have not only to inquire which of the goods already purchased will be brought to market, but to go deeper and inquire what cause determines the amounts of the things produced. We have found this cause in the Expenses of Production ; when the price is above them there is a tendency to increase supply and lower price ; when the price is below them there is a tendency to lessen supply and raise price ; when the price is equal to them there is equilibrium or rest. Thus the Law of Normal Value, or, as Adam Smith would have said, of Natural Value is :— The Normal value of a thing in any market, or, that value which would on the average be brought about by thè undisturbed action of free competition among its producers, is equal to its Expenses of production there. Whenever the value is below this level, forces are brought into play which tend to raise it ; whenever it ie 77 78 above this level, forces are brought into play which tend to lower it. The value of a commodity is in equilibrium and has no tendency either to rise or to fall when the amount produced can just be sold at a price equal to its Expenses of production. But this law is not complete because it takes no account of the fact that the Expenses of production are not fixed, but depend upon the amount produced. To this point we shall return1. It follows from this Law that things whose Expenses of production in any market are equal, tend to have the same exchange value ; and to exchange for one another in that market. For the sake of brevity it is usual to omit all mention of the place in which the exchange is supposed to be made ; but whenever we speak of the relative exchange values of two commodities we must have in our minds some one particular place in which we suppose them to be exchanged. If the value of salt is compared with that of tin, it makes a great difference whether the market to which reference is made is near the Cheshire salt mines or the Cornish tin mines. § 8. So far nothing has been said of the way in which Insurance against risk enters into Expenses of production. Insurance against the destruction and the depreciation of capital may perhaps best be included under the head of wear and tear of buildings, machinery, etc. Prudent manufacturers generally insure their premises against fire by paying to an Insurance Company a premium which they count as one of the necessary Expenses of their business : and they make a yearly allowance for the depreciation of their machinery through the constant tendency of invention to render obsolete that which is old. A prudent shipowner, if he does not pay a premium to underwriters to insure his ship on each voyage, puts by a fixed sum annually to form an insurance fund of his own. Secondly, there is the risk of having to sell the produce at a loss. For instance, iron and other things have often to be sold for less than their Expenses of production when trade is suddenly depressed ; and a change of fashion may have the same effect on articles of dress. But the Law of Normal Value makes allowance for this risk once, and if Insurance against it is made under a separate head, it is really counted twice over ; for the chance that the price of the commodity may fall below the Expenses of production is of the same kind as the chance that the price may rise very high. When the average price at which it is sold is computed, the low prices are reckoned in together with the high prices. The trade is therefore a remunerative one if this average price BOOK II. CHAP. II. §8—10. 1 See Book II. chap., v. § 3. LAW OF SUPPLY. h covers the Expenses of production) without any separate allowance being made for the insurance against the risk of a fall in price. § 9. It is important to remember that there is no necessary connection between the selling price of any individual thing and its Expenses of production ; the connection is between the Expenses of carrying on a certain process of production and the total sum that is received by selling the • products. There is no connection between the Expenses of a fishing boat on any k excursion and the price which is got for its hauL The haul may be bad, and the price got for if much below its Expenses of production, or the haul may be good and the price much above these Expenses. What the Law of Normal Value states is that the total price got for the fish must in the long run, taking good and bad hauls together, çpver the Expenses of the boat. The price of every individual thing is, as we saw in the last chapter, limited by its value in use to the purchaser. If a boat brings back only a few fish, only a small price can be got for the > catch. So again, the value in use of a bell with a flaw in it is very little; it can be used only as old metal and therefore its price is only that of the old metal in it. When it was being cast the same trouble and expense were incurred for it as for other bells which turned out sound. Its Expenses of production were the same as those of sound bells : but they have great value in use and are therefore sold at a high price. The price of each particular bell is limited by its value in use : what the Law of Normal Value states is that the price of cracked bells and sound bells together must in the long run cover the expenses ^ of making bells. ^ § 10. [The main outline of the Law of Normal Value was worked out by Adam Smith and Ricardo. They were careful to guard against implying that the price of each individual thing is equal to its Expenses of production ; but still this mistake has been made. This mistake has indeed led some people to approve a proposal made by Mr Carey, to say that the Value of a thing is equal to its Expenses (or, as he says, Cost) of reproduction. w It is quite true, as Mr Carey says, that when a new invention has very much diminished the difficulty of making, say, a steel i rail, no one will pay for an old steel rail a price equal to the expenses which were incurred in making it by the old method. He seems to think that they will always pay for it a price equal to the Expenses of its reproduction, that is of producing a similar rail by the new method. But this is not the case. I f trade has become suddenly bad, and iron-masters have many steel rails on hand, no one will pay for a steel rail a price equal to its Expenses of reproduction ; because rails are being sold for les? 79 So BOOK II. GHAP. II. § io. Again, no one will pay for a bell with a flaw in it, or for a dress that has gone out of fashion, the Expenses of its reproduction. When fashion is displacing broad ribbons by narrow, broad ribbons sell for less, and narrow ribbons sell for more than their Expenses of reproduction. Gunpowder in time of war and quinine in time of fever often fetch more than their Expenses of reproduction. If the phrase “ Expenses of reproduction” were substituted for the phrase “ Expenses of production” in our Law of Normal Value, the meaning of the Law would not be altered by the change. For the Expenses of reproduction of a thing are in the long run the same as its Expenses of production. The statements that value must be equal to Expenses of production, and that it must be equal to Expenses of reproduction, are equally false when they refer to the market value of any individual thing, and equally true when they refer to the Normal Value about which the market value oscillates. The advantages of the two phrases are so far about equally balanced : but the phrase Expenses of production has this very great advantage over its rival, that it calls attention to the way in which the difficulty of producing a thing determines supply in the first instance, and value in the second. Producers debating whether to increase their supply of a commodity do not inquire whether, the price they get for it will cover its Expenses of reproduction, but whether the price will cover its Expenses of production.] 4 C H A PT E R III. RENT. § i. IT remains to inquire how the Expenses of production of a commodity may be affected by the payment of rent for the land on which it is produced ; but before doing this it will be well to take up our account of the rent of land where we left it at the end of the first book. Rent was then defined as that payment for the use of land which the owner can obtain by free competition for lending out the use of it to others. We have now to inquire how the amount of this rent is determined in any particular case. Suppose a farmer to have ^500 which he is thinking of applying .in extra cultivation of his farm ; and to have calculated that it will only just answer his purpose to do so. He has calculated, that is, that if he applies this extra £500 he will, after paying for labour, seed, taxes, &c., get an extra net produce of the value of about say ^40 ; i. e. at the rate of. 8 per cent, on the extra outlay. This is, we suppose, just sufficient to remunerate him : so that if he expected to get less, the chance of the improvement turning out unsuccessful and the prospect of additional trouble in working it, would induce him to invest the money in railway stocks or some other securities. He hears at this time that a small adjacent farm of 50 acres is to be let, and he is asked what rent he would be willing to pay for it. His £ s ° ° would just be enough for working this farm, and he could work it with the same trouble that it would give him to apply the extra £500 to the farm he already has. He calculates that taking one year with another he may expect to get from it ,£ioo worth of net produce after paying for laboui, seed, taxes, &c. So he will pay just £60 rent for the use of this land. If he can get the land for this he will take it : but he will not give any more for it ; and it will not be likely to be worth any one else’s while to offer more. So the landlord cannot get more than this for it. If he puts up the farm to competition and plays off M. 0 82 one farmer against another, he may just get £60 ; and this is 4 then the competition rent, or as it is sometimes called the Economic rent of the farm. Many disturbing circumstances such as custom, the absence of an active spirit of competition on j the part of the farmers, generosity or sluggishness on the part of the landlord, may cause the actual rent to be less than this. But £60 is the rent that will be obtained, if there is a perfectly good market for the hire of land : that is, if on the one hand the landlord exerts himself to get the best rent he can for the land, and on the other hand there are competent men in the neighbourhood who are ready to rent farms. ’ § 2. This illustration shews us that (i) The Economic rent of a piece of land is found by subtracting from the value of its annual produce an amount sufficient to return the farmer’s outlay with profits. Of course allowance must be made for the risk of bad harvests : this is done by assuming that the harvest is an average one. It must also be supposed that the farmer has neither more ' nor less skill and enterprise than most others in his neighbourhood, or as we may say, that he is an average farmer1. The rent then is the surplus return which the land gives in an average harvest, after repaying the average farmer’s outlay with profits, provided he has applied so much capital to it as to make this surplus return as large as he can. If he has applied less than this amount of capital some one else who intends to apply more than he has done, and thus obtain a larger surplus return, may offer to pay a higher rent. Further the above illustration shews that : (ii} The amount of produce which a farmer must retain in order to be remunerated for his outlay, can be discovered by observing what amount of additional return is just sufficient to induce him, or another farmer in the same neighbourhood, to apply extra capital to his land2. 1 A farmer of more than ordinary ability ought in justice to retain for himself that portion of the net produce of the land which is due to his exceptional qualities. Such a farmer is almost certain to improve his land, and invest capital in it in various ways. His landlord has the 4 power of raising his rent above that which the land could have paid, if i it had remained in the hands of an ordinaiy farmer, by the threat of ] ejecting him without compensation for his improvements. But this is very seldom done in England. \* The first part of the theory was known to Economists before Adam i Smith’s time. It is commonly, though perhaps erroneously, supposed that he had no knowledge of this second part. Ricardo did more towards working out the second part than any one else : but he did not invent it, and the form in which he stated it is open to some exceptions. BOOK II. CHAP. III. §2. RENT. That is, it can be discovered by observing the Return to his last Dose of. capital. The return to this dose remunerates him, but only just remunerates him ; therefore if he retains for himself as many times this return as he applied doses, he will just be remunerated for applying them. Therefore to find his rent, this return must be multiplied by the number of doses he applies, and then subtracted from his total produce. The Law or Bent then is :— If a farmer applies as much capital to his land as he profitably can, his rent is what remains after deducting from his total produce the return to his last dose multiplied by the number of doses he applies1. 1 In Book I. Chap. iv. § 3 it was shewn that if the doses of capital which a farmer applies to a given piece of land are measured along the line Ox, the returns due to the several doses can be represented by the figure OPQM. It was also shewn that, on account of the Law of Diminishing Return, when a certain number of doses have been applied to the land, the returns to these doses must diminish ; that is the line PQ must gradually approach Ox. This diagram may be used to represent the theory of rent. After the farmer has applied a considerable number of doses to the land, the 83 returns will begin to diminish. A point will be reached at which the return due to the last dose only just repays the farmer for applying it. At this point he will stop. Let each dose be ^10, and let the last dose which it is worth his while to apply, be the fortieth. Suppose that this fortieth dose is represented on the figure by the point M; so that AfQ, the return due to it,, is just sufficient to repay the farmer for applying a dose of ^10. Since the return to this dose just remunerates the farmer, the returns to the other doses must more than remunerate him. For instance, SN being the return to the dose at N, draw QTR parallel to Ox cutting SN in T and OB\*in R, then N T is equal to QM, and therefore N T is that portion of the return due to the A^th dose which is required to remunerate the farmer for applying that dose ; TS is surplus produce which the landlord can claim as rent. So if we draw verticÿ lines from other points in Ox, the portions of them that lie between Ox and RQ will be those w 6—2 84 BO O K II. CHAP. III. §3, 4- § 3. It may happen that there is in the neighbourhood land for which no rent can be obtained, because the return to the capital applied to this land remunerates, but only just remunerates, the farmer. In this case we may say that The amount of produce which a farmer must retain in order to be remunerated for his outlay, is equal to the produce that could be raised by the same amount of capital from an adjacent piece of land that pays no rent, but yet is cultivated. The Law of Rent may therefore be stated thus :— The rent of a piece of land is the excess of its produce over the produce of an adjacent piece of land which is cultivated with an equal amount of capital, and which would not be cultivated at all if rent were demanded for it. Land that is cultivated, but is so poor as not to be able to pay any rent, is often spoken of as on the Margin o f cultivation. 4. Whenever the increased demand for food causes a rise in the value of agricultural produce, it causes a twofold rise in the rent of land. As the value of food rises, the amount of it which is necessary to remunerate the farmer for his outlay diminishes t and therefore the surplus portion of the produce, which the landlord can claim as rent, increases in amount. And as each part of this surplus rises in value the value of the rent is raised twice, once in proportion to the increase in the amount of the produce which goes as rent, and once in proportion to the rise in value of each part of it. In a country in which there is any uncultivated land, an increase in the demand for food causes an extension of the Margin of cultivation, at the same time that it causes a rise in the rent of land. In such cases it is sometimes said that the I rise in rent is caused by the extension of the Margin of culti- ! vation. But this is incorrect : the two changes are the effects of the same cause : viz. the increase in the demand for food. On the other hand when a diminution of the demand for food, or increased facilities for its importation cause a fall in its value, there is a twofold fall in rent For the amount of agricultural produce which is necessary to remunerate the farmer for his portions of the returns to the corresponding doses of capital that are required to remunerate the farmer. And those portions which lie between RQ and HQ will be surplus portions which the landlord can claim as rent. If now we suppose, as we did before, that these vertical lines are thick lines, so that altogether they fill up the space POMQ\ then POMQ represents the total produce of the land. In the same way ROMQ represents the farmer's share of the produce ; and PRQ represents the surplus portion o f the produce which the farmer has to pay to the landlord as rent. RENT. outlay increases; and therefore the surplus portion which the landlord can claim as rent diminishes. And as each part of this surplus falls in value, the value of the rent is lowered twice, once in proportion to the diminution in the amount of the produce which goes as rent, and once in proportion to the fall in value of each part of it. But the effect of increased facilities for the importation of food in lowering rents can only be temporary. The ease with which it enables labourers to obtain an ample supply of food increases the growth of population and wealth; so that before long the farmers will find the demand for milk, vegetables, hay, straw and other produce, which cannot easily be imported from distant CQuntries, much greater than it would otherwise have been ; and the demand for other kinds of their produce not much less. After the repeal of the Corn-laws in 1848, it was expected that the free importation of com would make rents fall ; but their repeal has caused an enormous increase in England’s wealth, and rents are very much higher than they were then. At the present time however, the great development of wheat fields and railways in the north-west of America, and the newly discovered methods of importing fresh meat, are lowering the prices of corn and meat ; and threatening to cause a fall of rents. It is difficult to estimate the effect of these permanent causes of depression because they are much overlaid by three temporary causes :— firstly a succession of bad harvests in England ; secondly (t railway-wars ” in America leading to extremely low rates of freight ; and thirdly a great falling off in the demand for meat, especially on the part of the working classes, occasioned by the present commercial depression. In a country which imports a great deal of food, an increase in its production at home will not much affect its price ; and therefore an improvement in the arts of production will make but little change in the amount of the produce which is required to return the farmer’s outlay with profits ; nearly the whole of the increase will go to the landlord in the form of higher rents. In a country which produces its own food, the immediate effect of a sudden improvement in the arts of cultivation would be to lower the price of agricultural produce, to make it unprofitable to apply so much capital as before to good land, and to throw some land out of cultivation. It can be shewn that if the fall in price did not at once lead to an increased demand for food, the part of the produce which goes to the landlords — their “ corn-rents ” — would probably fall1 ; and that 1 Mill, after Ricardo, says they would certainly fall. He assumes that the cultivated part of a country consists of three qualities of land, yielding at an equal expense 60, 80, and 100 bushels; and he then shews that an improvement which increased the return to each dose r 8S 86 their rents measured in money would certainly fall. After a time however the growth of population would cause such a demand for food that its value would rise to its old level : no more corn than before would be required to return with profits each dose of the farmer’s capital, more capital would be applied in cultivation than before the improvement, and there would be a very great increase of rents.] £§ 5. It has already been seen that the Law of Diminishing Return does not apply to mineral produce in the same way as it does to agricultural For this reason the theory of the rent of land cannot be applied directly to mines. It is true that the rent of a mine is the excess of the value of its produce over that amount which is required to return with profits the capital of those who rent the mine ; and we can determine what this amount is by discovering how much produce could be obtained by this capital if it were applied to a mine which is being worked but for which no rent is paid. (Just before a mine is abandoned it passes, as a rule, through the stage of being worked, but yielding no surplus from which rent can be paid.) But it is not true of a mine as it is of a farm that this rent is a fixed annual payment. There is a certain maximum net surplus which can be got from a farm by applying just the right amount of capital to it, and the rent is fixed so as to be equal to the value of this surplus in an average harvest. It is the business of the tenant to apply the capital requisite for obtaining this net surplus ; the landlord has no further concern in the matter, than to see that the tenant does not leave the land in a worse condition than he finds it in. And farming contracts are all subject to a general condition that the land is not to. be impoverished by the tenant. But contracts for the rent of mines cannot be subject to this condition, because a mine is impoverished whenever ore is taken from it. The owner must therefore take care that those to whom he leases it do not work it so as to raise a very large amount annually without paying him a rent equivalent to the deterioration of his property. This deterioration may be regarded as the diminution of the power that the mine has of returning to capital surplus profits over and above the ordinary trade profits : and corresponds therefore to the surplus profits which the lessees capital by one-third, would lower corn-rents in the ratio of 60 to -26$. But the distribution of fertility in a country might be such that we should rather suppose it to consist of three qualities yielding at an equal expense 60, 65, and 115 bushels; and in that case the improvement would raise corn-rents in the ratio 60 to 66|. Problems of this kind cannot be properly treated by numerical illustrations; they can be solved easily and completely by the aid of mathematics or of diagrams like that given in the present chapter. \* BOOK II. CHAP. III. § 5. \* can make by working the mine, that is to the rent which the owner can obtain from them. It is an extremely difficult problem 4o calculate -what bargain should be made in the case of any particular mine. It should secure the interests of the owner and yet leave the lessees proper freedom to decide for themselves how much capital they will apply in any year to the working of the mine, and therefore how much produce they will raise from it. In practice this is roughly done by his demanding a royalty of so much a ton on the gross produce raised, in addition to a fixed annual payment.] RENT. 87 C H A PT E R IV. RENT IN RELATION TO VALUE. § i. W e may now consider the relation of rent to Expenses of production and to value. We have seen that the Normal value of a thing, or that value which would on the average and in the long run be brought about by the undisturbed action of free competition 4 among its producers, is equal to its Expenses of production. . But this Law requires further explanation, because it often happens that different portions of the same commodity have different Expenses of production. We shall find that its price is then regulated by the Expenses of production of that portion which is raised under the most unfavourable existing circumstances. Suppose for instance that, with an average harvest, ten million quarters of com are raised in England, and that the Expenses of production of the last million quarters are at the rate of 50J. a quarter. If the farmers had expected to get less than 50s. a quarter they would not have raised these last quarters. And since they find it worth their while to raise the whole ten million, we know that they get 50J. for each of the last million. And in the same market there can only be one price for one and the same commodity. Therefore the average price of all the corn in the market must be 50s. The Expenses of production of some of the corn may have been only 30J. a quarter. The 50s. got for a quarter of this corn is divided into 3ar. which goes to the farmer, and 20s. which, as we have seen, goes to the landlord as rent. And if a person looks at this corn he may argue that its whole Expenses of production were 30?. to cover the fanner’s outlay and 20s. to pay his rent, and that therefore rent enters into the Expenses of production of this com. He would be right if he meant 89 quarter ges and thing it. But he would be wrong if he meant that the selling price of corn was\governed by the rent that has to be paid for the use\*\* of land, 'tie would then be mistaking cause for effect, and effect for cause. Rent is not the cause of a high price of com, v but its effect. The price of com must be on the average just high enough to cover the Expenses of production of that portion v of it which is raised under the most unfavourable conditions. The amount that is raised, and the price at which it is sold, are thus governed by the numbers of the population which demands com on the one hand, and by the amount of fertile v land, which is the source of supply, on the other. The price is determined by the Expenses of production of that which is raised under the most unfavourable conditions and which pays no rent. The rent is governed by the excess of this price over the Expenses of production of the other produce that the farmer raises more easily. Ricardo taught that Rent does not enter into Expenses of production. By this he meant that:— The Normal value of a commodity is equal to the Expenses of production of that portion of it which is raised under the most unfavourable existing circumstances, and which pays no rent Rent does not V determine the Normal value of produce ; but is determined by it. B 2. It has often been said that the doctrine, Rent does not enter into Expenses of production, applies to raw produce but not to manufactured. This statement requires to be interpreted carefully. For it is as true of manufactured as of agricultural produce that its price must equal the Expenses of production of that portion of it which is raised under the most unfavourable existing circumstances, and pays no rent. It may be said that every manufacturer in England pays ground-rent. But it is no less true that, every hop-grower in England pays rent. The price of hops must be equal to the Expenses of production of that portion of his crop which is due to his last doses of capital— » those doses which he is only just induced to apply to the V land,— and though other portions of his crop pay rent, this portion does not. As the hop-grower reckons how much capital it will answer his purpose to expend on an acre of land, so a man building a woollen factory reckons how much capital it .U will answer his purpose to expend on a yard of building ground. If he is building a new factory, he will weigh against one another the advantages of building it (say) three or four stories high. He may think that the advantages and disadvantages of the fourth story are nearly balanced. That is, he may think the extra cloth that he will obtain by having a fourth story, will on the whole just cover its Expenses of production including profits on the capital invested in building the fourth .story, but not including any ground-rent. He must pay the ground-rent whether he has three stories or four ; so that if he decides on having the fourth story, the extra cloth due to it will be cloth which he has only just been induced to produce, cloth produced Y under the most Unfavourable existing circumstances, cloth ^ which pays no ground-rent. It is true that a manufacturer when making up the profit and loss account of his business would count his rent among his expenses. If the ground-rent in, say, Leeds rises, a manufacturer finding his Expenses of production increased may move into the country, leaving the land on which he used to work to be built over with shops and warehouses, for which a town v situation is more valuable than it is for factories1. For he may think that the saving in ground-rent that he will make by moving into the country, together with other advantages of the change, will more than counterbalance its disadvantages. In a discussion as to whether it was worth his while to do so, the ground-rent of his factory would be reckoned among the Expenses of production of his cloth. This is true. But it is no less true that in making up the profit and loss account of the cultivation of land, the farmer’s rent must be reckoned among his Expenses. A hop-grower, for instance, may find that on account of the high rent whicluhe pays for his land, the price of his hops will not cover their Expenses of production where he is; and may abandon hop-growing, or seek other land for it; while the land that he leaves may perhaps be let to a market-gardener. After a while again the demand for land in the neighbourhood may become so great that the price which the market-gardener obtains for his produce will not pay its Expenses of production, including rent ; and so he in his turn makes room for, say, a building company.] 1 This process is actually going on. When steam came into use some industries that had been scattered throughout the country whereY ever water-power was to be had, were brought together into a tew large towns. Now there is a change in the opposite direction, and many of these towns are becoming the commercial centres of manufacturing districts. 90 BOOK II. CHAP. IV. § 2. C H A PT E R V. INFLUENCE OF DEMAND ON VALUE. § l. À CHANGE which affects the Expenses of production o f all commodities in like proportion does not alter the purchasing powei\* of one thing relatively to another; it does not alter values. For instance if interest on capital were an equally important element in the Expenses of production of all commodities, a rise or fall in interest would not affect values. But in fact interest is a very large element in the Expenses of production of some things, and a very small element in those of others. It is a large element in cases in which a great deal of Fitted capital is used, particularly if, as happens in some kinds of mining, much of it has to be sunk a long time before any return can be got for it1. Again the manufacture of lace by machinery requires much Fixed capital, and interest enters as a very important element into its Expenses of production. On the other hand the Expenses of production of hand-made lace consist almost exclusively of wages ; interest enters into them to a very insignificant extent. A fail in the rate of interest would then lower the value of machine-made relatively to hand-made lace. Another cause which may alter the Expenses of production, and therefore the Normal value of a thing, is the discovery of an improved method of making it. Another is a change in the Normal rate of wages of some class of workmen who are engaged in making it. Another is a change in the difficulty of producing it, occasioned by a change in the amount of it that is produced. The effects of this last cause will be examined in the present chapter. § 2. We have already seen that some commodities obey the Law of Increasing, and others that of Diminishing Return. For instance, the more capital and labour are applied to watchmaking, the smaller will be the difficulty of producing a watch. 1 Book IT. ch. ii. § 4. 92 BOOK IL CHAP. V. § 2, 3. A temporary increase in the demand for watches will no doubt , raise their price, thus raising the wages of labour and the profits | on capital in the trade above their Normal level. But a slow I and gradual increase in the demand will bring with it a corresponding supply of labour and capital in the trade, so that wages and profits will remain at about their Normal level. And meanwhile advantage will be taken of those various economies which arise, as we have seen, from increased division of labour and production on a larger scale. ^ Thus a gradual growth of the demand for watches will diminish J their Expenses of production and therefore their Normal value. ' On the other hand an increased supply of raw produce, however gradually it may be called forth, is almost sure to obey the Law of Diminishing Return. It may indeed happen that the new demand leads to Openings up of fresh sources of supply, and to improvements in the arts of production ; but the general progress of knowledge would in any case have done much , towards bringing about' such change^. And on the whole, an increase of demand for raw produce seems almost always to Àincrease the difficulty of producing it, and therefore to increase its Normal Expenses of production, and therefore its Normal value. Adam Smith pointed out that in an uncivilised country wood and meat and leather and milk are of low value relatively not only to manufactures, but also to com. fo r where there is an unlimited supply of land, trees can be cut down and animals can be reared without the steady work, so abhorrent to uncivilised races, that is required for the cultivation of com. Gradually, as civilisation progresses, the increasing scarcity of land, and the increase of capital and patient industry causes the Normal value of wood and \* animal produce to rise relatively to that of corn, at the same time that the Normal value of com rises relatively to that of manufactured commodities. Among some savage races and occasionally where there is a sparse civilised population, a pound of meat exchanges for less than a pound of corn. But in England even before the repeal of the Corn Laws a pound of meat was worth two or three pounds of wheat. And now that the artificial obstacles to the importation of corn are removed, 4 while the natural hindrances to the importation of fresh janimal food are not yet surmounted, the difficulty of obtaining Jwheat has fallen still further relatively to that of obtaining meat ; jso that now a pound of good meat will purchase about seven Ipounds of wheat. § 3. We are now in a position to see what is the nature of the influence which Demand exercises on Value. There is pne exceptional case in which value is determined entirely by demand. This is the case of a commodity which is fixed in I N F L U E N C E O F D E M A N D O N V A L U E . 93 amount, as for instance Raphael’s pictures. Their price depends on the desire that people have for such pictures, and the means at their disposal for purchasing them. According to the Law of Demand the price of a commodity “ measures its Final utility to each purchaser, that is, the value in use to him of that portion of it, which it is only just worth his while to buy.” The greater the number of pictures, the less will be the competition among purchasers for them, the less will be the price which measures their value in use to those who are only just induced to purchase them. In this exceptional case in which the supply is fixed, the price is determined solely by the utility of the thing ; Demand is the sole regulator of value. The opposite extreme case very rarely occurs; it is that in which the Normal Expenses of production are fixed ; that is, are the same whatever be the amount produced. In this case Normal value is determined by Expenses of production alone. The part which Demand plays here is to determine the amount produced. If no one will pay for the thing a price equal to its fixed Expenses of production, the thing will not be made ; so that even in this exceptional case Demand is a condition of value. And if there are any who will give this price, the amount produced is determined by the condition that its value in use to those purchasers who are only just induced to buy it, is measured by this price. But in the case of most commodities it is not true either that the amount produced is fixed, or that the Expenses of production are fixed ; and therefore as a rule, Demand is one, but only one of the causes that determine value. An increase of Demand increases the amount produced, and this alters the Expenses of production ; so that value depends partly on Demand, because Normal value is equal to Normal Expenses of production, and Demand is one of the determining causes of these Expenses. The Law of Normal Value1 requires then to be supplemented by the statement that:— The Normal supply of a commodity is such that its Normal Expenses of production equal the price which will call forth a demand for this amount ; and the price so determined is the Normal price. The Normal price of a commodity is therefore not fixed, but may rise or fall slowly. And we shall presently see how meanwhile the Market price oscillates with comparative rapidity up and down on either side of the Normal price ; just as a cork floating on the surface of the water oscillates quickly up and down with every passing wave on either side of the mean level of the water, while that mean level itself is rising or falling slowly with the flowing or ebbing tide. 1 Book 11. ch. ii. § 7. C H A P T E R VI. DISTRIBUTION. § I. W e have seen that the -Expenses of production of a commodity may be reckoned in such a way that rent does not enter into them. The remaining Expenses may be classed as wages, profits, the expense of raw material, and other Circulating capita], the wear and tear of Fixed capital, and taxes. For instance, the Expenses of making a box, putting aside rent and taxes, may be classed as firstly the wages of the carpenters who made it and the profits of the manufacturers by whom they were employed ; secondly the price of the wood, hinges, lock and other Circulating capital that was used up in making it ; and thirdly the wear and tear of the buildings, machinery and other Fixed capital employed. But all these may, as a rule, be classed as wages and profits. For the repairs of Fixed capital and the outlay of Circulating capital may be divided into wages, profits, the repairs of Fixed and the outlay of Circulating capital ; and these Expenses can again be analysed in the same way ; and so on. For instance, the price of the lock, or the expenses of repairing the planing machine used in making the box, may be resolved into the wages and profits of those who made the lock or repaired the machine, the price of the Circulating capital they used up, and the wear and tear of the Fixed capital they employed, and so on. It is true that some of the expenses incurred may be monopoly prices of certain things ; as for instance the price of the lock if it is patent : while others may be taxes ; and: these cannot be analysed into wages and profits. But though such exceptions must not be overlooked, and though in a few cases they are of great importance, we may yet say that the Expenses of production of a thing can generally be divided into wages and profits. These wages and profits include between them the remuneration of the efforts and abstinences required for producing it. We have now to consider how these efforts and abstinences are measured by their several remunerations ; that DISTRIBUTION. is how the Cost of production of a thing is measured by its Expenses of production. We shall thus see what is the influence that Cost of production exerts on value— not directly indeed, but indirectly— Cost of production affecting Expenses of production, and Expenses of production affecting value. § 2. That part of the net income of a country which is derived from a farm is not the whole value of the crops, but only what remains after deducting the value of the seed sown ih it, of the food of the horses who ploughed it, of the wear and tear of the plough, and so on. So the whole net annual income of a country consists of all those commodities and conveniences of life, which are produced during the year, after making allowance for replacing the Auxiliary capital that is consumed or worn out during the year. This net annual income is divided into :— firstly earnings of all kinds of work including business management, secondly interest on capital, thirdly rent obtained for the use of land or any other property that is naturally or artificially limited, fourthly taxes paid to the state. And a corresponding classification may be made of those among whom the income is divided : they may be classed as firstly those who earn an income by work, secondly capitalists, thirdly landlords, fourthly the State. Of course the same person may appear in two or more of these classes ; employers generally appear in the first two classes, and landlords who farm their own land in the first three. Again, the State is generally a landlord and a capitalist. In the first Book we have inquired into the Production of wealth ; and have seen how the real net annual income of the country is determined. We have seen how that share of it which the landlord can claim as rent, is fixed by definite \* economic laws ; and as the share which the State assumes to itself as taxes depends on causes which cannot be examined here, we must take it for granted. We may then regard the vamount which remains after deducting rent and taxes from the net annual income of the country, as a given Fund, and call it the Wages-and-profits Fund1. The problem of Distribution with which we shall be chiefly occupied during the rest of the present volume consists of an inquiry into the way in which this Fund is divided up. § 3. The term “ labourer” has been used by Economists to include all kinds of workers, and not merely unskilled labourers, to whom the term is confined in trade usage. And' the term " wages ” has been used to include the earnings of all kinds of work except that of business management 1 The way of looking at the source from which wages are paid is not quite the same as that adopted in what has been called “ the Wages-Funr\* theory.” This will be discussed at a later stage ; see Book in. ch. v 95 96 B O O K - IL C H A P . V I. § 3 , 4. The Earnings of Management are generally classed with interest by Economists as well as by men of business. And this is doubtless the best classification in many social and some economic inquiries ; because those who get the Earnings of Management are to some extent a separate class from those who get earnings of other kinds. Account will be taken of this fact when we come to discuss Market values, but in the theory of Normal values we are seeking for fundamental economic laws. And it is a universal rule of science that in seeking these fundamental laws, we should class together things that are Jsimilar in nature, and may be expected to obey similar laws. ^And therefore we shall find it best to class Earnings of Management with the earnings of other kinds of work ; because they jare similar in nature to other earnings, and are in the long run governed by the same laws. The earnings of a business man are uncertain, but so are those of a fisherman ; they are got by mental work, but so are those of the barrister and the physician, whose incomes have been classed by all Economists with the wages of skilled labour. And if instead of classing «Earnings of Management with other earnings, we were to class them with interest under the head of profits, we should be classing together two things entirely different in nature and governed by wholly dissimilar. laws. For though in the passing vicissitudes of trade it is sometimes difficult to draw a clear line between the interest on a business man’s capital and his Earnings of Management, yet we shall find that there is little in common between those fundamental laws which determine in the long run the Normal rate of interest, and those which determine the Normal Earnings of Management. It seems therefore best to class Earnings of Management with Vwages under the head of Earnings; to regard the het income of the country, after deducting rent and taxes, as divided not jinto wages and profits, but into interest and earnings ; and to call it the Eamings-and-interest Fund. We have then to seek for the laws which determine in the long run that remuneration of abstinence which is called interest, and those various remunerations of industry which are called earnings. This inquiry has two sides, that of Supply and that of Demand. The Law of Supply of capital has been already dis- 4jcussed ; and a general account has been given of the increase of population. But some account must be given of the Laws of jSupply of unskilled labour, of skilled labour and of business power. Afterwards we shall consider the relations in which Normal Supply and Demand stand to one another, and the Laws which determine Normal Interest and Normal Earnings. g 4. But before going further it will be well to guard against some misunderstandings. DISTRIBUTION. Firstly, some confusion has arisen from the use of the term '“ Cost of production” in two senses. It has been used to indicate what are here called the “ Expenses,” as well as what is here called the “ Cost of production1.” Thus the Law of (Normal) value has been stated to be that value tends to equal cost of production ; this of course does not mean that the value of a thing tends to equal what is called in this book its Cost of production, Le. the efforts and abstinences that have been required for making it. What is meant is that the value of the thing tends to equal the sum of those values which measure the y efforts and abstinences required for making it; that is, Cost of production is used to denote what we have expressed by the term Expenses of production. For an exchange value or price, though it may be equal to a set of exchange values or prices, cannot be equal to a set of things unlike in kind to it. There cannot even be any direct comparison between one set of efforts and abstinences and another. We cannot subtract the labour of a carpenter in making a box from the labour of a watchmaker in making a watch. But we can subtract the exchange measure or price of the work of a carpenter in making a box from the price of the work of a watchmaker in making a watch. It may happen that an hour’s work by a business manager, or two days’ work by a watchmaker, or three days’ work by a carpenter, or ten days’ work by an agricultural labourer, may all have the same exchange measure, say, a guinea. A guinea may also be the exchange measure of the abstinence or sacrifice involved in the loan of 20 guineas for a year. These various efforts and abstinences, these elements of Cost of production, are certainly not equal to one another. But they would all exert an equal influence upon value ; because their Economic measures, the expenses which would have to be incurred by anyone who would purchase them are all equal.] [§ 5. Again it is sometimes implied that besides the four classes who get respectively earnings, interest, rent and taxes, there is another distinct class of “ consumers” who may bear the burden or reap the benefit of economic changes. But there is no such class. Of course the consumers of any particular commodity will suffer as a body if anything should raise its price. If for instance the wages of the producers rise, there will be a gain to this branch of industry at the expense of the consumers ; that is, probably of some landlords, and some 1 Mill used the phrase in these two senses ; he himself distinguished clearly between them; and a careful examination of the context will show which sense he meant in each place. But the want of a formal distinction between them has confused many of his readers. See the Fortnightly Review for April, 1876. M. 97 7 - 98 BOOK II. CHAP. VI. § 5, 6. capitalists, and some labourers and the State. But there is no separate class of consumers of things in general, on whom the burden or the gain of any economic change can be thrown. Every such burden and every such gain must be distributed among the receivers of earnings, capitalists, landlords, and the State. There is of course a class of dependents— children, invalids, paupers and others— but what they consume is portioned out to them by the will of others who have an independent income ; and not according to econorilic laws. The consumption of those who depend on the support of another, is really a part of his consumption.] [§ 6. Again there are some intricate questions connected with the interpretation of the term “ the net annual income of a country.” They hâve but little bearing on the main argument of the book, but they cannot be entirely ignored. This net income is found by deducting from the total produce of land and capital and labour, whatever is required to replace the Auxiliary capital consumed or worn out in the year. But it must be remembered that this total produce includes all results of man’s work, which have market value, whether the work be manual or mental, whether it produces material wealth or not. It includes the benefit derived from the advice of a physician, the pleasure got from hearing a professional singer, and the enjoyment of all other services which one person may be hired to perform for another. It includes the services rendered not only by the omnibus driver, but also by the coachman who drives a private carriage. It includes the services of the domestic servant who makes or mends or cleans a carpet or a dress, as well as the results of the work of the upholsterer, the milliner and the dyer. We must however be careful not to count the same thing twice. If we have counted a carpet at its full value, we have already counted the values of the yarn and the labour that were used in making it ; and these must not be counted again. But if the carpet is cleaned by domestic servants or at steam scouring works, the value of the labour spent in cleaning it must be counted in separately ; for otherwise the results of this labour would be altogether omitted from the inventory of those newly-produced commodities and conveniences which constitute the real income of the country. Again, suppose a landowner with an annual income of ;£ 10,000 hires a private secretary at a salary of ^500, who hires a servant at wages of ^50. It may seem that if the incomes of all these three persons are counted in as part of the net income of the country, some of it will be counted twice over and some three times. But this is not the case. The landlord transfers to his secretary, in return for his assistance, part of the purchasing DISTRIBUTION, power derived from the produce of land ; and the secretary again transfers part of this to his servant in return for his assistance. The farm produce the value of which goes as rent to the landlord, the assistance which the landlord derives from the work of the secretary, and that which the secretary derives from the work of the servant are independent parts of the real net income of the country ; and therefore the j£ io,ooo and the ^500 and the £ 50 which are their money meàsures, must all be counted in when we are estimating the income of the country. The benefit which a man gets from living in a hired house must be reckoned in the total produce of the country’s land, labour and capital ; for the rent that he pays, after deductions have been made for Depreciation, is the interest on the capital of the owner of the house ; and the Earnings-and-interest Fund must therefore include the benefit which is got from living in it. If a man lives in his own house, the benefit he derives from it, after making deduction for Depreciation and repairs, is part of the real income of the country ; and the same may be said of the benefit he gets from the use of furniture, clothes, and other things which are not in ordinary usage reckoned among capital. This shows that our account of the Earnings-andinterest Fund requires a slight correction. We ought strictly to ' regard it as found by deducting from the net produce of the country, not only for rent and taxes, but also for the benefits that are derived from the use of furniture and other commodities Which we do not call capital, because they are in the hands of consumers. In estimating these benefits, allowance must of course be made for Depreciation and repairs. It may be noticed that if we had adopted the plan which Mr Jevons has advocated, of calling all commodities in the hands of consumers, capital1; that is of using the word capital as synonymous with Material wealth, this last difficulty would not have arisen, for we should have included the benefits derived from the use of these commodities as part of the Earnings-andinterest Fund, and regarded the owner of them as paying interest to himself ; just as in reckoning up the rents of a country we suppose a landowner who farms his own land to pay rent to himself. But Mr Jevons’ plan, though it would be convenient in some ways, is so opposed to ordinary usage, that it seems best to adhere to the old use of the word. In this sense it includes carriage horses that are let out on hire, and farm horses whoever owns them. It includes a doctor’s horses and 1 Theory of Political Economy, pp. 245— 253. Mr Giffen includes all material wealth under the £ 8,500,000,000 at, which he estimates “ the capital” of the country. 99 7—2 carriage if he hires them, in which case the revenue from them enters into the Eami ngs-and-interest Fund ; but if they are his own, and used partly for his own enjoyment, partly as a source of revenue, it is difficult to say how much of their value should be regarded as capital1.] to o BOOK IL CHAP. VI. § 6. 1 See Book I. ch. ill. § i. C H A P T E R V IL SUPPLY OF UNSKILLED LABOUR. § i. “ T h e Real wages of labour may be said to consist in the quantity of the necessaries and conveniences of life that are given for it; its Nominal wages in the quantity of money... The labourer is rich or poor, is well or ill rewarded, in proportion to the real, not to the nominal, wages of his labour1.” This distinction is very important ; but in the present book we have agreed to neglect variations in the purchasing power of money. And so long as we do this, a rise or fall in the Nominal or money wages of labour means just the same as a rise or fall in the general purchasing power of these money wages, and therefore just the same as a rise or fall of Real wages. . There is another distinction which turns upon the difference between two ways in which wages may be measured. The wages that a man earns in a day may be called his Time wages: the wages that are paid to him for doing a given amount of work of a given quality (whether the bargain made with him, is by the day or by the piece) may be called his Task wages. For instance, if a man is hired at 15J. a week and a boy at 5^. a week, and the man does three times as much work on the average as the boy, their Task wages are equal though their Time wages are unequal. The full importance of this distinction will be seen further on. The wages of which we speak in this chapter are Real Time wages. We have already seen that according to the Law of Population “ a rise in the rate of wages causes either a rise in the Standard of Comfort of the people, or an increase in the number of marriages and births. A rise in the Standard of Comfort is almost sure to increase the percentage of children who grow up to be efficient workers. Therefore a rise in wages almost always promotes, and a fall in wages almost always checks, the growth of population.” 1 Wealth of Nations, Book I. ch. v. 102 It is true that when the labourer’s wife and children can earn good wages, he may be able to bring up his family in comfort even when his own wages are low. But this correction is of much less importance than at first sight appears. For if the mother were not working for wages, she would be doing work at v/ home that would promote the health, and the moral, if not the / intellectual education of her children. While she is earning ^ wages, some of this work will be neglected, and she will have to spend part of her wages on hiring others to do the rest of it. The wages of the children themselves are perhaps a more important means of providing them with food and clothing ; but those children who are set to earn wages at an early age are often unhealthy as well as ignorant. So that we shall not go very far wrong if, in considering the effect of wages on the growth of population, we confine our attention to those earned by the labourer himself. Perhaps the chief exception to this rule is the case of agricultural families in new countries, who work together so that the family life is not broken up, and the physical, mental and moral well-being of the children is well cared for. § 2. The Standard of Comfort which young people are prudent enough to secure for themselves before they marry, varies from place to place and from time to time. When a high Standard of Comfort sets in, social opinion requires parents to provide wholesome rooms, nourishing food and a good education for their children. The next generation of workers is then healthy, intelligent and skilful, and of great value to employers, who by competing with one another keep up the rate of wages : the produce raised by them is so great that their share increases without diminishing the rate of profits. It is true that a rise in wages may lead to dissipation and extravagance, and fail to benefit the rising generation; but such cases are rare except in districts in which men, and, what is perhaps more important, women have been rendered hard, coarse and reckless by the nature of their occupations. When a rise in wages results in an improvement in the homes, the food, education and therefore in the working efficiency of the people, it may elevate them permanently. And so if wages are already very low, a further fall may cause a degradation of labour. “ With less food, which is the fuel of the human machine, less force will be generated ; with less clothing more force will be wasted by cold ; with scantier and meaner quarters, a fouler air and diminished access to the light will prevent the food from being duly digested in the stomach, and the blood from being duly oxydised in the lungs ; will lower the tone of the system, and expose the subject increasingly to the ravages of disease. Now in all these ways the labourer becomes less efficient simply through, the reduction of his BOOK It. CHAP. VII. § i, 2. SUPPLY OF UNSKILLED LABOUR. 103 w ages1.” The produce of his labour is less : the portion that he receives as his share is smaller, but the portion that capital receives as its share is not greater. “ What is taken from wages no man gains ; it is lost to the labourer and to the world. ” § 3. So far we have been considering the causes that determine the supply of unskilled labour generally ; but the supply of unskilled labour in any particular trade is determined b y other things besides the rate of wages. If equal wages were offered in ships going to the Mediterranean and the North Sea, th e former would be full and the latter empty ; so higher wages are offered on the latter ships to counterbalance the disadvantage of the ungenial climate to which they sail. Carters w ho work underground in mines are paid more than carters who work above ground ; and again, a night porter gets higher wages than a day porter. If a trade has any disadvantage, such a s unhealthiness, dirtiness, &c., higher wages will be necessary to induce men to seek the trade than would have been required in the absence of the disadvantage ; and the necessary addition to wages may be called the money value of this disadvantage. In the same way if a trade offers any exceptional personal advantage, such as a good social position, lower wages will attract men to it than would be necessary in a trade similar in other respects, but without the advantage : and this difference of wages is the measure or money value of the advantage. If the wages and the money equivalents of the other special advantages of a trade be added together into one sum, and the money values of its special disadvantages be subtracted from the sum, the balance that remains may be called the Net Advantages of the , trade. If any unskilled occupation offers higher Net Advantages than other occupations in the same place, additional supplies of labour will as a rule speedily drift into it and will force down its wages. To this general rule there are important exceptions : but they will be best discussed at a later stage. 1 Walker On Wages, ch. iv. C H A PT E R VIII. SUPPLY OF SKILLED LABOUR. § i. W e have seen that the supply of unskilled labour depends chiefly on the food, clothing and other necessaries of ^ life which labourers have been able to afford their children in preceding generations. The natural instincts which are common to all ranks of human life, and even among the lower animals, induce parents to exert themselves to provide food and shelter for their offspring. But it is otherwise with the special requirements of the skilled labourer; for his education demands an investment of capital, the fruits of which will appear in after time in his wages. Adam Smith says : “ When any expensive machine is erected, the extraordinary work to be performed by it before it is worn out, it must be expected, will replace the capital laid out upon it, with at least the ordinary profits. A man educated at the expense of much labour and time to any of those employments which require extraordinary dexterity and skill, may be compared to one of those expensive machines. The work which he learns to perform, it must be expected, over and above the usual wages of common labour, will replace to him the whole expense of his education, with at least the ordinary profits of an equally valuable capital. It must do this too in a reasonable time, regard being had to the very uncertain duration of human life, in the same manner as to the more certain duration of the machine.” Thus Adam Smith seems to imply that the Laws which determine the investment of capital in the education of the labourer are similar to those which govern the investment of capital in any vendible commodity. If this were the case, the Normal price or wage of skilled labour would be determined simply by its Expenses of production, just as the price of a commodity is determined. Take for instance the supply of cart-horses. There is some price which on an average just repays the expenses of rearing and breaking in horses of various kinds; and such a number of each kind are reared as will sell for prices SUPPLY OF SKILLED LABOUR. which will cover their respective Expenses of production. The calculations of horse dealers may fail through their having incomplete knowledge of the circumstances of the market, or I because of sudden changes in the demand for horses. Still if ! we take long periods, the average supply of horses of various kinds is such, that the prices at which they can be sold are I equal to the Expenses of rearing and preparing them for their several kinds of work. Again, the skill of a slave is vendible, and therefore the Law f of Normal Supply applies to it. A slave-owner who rears slaves for hire in different employments, calculates the expenses ’ "of each kind of training, and the probable future demand for ! each kind of work. He estimates the rate of hire and the length o f time during which it can be earned. He allows for inconstancy of employment and for uncertainty of success, but \* not necessarily for any discomforts except those which are likely to injure the health and shorten the life of the slave. He allows for the fact that the expenses of education will be incurred soon, and that the wages will not be received for many years ; after having made these calculations he can tell whether it will be worth his while to bring up the slave as an unskilled or as a skilled labourer. He will know whether the higher wages which the slave can earn for him will be worth more to him than the extra expenses involved in giving him the higher training. Suppose for instance he finds that by incurring an extra expense of ;£ioo for the slave’s education, he can secure, after allowing for the chance that the slave may die early, an increased receipt in wages which will only be equal to £ 100 spread over the rest of the slave’s life time ; he will then certainly not give him the education. He would not give it if this increased value would be ,£200. If it Were ^300 he might be in doubt, but if it were ^400 he would probably give it 1. Thus slave-owners adjust the supply of any kind of skilled slave labour to the demand for it, so that the pecuniary gain which is derived from a slave’s work in any Occupation corresponds to the trouble and expense of bringing him up to it2. § 2. But parents When selecting trades for their children differ from slave-ownefs when choosing occupations for their slaves in four important respects. Firstly, parents in estimating 1 Calculations of this kind are worked out in Engel’s Der Preis der Arbeit. 2 For this reason the population of a slave country might be better educated than the children of poor labourers in a free country; but their masters generally fear that education might lead them to rebel ;. and a slave, because he is a slave, is almost always limp and apathetic in character: and the attempt to educate him for difficult work, or for responsibility, is like trying to put a polish on soft wood. But in ancien\*- 105 106 BOOK II. CHAP. VIII. § 2. the Net Advantages of a trade for their son take account of everything that will add to his comfort and happiness, while the slave-owner may take account of such things only so far as they affect the slave’s efficiency. Secondly, parents who give their son an expensive education expect that he, not they, will reap its chief fruits ; but the slave’s skill is the property of his owner. Parents resemble the slaveowner in having to estimate on the one hand the Net Advantages of each occupation, and on the other the trouble and expense of preparing for it. But even when they have made these estimates they cannot decide on their course by a mere arithmetical calculation, as he can. For they must also decide this moral question : To what extent are we ready to sacrifice ourselves for our son’s benefit ? What is the value to us of an advantage that will be enjoyed by him ? Thirdly, few parents are capitalists ; most of them cannot borrow money on anything like the same terms as a capitalist can. To take an extreme case : some unskilled labourers, even with the help of the wages that their children bring in, cannot purchase as much food and clothing and house-room as the health of the family requires ; and it would be useless to ask one of them to calculate the advantage which his son might derive if £$o were spent in preparing him for a skilled occupation. It is likely that these advantages may include wages for his son when grown up at the rate of £yo a year more than those of an unskilled labourer. If he worked out this sum he would find that, allowing the market-rate of interest, the present value of these extra wages would be several hundred pounds. In such a case a slave-owner would make the outlay on business principles. The father does not because he cannot. Fourthly, in assuming that the slave-owner is a capitalist, we supposed him to manage his affairs as capable business men do. We supposed that he is always on the look out, as business men are, fqr profitable investments of his capital; that he watches the variatiqns of supply and demand with regard to each particular kind of slave’s skill ; and that he is continually making out the profit and loss accounts of various modes of educating his slaves. But parents do nqt generally carry on their inquiries in this business-like way : a poor and ignorant parent is not likely to think of obtaining for his son a lot in life very different'from his Greece and Rome, when the moral sense of the world had not outgrown slavery, and slaves worked without a feeling of degradation, masters used often to give a high literary or artistic education to their slaves as a matter of business. SUPPLY OF SKILLED LABOUR. 107 own. A man brought up wkh narrow surroundings is apt to acquiesce m them. His own start in life was a poor one, and it seems to him quite reasonable that he should make what he can out of the labour of his son j his wages almost imply that the world expects him to do it, and his neighbours do it; so he allows a small present gain to himself to outweigh a great future advantage for his son. The poor are moved as much as any other parents by the sight of the sufferings of their children, but they are careless about the distant future both of their children and of themselves ; for they have not a vivid imagination ; they are ruled by custom and not by the deliberate use of their reason. The lower we go in the social scale the less do parents seem to see the benefits that they may confer on their sons by investing trouble and money in their education ; and the smaller is their power of making such sacrifices. The rate of interest —) at which parents discount future advantages to their children ' increases with the narrowness of their education and the pressure of immediate want. § 3. Mill was so much impressed by the difficulties that beset a parent in the attempt to bring up his son to an occupation widely different in character from his own, that he said :— “ So complete, indeed, has hitherto been the separation, so strongly marked the line of demarcation, between the different grades of labourers, as to be almost equivalent to an hereditary distinction of caste ; each employment being chiefly recruited from the children of those already employed \n it, or in employments of the same rank with it in social estimation, or from the children of persons who, if originally of a lower rank, have succeeded in raising themselves by their exertions. The liberal professions are mostly supplied by the sons of either the professional or idle classes : the more highly skilled manual employments are filled up from the sons of skilled artizans, or the class of tradesmen who rank with them : the lower classes of skilled employments are in a similar case; and unskilled labourers, with occasional exceptions, remain from father to son in their pristine condition. Consequently the wages of each class have hitherto been Regulated by the increase ©f its own population ratber than that of the general population of the country.” But he goes on, “ The changes now so rapidly taking place in usages and ideas are undermining all these distinctions ; the habits or disabilities which chained peqple to their hereditary conditiqn are fast wearing away, and every class is exposed to increased and increasing competition froip a{ least t}ie class immediately below it1.” \* \* Political Economy, Book II. ch. xiv. § a. lo3 BO O K IL CHAP. V III. §3, 4. This movement has continued since he wrote, and the divisions between the various grades of English society are not so clearly marked in this generation as they were in the last. Each of Mill’s four grades is subdivided into a number of lesser grades, rising one above the other like the steps of a long staircase that is arranged in four flights, with a short landing at the .\*■end of each of them. Parents generally bring up their children to an occupation in the same social grade as their own. But when the supply o f labour in one grade is such that its Net Advantages are m u ch ^ i greater than those in the grade next below, the more thoughtful ” and self-denying parents in the grade below push their sons into that grade. Thus the Law of Supply, though its action is more slow with regard to labour of different grades than with regard to different kinds of commodities, yet acts in the same way in 4 the two cases. When the Net Advantages of a trade rise above• what is required to compensate for the trouble and expense of preparing for it, i.e. are abnormally high, forces are brought into play which tend to increase the supply of labour in that trade. The increase will come first and chiefly from other trades in the same grade, unless the Net Advantages of other trades in the same grade are also abnormally high. In that case the increase of supply will come chiefly from the grade immediately below, and will come more slowly. In any case the increase will be chiefly due to the action of parents in selecting an occupation for their children, but will be to some extent caused by a change of occupation on the part of adults. On the other hand, when the Net Advantages in a trade fall below what is required to compensate the trouble and expense of preparing for it, forces will be brought into play which tend to diminish the supply of labour. Foi: the most versatile men in the trade, and those who have the least taste and the least Specialised skill for the work, will begin to seek other occupations ; and parents will avoid bringing up their sons to it. J To complete our account of the causes which determine the supply of different kinds of skilled labour we must examine the various advantages and disadvantages that attract men to or repel them from different trades, and the various industrial qualities that may be required in a trade. § 4. Firstly, with regard to the various advantages and disadvantages of which account must be taken in estimating the Net Advantages of a trade. , Under this head must obviously be reckoned the constancy 4 or inconstancy o f employment in a trade, and the chances of success orfailure in it. To allow for these inequalities we must, when we are estimating the wages of a trade, look for the wages earned not in a day but in a year, or even in SUPPLY OF SKILLED LABOUR. r several years, by those who have an average success in it. We. cannot measure the wages of a fisherman by the price for 1 which he can sell his fish when he returns with a boat full, without taking account of the many days when he returns having taken nothing. W e cannot reckon the wages of a dock - > porter as those which are to be had when the docks are full of ships eager to be unloaded, without taking account of the fact v f that sometimes the harbour is nearly empty and most of the porters waiting in enforced idleness. In comparing the wages g ^ o f a bricklayer, a carpenter and a railway guard, we must remember that the railway guard’s employment is secure at all times, that the carpenter may be thrown out of work by dulness of trade, and that bricklayer’s work may be interrupted by ^ ' dulness of trade and by unfavourable weather. In estimating ’ ^ t h e earnings of barristers we must take account of the many who ® fail as well as of the few who succeed\* But independently of considerations such as these, which properly enter into our estimate of the average wages to be ' earned in a trade, there are other advantages and disadvantages which must be reckoned separately. These may be classed ► under the heads of healthiness, comfort and social position. Danger may be regarded as a kind of unhealthiness. Dirtiness, \* > physical and mental strain, anxiety and monotony are the chief discomforts to be considered. Adam Smith points out that the aversion which men have for the work of a butcher, and to some extent for the butcher himself, raises the earnings of butchers above those of bakers. If earnings were equal in the two trades the number of bakers would increase, and that of v butchers would diminish, until the wages of butchers had risen above those of bakers enough to compensate for the disagree- , ableness and the social disadvantages of the butcher’s work. Again, the wages of domestic servants, including their board and lodging, are much higher than are those of women who do work of equal difficulty in factories or in their own homes. For the servant must always submit to some loss of freedom, and if she happens to fall under the control of an ill-mannered mistress, to some loss of dignity. , When the disagreeableness of work takes the form of dirtiness it generally has the further disadvantage of involving a loss of social position; the most disagreeable of all work is chiefly done by those who cannot get any other employment ; , and its wages are therefore low, in consequence of the in- ^ competence of those who do it, and in spite of its repulsive character. But personal discomfort which arises from a sense of danger, often does not repel men. As Adam Smith says: “ The dangers and hairbreadth escapes of a life of adventures, instead of disheartening young people seem frequently to X09 no recommend a trade to them...the distant prospect of hazards from which we can hope to extricate ourselves by courage and address is not disagreeable to us, and does not raise the wages of labour in any employment. . It is otherwise with those in which courage and address can be of no avail. In trades which aie known to be very unwholesome, the wages of labour are always remarkably high.” The fact that an occupation offers some very high prizes, attracts men into it in two ways. Firstly\* young men form, fortunately for the world, high estimates of their own chancesj of attaining excellence, and thus are more attracted by the prospects of a great success than they are deterred by the fear of failure; and secondly, the social rank of an occupation depends more on the highest dignity and the best position which can be attained through it, than on the average good fortune of those engaged in it The learned professions offen such high attractions independently of the earnings to be got in them, that their average earnings are much lower than what could be earned by an equal amount of trained ability and industry in other occupâtions» § 5. Next, to consider the various industrial qualities, physical, mental and moral, that may be wanted in a trade. A man’s physical and moral qualities depend chiefly upon the character of his home in youth. If he was well fed and housed, if his father, and what is perhaps more important, if his mother had energy and kindness and honesty, he is pretty sure to have those physical and moral qualities which are a necessary condition of industrial efficiency. That part of a man’s wages which he owes to his education may be regarded as a kind of profit on the capital invested in it: that part which he owes to exceptional natural qualities may be regarded as a kind of rent; that is, it is the income derived from an agent of production the supply of which is determined by natural causes, and not by the deliberate outlay of human effort for the sake of future advantage. The wages in a trade that requires only physical strength, when they are paid by the day, are generally very low; but when they are paid by the piece, as in the case of navvies, an exceptionally strong man can sometimes earn high wages. Much work that requires great strength is also to some extent dirty and disagreeable ; and when it requires specially trained skill, and that intelligence and judgment which And a good market in any trade, it receives very high wages. For instance many of those engaged in iron and glass works must have great natural intelligence and a long special training. They are often thrown out of employment by depressions of trade, and their skill would be useless to them if for any reason they had to BOOK ' II. CHAP. VIII. § 4, s. ' SUPPLY OF SKILLED\* LABOUR. n r t > leave the trade. Their work is dirty and disagreeable, and it requires great physical strength} and a power, which many strong men have not, of enduring great and sudden variations of heat and cold. So it is not to be wondered at that in times of prosperity the best iron heaters and glass-blowers, working piece work, can earn wages many times as high as those of some to whom custom has given a higher social position. Many kinds of office work require a rare combination of high mental and moral qualities; but almost any one can be easily taught to do the work of a copying clerk, and probably there will soon be few men or wcftnen in England who cannot write fairly well. When all can write, the work of copying, which used to earn higher wages than almost any kind of manual labour, will rank among unskilled trades. In fact the better kinds of artisan work educate a man more, and will ) be better paid than those kinds of clerk’s work which call for neither judgment nor responsibility. The best thing that an artisan can do for his son is to bring him up to do thoroughly the work that lies at his hand, so that he may understand the mechanical, chemical or other scientific principles that bear upon it ; and may enter into the spirit of any new improvement that may be made in it. If his son should prove to have good natural abilities, he is far more likely to rise to a high position in the world from the bench of an artisan than from the desk of a clerk. § 6. Next, if we compare those industrial qualities which are Non-specialisecj, and can be turned to account in many trades, with those which are Specialised to one trade, we shall find that the former are rising in importance relatively to the latter. It has already been remarked that the division of labour sometimes enables a man to pass easily between trades which used to be totally distinct1. This remark may be extended. The great tendency of the growth of machinery is to supplant manual work which requires only physical strength, or skill that is got by the constant practice of one set of movements. Machinery can make uniform movements more accurately and effectively than man can ; and most of the work which was done by those who were specially skilful with their fingers a few generations ago, is now done by machinery; and since machinery does not incroach much on that manual work which requires judgment, while the management of machinery generally does require judgment, there is a much greater demand now than formerly for intelligence and resource. Those qualities which enable men to decide rightly and quickly in new and 1 Book I. ch. viii. § 9. tï2 BOOK IL CHAP. V III. §6,7. difficult cases, are the common property of the better class of I workmen in almost every trade ; and a person who has acquired them in one trade can easily transfer them to another. They are Non-Speciaiised. It is however true that there is much technical trade knowledge with regard to processes and the quality of materials, which is of little use save in the trade in ^ which it was acquired; and the progress of industry tends to increase the amount of the Specialised knowledge which is ! needed by the better class of workmen in almost every trade. § 7. We have seen that the supply of skilled labour is not adjusted to the demand for it as easily as the supply of a vendible commodity is adjusted to the demand for it. For a capitalist can devote his capital to producing any commodity which is in great demand, and obtain his reward in the high price that he gets for it. But if a capitalist goes to the expense r of increasing the supply of skilled labourers in any trade, the skill when acquired will belong, not to him, but to those whom he has educated. The only partial exception to this rule is met with in the system of apprenticeships. This system does enable capitalists to invest some capital profitably to them- I selves in the education of the sons of the poorer classes. In ' some cases the employer pays a lad considerable wages for I several years during which he is learning the trade ; but I during the last two or three years of his apprenticeship, the lad’s work is worth much more than the wages for which he is bound to serve by the terms of the apprenticeship ; and thus repays with interest the expense which taskmaster incurred for him, both in paying him wages and in loss from his spoiling material and his clumsy use of tools. But the system labours under some disadvantages. The spirit of the age in England, j and to an even greater extent in new countries, renders lads unwilling to be bound to a master for a long period of years. Moreover in Adam Smith’s time it was complained that masters often neglected to give proper instruction to their I apprentices, and this evil is greater now than it was then. For the lad used to be taught by the master, who had a direct interest in teaching him well; while now he is taught by a hired workman who generally has no such interest. But if the workman is the apprentice’s father he will take pains with the lad; and therefore employers like to have as apprentices the sons of those already in their employ; and fathers like to have their sons apprenticed under them. Something like a trade caste has grown up in this way in glass blowing and 4 some other trades. As we have seen, the rate of interest at which parents discount future advantages to their sons increases with the narrowness of their education and the pressure of immediate SUPPLY OF SKILLED LABOUR, II3 l want. This fact affords a strong argument in favour -of any , public or private action which may aid the poorer classes of/\^ parents in giving an improved education to their children. In this matter Government will find a direct pecuniary gain in doing what a private individual will not do unless from a > sense of duty. A Government which can borrow money, as ours can, at a little over three per cent, interest, may make a good investment by spending money on education, and thus advancing the capital which the poor man has no means of j^kborrowing at any tolerable rate of interest About a tenth of ^ ^ th e total income of the country is paid to Government in Imperial and Local taxes; so that with the present’ rate of taxes Government will gain about a tenth of whatever increase in the national wealth comes from an improved system of "^^ducation. And it is probable that this tenth would be sufflp c ie n t to repay with interest any outlay that Government may make on that general and technical education, which is required \* to enable Englishmen toehold their own in competition with those who have been taught in the admirable schools, that are to be found in some foreign countries. M. 8 C H A PT E R IX . SUPPLY OF BUSINESS POWER. § i. L e t us next inquire what are the causes which determine the supply of business power of different kinds ; and how the rate of Earnings of Management in a trade affects the supply of business power in it. The successful barrister or physician may obtain earnings equal to those of three hundred unskilled labourers ; but on the whole the most highly paid work is that of business management. It is only within the last few generations that this work has taken its present shape : it is only quite recently, and only in a few countries, that much intellect of the highest order has been required for, and given to, the task of conducting manufacturing and commercial enterprises. In old times there was little variation in the character of the things that were made, or in the manner of making them. Inventions came slowly; in some trades new processes were scarce discovered once in a century ; and after they had been once discovered, a generation or more might elapse before they 1 came into general use. The business man of former times required industry, sound judgment, and the power of dealing with men ; but he could very often get on without much faculty for originating new schemes. But now the manufacturer or merchant who controls a large capital cannot be sure even of holding his own unless he is quick to take advantage of new inventions, and has some power of striking out new lines for himself. The difficulty of this work of pioneering is increasing in many ways. For fashion and taste change more rapidly 4 than they used to do; and these changes affect the great masses of the population, and not only the upper classes as in the olden times. And again every change in the manner of carrying on any trade alters the character of the things that it wants to buy from other trades, and each trade buys from and \* sells to many more trades now than formerly. A manufacturer has to wàtch the progress of the trades to which he sells, as well as of his own. Again, those nations which had no manufactures were content till lately to buy any of the ordinary English goods that would at all serve their purpose. But now the English manufacturer meets the competition of Americans and others, who make a special study of the needs of each backward country, To hold his own, he must vary the machines and implements and other things, which he offers to other countries so as to suit various climates and various soils ; he must meet the special requirements of races of different temperaments and habits, and in different phases of civilization. This requires wide knowledge and a constant activity of intellect. Much of the work of business is then so difficult, and requires so much special training and such a rare combination of natural qualities, that the Earnings of Management got by it may be very high, without there being many men who can do the work and get these high earnings. § 2. We have already seen that the faculties required by the skilled workman,— his Personal capital— may be classed as Specialised and Non-specialised. The skill and knowledge which are of little use save in the trade in which they have been acquired are Specialised capital ; the general ability and resource, the energy and strength of character, the honesty and steadiness, which can be easily transferred from one trade to another, are Non-specialised capital. And we have seen that the progress of invention tends to diminish the importance of mere manual skill, and of the knowledge of the rules of thumb that have been handed down from earlier generations, and that therefore in many, though not in all trades, the impdrtance of the workman’s Specialised qualities has diminished relatively to his Non-specialised. The same change is going on with greater rapidity in the case of the employer. The employer who was “ master of his men” in this sense that, if any of them were doing their work badly, he could shew them how to do it better, is becoming rare in many trades ; and much work that used to be done by the head of a business is now done by foremen, overlookers, and sub-contractors. This change is to be regretted ; it tends to impair the thoroughness of work, and it causes estrangement between the employer and the employed by lessening his personal influence over them. But evil as the change is in some respects, it has the advantage of leaving the whole time and energy of the head of the business free for what has become, in this modern phase of the division of labour, his chief work. SUPPLY OF BUSINESS POWER. 115 8— 2 n 6 BOOK II. CHAP. IX. § 3. Bagehot compares him to the military commander of modern times who, instead of mixing in the fray himself, sits at the far end of a telegraph-wire with his head over some papers,, and directs and organizes from a distance. It is his work to study changes in the markets in which he buys, and in those in which he sells ; to be on the alert for new wants and new inventions, ^and to devise new modes of getting over new difficulties\* And most of the qualities that are required for this work are N on-specialised. They depend partly on early training, partly on the education of business ; but if educated in one trade theÿ^H jfcan be transferred to another. It is true that a man who is not well acquainted with the technical details of the trade in which he is engaged, is at some disadvantage, however great be his general or Non-specialised business power. But, as Bagehot points out, the disadvantage need not be very great if he has a competent staff of s u b - f li ordinates who possess the requisite detailed knowledge. So that the head or managing mind of a business may lose less than those who work under him would by changing his trade. A minister of state may move from the Indian Office to the Foreign Office, or from the Poor Law Board to the Admiralty, without any great loss of efficiency. He obtains information on technical details from the permanent Secretaries and clerks under him. His judgment and sagacity are as useful in one office as in another, while much of the technical knowledge of his subordinates is of little value save in that office in which it has been acquired. It is true that “ Little good for the most part comes of people who have been brought up on one side of the business world going quite to the other side, of farmers\* sons going to cotton spinning, or of lawmakers’ sons going into shipping. Each sort of trade has a tradition of its own, which is never written, probably could not be written, which can only be learnt in fragments, and which is best taken in early life, before the mind is shaped and the ideas fixed. But each trade in modern commerce is surrounded by subsidiary and kindred trades, which familiarise the imagination with it, and make its state known1.” And when high profits are being made in that trade, business power comes into it from the surrounding trades. § 3. We have seen that the supply of skilled workers in any. trade depends on the estimate that men form for themselves\* and parents form for their children, of its Net Advantages on the one hand, and of the difficulties of preparing for and enters ing it on the other. We saw that if they made their calculations correctly, and if there were no natural or artificial barriers against entering a trade, the supply would be so adjusted to the 1 In Fortnightly Review, vol. xix. SUPPLY. OF ' BUSINESS POWER. fi7 demand that the Net Advantages of being in that trade would correspond to the trouble and expense of preparing for it. And we have now seen hofcr the supply of business power in a trade is increased by a rise in the Earnings of Management that can be got in it. But the case of business power differs from that of skilled labour in several respects. Firstly, the average Earnings of Management cannot be easily ascertained. It is comparatively easy to find out the ges of bricklayers or puddlers by striking an average between wages that are earned by men of various degrees of efficiency, allowing for the inconstancy of their employment. But the ings of Management which a man is getting, can only be after making up a careful account of the net profits of his business, and deducting interest on his capital. The exact state of his affairs is often not known by himself ; and it can seldom be guessed at all accurately even by those who are in the same trade with himself. It is not true even in a little village at the present day that every one knows all his neighbours affairs. “ The village innkeeper, publican or shopkeeper, who is making a small fortune does not invite competition by telling his neighbours of his profits, and the man who is not doing well does not alarm his creditors by exposing the state of his affairs1.” But what we are now discussing are the causes that determine the supply of business power in each trade in the long run; and the average rate of profits in a trade cannot rise or fall much without general attention being attracted to the change before long. And though it is a more difficult task for a business man than fo r. a skilled labourer .to find out whether he could improve his prospects by changing his trade, yet the business man has great opportunities for discovering whatever can be found out about the present and future of other trades ; if he wishes to change his trade, he can, as we have seen, generally do so more easily than the skilled workman can ; and he can choose any trade for his son. Secondly, the supply of business power in a trade differs from the supply of skilled labour in being partly dependent on the supply of capital necessary to give it scope. But this difference again is less important than it appears. We cannot properly examine here the banking and other modern agencies by which capital is transferred from where it is not wanted to where it is wanted. For our present purpose it is sufficient to assume that “ A very great many of the strongest heads in England spend their minds on little else than thinking whether other people will pay their debts. And the combined aggregate of these persons is a prepared machine ready to carry capital in 1 Cliffe Leslie in Fortnightly Revtetv, vol. xxv. nS BOOK IL CHAP. IX. §3. any direction. The moment any set of traders want capital, the best of them, those whose promises are known to be good, get it in a minute, because it is lying ready in the hands of those who know, and who live by knowing, that they are fit to have it V ’ A man who has business power and a little capital can get more capital. If he turns this to good account his improved position will enable him to borrow still more ; and thus before long he may wield a capital so large that his profits leave him a rich income after paying interest on what he has borrow Again, a man without capital may be taken into private partn ship by others ; or he may become the manager of a joint-s" company. “ It is no longer true that a man becom employer because he is a capitalist. Men command because they have the qualifications to profitably employ laoour. To these captains of industry (or organizers of industry) capital and labour alike resort for the opportunity to perform their several functions2.” We may conclude then that though there are several differences between the case of business power and that of skilled labour, yet these differences do not prevent the supply of business power in a trade from being determined in all essential respects in the same way as the supply of skilled labour. We shall however see further on that the Earnings of Management of a man working with borrowed capital fall short of the income of a man of equal ability working with his own capital by more than the mere interest on it. 1 1 Bagehot, l.c. \* Walker on Wages, ch. xiv. C H A P T E R X. INTEREST. § I. T he Laws of Normal Supply of capital and of different kinds of industry have now been examined. Our next step is to investigate the relation between Normal Supply and Demand, and the way in which the Normal interest on capital, and the Normal earnings of each kind of industry, are determined. W e have seen1 that the total net annual produce of a country’s capital and industry, after rent and taxes have been deducted from it, consists of interest on capital, and the earnings of different kinds of industry ; and we have called it the Earnings-and-interest Fund, The share which capital obtains of this Fund depends firstly upon the amount of the Fund, and secondly on the manner in which it is divided. Firstly, with regard to the amount of the Earnings-andinterest Fund that is obtained by a given amount of capital and labour. This depends on the extent and the richness of the natural sources of agricultural and mineral wealth, on the progress that has been made in the arts of agriculture, mining, and manufacture, and on the means that nature and art have provided for conveying men, goods and news rapidly and inexpensively from one place to another. A country which has abundant and easy communication with distant places can often obtain a very high return to her capital and labour even though her own soil be poor ; for the concentration of her v industry may give her such advantages in manufacture and in the carrying trade that she can purchase much of her food and raw material from abroad at a moderate cost. But the highest returns of all are obtained in new countries where the arts of civilization are used in turning to account the rich resources of Nature ; this condition however cannot be fully satisfied unless the population is sufficiently dense to enable . them to organize easy means of communication with one ' another and avail themselves of the Division of labour. 120 Such then are the causes which determine the amount of the Eamings-and-interest Fund which Nature returns to a given amount of capital and industry. Our next step is to inquire how this Fund is divided into the share which capital takes as interest, and that which industry takes as earnings. Afterwards we shall have to inquire how the share which goes to industry is divided among the various ranks of unskilled and skilled labour, and business power. When a capitalist employs his own capital in business, Jie must be reckoned both among those who supply the capital which aids industry in production, and among those whose industry forms the demand for the aid of capital. The interest which he gets for his capital is not clearly marked off in practice \_ from the Earnings of Management which he gets by his work ; the two are reckoned together under the name of profits. But for the purposes of theory we must distinguish between the two : we must regard as interest that part of his income which he could get by lending his capital to be used by others ; just as, if a man cultivates his own land, we must regard as rent that portion of his income which he could get by letting the land to others. § 2. The rate of interest which results from the division of the Eamings-and-interest Fund into the shares of capital and industry will be found to depend upon the urgency of the demand of industry for the aid of capital. In a civilized £tate all production requires both capital and industry ; each of these demand^ the aid of the other, but not to the same extent at different times and places. If in any place there is an abundant supply of industry and a scarce supply of capital, while such methods of production are used as to give those who are working with much Auxiliary capital a great advantage over those who are working with little ; then the demand of industry for the aid of capital will be urgent. That is to say, a given amount of capital employed in production will be able to get a large share of the produce which it aids industry in raising ; because industry is in urgent need of the assistance of a great deal of capital, and the scarcity of capital enables the owners of it to hold out for very advantageous terms without running any risk of its being left unemployed. Conversely, the demand of industry for the aid of capital will not be urgent if there is a large supply of capital in proportion to the population, and the methods of production are such that no great gain would arise from an increase in the amount of Auxiliary capital used. Industry will not then be compelled to resign to capital enough of the produce to afford a high rate of interest. Thus we see that, other things being equal, an increase of BOOK II. CHAP. X. § 2. IN T E R E ST . I2i capital will diminish the competition of industry for the aid of capital; and will tend to raise earnings at the expense of interest In just the same way, other things being equal, an increase of population, capital being stationary, will increase the competition of industry for the aid of capital, and will raise interest at the expense of earnings. But other things do not generally remain equal. The progress of civilization increases the demand of industry for the aid of capital independently of any increase in the population of the country. For it causes a continual increase in the amount and expensiveness of the machinery and other things which men use as means to the attainment of their ends. Let us look at some instances of this change. Formerly men pumped up by hand, or brought in buckets whatever water they wanted. Now water-companies erect expensive works, and the water flows where it is wanted. Every Such water-supply represents effort economised by being capitalised. So all the modem contrivances for lighting and draining towns, for carrying men and goods by railways and Canals, and for carrying news by telegraph, enable men to attain their ends with much less total effort than they otherwise could, on condition that they capitalise a large part of their effort ; that is on condition that they make a great outlay at starting in the expectation of being rewarded by the benefits they will derive from it in the course of years. Again, the more durable the Fixed capital that is used, the larger the total amount of the capital that can l>e employed. For instance the amount of capital which can profitably be invested in the buildings of a country increases rapidly when massive stone buildings begin to displace wooden fabrics that are quickly built and that quickly perish. This change is now going on in America. Almost every important invention leads to an increase of the scope for the profitable employment of capital. Whenever a machine is made to do work that used to be done by hand, the scope for the employment of capital is increased. The Net Return of a machine may be defined as the value of the work that it does after allowing for Depreciation and the expenses of working it, including the Earnings of Management. Machines are often invented which cannot be used because their Net Return would not afford the market rate of interest on the capital invested in them. In course of time, when the machine is improved so that its Net Return would afford this interest, it is brought into use, and offers a fresh field for capital. Thus every increase in the advantages that can be obtained by the use of expensive machinery increases the scope for the profitable employment of capital, and thus raises the rate of interest. We see then that an increase in the durability of Fixed 122 capital, and an extended use of machinery and other Auxiliary capita], increase the amount of capital which is required to assist and give employment to industry; conversely every such change diminishes the amount of industry which is required to assist and give employment to capital. Every such change increases the demand of industry for the. aid of capital § 3. Next, the amount of capital, of which the aid in production is demanded by a given population in a given state of the arts of production, depends on the rate of interest at which they can obtain its aid. To fix the ideas, let us take some particular trade, say that of hatmaking, and inquire what determines the amount of capital which it absorbs. Suppose that the rate of interest is 4 per cent per annum on perfectly good security ; and that the hatmaking trade absorbs a capital of one million pounds. This implies that there is a million pounds worth of capital which the hatmaking trade can turn to so good account that they would pay 4 per cent per annum net for the use of it rather than go without it : net, that is exclusive of all allowances for Trade and Personal Risks and Depreciation. Some things are necessary to them ; they must have not only some food, clothing, and houseroom, but also some Circulating capital such as raw material and some Fixed capital such as tools and perhaps a little machinery. O f course competition prevents anything more than the ordinary trade profit being got by the use of this capital ; but the loss of it would be so injurious that those in the trade would have been willing to pay 50 per cent, on the capital, if they could not have got the use of it on easier terms. There may be other machinery which the trade would have refused to dispense with if the rate of interest had been 20 per cent, per annum, but not if it had been higher. If the rate had been 10 per cent, still more would have been used ; if it had been 7 per cent., still more ; if 5 per cent, still more ; and finally the rate being 4 per cent they use more stilL When they have this amount the Final Utility, as we may say, of the machinery, i.e. the utility of that machinery which it is only just worth their while to employ, is measured by 4 per cent. A rise in- the rate of interest would check their use of machineiy ; for they would avoid the use of all that did not give a net annual surplus of more than 4 per cent, on its value. And a fall in the rate of interest would lead them to demand the aid of more capital, and to introduce machinery which gave a net annual surplus of something less than 4 per cent, on its value. Again, the lower the rate of interest the more substantial will be the style of building used for the hatmaking factories and the homes of the hatmakers. And a fall in the rate of interest will lead to the BOOK II. CHAP. X. §§ 3, 4. employment of more capital in the hatmaking trade in the form of larger stocks of raw material and of the finished commodity in the hands of retail dealers. A fall in the rate of interest would increase the demand for capital on the part of some trades more than of others. A century ago it would have had little effect upon agriculture ; but now it would much promote the use of farming machinery and of improvements the return to which is expected to be spread over a long period of years. In cotton manufacture, again, it would have but little effect, because the work of that trade is of so uniform a character, and on so large a scale, as to give continuous employment to almost any machine that can be invented for it ; though even in that trade an increased command over capital will quicken the adoption of improvements in machinery. But in the wood and iron manufacturing trades there are a great many machines which it would be economical to use if they could be kept continually at work ; but for which small factories can only find occasional employment. The use of such machinery would be increased by a fall in the rate of interest Of course there are some years of depression of trade in which the Net Return of a good deal of machinery is nothing at all ; and again some machinery is made which never gives any Net Return. But if we look at average results, we find that the Net annual Return of the machinery and other capital in use in a country just affords interest on its value at the rate current there. If for instance we supposed, what is about the fact, that the rate of interest in England is 4 per cent., and that all the various trades of England employ between them in different ways a capital of ^4,000,000,000 ; this would show that ^4,ooo,poo,ooo is just that amount of capital which can be employed in England at the present time without the use of any which is thought likely, after allowing for risks, to give an annual Net Return of less than 4 per cent on its value. § 4. Thus the Demand for the loan of capital obeys a law similar to that which holds for the sale of commodities. Just as there is a certain amount of a commodity which can find purchasers at any given price, and when the price rises the amount that can be sold diminishes, so it is with regard to the use of capital. In any given state of the arts of production in a country, there is a certain amount of capital which it would be worth while for the various trades to employ in industry if they have to allow capital’s share of the year’s Earnings-andinterest Fund to be 7 per cent, on the capital ; or, as we may say, if they have to pay 7 per cent, per annum for the use of capital. If they have to pay 6 per cent, for its use, it will be worth their while to employ a larger amount; if 5 per cent INTEREST. 123 i n BO O K Ii; CHAP. X. §§ 5— 7- larger amount still; if 4 per cent, a larger amount still; and so on1. There is also the demand for the loan pf wealth by persons or States who do not intend to use it productively, but who mortgage their future incomes to enable them to increase their expenditure in the present. This part also of the demand for capital will be the greater the lower the rate of interest at which loans can be obtained. W e see then that the demand for capital depends on the numbers of the population, the natural resources of the country, the scopeTthat the arts of production afford for the employment of Auxiliary capital, and the needs of unproductive consumers. The amount of capital which can find borrowers in a market in a given time depends on the rate of interest at which it is offered for loan, and varies so that the amount demanded is increased by a fall and diminished by a rise in this rate. The current rate of interest measures the Final Utility to each borrower ; that is, the advantage to him of that capital which he is only just induced to employ. § 5. The annual addition to the capital of a country is not any considerable part of the whole, so that if we consider only short periods of time we may, without any great error, regard the supply as fixed during that time. On this supposition, the Law of the rate of interest becomes similar to that of the value of a commodity, the amount of which cannot be increased : Demand is the sole regulator of value. The rate of interest is then simply determined as just that which will call forth a demand for the existing capital § 6. But when we are considering long periods of time we cannot neglect the influence which the rate of interest exercises on the increase of capital : and the problem becomes difficult. To simplify it, let us begin by making the opposite assumption to that which we made when we supposed the supply of capital .to be fixed independently of the rate of interest. Let us now suppose that the rate of interest,— the price that can be obtained for the use of capital— exercises an overwhelming influence on the accumulation of capital. Let us imagine for instance that people would save rapidly if they could obtain 5 per cent, per annum interest for their capital ; but that if they could not obtain as high a rate as this, many of them would cease to save and begin to consume their capitaL In this case the Normal rate of interest for secure investments would be fixed at 5 per cent. For so long as the rate of interest was more than 5 per cent., capital would accumulate -1 Comp. Jevons, “ The advantage of capital to industry,” Theory of Political Economy, ch. vii. rapidly. The growth of capital would make the division of the produce more favourable to industry and less favourable tq capital ; earnings would rise and interest would fall. If it fell below 5 per cent., the accumulation of capital would be checked, and many would consume their capital The demand of industry for the aid of capital would thus be increased, the division of the produce would become more favourable to capital ; -earnings would fall, and the rate of interest would rise again to 5 per cent, and so on. Thus a rate of interest of 5, per cent per annum would be the Centre or Normal value toward which the remuneration of abstinence continually gravitated, and any deviation from which would be but a temporaiy irregularity, which the moment it exists, sets forces in motion tending to correct it. On this assumption the Normal rate of interest would be fixed at 5 per cent, independently of all changes in the field for the employment of capital ; such as the opening up of rich new countries, or inventions that gave room for vast extensions of Fixed capital. Of course such a change might raise the rate of interest temporarily ; but since this rise would, on our present hypothesis, call rapidly into existence a vast increase of capital, interest would not remain for any considerable time higher than 5 per cent. On this supposition the Normal rate of interest would be rigidly fixed in the same way as is the Normal value of a commodity, the Expenses of production of which are independent of the amount produced1. § 7. But in fact the influence which the rate of interest exerts on the accumulation of capital is much weaker than we have supposed it to be in this imaginary case. The motives which men have for saving are various, and their characters differ widely ; some will be improvident however high be the rate of interest; and however low it be, others will save for their families and for their own old age. On the whole a fall in the rate of interest in a country is likely to check the growth of capital in some ways, and to promote it in others, but the latter effects are on a smaller scale than the former, so that a fall in the rate of interest will diminish the rate of accumulation of capital to some extent, though often only to a small extent2. Resuming then, we find that the demand for capital, other than that which comes from unproductive consumers, depends on the natural resources., of the country, the numbers of its industrial population, and the state of the arts of production : this demand determines the rate of interest at which any given amount of capital can be employed in the country. The supply 1 See Book II. ch. v. § 3. 8 Book 1. ch. vi. § 4 .. INTEREST. 125 Ï2Ô depends upon the slow operation of many causes, one of which I is the rate of interest. \*1Combining these Laws of Supply and Demand we get the | Law of the Normal Bate of Interest, which is :— (When the economic conditions of a country have been nearly uniform for a long period of time, the supply of v capital is such, that the rate of interest which can be ]obtained for it is that which has been required to cause this supply to be forthcoming ; and the rate thus deter- I mined is the Normal rate. The rate is in equilibrium A when it is just that at which the whole supply of capital 1 can find employment. I§ 8. The greater part of the capital in England has | been accumulated since the country entered on the economic 1phase in which it is now ; and the Normal rate here is \*\*j about four per cent, a year on good security. In the case of England the influence which foreign markets exert on the rate i of interest is of primary importance. If the rate were to rise Jmuch above four per cent., a great deal of capital that is now j sent abroad for investment would be retained at home. More capital would be employed here ; the demand of capital for the I aid of industry would rise; and this would make wages rise at the expense of interest. On the other hand\*-tf the rate\* were ^ to fall much below four per cent., the amount of capital seeking j investment here would diminish; there would be less competition on the part of capital for the aid of industry ; the division of the produce of capital and industry would be more favourable to capital ; and the rate of interest would rise. The Normal rate of interest in England does not seem likely to deviate much from four per cent, for some time to come ; but it may be slowly altered by changes in the field of employment, while the market rate of interest is oscillating rapidly up and down on either side of the Normal rate as a centre. § 9. Before the invention of the steam-engine it seemed { likely that the Normal rate of interest in Western Europe would j soon become lower than it is now. The changes that have come . over the face of modern industry have given room foç the profitable employment of a vast amount of Auxiliary capital of which ^700,000,000 have been invested in English railways. Successive inventions together with the development of foreign commerce have enabled capital to increase much faster than population, without causing any fall in the Tate of interest And history records several other periods during which the wealth and capital of a nation increased very rapidly, while at the same time the opening out of new fields for the employment of capital was causing a rise in the rate of interest. In new countries the return which nature gives to man's BOOK II. CHAP. X. §§ 8—IO. INTEREST. work is often so rich that, though the wages of labour are high, ten per cent, or more can be got for the use of capital. But capital rushes in from older countries; and before a district has been occupied for many generations, the richest natural sources of wealth become private property for which a high rent is demanded. And then the produce which remains to be divided between capital and labour, the Eamings-and-interest Fund, obeys the Law of Diminishing Return and does not increase as fast as the capital increases : so that it no longer affords a very much higher rate of interest than that which can be had in older countries. A rate of eight per cent, on sound investments has spread like a wave steadily over the greater part of the North American Continent; and this is being followed by waves of seven and of six and even five per cent, interest that have already started on their way westward and southward from the Northern Atlantic States. § 10. It is difficult to forecast the distant future of the rate of interest. Hitherto the progress of civilization has increased the willingness to save at a low rate. In old countries, in which men are accustomed to work patiently for small gains and to value highly the possession of a secure income, a low rate of interest seems to have little effect in checking the accumulation of capital. In England for instance, in spite of the low rate of interest, the capital of the country is increasing at the average rate of about £200,000,000 annually, that is by a little more than a thirtieth of its total amount. If this rate of increase were sustained for four hundred years, the capital owned by Englishmen wohld be multiplied a million fold, and in eight hundred years a billion fold. But however high the hopes we may have of the future progress of the arts of production, we cannot suppose that there will ever be a field for the profitable employment of as much capital as this. Sooner or later the rapid growth of capital must increase the competition of capital for the aid of labour, and diminish the competition of labour for the aid of capital; so that capital’s share of the total net produce will cease to be proportionately as large as before. And, at the same time, the total net produce that can be obtained by a given amount of capital and labour will diminish according to the Law of Diminishing Return. So that ultimately the Normal rate of interest will fall. ' There is no reason to think that it will fell rapidly down to a minimum and then remain stationary. Rather should we expect that, with some slight oscillations, the Normal rate of interest will keep on falling, but that the rate of its fall will become continually slower and slower. It is thus likely never to attain but always to be approaching its nrininum. But we have no means of guessing what that minimum will be. 127 C H A P T E R XL WAGES. § l. We have seen how the produce of land, capital and industry, exclusive of rent and taxes, is divided into Interest and Earnings ; into the share which remunerates abstinence, and the share which remunerates work whether bodily or mental. Let us now look at the way in which this latter part is Normally subdivided among unskilled labour and the different kinds of skilled labour and of business power. We are not now concerned with the advantages which custom or social opportunities or trade organizations give to the various ranks of industry in bargaining for their several shares of the Earnings-and-interest Fund; but when we come to discuss the theory of Market wages in the next Book, we shall see how such advantages may cause the wages of a trade to diverge for a considerable time from their Normal level. Let us first inquire what constitutes the Normal demand for the work of each trade. This demand depends partly upon the desire of consumers to obtain the things which that trade produces or helps to produce, and partly upon the extent to which its aid is wanted by other industrial classes and the owners of capital who take part in making these things. Thus the demand for the work of a trade may be said to depend on the competition fo r its aid in production. The meaning of this term may be made more clear by an illustration. The recent advance in England’s wealth has caused a great demand for building ; and those who produce other things nave had to give more of them than before for the purchase or hire of a house. There has been an increased competition for the aid of the building trades, which has raised their wages and enabled them to obtain a larger share of the wealth of the country than before. Now suppose that during such a rise in the price of houses, there is a sudden check to the supply of (say) house carpenters. The rest of the building trades will then find it difficult to obtain the aid of carpenters to supply roofs, floors, WAGES. &c. And since the work of masons, plasterers, and master builders will be of little use without such aid from the carpenters, the competition of the other building trades for the aid of carpenters will force up the wages of carpenters, and enable each of them to obtain an exceptionally large share of the Earnings-and-interest Fund. This competition will not act directly ; the masons will not ask the carpenters to assist them in their work. It will act indirectly through the master builders ; for, as we shall see presently more fully, in all conflicts between the different industrial classes the employers of labour act as buffers, which absorb part of the force of a blow, but pass on most of it to others. The scarcity of carpenters will compel employers to offer them higher and higher wages ; the consequent rise in the price of houses will check the demand for them; the employers will have an over-abundant supply of masons and other labourers, and will lower their wages. Meanwhile the check to the building trade will increase the competition among master builders, and so diminish their own Earnings of Management. Thus the increased competition for the aid of carpenters will raise their wages, and this rise will be obtained partly at the cost of those who require houses, partly, at the cost of the rest of the building trades including the master builders. § 2. The demand for unskilled labour depends on the competition there is for its aid, whether in producing things or in ministering directly to people’s wants. It is increased firstly by every increase in the amount of capital that is ready to support and assist industry, and secondly by every increase in the amount of business power and of skilled labour of various kinds, that are competing for the aid of unskilled labour in the work of production. When the wages of unskilled labourers exceed that amount which enables them to maintain the Standard of Comfort to which they are accustomed, population increases fast; but when they are less, the growth of population is checked. For instance, the daily wages of unskilled labourers in England have seldom been less than what would purchase half a peck of wheat ; and they have seldom risen above what would purchase two pecks ; they have oscillated up and down between these two limits. If it were a fact that whenever the day wages of unskilled labour were less than a peck of wheat, population diminished, and that whenever they were greater than a peck of wheat, population increased rapidly; then a peck of wheat would be the Normal day wages of unskilled labour. For when a high rate of wages caused a rapid growth of unskilled labour, it would increase the competition of unskilled labour for the aid of capital and of other classes of industry ; and this woulr1 9 129 M. 130 BOOK II. CHAP. X L § 3. lead to a fall of wages : conversely when a low rate of wages checked the supply of Unskilled labour, it would lead to an increase in the competition on the part of capital and other classes of industry for the aid of unskilled labour; and this would lead to a rise of its wages. Thus a peck of wheat would be the centre or Normal value towards which the remuneration of unskilled labour constantly gravitated, and any deviation from which would be a temporary irregularity, which, the moment it exists, sets forces in action tending to correct it. The Standard of Comfort is not in fact rigidly fixed. But yet it is, at any place and time, so nearly fixed, and does exercise so great an influence on the growth of population, that the wages which afford the means of maintaining this Standard may fairly be called the Normal wages of unskilled labour there and then. When wages are at this level they are in equilibrium, unless there happens to be at the time a great change in the field of employment for labour. This is the Normal or Centre value about which fluctuations of Siipply and fluctuations of Demand cause the wages of unskilled labour to oscillate. But the Normal value itself varies from place to place and from time to time with changes in the Standard of Comfort of the people. A rise of wages caused by an increased demand for labour will be temporary, unless it lead to a rise in the Standard of Comfort ; in which case it will be permanent, and Normal wages will be raised. § 3. Next with regard to various kinds of skilled labour. The demand for each class of skilled labour depends on the competition there is for its aid. It is increased firstly by every increase in the capital that is ready to support qnd assist industry, and secondly by every increase in the unskilled labour, in the skilled labour of other classes, and in the business power that are competing for the aid of labour of this class. The causes which determine the supply of skill of any kind in one generation are, as we have seen, chiefly to be sought in the opportunities and the habits of the previous generations. The trouble and expense which parents will undergo in preparing their son for his work depend on their means, their habits of forethought and self-denial, and their knowledge of and access to various trades. The poorer and the more ignorant parents are, the higher is the rate at which they are likely to discount the wages which their son will receive at a future time. The lower parents are in the social scale, the greater must be the advantages which they can procure for their son by a given outlay on his education in order that they may be induced to make it. Parents generally bring up their sons in the same industrial grade as their own ; there is so much freedom of intercourse \*tween different trades of the same grade in the same place WAGES. that their wages (or more strictly their Net Advantages) seldom differ much for any long period together. Any increase in the inducements to parents to bring up their children to a trade, or to adults to enter it, increases the supply of labour in it ; the chief part of this supply being drawn from trades in the same grade with it. Thus we are brought to theLawof Normal Wages, which is The amount of the Earnings-and-interest Fund, and the way in which it is divided fnto the shares of interest and earnings being already known, the wages of each trade depend on the way in which this latter share is sub-divided. The Normal wages of a trade are therefore determined by the relation in which its wages (or more strictly, its Net Advantages), must stand to those of other trades in order that the supply of labour in it may be kept up, and this depends on the difficulty of the work to be done in it, on the expensiveness of the general and special education, and on the natural qualities, physical, mental and moral, required in it. Trades in the same industrial grade generally require an equally difficult and expensive education, and have equal wages. The lower the grade of a skilled occupation, the higher is the ratio which its wages bear to the expenses of preparing for it. When a rise in the wages of a trade above the Normal value causes a rapid increase of the numbers in it, its members find a diminished competition on the part of others for their aid ; they are at a disadvantage in bargaining for their share of the produce of land, capital and industry, and their wages fall. Conversely when a fall in the wages of this grade below the Normal value checks the increase of the numbers in it, the competition on the part of capital and the other classes of industry for the aid of this trade is increased, and its wages rise. But when the wages are at the Normal value they are in equilibrium; the growth of numbers is neither so fast as to lower them, nor so slow as to raise them. Toward this value the wages of the trade continually gravitate, and any deviation from this rate is a temporary irregularity, which, the moment it exists, sets in motion forces tending to correct it. But these forces often act very slowly because of the friction of various social and economic obstacles that hinder men’s passing from one trade to another, especially if they are in different grades; and the “ temporary irregularities” may extend over very long periods of time. Fluctuations of supply and demand cause the wages of each trade to oscillate above and below their Normal value ; just as a floating cork oscillates with each passing wave, above and below the surface of the sea. The 9 - 2 \*31 ly i BOOK II. CHAP. XI. § 3, 5 Normal wages of skill of various grades rise or fall slowly, as the tides rise or falL They change with those slow changes in the social condition of various ranks of the people which work themselves out in the course of many generations. v The effect of progress is as a rule to increase the supply of intelligence and ability, so that if the difficulty of the work to be done did not increase, the wages of skilled labour would be likely to fall. And in fact the Task-wages, that is the wages that are paid for skilled labour of a given efficiency, are falling in many branches of manufacture ; though they are on the whole rising in some trades which are under specially favourable circumstances, as for instance the building trades. But the difficulty of the work to be done, and the intelligence required for doing it are increasing, and the average earnings of the workers — their Time wages— are generally rising, even where their Taskwages are falling. § 4. A rise in the Time-wages of any trade tends to diminish profits. But if the wages that are paid for work vary according to its efficiency— if Task-wages are unaltered— the share of the produce of industry that is left' for others will be the same whether Time-wages are high or low. It is only where the rise in Time-wages is not accompanied by a corresponding increase in efficiency, and therefore Task-wages rise, that the change is injurious to capital. In fact when a labourer has to be supplied with costly machinery, a rise in Time-wages is a great benefit to capital, if it lead to such an increase of efficiency as to keep Task-wages unchanged. Of course the machine itself sometimes “ sets the pace,” and an indifferent worker may be able to do all that is required. But such cases are much more rare than is generally thought. Of two weavers or two turners working side by side at similar looms or similar lathes driven by the same machinery, one will often do twenty per cent more in a day than the other; and it would be to the advantage of the employer to secure the energetic man by paying him more than twenty per cent, more Time-wages, that is, by paying him Task- 1 wages at a little higher rate than he pays the inefficient man. I For in this way he will turn his machinery and the space in his j factory to the best account. I Even in the case of unskilled labour it is to the advantage of capital that Time wages should rise, provided Task wages do | not rise too. If two labourers at i 8j. a week will do as much work as three at 12j., the former are in the long run cheaper. 1 For they are likely to remain longer in full health and strength and to have healthier and stronger children than the others ; and a much heavier burden will be imposed on capital through the poor rates by three low-waged than by two high-waged labourers. WAGES. There is however a limit beyond which a rise in the Timewages of unskilled labourers will not cause a proportionate increase of their efficiency, and when this limit is passed any further rise in Time-wages will raise Task-wages too ; but even then the rise of Task-wages will be slower than that of Timewages. The growth of capital and the progress of the arts of manufacture will, if wars can be kept down, make the world rich enough to afford high wages for that little unskilled labour which cannot be supplanted by machinery. This increase of wealth will not indeed raise the wages of unskilled labourers if their Standard of Comfort remains low and they marry imprudently. But a time -may come when the Standard of Comfort of unskilled labourers will be such as to keep them from rapidly increasing their numbers, even though their wages have become as high in old countries as they are now in new countries. £§ 5. The influence which demand exerts on wages is sometimes expressed by saying that under a system of free Competition every man’s wages tend to be equal to the discounted value of the produce of his labour. Let us for instance imagine a thing to be made by unskilled labour alone without any superintendence, and without the aid of any capital except that which was advanced in the payment of wages. Suppose that this capital has been advanced gradually, some of it a short time, some a long time, on the average half-ayear before the thing is ready for sale. Let the rate of interest for six months, allowing for risk, be five per cent. Then if the thing can be sold for^io5,the Discounted value of thishalf-a-year beforehand will be ;£ioo. And competition will tend to make the wages of those who made it equal to this Discounted value of j£ioo. But a case as simple as this never occurs in practice. The earnings of many different kinds of industry, one of which is almost always that of Superintendence or Management, enter into the Expenses of production, and therefore into the price of almost everything that is sold. And in order to deduce from he price the earnings of one of these kinds, we must find out .ot only the interest on the capital employed but also the earnings of the other kinds of industry, and deduct them all from the value of the produce raised. We cannot then speak of the Discounted value of the work of any one of these classes. But we may still speak of the “ Net Return” of that labour. The N et Return of a machine was defined as the value of the work that it does after allowing for Depreciation and the Expenses of working it, including the Earnings of Management. The N et R etu rn of a man’s labour is the value of the produce which he takes part in producing after deducting all the other Expenses o f producing it \ 1 A statement of this kind has been mistaken by some writers fo\* 133 134 BOOK II. CHAP. XL § 5. This phrase is very useful when we are examining the part which Demand takes in determining wages. For instance, if the demand for houses is rising, the Net Return of the labour of some or all of the various building trades must be increasing. There may be signs of a greater scarcity of labour in one of the building trades, as for instance that of the carpenters, than in others ; and if we believe that the earnings of other branches of the building trades (including the Earnings of Management of the master builders) are not likely to rise, we may say that the competition for the aid of carpenters’ labour will increase, that the Net Return of their labour will rise in value, that they therefore will get an increased share of the Earnings-and-interest F und ; and that their wages will rise. Again the phrase Net Return of labour can be usefully applied in explaining the influence of Demand in equalising Task-wages in the same occupation. This influence may be described by saying that under a system of free competition every man’s wages, or more generally every man’s earnings, tend to equal the Net .Return of his industry.] theory of wages. But really it is only the Law:—“ Value tends to equal Expenses of production”— written in a new form. CH A PTE R XII. , § i. We have seen that the supply of business power is determined in all essential respects in the same way as the supply of skilled labour ; we have now to examine the Laws that govern the Normal Earnings of Management. These Laws appear at first sight to differ much, and they really do differ a little from those which govern the Normal wages of skilled labour; the chief difference arising from the fact that the Earnings of Management can be obtained only by those who have the control of capital. Let us then begin by comparing the Earnings of Management of two men carrying on similar business, the one with his own capital, the other with'borrowed. The man who works with his own capital considers that his Earnings of Management are the whole net profits of his business after deducting the interest that he could obtain by letting out his money on good security. But interest at a much higher rate than this must be paid by a man who borrows capital for his business, at all events unless his own property is sufficient to £ive good security for it; and interest at this high rate must be deducted from the profits of his business in order to find his Earnings of Management. The rate of interest which he has to pay is high, because in his case a new set of risks is introduced in addition to those unavoidable risks which exist in every business. Those risks which arise from such causes as the chance of destruction or depreciation in value of the capital employed, or of the goods produced in it, are inseparable from business, and maybe called Trade Risks. Those further risks which are introduced when the capital of one man is under the control of another, may be called Personal Risks. These are due to the mistakes that the lender may make with regard to the borrower’s business ability, and honesty. A man trading with his own capital has every motive for exerting himself to discover whether he is carrying on his business at a loss. But the man working with borrowed capital EARNINGS OF MANAGEMENT. 136 BOOK II. CHAP. XII. § 2. has not such strong motives. If his moral sense is not very active he may, without intending any deliberate fraud, carry on a losing business so long as to cause heavy losses to his creditors. If he has not a strict sense of honour, and finds himself in difficulty, he may plunge into rash speculations : for if they succeed, the gain will be his ; and he may not care whether he fails for a large or a small sum. One way in which the lender can insure himself against these various risks is by charging a high rate of interest for his loans. But a very high rate would be required to cover the risk of loans made for a long period of time ; and therefore such loans are generally made for short periods. The shorter these periods are the less is the risk which the lender runs, and the sooner can he recover the use of his capital for himself, if the course of his own trade should make him wish to do so. Thus Bankers and others are willing to lend money for a few months1 at a rate of interest sometimes not exceeding three or four per cent, a year, even when the borrower cannot offer security which would induce them to lend him capital for a long period of time at any moderate rate. But a man who is much dependent on such short loans labours under great disadvantages. F or if any misfortune should injure his credit, or if a disturbance of the money market should cause a temporary scarcity of loanable capital, he maybe quickly brought into great straits. He may not be able to obtain a renewal of the loans on moderate terms, or even on any terms, and may thus be cut short in his most hopeful enterprises. One of the chief symptoms of an impending commercial crisis is a rapid succession of forced sales at a loss by those who have been trading with capital borrowed for short periods. Thus it appears that a trader who works on borrowed capital has in one form or another to pay a high rate of interest. But though high, it is not sufficiently high to prevent him from competing with those who trade with their own capital. On the contrary men trading with borrowed capital seem likely to displace to a great extent those trading with their-own. The reason of this is not far to seek. A man who has a capital of £50,000 can easily obtain a secure income of £2500 by lending it out. And very likely he may not care to undergo the labours and anxieties of a business life unless he can get Earnings of Management of £2500, or even £5,000 a year, exclusive of course of Insurance against Trade Risks. But a man of equal ability who owns little capital, and who therefore cannot live in comfort without working, will be content with lower Earnings of Management. He may be willing to employ ,£50,000 of bor1 This is done chiefly by “ discounting bills.” EARNINGS OF MANAGEMENT. 137 rowed capital, in addition to his own, even though after allowing for the interest that he actually pays and the indirect risks that he runs through working with borrowed capital, he does not clear more than ,£1000 a year by the work. He can therefore afford to sell at a price too low to give that rate of. profits which the man of independent means requires. Thus those who depend on their business for a livelihood, undersell and drive out of trade those who are not so dependent1. § 2. Again a man may obtain Earnings of Management by carrying on a business with capital, the owners of which take a part at least of the risks of the business. The simplest way of doing this is the old plan of partnership. In former times a man of little capital had small chance of getting high Earnings of Management unless he could obcain the confidence of some wealthy man or private firm, who admitted him as a partner. Again, if a man thinks that he can profitably employ more capital than his own in his business, he often converts it into a joint-stock company. That is he admits others to shares in his business : they take a share of the risk and a corresponding share of the net proceeds that remain after paying him Earnings of Management according to some plan agreed upon between them2. A joint-stock company of this kind which is managed by its chief shareholder or shareholders, may act almost as freely and promptly as a private firm can. It has some special disadvantages ; but under favourable circumstances, it may hold its own even in trades which require ready enterprise and quick action. The business of a large joint-stock company is however often carried on by Directors, who give it only a little of their time, and General Managers, who give their whole time. The General Managers are seldom men of much capital, and are contented to work for moderate salaries. The business of such a company will almost always be managed with less energy and economy than a similar private business in able hands. As Mill says, it may be possible to secure in hired managers that fidelity which shrinks from a deliberate neglect of duty, but not that zeal which is continually laying schemes by which greater profit may be obtained, or expense saved, and which is ever anxious about small gains and small savings. If however the 1 This is making English commerce increasingly democratic, and docs much harm in preventing “ the long duration of great families of merchant princes....But the propensity to variation in the social as in the animal kingdom is the principle of progress. ” See Bagehot’s Lombard Street, Introductory chapter. \* This is similar in some respects to the old plan of introducing “ sleeping partners” into a private firm. The sleeping partners supply Capital ana take a share of the risk, but have no part in the Management. company is large, and it can afford to pay fairly good salaries to its officers, its affairs are likely to be in many ways better managed than those of a private business in the hands of men of second rate ability. The directors are usually men who can bring a wide and varied business experience to bear in laying down the broad principles on which the affairs of the company are conducted, and in judging the ability and industry o f the chief officials under them. So that the management of a large joint-stock company, though generally far from perfect, is seldom very bad except where there is wilful wrong doing. The publicity of joint-stock companies helps more than it hinders them in trades in which it is necessary to obtain public confidence, as for instance in banking and insurance. They have a monopoly of railways and other undertakings which require enormous capitals, and they are fast pushing their way in all businesses in which large capitals can be managed chiefly by routine and in which there is little need for bold and speculative enterprise. For they can thrive with a much lower rate of profits than will remunerate a wealthy capitalist for undergoing the worry and fatigue of business. The growth of Joint Stock Companies offers great opportunities to those who have business power, to obtain the control over capital. § 3. We may next inquire how the Earnings of Management of a business are related to the capital employed in it. Is the Normal rate of profits for all capitals employed in trades of the same difficulty the same whether the capitals be large or sinall ? If two businesses in different trades are equally difficult and disagreeable, and require equal capitals, there will of course be a constant tendency to equality of their Earnings of Management. There may indeed be great differences between the Earnings of Management of two men with the same capital in the two trades ; but so there may in one and the same trade. These differences arise from inequalities in ability or good fortune, just as do those between the earnings of successful and unsuccessful medical men or barristers. Again it is true that an able business man who starts in life with a great deal of capital and a good business connexion is likely to obtain higher Earnings of Management than an equally able man who starts without these advantages. But there are similar, though smaller, inequalities between the earnings of professional men of equal abilities who start with unequal social advantages. What is meant then is that competition tends to equalise the Earnings of Management of men of average ability and good fortune in two occupations in which equal capitals are employed, and which are equally difficult and disagreeable. The profits in each case are to be found by adding the Earnings of Manage133 BOOK II. CHAP. XII. § 3, 4. EARNINGS OF MANAGEMENT. 139 ment to the interest on the capital employed. And since the amount of capital employed is the same, and the rate of interest to be allowed is, under the modem system of banking, practically the same for all trades, the amount of interest is the same in the two cases. Therefore when the Earnings of Management tend to equality in the two cases, the profits of the businesses tend to equality. So that :— The profits on equal capitals tend continually to equality in trades which involve equal risks, discomforts and exertions} and which require equally rare natural abilities and an equally expensive training. It has however already been noticed that if ^ioo has been invested in the production of a thing two years before it is ready for sale, we must allow twice, or rather more than twice as much under the head of interest as if the ;£ 100 had been invested only one year before it was ready for sale : but the total amount to be allowed as Earnings of Management on the ;£ioo will be nearly the same for the long period and for the short : so that the annual rate of profits will be much lower in the former case than in the latter. For this reason the annual rate of profits on the total capital employed is, as a general rule, lower in trades which make great use of Fixed capital, than in trades in which nearly all the capital is Circulating. §4. We have next to examine the consequences of the fact that the management of a large capital in any trade almost always requires rarer natural abilities and a more expensive training than the management of a small capital requires. We have seen that a man who conducts a large business must look far ahead, and wide around him ; and that he must be continually on the look out for improved methods of carrying on his business, while the man who manages a small business may be content to follow the lead that is given to him by his neighbours. The former pays subordinates to do the work on which the latter spends the greater part of his time ; and devotes all his energies to planning, and organizing, to forecasting the future and preparing for it He must have a knowledge of men and the power of managing them. He must select his subordinates well, and while keeping the control of the business in his own hands, he must give them the freedom which will call forth their energy and sense of responsibility. Those who cannot do this, are incapable of building up a large business, or even of keeping one together, if inheritance or other accident should put them in possession of it. A man who has all the rare qualities that are required for managing a large business will, unless he is specially unlucky, make a high rate of profits on his capital. These profits will increase his capital, and will encourage him to devise and cany 140 BOOK H. CHAP. XII. § 5. out bold plans on a broad basis. The confidence that others have in him will enable him to borrow capital easily ; and thus, because he has the faculties which are one condition of getting high Earnings of Management, he will rapidly acquire that control of a large capital which is the other condition. W e see then, firstly, that higher faculties are required for the management of a large than of a small capital ; and secondly ; that there is a process of selection continually going on by which V/ those who have some capital and great business power, soon get v' control over a large capital ; while on the other hand those who have not business power will speedily dissipate a large capital if they happen to get control over it. These facts shew that - the Earnings of Management in large businesses must be on the ^average higher than those in small; and they even give some reason for thinking that the average Earnings of Management în different businesses in the same trade vary almost in proportion to the capital employed. There is however an independent and stronger reason for believing that there is often the same average rate of profits on different capitals in the same trade. Let us suppose for instance that A and B are proprietors of neighbouring cotton factories which are alike in every respect excepting that A ’s is twice as large as B’s. They hire labour and they buy their raw cotton, machinery, building materials, &c., in the same market and at the same price. There may indeed be a few slight economies in A ’s business of which B cannot avail himself ; and on the other hand A may have to pay subordinates for doing some of the work that B with his smaller business finds time for doing himself. But if these differences be neglected, all the Expenses of production, other than profits, of a yard of A ’s calico will be the same as those of a yard of B’s ; and since they sell in the same market at the same price, the profits made on each yard of calico will be the same for A as for B. The rate of A ’s profits will be the same as that of B’s : A’s Earnings of Management will be twice those of B. The results which theory thus indicates are confirmed by experience. Experience and theory alike tell us that as a general rule there is a constant tendency to equality of the rate of profits not only on equal capitals, but also on unequal capitals in the same trade, and in trades that are equally disagreeable and difficult. But there are three important exceptions to this rule. § 5. The first exception arises from the fact that the head of a large business often pays wages to subordinates to do a great deal of work that the head of a small business does for nimself, and the payment for which is reckoned among his profits. For instance the average rate of profits made by small farmers is higher than that made by large ; because the small EARNINGS OF MANAGEMENT. 141 farmer’s profits include the wages of the work of his own hands, and of supervising hired labour more closely than the large farmer can. Again the rate of profits on a shopkeeper’s capital, particularly in some of the clothing trades, is generally higher for a small than a large capital, even where the two sell to the same class of customers. For the small shopkeeper includes among his profits the earnings that he gets by attending care\* fully to the special wants of each customer ; but the large shopkeeper has to pay high wages for this work. The second exception is closely connected with the first. There are many trades in which small makers and dealers are able to sell at a higher price than the large dealers can, because they get access to a different class of customers. One familiar instance of this is the fact that village shopkeepers generally get a very high price for their goods. Their capital is very small ; and their profits, though at a high rate on their capital, are so small in amount as not to attract competition. Again in money lending, the smaller the scale on which the business is transacted, the greater is the charge that is made for the loan of money. A man who troubles himself to lend money by a few pounds at a time, can often obtain a very high rate of interest for it. To take an extreme case, there are men in London and Paris and probably elsewhere, who make a living by lending money to costermongers. The money is often lent at the beginning of the day for the purchase of fruit, &c., and returned at the end of the day, when the sales are over, at a profit of ten per cent.; there is little risk in the trade, the money so lent is seldom lost Now a farthing invested at ten per cent, a day would amount to a billion pounds at the end of a year. But no one can become rich by lending money in this way ; because no one can lend much money in this way. The profits on the capital really consist almost entirely of the wages of work for which few capitalists have a taste. The third exception arises from the influence of the Law of ^ / Increasing Return. In many industries a large capital can avail itself of great economies that are out of the reach of a small V capital; and the large manufacturer can make higher profits than the small manufacturer. These industries would rapidly be concentrated in the hands of a few wealthy firms, if a man whose E ractical genius has created a large business, could ensure that is successors for several generations should have a like genius. But in the whole course of history we meet with but very few instances of private firms which have been managed with eminent genius for three generations in succession. The sons and grandsons of a successful man of business have seldom that rare combination of ability and assiduity which would enable them to carry on his work. And there are many instances h' 142 which a vast inherited business has been quickly destroyed by men who could have managed a small business well. In some industries large capitals have completely driven their smaller rivals from the field, and afterwards their competition among themselves has reduced the rate of profits very low. In rolling mills for instance there is little detail which cannot be reduced to routine, and a capital of ;£ 1,000,000 invested in them can be controlled by one able man. A rate of profits of 20 per cent, which is not a very high average rate for some parts of the iron trade, would give the owner of such works Earnings of Management amounting to more than £ 150,000 a year. And since iron-masters can with so little additional effort get the Earnings of Management on an increased capital, wealthy men remain in the trade longer than in most others ; and the competition of the great iron-masters with one another is said to have reduced the average rate of profits in their trade below the ordinary level § 6. We are now in a position to sum up our inquiry as to the way in which Normal Earnings of Management are determined. Firstly, with regard to the demand for business power. It is true that this demand is not measured by any definite market price list of Earnings of Management, such as that which states that carpenters\* wages in a certain'town are ninepence or tenpence an hour; and it is true that the fluctuations of Earnings of Management are greater than those of wages, because the fluctuations of trade prosperity exert a more direct and a greater influence on the incomes of employers than on those of the employed. But yet the demand for the aid of business power in production is fundamentally of the same kind as the demand for the aid of skilled labour. For instance if a manufacturer can improve the method of carrying on his business so that the work of four hundred men produces as much as that of five hundred men did previously, then he will gain an addition to his Earnings of Management equal to the wages of a hundred men. Thus the Earnings of Management of a manufacturer represent the value of the addition which his work makes to the total produce of capital and industry : they correspond to the effective demand that there is for the aid of his labour in production, just as the wages of a hired labourer correspond to the effective demand for his labour. The Law of Demand tells us that the value in exchange of anything is the measure of its Final value in use ; that is, of its value in use to those who are only just induced to purchase it ; and that this Final value in use diminishes as the supply of the thing increases. So it is with regard to skilled labour of any kind; every increase in the supply of it tends to diminish the Final value in use of the work it does, and therefore to lower its wages. And so it is with BOOK II. CHAP. XII. § 6, 7. EARNINGS OF MANAGEMENT. M3 regard to any order of business power ; every increase in the supply of it tends to diminish the Final value in use of the work it does, and therefore to lower its Earnings of Management. Secondly, with regard to the supply of business power. Returning to the case of the manufacturer who obtained high Earnings by an improvement in his methods of manufacture, we see that his success will induce others to follow in his steps, and that their competition will force down his Earnings. The extent to which they will be forced down depends upon the number of those who are able to do the work ; and this will depend on the Earnings to be got by it on the one hand and on its difficulty on the other. Thus the rarity of the natural abilities and the expensiveness of the special training required for the work play the same part in determining Normal Earnings of Management that they do in determining the Normal wages of skilled labour. In either case a rise in the income to be earned sets in operation forces tending to increase the supply of those capable of earning it ; and in either case the extent to which the supply will be increased by a given rise of income, depends upon the social and economic condition of those from whom the supply is drawn. The conditions which determine the Normal demand for, and the Normal supply of each kind of business power being known, its Normal Earnings of Management are determined as those which will equate supply and demand in the long run. Thus the Law of Normal Earnings of Management is similar to that of the Normal wages of skilled labour, and similar also to the Law of Normal value for commodities1 ; and is :— The Normal supply of each kind of business power is that to y which the field of employment will just afford the Earnings L of Management which are required to call forth this supply : 7 and the rate of the Earnings of Management so determined is \ the Normal rate for this order of business power. J The conditions which thus determine this Normal rate vary from place to place and from age to age ; but since business power is easily transferred from place to place, variations of the Normal Earnings of Management between different places are less important than those which occur from age to age. § 7. But in spite of the fundamental similarity between the Law of Normal Earnings of Management and that of Normal wages of skilled labour, there are several important differences between the two cases. Firstly, the Earnings of Management which a business man gets, depend upon the capital with which he has started, in the same way as, but to a greater extent than, the income of a professional man depends on the start in life which the social position of his parents gives him. And a 1 Comp. Book II., ch. v., § 3. 144 business man working with his own capital includes among his Earnings the equivalent of that Insurance against Persona] Risks which must be allowed for, in some form or other, by those who work with borrowed capitaL Secondly, business men are chosen by a process of natural selection from among many millions of competitors. For many employers of labour, in soma parts of England more than half, have risen from the ranks of hired labour. Every artisan who has exceptional natural abilities has a chance of raising himself to a post of command, and is in fact a candidate for the prizes that may be earned by success in business ; and the average of these Earnings of Management is high, partly because the class of employers contains, in addition to the able men that have been born within its ranks, a large share of the best natural abilities that have arisen among the lower ranks of industry. While Profits on capital invested in education is a specially important element in the incomes of professional men, Rent of rare natural abilities is a specially important element in the incomes of business men. The total amount of the Earrings of Management got by business men in a country may be found by subtracting interest on the whole amount of their own capital from their total net incomes after allowing for all expenses and losses : and, even after allowance has been made for Insurance against Personal Risks, this amount certainly gives a very high rate of wage for the skill and ability of business men. But this rate is not so high as at first sight appears ; for great deductions must be made on account of those who have lost their capital in trade. The earnings of the labour that these men have wasted, together with all the capital that they have lost, must be deducted from the Earnings of Management got by successful men before the average Earnings of Management can be found. Those who fail are quickly lost from sight and memory ; but their number is very great. It is said that in America three-fourths of those who engage in trade become insolvent in the course of the first five years1. § 8. The supply of skilled labour is increasing faster than that of unskilled labour, and the supply of business power is increasing faster than that of the lower kinds of skilled labour. Thus the competition of business power for the aid "of the lower orders of labour in production is increasing. And as a consequence the Earnings of Management that can be got by doing work of a given order of difficulty— the TaskEarnings of Management— are diminishing. The continual increase in the complexity of business, and the continual increase in the amount of capital that can be employed in business 1 Bowen, American Political Economy% ch. X. BOOK II. CHAP. XII. § 8. under a single management, are indeed givi P » business ability of the highest order the opportunities^of-iobtaining greater Earnings of Management than were evefcjfeard of in earlier generations. But the total amount of the Earnings of Management is not so high in proportion to the amount of capital employed as it used to be. And the ratio which the Earnings of Management of a business bear on the average to the capital employed goes on diminishing; and will probably continue to diminish. For the growth of education will increase rapidly the supply of business power that is competing for the aid of hired labour in production ; and this competition will prevent the Earnings of Management from growing as fast as capital is likely to grow. Since the Normal rate of Interest is likely to fall, and the ratio which Normal Earnings of Management bear to capital is likely to fall, and since profits are composed of interest and ") Earnings of Management, therefore the Normal rate of profits J is likely to fall, it will not fall rapidly for a time, and then remain stationary at a minimum. But subject to some oscillations, its fall will probably be continuous, though increasingly slow, so long as the world is inhabited by men of the v same nature with ourselves. EARNINGS OF MANAGEMENÿtf^ ' ' « 4 5 ' M. i ! C H A PT E R X III. RELATION OF NORMAL TO MARKET VALUE. § I. W e have seen that “ every producer of a commodity calculates the price which he will be able to obtain for it, and the Expenses of producing it, and determines by this means to what extent it is his own interest to increase or diminish his production. If there is free competition, his interest leads him to act in the same way as he would if his only object were to regulate the amount produced so that it could just be sold off at a price equal to its Expenses of production.” Thus the I Normal value of a thing— that toward which the Market value I continually tends— is equal to its Expenses of production. 1 These Expenses of production may ultimately be resolved into wages and profits, or rather into earnings and interest. We have seen how each Expense of production measures that effort or abstinence which is the corresponding element i/ of Cost of production. We have seen that the interest which can be got by abstaining from the immediate consumption of ;£ioo worth of wealth and saving it to be used as capital, is a practically fixed and uniform amount at any given time and in any given country. This Normal rate of interest depends on the one hand on the field that there is in that time and country for the employment of capital, and on the other hand on the supply of capital ; this supply depends on many causes, one of which is the rate of interest that has hitherto prevailed in the country. The Normal wages of unskilled labour in any time and country depend on the scope for its employment on the one hand and on its supply on the other. The scope for its employment depends partly on the natural resources of the country, partly on the amounts of capital, of skilled labour and-of business power that are seeking its aid in production. The supply of labour depends, as regards both quantity and quality, on many causes, the chief of which is the rate of wage>. "it true that things which are luxuries in one stage of cb’i/.i :ic\ are regarded as necessaries in another, and that all sud'. cIil R E LA TIO N OF NORM AL TO M ARKET. V ALU E. 147 affect the Normal wages of unskilled labour. But in any given phase of civilization the Normal Task Wages of unskilled labour are nearly constant. So with regard to the earnings of skilled labour, including the incomes of professional men, and of all others who render skilled service for payment. The earnings of each kind of skilled labour depend on the scope for its employment on the one hand, and on the supply of it on the other. The expensiveness of the education, and the rarity of the natural qualities required for it are the chief of the conditions which determine the supply that will be called forth by any given rate of wages ; so that the supply is governed by laws similar in many respects to the Law of Normal Supply of commodities. The Normal wages of skilled labour of any given degree of difficulty, may vary slowly. But at any time and place they are determined by the social and economic condition of the people, and they may be said to measure the efforts involved in the work. The fact that Earnings of Management can be obtained only by those who have the control of capital, does not prevent their Normal value from being determined substantially in the same way as the wages of skilled labour. That part of the Normal Earnings of Management which can be got by a man who works with borrowed capital, measures the difficulty of his business. A man who conducts a similar business with his own capital obtains in addition th.e equivalent of Personal Risks. Thus the Cost of production of a thing is measured by its Normal Expenses of production. If the difficulty of producing a thing, or its Cost of production is independent of the amount produced, Cost of production determines Expenses of production and therefore determines Normal value. But when the amount produced is increased, the Cost of production may increase according to the Law of Diminishing Return, or may diminish according to that of Increasing Return. In order to cover these cases, the Law of Normal Value must be stated thus:— “ The Normal supply of the commodity is such, that its Normal Expenses of production equal the value, which will call forth a demand for this amount ; and the price so determined is the Normal value.” Normal value still measures Cost of production, but is not determined simply by it. [§ 2. This is an instance of the rule that in Nature changes generally react on one another. For instance it is not true that the state of a man's lungs is determined by that of his heart, or vice versa; but subject to external influences, the conditions of it, lungs and other parts of his body determine one i.'i. So when two unequal balls A and B are put into a smooth basin, it is not right to regard A ’s position as determining the position of B. For though it is true that if v 10— 2 148 BOOK II. CHAP. X III. §3. know exactly where A is, we can tell at once where B is, it is equally true that if we know where B is, we can tell where A is. The positions of A and B are determined simultaneously by the action of the Law of Gravitation. So it is with regard to Normal value. It is true that in the exceptional case in which the difficulty of production of a thing is fixed independently of amount produced, Cost of production determines Normal value. But as a rule the Cost of production of a thing is not fixed : the amount produced and its Normal value are to be regarded as determined simultaneously under the action of Economic Laws. It is then incorrect to say, as Ricardo did, that Cost of production alone determines value : but it is no less incorrect to make utility alone, as others have done, the basis of value. It is certainly true that utility is a condition of value always; and that in cases in which the supply of the commodity is fixed, utility determines price. It is true that the price of every commodity must be the measure of its Final utility; that is of its value in use to those who are only just induced to purchase ] it. But it is not true that this Final utility determines value : for it changes itself, according to the Law of Demand, with every change in the amount of the commodity that is offered for sale. This amount, and therefore the Final utility of the commodity, depend upon the relation between the circumstances of supply and those of demand.] § 3. W e have now to pass from the theory of Normal value to that of Market value. Normal results are those which would be brought about by competition if it acted freely, and always had time to cause those effects which it has a tendency to cause. Market results are those which actually are brought about by the complex social and economic forces of the world in which we live. We have compared Normal value to the Normal growth of a tree. Let us now compare it to the Normal tides which there would be if there were no disturbing wind, and no irregularities of coast line. Observation tells us that waves driven by the wind over the sea make its surface rapidly rise and fall ; and that the irregularities of coast line pile up the tidal waves in some places ten times as high as they are in the middle of the ocean. The theory of the Normal tides does not tell us what is the highest point that the tide leaches at any place in (say) the Bristol Channel We cannot find out this without examining the special circumstances of the case, and allowing for the influence of the winds, and the peculiar nature of the shores. But on the other hand we can make no progress in explaining the movements of the sea unless we first understand which of ^em are due to local or transitional causes, and which to the RELA TIO N OF N ORM AL TO M A R K E T V ALU E. 149 Normal influence of the attractions of the Moon and Sun. And we cannot do this until we have first worked out the abstract theory of the tides that would be formed in a world in which there were no disturbing winds and no irregular shores. So with regard to the theory of Normal value. It does not tell us what will be the wages of a certain work, or the price of a certain thing at any particular time. We cannot discover the Market value of a thing without allowing for the fluctuations of supply and demand, and for the resistance which local obstacles oppose to the free movement of the stream of competition. But on the other hand we can make no progress in explaining the movements of wages and prices, unless we first understand which of them are due to local or transitional causes, and which to the Normal action of free competition. Tfte theory of Normal value is the starting point from which we must set out to explore all the various irregularities and unevennesses of Market values. It teaches how the great tidal waves of wages and prices would move if every one were careful to forecast the future, and deliberately to shape his course so as to obtain the greatest economic advantages for himself and his family. It puts us in the right position for examining how man’s action is modified by custom, or apathy, or generally by motives other than the desire for wealth. \* W e can then apply our theory to explain facts, so far as it will go; and those facts which cannot be explained by our theory are “ light-giving” facts, and shew us how to correct and enlarge our theory. Thus the science of Economics progresses step by step, alternately applying theory in the search for and explanation of new facts, and applying new facts in correcting and broadening and strengthening theory. BOOK III. M A R K E T VALU E. C H A PT E R I. CHANGES IN THE PURCHASING POWER OF MONEY. § i. THROUGHOUT the discussion of the Theory of Normal value it was assumed that the purchasing power of money remained unchanged1 ; so that a rise or fall in the exchange value or general purchasing power of a thing could always be shortly expressed as a rise or fall of its price. W e must now inquire briefly how the value or general purchasing power of money changes from time to time. But a full discussion of the theory of the value of money belongs to the “ Economics of Trade and finance.” The most obvious of the causes that affect the purchasing power of the precious metals in a country is the quantity of them that is available for use as money. If this increases very fast, there will be more than is wanted to carry on the business of the country at the old prices, and prices will rise. On the other hand, if the amount of the precious metals remains stationary while the population and wealth of the country increases, there will be a great demand for money to carry on the business of the country; the purchasing power of the precious metals will rise, and prices will fall. For instance at the beginning of the sixteenth century when the new supplies from the American silver mines made themselves felt, the purchasing power of silver began to fall; and early in the seventeenth century prices in London were on the average three times as high as they were in 1500. Again prices were high at the beginning of the present century. But no 1 Bk. 11. ch. i. § 4. important fresh supplies of metals came from the mines till 1850. During that time the stock of precious metals was being diminished by their use in the industrial arts and by wear and tear; and meanwhile population and wealth were increasing rapidly. So the purchasing power of gold rose ; and prices fell to about half what they were in 1800— 10. About 1850 the gold mines of California and Australia were discovered, there was a great increase in the supply of the precious metals, and prices rose again. § 2. But though the amount of the precious metals in circulation is the most obvious of the causes that affeçt the purchasing power of money, a no less important cause is the growth of artificial substitutes for the precious metals as a medium of exchange. The most familiar of these substitutes are bank notes. They pass freely from hand to hand, and exert nearly the same influence over prices as an equivalent amount of coined money does. But in England this influence is not as important as that exercised by cheques, which have displaced both coin and bank notes in nearly all wholesale and in many retail transactions. A cheque does not circulate freely, but is generally given by the person who receives it to his banker, who demands payment of it for him. But though cheques do not act as substitutes for coin in the same way as bank notes do, yet the total amount of them is so great as to exercise a very powerful influence over C rices. Again the modern system of credit enables a man who as neither money nor anything that immediately represents money, to obtain from a banker or other money dealer the means of purchasing goods. He can do this not only on his own credit (as when a bank allows him a “ Book credit”], but on the credit of others who have undertaken to pay him money at a future date (as when he “ discounts a Bill”). The business of the civilized world has increased very rapidly during the present century ; and an enormous amount of coin would have been required to carry it on with the present prices. If credit had not found substitutes for coin, there would have been so great a demand for the precious metals, that their purchasing power would have become many times as great as it actually is; prices would have been very low. The growth of credit supplies a permanent substitute for the precious metals, and therefore affects their Normal values. But credit fluctuates, and each fluctuation alters their Market values. For instance, an expansion of credit coincided with the influx o f precious metals consequent on the discovery of the Californian and Australian mines, and increased the upward tendency o f prices. But in 1857 there was a crisis; that is, many trading CHANGES IN THE VALUE OF MONEY. 151 152 BOOK III. CHAP. I. §3. firms were unable to pay their debts, credit was violently contracted, and prices fell, although the store of precious metals in the country was growing as rapidly as ever. After a time credit began to expand again, and prices rose till 1866 when there was another crisis^ and prices fell. Again credit expanded, and prices rose till 1873; when, though there was no crisis, a gradual contraction of credit set in which has continued till 1879. The lowest point which prices reached between 1857 and 1866 was much higher than the level of 1850 ; and the lowest point between 1866 and 1873 was higher still. But since then there has been a slight check in the supply of gold; and a great deal of gold has been absorbed by the adoption 0/ a gold currency in Germany and other causes; and prices, measured in gold, are now\* (1879) as low as they were in 1850. § 3. The beginning of a period of rising credit is often a series of good harvests. Less having to be spent in food, there is a better demand for other commodities. Producers find that the demand for their goods is increasing ; they expect to sell at a profit, and are willing to pay good prices for the prompt delivery of what they want Employers compete with one another for labour; wages rise; and the employed in spending their wages increase the demand for all kinds of commodities. New public and private Companies are started to take advantage of the promising openings which shew them selves among the general activity. Thus the desire to buy and the willingness to pay increased prices grow together ; Credit is jubilant, and readily accepts paper promises to pay. Prices, wages and profits go on rising : there is a general rise in the incomes of those engaged in trade: they spend freely, increase the demand for goods and raise prices still higher. Many speculators seeing the rise, and thinking it will continue buy goods with the expectation of selling them at a profit. At such a time a man, who has only a few hundred pounds, can often borrow from bankers and others the means of buying many thousand pound»\* worth of goods ; and every one who thus enters into the market as a buyer, adds to the upward tendency of prices, whether he buys with his own or with borrowed money. This movement goes on for some time, till at last an enormous amount of trading is being carried on by credit and with borrowed money. Old firms are borrowing in order to extend their business ; new firms are borrowing in order to start their business, and’ speculators are borrowing in order to buy and hold goods : trade is in a dangerous condition. Those whose business it is to lend money are among the first to read the signs of the times ; and they begin to think about contracting their loans. But they cannot do this without much disturbing trade. If they had been more chary of lending at an earlier stage, they would simply have prevented some new business from being undertaken ; but when it is once undertaken, it cannot be abandoned without a loss of much of the capital that has been invested in it. Trading companies of all kinds , have borrowed vast sums with which they have begun to build railways and docks and ironworks and factories ; prices being high they do not get much building done for their outlay, and ' though they are not yet ready to reap profits on their investment, they have to come again into the market to borrow more capital. The lenders of capital already wish to contract their loans ; and the demand for more loans raises the rate of interest very high. Distrust increases, those who have lent become eager to secure themselves; and refuse to renew their loans on easy or even on any terms. Some speculators have to sell goods in order to pay their debts ; and by so doing they y check the rise of prices. This check makes all other speculators anxious, and many rush in to sell. For a speculator who has borrowed money at interest to buy goods may be ruined if he holds them a long time even while their price remains stationary; he is almost sure to be ruined if he holds them while their price falls. When a large speculator fails, his failure generally causes that of others who have lent their credit to.him ; and their failure again that of others. Many of those who fail may be really “ sound,” that is their assets \* may exceed their debts. But though a man is sound, some untoward event, such as the failure of others who are known to be indebted to him, may make his creditors suspect him. They may be able to demand immediate payment from him, while he cannot collect quickly what is owing to him; and the market being disturbed he is distrusted ; he cannot borrow, and he fails. As credit by growing makes itself grow, so when distrust has taken the place of confidence, failure and panic breed panic and failure. The commercial storm leaves its path strewn with ruin. When it is over there is a calm, but a dull heavy calm. Those who have saved themselves are in no mood to venture again; companies, wdiose success is doubtful, are wound up; new companies cannot be formed. Coal, iron, and the other materials for making Fixed capital fall in price as rapidly as they rose. Iron works and ships are for sale, but there are no buyers at any moderate price. Thus the state of trade, to use the famous words of Lord Overstone, “ revolves apparently in an established cycle. First we find it in a state of quiescence,— next, improvement,— growing confidence,— prosperity,— excitement,— over-trading,— convulsion,— pressure,— stagnation,— distress,— ending again in quiescence.” CHANGES IN THE VALUE OF MONEY. 153 154 BO O K III. CHAP. I. §4. c§ 4. After every crisis, in every period of commercial depression, it is said that supply is in excess of demand. Of course there may easily be an excessive supply of some particu- | lar commodities ; so much cloth and furniture and cutlery may raises the price of com \* Mill, Book iv. ch. ii. § 4. l6o BO O K III. CHAP. II. § 3. In recent times the action of merchants and dealers in equalising prices has been much aided by the steamship, the railroad, and the telegraph. Not long ago news from distant markets was many months in coming, now it comes in a few ^ minutes. Merchants can even telegraph instructions to ports at which their ships are to call ; so that a scarcity in a market many thousands of miles away can be met in a few days by the arrival of English ships1. § 3. The market price of many things is settled from day to day by the action of dealers rather than by that of producers. Many kinds of raw produce can only be produced at certain times of the year; and the immediate effect of a rise in the 4 price of such things is not to increase the production of them, but simply to induce dealers to bring forward larger quantities for sale, and perhaps to import fresh supplies from distant places. If we go into any com, or wool, or cotton market, we shall see dealers selling readily on one day, and holding back on another. The amount which each of them offers for sale at any price is governed by his calculations of the present and future conditions of the markets with which he is connected. There are some offers which no dealer would accept; some J which no one would refuse. There is some price which will be accepted by those who can least afford to wait, and by those whose expectations of the future condition of the market are least sanguine ; but not by others. The higher the price that is bid, the larger will be the sales. For instance, the conditions of a certain com market may be such that a price of 50?. would induce dealers in it to sell 500 quarters during the day ; while they would be induced to sell 700 quarters by a price of 51J., 1000 quarters by a price of 52J., and so on. Thus in any market, at any time, there is some price at which each particular amount will be offered 1 for sale. And in the same way there is some price at which each particular amount will find purchasers. Perhaps the millers and the speculators in com who attend the market would between them buy 900 quarters if they could be got for 50J. each, 700 if they could be got for 51J., but only 600 if they could not be got for less than 52s. If everyone knew exactly the state of the market, exactly how eager buyers were to buy and sellers to sell, the price would be fixed at once at 51J., and 700 quarters would be sold off during the day at this price : this-^ would be the price which would equate supply and demand. But in fact the price would oscillate up ami down during the day, and even at the same moment bargains would be struck at slightly different prices in different parts of the same çom-exchange ; 1 Comp. Crump’s New departure in Political Economy; the average price for the day would be about 51 s., and the total amount sold would not differ far from 700 quarters. Thus the Normal price of corn varies from one age to another under the slow action of economic changes : meanwhile in each age the average price between successive harvests varies from year to year\* owing to the failure of producers to adjust the supply of corn to the demand for it : and the daily price in each market is swayed backwards and forwards on either side of the \* average price for the year by the calculations and bargainings of dealers. In the case of com the demand is pretty well known beforehand 5 and the rapid fluctuations that occur in its price from week to week and from day to day, are chiefly due to imperfections in the estimates that dealers form of the stocks in existence at any time, and to changes in their forecasts of the coming harvest. There are many other things, such as coal and iron, the prices of which fluctuate from day to day chiefly in consequence of the estimates that dealers form of the present and coming demand for them. But whatever be the nature of the calculation by which the bargainings of dealers are chiefly governed, these bargainings, except when dealers combine to keep prices artificially high, tend to make the market price such as to equate supply and demand in the market. That is they tend to make the price such, that the amount which people are willing to offer for sale in the market at that price is just equal to the amount which can find purchasers at that price in the market. § 4. Let us next consider some fluctuations of price that arise from the failure to forecast changes in demand. A change of fashion often makes the Market price of some kinds of materials very much higher or very much lower than their Normal price. A manufacturer who is quick in reading the signs of the times, and anticipates the coming demand for stuffs of a particular kind, makes large profits. But in this case supply can be adjusted to the demand very rapidly ; and therefore the price cannot easily be raised much by an increase in demand unless it is very great and sudden. But it is otherwise with houses, the supply of which cannot be quickly increased to meet a new demand; their value rises and falls with changes in the prosperity of the place in which they are. When Berlin became an imperial city, there was a great demand for house room : house rents rose extrayagantly, and hod-men earned in a day more wages than agricultural labourers earned in a week. On the other hand the value of houses falls very low in places from which population is receding. There is more than one place in America in which a town of 20,000 inhabitants has grown up in a year, bu\* MARKET FLUCTUATIONS. 161 I I 1Ô2 in a few more years has been deserted, and the houses in it been left valueless. In some Cornish villages which have been deserted by their mining population, and in some other parts of England, houses can be bought for less than half of what it would cost to build them. § 5. Prices are liable to great fluctuations in trades in which there \*is a great use of Fixed capital. For when prices are at their Normal level, the price of goods not only returns with interest the manufacturer’s outlay for raw material and labour, but also gives interest on his Fixed capital with an allowance for its Depreciation, and pays his own Earnings of Management As long as the price pays him back what he spends in raw material and labour with interest, his desire to keep his workmen together, and to keep up his trade connexions may induce him to go on. Therefore, the more Fixed capital a manufacturer uses, the stronger are his inducements to keep his factory at work even after the price of his goods has fallen a long way below their Normal value, and thus to force down prices further still. Again, most of the capital that is Fixed is also Specialised. The capital for instance that is sunk in opening a coal mine, cannot be withdrawn from it when the price of coal falls ; and a prudent capitalist will not be induced by a high price of coal to invest money in coal mines, unless there is reason to expect that the price will be sustained for a long while. The price o f coal may therefore rise far before any attempt is made to increase the supply by opening up new mines ; and even after that, a considerable time will elapse before the supplies from the new mines can exert any influence. Coal mining requires also a great deal of Specialised Personal capital ; and a sudden increase in the supply of coal cannot be got without suddenly increasing the number of miners. Men who are not fam iliar with the dangers and discomforts of underground life, cannot be attracted to it except by high wages; and since their labour is at first very inefficient, the wages paid for raising a ton of co a l are very much higher than the Normal wages of the work. And therefore the increased demand is able to raise the price o f coal very high without calling forth such an increase of supply as to force the price quickly down again. On the other h a n d when the demand for coal falls off, the Material and Personal capital that is specialised in coal mining keep up the su p p ly, and the price is forced down very low. The use of coal fo r domestic purposes varies but little from year to year ; the a m ou n t of coal that is wanted for making steam varies a little m o re ; but by far the most important variations in the demand for c o a l are due to its use in the manufacture of iron. § 6. There are great fluctuations in the price of iron, p a r t ly BOOK III. CHAP. II. § 5, 6. because a rise in price does not increase quickly the supply of the coal and iron ore, and of the Fixed capital that are wanted in its manufacture ; partly because the demand for iron is subject to violent changes, in consequence of its being chiefly used in making machines, railroads, and other forms of Fixed capital. The demand for Fixed capital is liable to more extreme fluctuations than the demand for commodities that are wanted for immediate consumption, and the trades which make Fixed capital are more affected than any others by alternations of commercial prosperity and adversity. For while credit is expanding, the extra purchasing power which credit gains goes chiefly to traders and trading companies, who, whether they want it to begin or to extend their business, are sure to spend a great part of it on machinery, buildings, ships, railway material, and other forms of Fixed capital. On the other hand where credit is contracting many find their means of purchasing J altogether cut off, while those whose means are not straitened do not care to invest in Fixed capital until theyJfhink prices have nearly reached their turning point. The price of pig iron was doubled, and the price of some kinds of coal was quadrupled in the years 1870— 1873, but now their prices are lower than in 1870 ; the wages of labour in the iron trades rose fifty per cent, in the years 1870— 1873, and have again fallen ; and similar changes have occurred with each fluctuation of general prosperity. The price of iron in England and other countries rose more rapidly than almost any other prices in the years preceding each of the crises of 1837, 1847, 1857, and 1866 ; and it fell more rapidly than most other prices in the years following each of these crises ; so that not only the price, but also the value or general purchasing power of iron rose and fell at each of these periods. Nearly the same may be said of the building trades as of the iron trades, the wages of navvies and of masons, bricklayers and carpenters have risen rapidly before, and fallen rapidly after, almost every crisis. In some periods of commercial depression almost the only demand for Fixed capital comes from manufacturers who do not like to close their works for alterations while trade is brisk ; but who, when trade is slack, and the prices of building and machinery are low, take the opportunity of making such extensions and repairs as may enable them to profit by the revival of trad e when it comes. But this demand is not nearly sufficient to m ake up the deficiency that arises from the general contractio n of credit, from the failure of old firms and public companies, a n d from the absence of new companies. T h e present depression of trade has been accompanied by 2. g r e a t falling off of England’s exports, but not by a corresponding dim inution of her imports. This is partly due to the fact that MARKET FLUCTUATIONS. 163 11— 2 i64 BOOK III. CHAP. II. § 7, 8. Englishmen, having suffered great losses in foreign investments, have been bringing home a great deal of their capital from abroad. When English capital is lent to foreign countries much of it is generally spent on railway lines and machinery ^ that are bought in England. This was especially the case in the years of prosperity that preceded 1873; and the subsequent decline of England’s exports injured the iron trade more than other English trades. But it did not affect the building trades, j because they do not make things for exportation. On the other hand the capital-which Englishmen brought home could not be invested here without leading to a demand for building in various forms, and so somewhat sustaining the building trades. $ This explains why it was that after 1873 prices fell very rapidly in the iron trade, but were kept lip for a long time in the building trades. \* § 7. One cause of an increase in the demand for a commodity is a failure in the supply of something for which it can be substituted. Thus the failure of the supply of cotton during the American war increased the demand for wool. Again, one cause of an increase in the supply, and consequent fall in the price of anything, is an increase in the demand for sdmething \*| that is produced with it. The prices of the gas and the coke that are got from a ton of coal, must together be enough to \* cover their joint Expenses of production. If the demand for gas rises, more coke will he produced, and its price must fall, so that the increased supply may be taken off the market. The rise in the price of gas must be sufficient to cover this fall in the price of coke, and also to cover the increase, if there is any, in the joint Expenses of production of gas and coke. Again, since the repeal of the Corn Laws much of the wheat consumed in England has been imported, of course without any straw. This has caused a scarcity and a consequent rise in the price o f straw, and the farmer who grows corn looks to the straw for a great part of the value of the crop. The value of straw then is high in countries which import com, and low in those which export com; In the same way the price of mutton in the wool-producing districts of Australia was at one time very low. The wool was exported, the meat had to be consumed at home; and as there was no great demand for it, the price of the wool had to defray almost the whole of the joint Expenses of production of the wool and the meat. A fte r ­ wards the low price of meat gave a stimulus to the trade o f preserving meat for exportation, and now its price in A ustralia is higher. Similar remarks apply to that part of the Expenses of p ro ­ duction of a thing which depends upon transport, whether o f the raw material to the place o f,manu facture, or of the finish ed MARKET FLUCTUATIONS. commodity to the place of sale. The expense of making and keeping in repair roads and railways is shared among the different things that pass over them. When the discovery of mines in a district leads to the making of railways to it, the inhabitants at once find that they get a higher price for everything that they produce to be sent away to distant markets, and that they have to pay a lower price for things which they buy from a distance. Again the exports from England being less heavy and bulky than those from America the competition for freights from England is so great as to keep them very low, and thus to throw the greater part of expense of working the ships on the freight of goods from America. Any increase in the American demand for the heavier and bulkier kinds of English goods increases the competition of ships for freights from America, and so tends to lower the price 01 American goods in England. Again, if a commodity which is heavy but not bulky, is exported in the same ships with another which is bulky but not heavy, the freights which either of them has to pay may be very small or very great according as the demand for the other of them is very great or very small. Fdr instance, ships whose chief cargo consists of iron rails, often carry light Staffordshire potteries for very low freights. And tin is brought from Australia in wool ships as ballast, almost the whole expenses of the ships for the double journey being borne by the wool : but if the production of minerals in Australia were to increase, and that of wool to decline, the wool might be carried for very low freights. § 8. When an increase in the demand for a commodity raises its price, the gain at first goes almost entirely into the hands of the manufacturers\* But soon their eagerness to extend their business leads them to compete with one another for the hire of labour, and gradually wages rise till a large part of the gain is transferred from the employers to the employed. Conversely when the demand slackens and the price of the commodity falls, the burden of the fall is at first borne entirely by employers ; but gradually it is in a great measure transferred from them to the employed. The manner in which wages rise in the one case and fall in the other, and the distribution of the rise or fall among the various classes of workmen who are employed in producing the commodity, are a good deal influenced in England at the present time by trade combinations ; and they will be most conveniently discussed at a later stage. But it may be noticed here that movements in wages almost always follow, and scarcely ever occasion, movements in prices. A rise in price is occasioned by an increased demand : after a time wages rise ; still the demand increases, and still the 1 6 5 price rises ; but the further rise in price is occasioned as the first rise was by the fact that demand has increased more rapidly than supply. It is true that the rise in wages seemed to play some part in sustaining the upward movement of prices ; ^ for if wages had remained low, and the whole gain of the high prices had gone to profits, capital would have been attracted into the trade more rapidly, the supply would have increased 1faster, and therefore prices would not have risen so much. ! But still each rise in wages is caused directly by a rise in price. I The upward movement in wages is scarcely ever so great in proportion as the upward movement of prices, and therefore scarcely ever so great as the upward movement of the Earnings | of Management. One cause of this is that the rise in price goes in the first instance, as has just been said, entirely into the ! pockets of the employer ; and the rise in wages seldom begins until the rise in profits has called forth increased competition among employers. But another cause is that a great deal of risk is involved in bringing additional capital into a trade to meet what is not unlikely to be only a temporary increase in I demand ; and part of what appears as a rise in profits is really | only insurance against this risk. «• In just the same way when the demand slackens, prices begin their downward movement first ; the fall of wages begins ] later on ; it is occasioned by and is not the cause of the fall in I prices; and the fall in price and the consequent fall in profits is } much greater in proportion than the fall in wages. At such times indeed there is often no profit at all, but a great loss of capital. This loss may be unequally distributed among the various producers, but taking trade as a whole, it may be regarded as paid out of that insurance fund which was got together in times of high prices and high profits. C §9- When the wages of any class of labourers have been I raised by a rise in the price of the commodity produced by | them, the rapidity with which an additional supply of labour 1 comes into the trade depends on the relation in which these wages stand to those of other trades. If the wages in this trade . are abnormally high relatively to others, the rise is likely to attract so much additional labour, as to prevent the upward movement of wages from going very far, and to make them fall | fast and far when the time of depression comes. On the other 4 hand, if, before the rise came, the wages were below their I Normal level relatively to other trades, the rise may go on for I a long time without bringing in much additional labour, and | whatever rise is gained will probably be maintained. Thus we see how the Law that Normal value is determined by Normal Expenses of production is consistent with the fact iC 6 BOOK III. CHAP. II. § 9. ! I that market fluctuations of value are the cause and not the consequence of market fluctuations of Expenses of production. If Ricardo and Mill had taken more pains to make clear the distinctions between the theory of Normal value and that of Market value, there could not have been as much controversy as there has been on the question whether value is governed by Expenses of production, or Expenses of production by value.] MARKET FLUCTUATIONS. 167 C H A PT E R III. LOCAL VARIATIONS OF PRICES AND WAGES. INFLUENCE OF CUSTOM. § i. W h e n seeking for the Law of Normal value of a commodity, we noticed that the same commodity may have different Normal values in different markets. Its Normal value in any market is equal to its Expenses of production there, and among these is to be reckoned the Expense of carrying it there, from the place at which it is made, including of course any customs duties that have to be paid on the way. Thus the Law of Normal value contains in itself the following Laws of Local Variation of Normal prices : If two markets are supplied with a commodity from the same source, its Normal value is higher in the more distant (or less accessible) market by the difference between the expenses of carrying it from the place where it is made to those markets. If there are two places in which a commodity is made for sale in the same market, its Normal value is lower in the more distant (or less accessible) of these two places by the difference between the expenses of carrying it to this market from the two places. These laws take it for granted that if the difference between the prices of a commodity in two markets is greater than the expense of carrying it from one to the other, some one will se% to work to bring it from the cheaper to the dearer marketThey assume that there are men connected with these markets who have the capital and the business habits required for the work; and that the demand for the commodity is on so great a scale as to make it worth while to organize a traffic in it. These conditions are fulfilled when the commodity is one in general use, and when the markets are large towns in close commercial intercourse with each other; local variations in the wholesale prices of the staple wares of commerce are, with a few exceptions, to be attributed to differences in the expense LOCAL V A R IA T IO N S OF VALU E. 169 of transport The most important of these exceptions occur when some producers are very anxious to force their way into a distant market. The competition of others who are more advantageously placed for supplying this market may induce them to sell at a price so low that, after allowing for the expenses of carriage, it does not afford them the ordinary rate of profits ; while perhaps they make up this deficiency by combining with one another to sfell at a high price in their own markets. Thus for instance English manufacturers sometimes sell goods in America at a price which, after allowing for taxes and expense of transport, is lower than that at which they sell in England ; and some American goods, such as sewing machines, are sold in Canada at a lower price, allowance being made for the customs duty, than in the United States. But the tendency of competition is to remove anomalies of this kind by breaking up local combinations, and compelling the producers to sell in the home market at a price equal to the Expenses of production of the commodity there. Where there is no organized traffic, prices are not determined by free competition : their local variations are not governed simply by the above laws. Some account then must be given of the local variations of the prices of things that cannot easily be sent long distances to market, and of retail prices generally. § 2. First among the things which cannot be sent to market is land. But railways now enable many kinds of agricultural produce to be sent gregt distances to market ; and thus bring them under the Laws of Local Variations of Normal prices. And since the rent of land is the excess of the price which its produce fetches, over the expenses of raising it and sending it to market, it follows that, if land were always let by free competition for agricultural purposes, rent would be brought under the indirect influence of these laws. And the value of land is in a great measure determined by its rent. For suppose the rate of interest for safe investment were four per cent., so that people could obtain a secure income of ;£ioo a year from the investment of ^2500. Then if rent were the only advantage which land gave to its owner, and if rent were not likely to rise, people would be willing to pay just ,£2500 for land that would yield them a secure rent of ^ ioq a year. But when people buy land they often look forward to a rise in its rent. The opening up new fields for agriculture in America and increased facilities of transport may indeed check the rise in the value of agricultural produce, and so check the rise in the rent of English farms; but the rise in the rent of land near large towns and in manufacturing districts seems likely to continue without much interruption. This rise 170 BOOK III. CHAP. III. §3. jin as far as it is caused by the growth of population and independently of any action on the part of the landlord, has been called an “ unearned increment ” of the rent of land. When this has been estimated, the amount that is added to the value < of the land on account of it, can be found by a simple arithmetical calculation. But this is an incomplete account of the causes which 1 determine the value of land. For firstly, even where competi- ' tion is perfectly free, allowance must be made for the other Iadvantages which land affords besides the right to receive its !rent Some people derive a peculiar pleasure from the owner- 1 ship of land ; they love their land as they love their dogs, and they are as willing to pay for the gratification of their affection ^ for the one as for the other. This feeling, and the social position !which land gives, raise its value further still. But the amount of this last addition depends on national character and on social arrangements which vary from one country to another and from one time to another ; and these variations obey no law. Secondly, the rent of land is seldom determined by perfectly | free competition. The comfort of the landlord, the social posi- I tion which counts for much in the value of his land, require 1him to live on cordial terms with his tenants ; and he is seldom | anxious to drive hard bargains with them. The competition for a farm is often practically limited to a few families in its | neighbourhood, or rather, to those who get on well with the j landlord in personal and social and political matters ; and the !landlord would often find it difficult, even if he were inclined ' to do so, to exact the highest rent which the land could be made to pay. On the other hand it is true now, as it was in Adam Smith’s time, that “ upon equal dr iieârly equal profits men will choose |to employ their capitals rather in the improvement and culti- $ vation of land than in manufactures or foreign trade.” In |some parts of the country, in which landlords grant long leases to their tenants, and enable them to invest their capital securely Iin the land, competition raises rents so high as to make the 1farmer’s rate of profits lower than that in almost any other Itrade. ' § 3. The hindrances to the free play of competition in deter- Imining the rent of land are relics of the past. In the village < \* Rules of the Co-operative Union. CO-OPERATION. of their wages or of their respective purchases during the quarter. But a movement begun with high aims often suffers from its own success ; and even at Rochdale the new comers, who have been attracted to the store by the desire of gain, have outvoted the older members, and divide the whole net profits among the members without giving any share to labour. Much of the good done by these stores is due to their plan of charging a full price at first, and returning to the consumer a share of the profits in a lump sum at the end of the quarter : for the workman is thus induced to put away week by week part of his earnings almost without knowing it. He may then withdraw his savings, perhaps to buy a sewing-machine, or some important piece of furniture ; and thus he is often led to aim at having a well-ordered home in which to take pride. If he has already such a home, the best thing he can do with his a dividend ” is not to withdraw it, but to take his part in “ saving joint capital by joint action for joint purposes.” In such stores there is much real co-operation : many of the members take an active interest in the way in which the business is carried on, are ready with help and advice whenever it can be useful, and take part in the selection and the supervision of the officers of the society. They themselves are educated by this work ; and are led on to discuss and undertake bolder co-operative enterprises partly by the setise that they have funds at their disposal which may fitly be used for the purpose, partly by their intercourse at the stores with others m whom the Co-operative Faith is strong. Their zeal is stimulated in annual Co-operative Congresses ; a Co-operative Board, Union, and Gild are continually at work, endeavouring at once to consolidate and to broaden the great movement ; and co-operators are learning that if “ education is desirable for all mankind, it is life’s necessity for co-operators1.’' § 9. The advantages which co-operative stores have in competition with shopkeepers may be classed thus : i. They have adopted cash payments as a principle. No customer is offended at being refused credit at a store : but shopkeepers find that even if they start with the cash system, they can hardly avoid meeting exceptional cases by giving credit, and soon the exceptional cases become the rule. The credit system not only leads to many bad debts, but it inverts the natural order of loans. It is reasonable that the trader who expects trade profits on his capital should borrow directly or indirectly from the private individual who can only get interest on it. But on the credit system the customer borrows the use of capital from the shopkeeper. 225 1 This truth is ably worked out in Professor Stuart’s address. M. IS 226 ii. The stores in which the co-operative element is strong, sell unadulterated goods ; and others are believed to do the same. iii. A store can do a large business without spending much on advertisements, or on an expensive site. A large store likes to have an imposing front : but this answers for a great many departments one above the other. iv. When competing with the small shopkeeper it has the advantage of buying in larger quantities, and therefore more nearly direct from the producer than he can. v. Those who have an interest in the success of the store, will wait more patiently to be served than the customers at a private shop : and therefore the amount of business done is larger in proportion to the number of those who stand behind the counter in the store than in the shop. The great difficulty that the stores have to contend with in seeking the custom of the working classes, arises from the fact that a working man who is not a unionist generally wants to borrow from shopkeepers when he is out of employment ; and that even unionists, when excited by a struggle for victory, are sorely tempted to accept the offer of friendly shopkeepers to advance the means of sustaining the strike for some time longer. And there are many in other ranks of life who cannot, or who think they cannot, avoid spending their income before they receive it. § io. It is probable that there will always be some shopkeepers who will carry on their business on the old plan, and who will retain those customers who are willing to pay high prices, on condition of receiving credit, of having assistants always ready to attend on them without delay and to shew them innumerable things which they do not want to buy, and of having every little trifle which they buy delivered at their houses at once. But all these services cost more to the shopkeepers than they are really worth to the great majority of customers. And the success of co-operative stores has proved that there is a great demand for traders who will act on the business principle of not doing any thing for the customer which must in the long run cost the customer more than it is worth his while to j>ay for it. But there is no reason why shopkeepers who adopt this principle should not be able to hold their own against all stores except those in which there is so much true co-operation as to be a great source of strength. With equal advantages the professional trader with trained skill and, specialised taste can hold his own against joint-stock companies managed by amateurs. And since in the long run the Earnings of Management of shopkeepers are governed by the Laws which were discussed in the second Book, they will, when the new system is fairly in BOOK III. CHAP. IX. § io, ii. operation, obtain as good incomes as they would have done if the old system had continued. The change will benefit shopkeepers by freeing them in a great measure from their dependence on small bodies of customers whose humours they have to consult ; they will have the sound position that comes of doing business in a business-like way. But the change will of course diminish the number of shopkeepers whose work is required1 ; and the process of thinning out must be painful ; the suffering need not however be great if shopkeepers adapt themselves quickly to the requirements of the new age, and urge all young men who have not special reasons for becoming shopkeepers to choose some other occupation. The general prosperity of the country will be much increased when the capital and labour that are now wastefully employed in the retail trade, are set free for other work. § 11. The third work of co-operators is to diminish the discords between retail and wholesale dealers and producers. For this purpose they have started a Wholesale Co-operative Society, which aims at doing for the retail stores what the retail stores aim at doing for the individual consumer. It buys directly from the producer, or importer ; it imports some things itself, and it has its own works for making biscuits, shoes, and soap. The retail stores who deal with it share in its profits ; and however little they buy, however little they understand the business of buying, they can, it is asserted, buy from it on the most favourable terms. There is a plan for making this society the centre of a grand Federation of Co-operators. It is proposed that co-operative consumers should guarantee to co-operative producers a steady market for their wares\*. Having uninterrupted employment, the co-operative producers would, it is argued, be free from anxiety, and be able to get high wages for themselves, while they produced at a low price on the average ; and the co-operative purchasers would gain by buying their goods cheaply in the first instance, and also by receiving a share of the net profits of the co-operative workshops. For it is assumed that the net profits of the co-operative workshops would be divided between the labour employed in it and the customers. The total demand of co-operative consumers is of course not constant from year to 1 It has been calculated by Mr Neale (.Economics of Co-operation) that there are 41,735 separate establishments for 22 of the principal retail trades in London. If for each of these trades there were 648 shops, that is 9 the square mile, no one would have to go more than a quarter of a mile to the nearest shop ; there would be 14,256 shops in all. Assuming that this supply would be sufficient, there are in London 251 shops for every 100 that are really wanted. \* See Book ill. ch. i. § 4. CO-OPERATION. 227 2 2 8 year; and therefore perfectly uninterrupted employment could not be guaranteed to the workshops if their annual produce exceeded the amount which the Stores were willing to purchase in the most unfavourable years. And the prices and wages in the workshops would have to fall whenever they fell in the outside markets : for otherwise the trust of co-operators would be put to Jtoo great a strain. The plan will require much exercise of the Co-operative Faith ; but it will be a gain to the world if many wish for it, and a great gain if they achieve it BOOK III. CHAP. IX. § ii. INDEX Apprenticeships, 112^104.^' Art, old use of the word, 3 Auxiliary capital, 19 Babbage, on the economy of skill, 5° Bagehot, functions of the employer, 51, 116; on the modem facilities for borrowing capital, 117, 137 Brassey, on the uniformity of Taskwages, 174 ; province of Co-operative production, 223 Brentano, on the difference between the price of labour, and that of commodities, 172 Buyer, 68 • Capital, 13, 99 ; Personal, 20 Carey, on the Law of Diminishing Return, 24 ; Cost of reproduction, 79 Circulating capital, 19 Community, village, 44, 60 Consumer, 97 Co-operation, principle of, 218 Co-operative distribution, 224, 225 production, 220, 221 Cost of production, 73 ; its relation to value, 95, 147 Cottier, 63 Credit, influence of, 151 Crisis, 151—.155 Crompton, on Industrial Conciliation, Book in. ch. viii. Custom, influence of, 44, 170 Deduction, 3 Degradation of labour, 10a Depreciation, 75 Depression, of trade, 154— 157, 163 Discounted value of labour, 133 Distribution, 94, 95 Dose of capital, 22 Eamings-and-interest Fund, 96 Earnings of Management, 75, 96, 137» 221 Economics, 2, 5 Expenses of production, 73, 74, 97 Feudalism, 61 Final utility, 69 Fixed capital, 19 Gilds, 46, 187 Grades of labour, 107 Hill, Octavia, on volunteer work in poor-relief, 34, 35 Holyoake, on Co-operation, Book III. ch. ix. passim Howell, on unions, Bk. in. ch. v. passim Improvements, influence on rent, 85 Induction, 3 Jevons, on capital, 20, 995 market, 67 ; Final utility, 6\* wages, 205 INDEX. oint Expenses of production, 164 oint Stock company, 35, 137 Labourer, 95 Law, 13 Law of Demand, 71 Diminishing Return, 22 Division of labour, 57 Increasing Return, 57 Local Variations of value, 168 Normal Earnings of Management, 143 Normal rate of interest, 126 Normal value, 77, 80, 93, cf.also 146,149,106,167 Normal wages, 131 Population, 29 Rent, 83 Supply, 76 Leslie, Cline, on the limitation of competition, 117, 171, 178, 185; on the irregular influence of new supplies of gold, 172 Localization of industry, 47, 53 Malthus, 29 Margin of cultivation, 84 Market, 67 Measure, of value in use by value ' in exchange, 69 ; of Cost of production by Expenses of production, 73, 74, 97, 147 Mill, on the relations of Economics to other sciences and to practice, 5, 2115 on the influence of improvements on rent, 85 ; on cost of production, 97 5 on the four grades of labour, 107 ; regards wages as paid out of the product of industry, 204 Mines, 25, 86 Monopoly value, 180 Nasmyth, use of machinery elevates the working classes, 50 Necessaries, vagueness of the term, 29 Net advantages of a trade, 103,108 Net income'of a country, 95, 98 230 Net return of a machine, 121 Net return of labour, 133 Nominal wages, roi Normal, 65, 66 Normal theory necessary introduction to Market theory, 148 Organized, 45 Over-production, 154 Peasant Proprietors, 40 Personal capital, 58, 61 Risks, 135 Wealth, 6 Piece-work, 195 Price, 68, 150, 158 Productive, 6 Profits must ultimately diminish, 145 on equal capitals, 138 on unequal capitals, 139 Real wages, 101 Remuneratory capital, 19 Rent, 64 ; of natural qualities, 1 ro, 144 . in relation to wholesale prices, 88—90; to retail prices, 179 Reproduction, cost of, 79 Retail profits, 177, 224 Ricardo, on rent, 82 n., 85 n., 89 Sargant, on the causes that determine saving, 41 Schulze-Delitzsch, 223 Science, 3 Seller, 68 Shifts, double, 196 Smith, Adam, his use of the word natural, 66 Specialised capital, 20 Speculators, 158 Standard of Comfort, 28 Strikes, cost of, 193 ; conditions of success, 209— 211 Subsidiary industries, 52 Superintendence, see Earnings of Management Task-wages, 101 Time-wages, 101 L INDEX. 231 Thornton, on the labourers\* disadvantage in bargaining, 200 Trade Risks, 78, 135 Value in exchange, 67 in use, 67 Wages, 128 Real and Nominal, 101 Time and Task, 101 Wages-and-profits Fund, 95 Wages-Fund, 203 Walker, on degradation of labour, 102 ; on the functions of the employer, 118 ; land the best savings bank, 40 Wealth, material, 6 Personal, 6 Women’s wages, 175 C A M B R I D G E : P R IN T E D BY C . J. C L A Y , M .A . & SO N , A T T H E U N IV E R SITY PRESS.

A Great Transition Initiative Essay Economics for a Full World Because of the exponential economic growth since World War II, we now live in a full world, but we still behave as if it were empty, with ample space and resources for the indefinite future. The founding assumptions of neoclassical economics, developed in the empty world, no longer hold, as the aggregate burden of the human species is reaching—or, in some cases, exceeding—the limits of nature at the local, regional, and planetary levels. The prevailing obsession with economic growth puts us on the path to ecological collapse, sacrificing the very sustenance of our well-being and survival. To reverse this ominous trajectory, we must transition toward a steadystate economy focused on qualitative development, as opposed to quantitative growth, and the interdependence of the human economy and global ecosphere. Developing policies and institutions for a steady-state economy will require us to revisit the question of the purpose and ends of the economy. Herman Daly June 2015 This essay has been adapted from a speech delivered on the occasion of the Blue Planet Prize, Tokyo, November 2014. 1 | Economics for a Full World | A Great Transition Initiative Essay The Economy as Subsystem of the Ecosphere When I worked at the World Bank, I often heard the statement, “There is no conflict between economics and ecology. We can and must grow the economy and protect the environment at the same time.” I still hear it a lot today. Although it is a comforting idea, it is at most half true. The “true” part stems from a confusion of reallocation with aggregate growth. Possibilities of better allocation almost always exist—more of something desired in exchange for a reduction in something less desired. However, aggregate growth, what macro-economists mean by the term “growth” (and the meaning in this essay), is that the total market value of all final goods and services (GDP) is expanding. The economy, as shown in Figure 1, is an open subsystem of the larger ecosphere, which is finite, non-growing, and materially closed, although open to a continual, but non-growing, throughput of solar energy. When the economy grows in physical dimensions, it incorporates matter and energy from the rest of the ecosystem into itself. It must, by the law of conservation of matter and energy (First Law of Thermodynamics), encroach on the ecosystem, diverting matter from previous natural uses. More human economy (more people and commodities) means less natural ecosystem. In this sense, the statement that there is “no conflict” is false. There is an obvious physical conflict between the growth of the economy and the preservation of the environment. That the economy is a subsystem of the ecosphere seems perhaps too obvious to emphasize. Yet the opposite view is common in high places. For example, a recent study by the British government’s Natural Capital Committee asserted, “The environment is part of the economy and needs to be properly integrated into it so that growth opportunities will not be missed.” To the contrary, it is the economy that is the part and needs to be integrated into the whole of the finite ecosphere so that growth limits will not be missed.1 But is this physical conflict economically important? Some believe that we still live in an “empty” world. In the empty world, the economy was small relative to the containing ecosystem, our technologies of extraction and harvesting were not very powerful, and our numbers were small. Fish reproduced faster than we could catch them, trees grew faster than we could harvest them, and minerals in the Earth’s crust were abundant. In other words, natural resources were not really scarce. In the empty world, it made economic sense to say that there was no conflict between economic growth and the ecosystem, even if it were not strictly true in a physical sense. It is the economy that is the part and needs to be integrated into the whole of the finite ecosphere. 2 | Economics for a Full World | A Great Transition Initiative Essay Figure 1: Welfare in a Full vs. Empty World Neoclassical economic theory developed during this era and still embodies many assumptions from it. But the empty world has rapidly turned into a “full” world thanks to growth, the number one goal of all countries—capitalist, communist, or in-between. Since the mid-twentieth century, the world population has more than tripled—from two billion to over seven billion. The populations of cattle, chickens, pigs, and soybean plants and corn stalks have as well. The non-living populations of cars, buildings, refrigerators, and cell phones have grown even more rapidly. All these populations, both living and non-living, are what physicists call “dissipative structures”—that is, their maintenance and reproduction require a metabolic flow, a throughput that begins with depletion of low-entropy resources from the ecosphere and ends with the return of polluting, high-entropy waste back to the ecosphere. This disrupts the ecosphere at both ends, an unavoidable cost necessary for the production, maintenance, and reproduction of the stock of both people and wealth. Until recently, standard economic theory ignored the concept of metabolic throughput, and, even now, its importance is greatly downplayed.2 Neoclassical economic theory developed in the empty world and still embodies many assumptions from it. Economy Matter Matter Energy Energy Recycle Ecosystem Heat Solar Energy Welfare economic services ecosystem services Empty World Economy Matter Matter Energy Energy Recycle Heat Solar Energy Welfare economic services ecosystem services Full World Ecosystem 3 | Economics for a Full World | A Great Transition Initiative Essay The concept of metabolic throughput in economics brings with it the laws of thermodynamics, which are inconvenient to growthist ideology. The First Law, as noted above, imposes a quantitative trade-off of matter/energy between the environment and the economy. The Second Law, that the entropy (or disorder) of the universe is always increasing, imposes a qualitative degradation of the environment— by extracting low-entropy resources and returning high-entropy wastes. The Second Law of Thermodynamics thus imposes an additional conflict between expansion of the economy and preservation of the environment, namely that the order and structure of the economy is paid for by imposing disorder in the sustaining ecosphere. Furthermore, this disorder, exported from the economy, disrupts the complex ecological interdependencies of our life-supporting ecosystem. Those who deny the conflict between growth and environment often claim that since GDP is measured in value units, it has no necessary physical impact on the environment. But one must remember that a dollar’s worth of gasoline is a physical quantity—recently about one fourth of a gallon in the United States. GDP is an aggregate of all such “dollar’s worth” quantities bought for final use, and is consequently a value-weighted index of physical quantities. GDP is certainly not perfectly correlated with resource throughput. Nevertheless, prospects for absolute “decoupling” of resource throughput from GDP are quite limited, even though much discussed and wished for.3 These limits are made visible by considering an input-output matrix for an economy. Nearly every sector requires inputs from, and provides outputs to, nearly every other sector. And these inputs require a further round of inputs for their production, etc. The economy grows as an integrated whole, not as a loose mix of sectors. Even the information and service sectors require substantial physical resource inputs. In addition to the supply side limit reflected in the input-output interdependence of production sectors, there is the demand side limit of what has been called the “lexicographic ordering of wants”—unless we first have sufficient food on the plate, we are just not interested in the information contained in a million recipes on the Internet. And, of course, the Jevons Paradox—the idea that, as technology progresses, the increase in efficiency with which a resource is used tends to increase the rate of consumption of that resource—negates much of the benefits of such progress. This does not deny real possibilities of improved technical efficiency in the use of resources, or ethical improvement in the ordering of our priorities. But these represent qualitative development and are frequently not captured in GDP, which mainly reflects quantitative growth. Since GDP reflects both harmful and beneficial activity, ecological economists have not considered it to be a desideratum in itself. Instead, they have distinguished growth (quantitative increase in size by accretion or assimilation of matter) from development (qualitative improvement in design, technology, or ethical priorities). Ecological economists advocate development without growth—qualitative improvement without Ecological economists have distinguished growth from development. 4 | Economics for a Full World | A Great Transition Initiative Essay quantitative increase in resource throughput beyond an ecologically sustainable scale. Given this distinction, one could indeed say that there is no necessary conflict between qualitative development and the environment. GDP accounting mixes together both growth and development, as well as costs and benefits. It thus confuses more than it clarifies. From Empty World to Full World: The Limiting Factor Has Changed When the entropic throughput becomes too large, it overwhelms either the regenerative capacity of nature’s sources or the assimilative capacity of nature’s sinks. This tells us that we no longer live in the empty world, but instead inhabit a full world. Natural resource flows are now the scarce factor, and labor and capital stocks are now relatively abundant. This basic pattern of scarcity has been reversed by a century of growth. Figure 2: Change in Limiting Factors This simple picture is instructive. In the past, the fish catch was limited by the number of fishing boats and fishermen. Now, it is limited by the number of fish and their capacity to reproduce. More fishing boats will not result in more caught fish. The limiting factor is no longer the manmade capital of boats, but the remaining natural capital of fish populations and their aquatic habitat. Economic logic would tell us to invest in the limiting factor. The old economic policy of building more fishing boats is now uneconomic, so we need to invest in natural capital, the new limiting factor. How do we do that? For one, we can do so by reducing the catch to allow fish populations to increase to their previous levels, and by other measures such as fallowing agricultural land to refresh its fertility. More generally, we can do so through restoration ecology, biodiversity conservation, and sustainable use practices. One could draw similar pictures for other natural resources. What ultimately limits the production of cut timber? Is it the number of chainsaws, sawmills, and lumberjacks, or the remaining forests and the growth rate of new trees? What limits the crops from irrigated agriculture? Is it the number of pipes, sprinklers, and pumps, or the stock Natural resource flows are now the scarce factor, and labor and capital stocks are now relatively abundant. Empty Full 5 | Economics for a Full World | A Great Transition Initiative Essay of water in aquifers, their recharge rate, and the flow of surface water in rivers? What limits the number of barrels of pumped crude oil: the number of drilling rigs or the remaining accessible deposits of petroleum? What limits the use of all fossil fuels: our mining equipment and combustion engines, or the capacity of the atmosphere to absorb the resulting greenhouse gasses without causing drastic climate change? In all cases, it is the latter, the natural capital (whether source or sink), rather than the manmade capital. Traditional economists reacted to this change in the identity of the limiting factor in three ways. First, they ignored it—by continuing to believe that we live in the empty world of the past. Second, they pretended that GDP is an ethereal, angelic number rather than a physical aggregate. Third, they claimed that natural capital has not, in fact, replaced manmade capital as the limiting factor because manmade and natural capital are interchangeable substitutes, at least according to neoclassical production functions. Only if factors of production are complements can the one in short supply be limiting. So even if natural capital is now scarcer than before, this would not be a problem, neoclassical economists say, because manmade capital is a “near perfect” substitute for natural resources. It is represented as such in multiplicative production functions such as the widely used Cobb-Douglas. But multiplying “factors” of production to get a “product” is mathematics, not economics. In the real world, what we call “production” is in fact transformation, not multiplication. Natural resources are transformed by capital and labor inputs into useful products and waste. While improved technologies can certainly reduce waste and facilitate recycling, agents of transformation (capital and labor) cannot serve as direct substitutes for the material and energy being transformed (natural resources). Can we produce a ten-pound cake with only one pound of ingredients, simply by using more cooks and ovens? And further, how could we make more capital (or labor) without also using more natural resources? While a capital investment in sonar may help locate those remaining fish, it is hardly a good substitute for more fish in the sea. And what happens to the capital value of fishing boats, including their sonar, as the fish disappear? Limits to Growth and the Optimal Scale of the Economy in a Full World It is clear from Figure 1 that the transition from empty to full world involves both costs and benefits. The brown arrow from Economy to Welfare represents economic services (benefits from the economy). It is small in the empty world but large in the full world. It grows at a diminishing rate because, as rational beings, we satisfy our most important wants first—the law of diminishing marginal utility. The costs of growth are represented by the shrinking ecosystem services (green arrow) that are large in the empty world but small in the full world. It diminishes at an increasing rate Technologies cannot serve as direct substitutes for natural resources. 6 | Economics for a Full World | A Great Transition Initiative Essay as the ecosystem is displaced by the economy because we—in theory—sacrifice the least important ecosystem services first—the law of increasing marginal costs. We can restate this in terms of Figure 3, showing the declining marginal benefit of growth of the economy and the increasing marginal cost of the resulting environmental sacrifice: Figure 3: The Limits to Growth From the diagram, we can distinguish three concepts of limits to growth: 1. The futility limit occurs when the marginal utility of production falls to zero. Even with no cost of production, there is a limit to how much we can consume and still enjoy it. There is a limit to how many goods we can enjoy in a given time period, as well as a limit to our stomachs and the sensory capacity of our nervous systems. In a world with considerable poverty, and in which the poor observe the very rich still enjoying their extra wealth, many view this futility limit as far away, not only for the poor, but for everyone. By its “non-satiety” postulate, neoclassical economics formally denies the concept of the futility limit. However, studies have shown that, beyond a “sufficiency threshold,” both self-evaluated happiness and objective indices of welfare cease to increase with GDP.4 2. The ecological catastrophe limit is represented by a sharp increase to the vertical of the marginal cost curve. Some human activity, or novel combination of activities, may induce a chain reaction, or tipping point, and collapse our ecological niche. The leading candidate for the catastrophe limit at present is runaway climate change induced by greenhouse gases emitted in pursuit of economic growth. Where along the horizontal axis it might occur is uncertain. The assumption of a continuously and smoothly increasing marginal cost curve is quite optimistic. Given our limited understanding of how the ecosystem functions, we cannot be sure that we have correctly sequenced our sacrifices of ecological services from least to most important. In making way for growth, we may ignorantly sacrifice a vital ecosystem service ahead of a trivial one. Thus, the marginal cost curve might in reality zigzag up and Even with no cost of production, there is a limit to how much we can consume and still enjoy it. 7 | Economics for a Full World | A Great Transition Initiative Essay down discontinuously, making it difficult to define the third and most important limit, namely the economic limit. 3. The economic limit is defined by the equality of marginal cost and marginal benefit and the corresponding maximization of net benefit. The economic limit would appear to be the first limit encountered. It certainly occurs before the futility limit, and likely before the catastrophe limit. At worst, the catastrophe limit might coincide with and discontinuously determine the economic limit. Therefore, it is very important to estimate the risks of catastrophe and include them as costs counted in the disutility curve as far as possible. From the graph, it is evident that increasing aggregate production and consumption is rightly called economic growth only up to the economic limit. Beyond that point, it becomes uneconomic growth because it increases costs by more than benefits, making us poorer, not richer. Nonetheless, we perversely continue to call it economic growth. Indeed, you will not find the term “uneconomic growth” in any macroeconomics textbook. Any increase in real GDP is called “economic growth” even if it increases costs faster than benefits. That richer (more net wealth) is better than poorer is a truism. The relevant question, though, is, does growth still make us richer, or has it begun to make us poorer by increasing “illth” faster than wealth? Examples of “illth” are everywhere, even if they are still unmeasured in national accounts. They include things like nuclear wastes, climate change from excess carbon in the atmosphere, biodiversity loss, depleted mines, deforestation, eroded topsoil, dry wells and rivers, sea level rise, the dead zone in the Gulf of Mexico, gyres of plastic trash in the oceans, and the ozone hole. They also include exhausting and dangerous labor and the un-repayable debt from trying to push growth in the symbolic financial sector beyond what is possible in the real sector. Economists will note that the logic employed in Figure 3 is familiar in microeconomics—the optimal size of a microeconomic unit, be it a firm or a household, occurs where the marginal cost is equal to the marginal benefit. That logic is not applied to the macro-economy, however, because the latter is thought to be the Whole rather than a Part. When a Part expands into the finite Whole, it imposes an opportunity cost on other Parts that must shrink to make room for it. When the Whole itself expands, it is thought to impose no opportunity cost because it displaces nothing, presumably expanding into the void. But as seen in Figure 1, the macroeconomy is not the Whole. It, too, is a Part, a part of the larger natural economy, the ecosphere, and its growth does inflict opportunity costs on the finite Whole that must be counted. Their refusal to acknowledge this is why many economists cannot conceive of the possibility that growth in GDP could ever be uneconomic. Standard economists might accept Figure 3 as a static picture but then argue that, in a dynamic world, technology will shift the marginal benefit curve upward and the marginal cost curve downward, moving their intersection (economic limit) ever to Has growth begun to make us poorer by increasing “illth” faster than wealth? 8 | Economics for a Full World | A Great Transition Initiative Essay the right, so that continual growth remains both desirable and possible. However, the macroeconomic curve-shifters need to remember three things. First, the physically growing macro-economy is still limited by its displacement of the finite ecosphere and by the entropic nature of its maintenance throughput. Second, the timing of new technology is uncertain. The expected technology may not be invented or come online until after we have passed the economic limit. Do we then endure uneconomic growth while waiting and hoping for the curves to shift? Third, the curves can also shift in the wrong directions, moving the economic limit back to the left. Did the technological “advances” of tetraethyl lead and chlorofluorocarbons shift the cost curve down or up? How about nuclear power? Or “fracking”? Adopting a steady-state economy at the macro level (while, of course, allowing for improvements in allocation at the micro level) helps us to avoid being shoved past the economic limit. We could take our time to evaluate new technologies rather than blindly adopting them in the interest of aggregate growth that may well be uneconomic. And the steady state gives us some insurance against the risks of ecological catastrophe that increase with growthism and technological impatience. Three Perspectives on Integrating Economy and Ecosystem Our vision and policies should be based on an integrated view of the economy as a subsystem of the finite and non-growing ecosphere. Three different theoretical understandings have grounded such attempts at integration, and all three start from the vision of the economy as a subsystem of the ecosphere and thus recognize limits to growth. They differ, however, in the way they each treat the boundary between the economy and the rest of the ecosystem, and these differences have large policy consequences for how we adjust to limits. Figure 4: Approaches to Integrating Economy and Ecosystem Economic imperialism seeks to expand the boundary of the economic subsystem until it encompasses the entire ecosphere. The goal is one system, the macroeconomy as the Whole. This is accomplished by the complete internalization of all Adopting a steady-state economy at the macro level helps us to avoid being shoved past the economic limit. 9 | Economics for a Full World | A Great Transition Initiative Essay external costs and benefits into prices. Those myriad aspects of the biosphere not customarily traded in markets are treated as if they were by imputation of “shadow prices”—the economist’s best estimate of what the price of the function or thing would be if it were traded in a competitive market. Everything in the ecosphere is theoretically rendered comparable in terms of its priced ability to help or hinder individuals in satisfying their wants. Implicitly, the end pursued is an ever-greater level of consumption, and the way to effectively achieve this end is growth in the aggregate exchange value of marketed final goods and services (GDP). Economic imperialism is essentially the neoclassical approach. Subjective individual preferences, however whimsical or uninstructed, are taken as the ultimate source of value. This is a perverse value judgment, not the absence of value judgments, as economists normally treat it. Since subjective wants are thought to be infinite in the aggregate, as well as sovereign, the scale of activities devoted to satisfying them tends to expand. The expansion is considered legitimate as long as “all costs are internalized into prices.” While costs should certainly be internalized into prices, this should not become an excuse for allowing excessive takeover of the ecosphere by economic growth. Unfortunately, many of the costs of growth that we have experienced have come as surprises. We cannot internalize them if we cannot first imagine and foresee them. Furthermore, even after some external costs have become quite visible (e.g., climate change), internalization has been very slow, partial, and much resisted. Profitmaximizing firms have an incentive to externalize costs. As long as the evolutionary fitness of the environment to support life is not perceived by economists as a value, it is likely to be destroyed in the imperialistic quest to subject every molecule and photon in creation to the pecuniary rules of present value maximization. There is no doubt that once the scale of the economy has grown to the point that formerly free environmental goods and services become scarce, it is better that they should have a positive price reflecting their scarcity than to continue to be priced at zero. But the prior question remains: Are we better off at the new larger scale with formerly free goods correctly priced, or at the old smaller scale with free goods also correctly priced (at zero)? In both cases, the prices are right. This question of optimal macro scale is neither answered nor even asked by either neoclassical or Keynesian economics in their blind quest for growth. Ecological reductionism begins with the true insight that humans and markets are not exempt from the laws of nature. It then proceeds to the false inference that human action is totally explainable by and reducible to the laws of nature. It seeks to explain whatever happens within the economic subsystem by exactly the same natural laws that it applies to the rest of the ecosystem. It subsumes the economic subsystem indifferently into the natural system, erasing its boundary. Taken to the extreme, this view purports to explain everything by a materialist deterministic system that has no room for purpose or will. This is a sensible vision from which to study the Many of the costs of growth have come as surprises. 10 | Economics for a Full World | A Great Transition Initiative Essay ecology of a coral reef or a rainforest. But if one adopts it for studying the human economy, one is stuck with the inconvenient policy implication that policy can make no difference. Ecology has inherited from its parent discipline, biology, a measure of modern biology’s mechanistic philosophy. This stems from a neo-Darwinian fundamentalism that is often uncritically accepted by many leading biologists as a deterministic metaphysics validated by science, rather than as a fruitful working hypothesis for doing science. Determinism is totally at odds with purposeful policy of any kind, and consequently with any economic thought aiming at policy. A happy marriage between economics and ecology, as in “ecological economics,” must overcome this latent incompatibility. Economic imperialism reduces everything to human will and utility, neglecting objective constraints of the natural world. Ecological reductionism sees only deterministic natural laws, and imperiously extends these into materialist “explanations” of human will and consciousness as mere illusions. It is a tragic irony that the discipline whose scientific findings have done most to awaken us to the environmental dangers we face is also the discipline whose metaphysical presuppositions have done most to weaken our will to respond to these dangers through purposeful policy.5 Economic imperialism and ecological reductionism are both monistic visions, albeit rather opposite monisms. The monistic quest for a single entity or principle by which to explain everything leads to excessive reductionism on both sides. Certainly, science should strive for the most reduced or parsimonious explanation possible without ignoring the facts. But respect for the basic empirical facts of natural laws on the one hand, and self-conscious purpose and will on the other hand, should lead us to a kind of practical dualism. After all, that our world should consist of two fundamental features offers no greater inherent improbability than that it should rest on only one. How these two fundamental features of our world (material cause and final cause) interact is a venerable mystery—precisely the mystery that the monists of both kinds are seeking to avoid. But economists are too much in the middle of things to adopt either extreme. They are better off denying the tidy-mindedness of either monism than denying the facts that point to an untidy dualism. The remaining perspective is the steady-state subsystem. It does not attempt to eliminate the subsystem boundary, either by expanding it to coincide with the whole system or by reducing it to nothing. Rather, it affirms both the interdependence and the qualitative difference between the human economy and the natural ecosystem. The boundary must be recognized and drawn in the right place. The scale of the human subsystem defined by the boundary has an optimum, and the throughput by which the ecosphere physically maintains and replenishes the economic subsystem must be ecologically sustainable. The goal of the economy is to minimize the lowentropy used up to attain a sufficient standard of living—by sifting it slowly and carefully through efficient technologies aimed at important purposes. The economy The steady-state perspective affirms both the difference and the interdependence between the human economy and natural ecosystem. 11 | Economics for a Full World | A Great Transition Initiative Essay should not be viewed as an idiot machine dedicated to maximizing waste. Its ultimate purpose is the maintenance and enjoyment of life for a long time (not forever) at a sufficient level of wealth for a good (not luxurious) life. The idea of a steady-state economy comes from classical economics, and was most developed by John Stuart Mill (1857), who referred to it as the “stationary state.”6 In such a state, the population and the capital stock would no longer grow, although the art of living would continue to improve. The constancy of these two physical stocks defined the scale of the economic subsystem. Birth rates would be equal to death rates and production rates equal to depreciation rates. Today, we add that both rates should be equal at low levels rather than high levels because we value longevity of people and durability of artifacts, and wish to minimize throughput, subject to maintenance of sufficient stocks for a good life. Policies for a Steady-State Economy Ecological economics should seek to develop the steady-state vision and get beyond the dead ends of both economic imperialism and ecological reductionism. Ten policies for moving toward a steady-state economy appear below. Many could be adopted independently and gradually, although they cohere in the sense that some compensate for the shortcomings of others. Of course, the question of the desired level of steady-state economy is crucial, and local, regional, and global ecological limits must be considered in fashioning effective policies. (1) Developing Cap-Auction-Trade systems for basic resources (especially fossil fuels): Set caps for natural resource according to three key rules: (1) renewable resources should not be depleted faster than they regenerate, (2) nonrenewable resources should not be depleted faster than renewable substitutes are developed, and (3) wastes from all resource use should not be returned to the ecosystem faster than they can be absorbed and reconstituted by natural systems. This approach achieves sustainable scale and market efficiency, avoids rebound effects, and raises auction revenue for replacing regressive taxes. (2) Tax shifting: Shift the tax base from “value added” (labor and capital) to that to which value is added, i.e., natural resource throughput, the source of social costs such as pollution and adverse public health effects. Such taxes will also encourage efficient resource use. (3) Limiting inequality: Establish minimum and maximum income limits, maintaining differences large enough to preserve incentives but small enough to suppress the plutocratic tendencies of market economies. (4) Reforming the banking sector: Move from a fractional reserve banking system to 100% reserve requirements on demand deposits. Money would no longer be mainly interest-bearing debt created by private banks, but non-interest-bearing government debt issued by the Treasury. Every dollar loaned for investment would What policies can move us toward a steady-state economy? 12 | Economics for a Full World | A Great Transition Initiative Essay be a dollar previously saved by someone else, restoring the classical balance between investment and abstinence from consumption, and dampening boom and bust cycles. (5) Managing trade for the public good: Move from free trade and free capital mobility to balanced and regulated international trade. While the interdependence of national economies is inevitable, their integration into one global economy is not. Free trade undercuts domestic cost-internalization policies, leading to a race to the bottom. Free capital mobility invalidates the basic comparative advantage argument for free trade in goods.7 (6) Expanding leisure time: Reduce conventional work time in favor of part-time work, personal work, and leisure, thereby embracing well-being as a core metric of prosperity while reducing the drive for limitless production. (7) Stabilizing population: Work toward a balance in which births plus in-migrants equals deaths plus out-migrants, and in which every birth is a wanted birth. (8) Reforming national accounts: Separate GDP into a cost account and a benefits account so that throughput growth can be stopped when marginal costs equal marginal benefits. (9) Restoring full employment: Restore the US Full Employment Act of 1945 and its equivalent in other nations in order to make full employment once again the end, and economic growth the temporary means. Un/under-employment is the price we pay for growth from automation, off-shoring, deregulated trade, and a cheap-labor immigration policy. Under steady-state conditions, productivity improvements would lead to expanded leisure time rather than unemployment. (10) Advancing just global governance: Seek world community as a federation of national communities, not the dissolution of nations into a single “world without borders.” Globalization by free trade, free capital mobility, and free migration dissolves national community, leaving nothing to federate. Such globalization is individualism writ large—a post-national corporate feudalism in a global commons. Instead, strengthen the original Bretton Woods vision of interdependent national economies, and resist the WTO vision of a single integrated global economy. Respect the principle of subsidiarity: although climate change and arms control require global institutions, basic law enforcement and infrastructure maintenance remain local issues. Focus our limited capacity for global cooperation on those needs and functions that truly require it. Larger Ethical and Ecological Context of Economics It is one thing to suggest a general outline of policies, but it is something else entirely to say how we will secure the will, strength, and clarity of purpose to carry out these policies—especially when we have treated growth as the summum bonum for the We should focus our limited capacity for global cooperation on those needs that truly require it. 13 | Economics for a Full World | A Great Transition Initiative Essay past century. Such will requires a major change in philosophical vision and ethical practice, a shift that is hardly guaranteed even in light of the increasingly perilous circumstances in which the planet finds itself. As a way to contemplate such a change, consider the “ends-means pyramid” in Figure 5. The policies suggested above belong in the middle, under “Political Economy.” At the base of the pyramid are our ultimate means (low-entropy matter-energy)—that which we require to satisfy our wants, but which we cannot make, only use up. We use these ultimate means directly, guided by technology, to produce intermediate means (e.g., artifacts, commodities, services) that directly satisfy our needs. These intermediate means are allocated by political economy to serve our intermediate ends (e.g., health, comfort, education), ethically ranked by how strongly they contribute to the Ultimate End under existing circumstances. We can perceive the Ultimate End only vaguely, but in order to ethically rank our intermediate ends, we must compare them to some ultimate criterion. We cannot avoid philosophical and theological inquiry into the Ultimate End just because it is difficult. To prioritize requires that something go in first place. Figure 5: An Ends-Means Pyramid of Human Activity The middle position of economics is significant. Economics traditionally deals with the allocation of given intermediate means to satisfy a given hierarchy of intermediate ends. It takes the technological problem of converting ultimate means into intermediate means and the ethical problem of ranking intermediate ends with reference to an Ultimate End as solved. All economics has to do, then, is efficiently allocate given means among a given hierarchy of ends. In neglecting the Ultimate End and ethics, economics has been too materialistic; in neglecting ultimate physical means and technology, it has not been materialistic enough. Ultimate political economy (stewardship) is the total problem of using ultimate means to best serve the Ultimate End, no longer taking technology and ethics as given, We can perceive the Ultimate End only vaguely. 14 | Economics for a Full World | A Great Transition Initiative Essay but as steps in the total problem to be solved. The overall problem is too large to be tackled without breaking it down into its pieces. But without a vision of the total problem, the pieces do not fit together. The dark base of the pyramid represents the relatively solid and consensual knowledge of various sources of low-entropy matter-energy. The light apex of the pyramid represents the fact that our knowledge of the Ultimate End is uncertain and not nearly as consensual as physics. The single apex will annoy pluralists who think that there are many “ultimate ends.” Grammatically and logically, however, “ultimate” requires the singular. Yet there is certainly room for more than one perception of the nature of the singular Ultimate End, and much need for tolerance and patience in reasoning together about it. The Ultimate End, whatever it may be, cannot be growth. A better starting point for reasoning together is John Ruskin’s aphorism that “there is no wealth but life.” How might that insight be restated as an economic policy goal? I would suggest the following: maximizing the cumulative number of lives ever to be lived over time at a level of per capita wealth sufficient for a good life. This leaves open the traditional ethical question of what is a good life, while conditioning its answer to the realities of ecology and the economics of sufficiency. At a minimum, it seems a more reasonable approximation than the current impossible goal of “ever more things for ever more people forever.” Endnotes 1. Dieter Helm, The State of Natural Capital: Restoring our Natural Assets (London: UK Natural Capital Committee, 2014). 2. This is despite notable contributions from Nicholas Georgescu-Roegen and Kenneth Boulding. See Nicholas Georgescu-Roegen, The Entropy Law and the Economic Process (Cambridge, MA: Harvard University Press, 1971); Kenneth Boulding, “The Economics of the Coming Spaceship Earth,” in Environmental Quality in a Growing Economy, ed. H. Jarrett (Baltimore: Johns Hopkins University Press, 1966), 3-14. 3. Tim Jackson, Prosperity without Growth: Economics for a Finite Planet (London: Earthscan, 2009), 67–71. 4. As indicated by the GPI (Genuine Progress Indicator) and its forerunner the ISEW (Index of Sustainable Economic Welfare). For an informative survey, see Ida Kubiszewski, Robert Costanza, Carol Franco, Philip Lawn, John Talberth, Tim Jackson, and Camille Aylmer, “Beyond GDP: Measuring and Achieving Global Genuine Progress,” Ecological Economics 93 (September 2013): 57-68. 5. This contradiction is most apparent in the work of acclaimed naturalist and environmentalist Edward O. Wilson, who strongly affirms both materialistic determinism and environmental activism. He recognizes the contradiction and, unable to resolve it, has simply chosen to live with it. See Wendell Berry, Life is a Miracle (An Essay Against Modern Superstition) (Washington, DC: Counterpoint Press, 2000), 26. See also Chapter 23 in Herman Daly, Ecological Economics and Sustainable Development (Cheltenham, UK: Edward Elgar, 2007). 6. John Stuart Mill, Principles of Political Economy IV.VII.I (London, 1848). 7. Capitalists are interested in maximizing absolute profits and therefore seek to minimize absolute costs. If capital is mobile between nations, it will move to the nation with lowest absolute costs. Only if capital is internationally immobile will capitalists bother to compare internal cost ratios of countries and choose to specialize in the domestic products having the lowest relative cost compared to other nations, and to trade that good (in which they have a comparative advantage) for other goods. In other words, comparative advantage is a second-best policy that capitalists will follow only when the first-best policy of following absolute advantage is blocked by international capital immobility. For more on this, see Chapter 18 in Herman Daly and Joshua Farley, Ecological Economics (Washington, DC: Island Press, 2004). A better starting point for reasoning together is that “there is no wealth but life.” 15 | Economics for a Full World | A Great Transition Initiative Essay About the Author Herman Daly is an ecological economist and Emeritus Professor at the University of Maryland, School of Public Policy. From 1988 to 1994, he was a Senior Economist in the Environment Department of the World Bank. Prior to that, he was a professor of economics at Louisiana State University, where he taught for twenty years. He was co-founder and associate editor of the journal Ecological Economics. He has written extensively on theorizing the steady-state economy and co-developed the Index of Sustainable Welfare. He holds a Ph.D. from Vanderbilt University. About the Publication Published as a Essay by the Great Transition Initiative. This work is licensed under a Creative Commons Attribution-NonCommercialShareAlike 4.0 International License. Cite as Herman Daly, “Economics for a Full World,” Great Transition Initiative (June 2015), https:// greattransition.org/publication/economics-for-a-full-world. About the Great Transition Initiative The Great Transition Initiative is an international collaboration for charting pathways to a planetary civilization rooted in solidarity, sustainability, and human well-being. As a forum for collectively understanding and shaping the global future, GTI welcomes diverse ideas. Thus, the opinions expressed in our publications do not necessarily reflect the views of GTI or the Tellus Institute.

ID28 M414 :HJ WORKING PAPER ALFRED P. SLOAN SCHOOL OF MANAGEMENT Industrial Economics: An Overview Richard L. Schmalensee Sloan School of Management Massachusetts Institute of Technology WP No. 1997-88 March 7, 1988 MASSACHUSETTS INSTITUTE OF TECHNOLOGY 50 MEMORIAL DRIVE CAMBRIDGE, MASSACHUSETTS 02139 Industrial Economics: An Overview Richard L. Schmalensee Sloan School of Management Massachusetts Institute of Technology WP No. 1997-88 March 7, 1988 Two decades ago George Stigler (1968, p. 1) described the boundaries of industrial economics or, as the field is frequently labeled, industrial organization: . . . there is no such subject as industrial organization. The courses taught under this heading have for their purpose the understanding of the structure and behavior of the industries . . . of an economy. These courses deal with the size structure of firms (one or many, "concentrated" or not), the causes ... of this size structure, the effects of concentration on competition, the effects of competition upon prices, investment, innovation, and so on. But this is precisely the content of economic theory -- price or resource allocation theory . . . Stigler went on to note that industrial economics deals not only with theory but also with measurement and hypothesis testing and with the analysis of public policies toward business. A fairly accurate capsule description is that industrial economics is the study of the supply side of the economy, particularly those markets in which business firms are sellers. Industrial economics emerged as a distinct field after the rise of the large modern manufacturing corporation around the turn of the century (Chandler (1977), Hay and Morris (1979, ch. 1)). For many years it was generally viewed as an intellectually isolated empirical field without much scope for formal theory or non-routine econometrics. But in the last two decades much of the significant work in industrial economics has been theoretical, and much of it has been produced and consumed by nonspecialists. The game- theoretic tools now generally used in this research (and regularly sharpened thereby) are described in Section I. The late 1980 's seems to have witnessed a shift of interest back to empirical - 2 studies, and Section I also provides a brief overview of the methods that have been developed and employed in this work. Industrial economics is now best defined by three main topical foci, two of which are discussed systematically in what follows. Section II considers determinants of the behavior, scale, scope, and organization of business firms Research in this broad area has spilled over into labor economics and corporate finance and has informed the study of the relation between corporate strategies and organizational structures (Caves (1980)). The second focus is imperfect competition . When the structural prerequisites of perfect competition are not satisfied, how do market conduct and performance depend on relatively stable observable variables - - that is, on market structure, broadly defined? When will rivalry be intense, so that rents are dissipated, and when will it be restrained, so that performance is nearlv monopolistic? Work on these core questions is discussed in Sections III and IV. Section III considers choices of price, output, and capacity, while Section IV deals with non-price rivalry: product selection, advertising, and technical change. Models of imperfect competition developed in industrial economics have been imported into the scientific and policy sides of international economics (Krugman (1986)) and, recently, macroeconomics, and have been employed prescriptively to determine business strategies (Porter (1980)). The third focus is public policy toward business . One normative question and two positive questions arise here. The normative question has been central to the field since it began: What policies are optimal? Historically industrial economists have concentrated on antitrust (or competition) policy, regulation, and government enterprise. In recent years they have paid increasing attention deregulation, liberalization of entry restrictions, privatization (Vickers and Yarrow (1988)), and industrial policies aimed at affecting technical progress and international competitiveness (Krugman (1986), Yarrow (1985)). Length restrictions preclude systematic discussions of all these policy domains. Instead, I discuss policy implications of research findings at several points in Sections II-IV and offer a few general observations in Section V. Length restrictions also preclude a systematic review of attempts to answer the two related positive questions. The first is the natural complement to normative studies: What are the effects of actual policies? Specific answers of course vary considerably, but it is clear that governments often intervene in markets in ways that do not enhance efficiency. Stimulated in large part by Stigler's (1971) discussion of the discrepancy between economic theory and political practice, industrial economists have addressed a second positive question: What determines actual policies? These scholars have encountered a substantial number of political scientists also using rational actor models to study policy formation. The topical and methodological breadth of industrial economics, the pace at which it has developed in recent years, and limits on the length of this essay confine me to a broad overview of research on some central 2 topics. Comprehensive expository surveys are provided in Schmalensee and Willig (1989) and Tirole (1988); I have drawn heavily on those works and recommend both to the reader who wishes to learn what lies behind the many facades displayed in this essay. I. TOOLS AND METHODS This section provides a brief overview of research methods in industrial economics that is designed to complement the substantive discussions that follow. The tools employed in theoretical research, which are considered first, have become more uniform in the last decade, while the methods used in empirical work have become more diverse. Theory of Strategic Behavior Except where monopoly is assumed and the possibility of entry is assumed away, theoretical research in industrial economics today employs the tools 3 of noncooperative game theory. Modeling typically begins with the specification of the extensive form of a game: a description of which players move when, the actions and information available at each move, the probabilities of any random events to be chosen by "nature," and the functions determining each player's payoff. Some information may be private; each firm may know only its own cost function, for instance. Other information may be common knowledge; all firms may know the market demand function, for instance, and also know that all other firms have this same information as well. It is then assumed that observed behavior will correspond to a Nash equilibrium of the specified game, a situation in which each firm's strategy (a list of the moves it will make in all possible situations) is a best response to the strategies of its rivals. Nash equilibria can involve pure strategies (conditional choices of actions) or mixed strategies (conditional choices of probability distributions over actions). Equilibria involving only pure strategies seem generally to have more intuitive appeal. 5 - In game- theoretic terms, the familiar Cournot model is a one-period game in which N firms (N > 2) simultaneously choose output levels of identical products. If Q is total output, and P(Q) is the market inverse demand function, which is assumed to be common knowledge, ellers' payoffs are given by their profit functions: n i = q^Cq^q^) - C^q^, i-1 N, (1) where q- is firm i's output and q • = Q - q- is the total output of its rivals. The first-order necessary conditions for each q- to be a best response to the corresponding q - are as follows: 3n i /aq i = P(Q) + qi P'(Q) - MCi (qi ) =0, i = 1,..., N, (2) where MC^ is firm i's marginal cost, dC i /dq i , and P'(Q) = dP(Q)/dQ. A Nash equilibrium in pure must be a solution to equations (2) . Perhaps the greatest merit of the game-theoretic approach is that it disciplines theoretical discussion by, in effect, forcing theorists to specify and then abide by the rules of the games they analyze. There is no place in the analysis of the basic Cournot game for discussions of conjectured rival response, for instance; the (unbounded) rationality of all players is common knowledge in this game and most others. Similarly, dynamic adjustment paths can only arise in more complex games that unfold over time. And imperfections or asymmetries in information give rise to different games and thus cannot be handled informally. Since market behavior often involves making decisions over time or without full information, much recent work has involved games with these features. - 6 Dynamics While time is most naturally treated as continuous in many situations, continuous- time dynamic games (usually called differential games) are technically more challenging than discrete- time games, and the latter dominate the literature. These can be solved by working backwards from the last period if there is a last period. Infinite-horizon games are often more appealing in principle and, where stationarity can be exploited, simpler to analyze. But they typically have equilibria that do not appear even in the limit of the corresponding finite-horizon games. The analysis of multi-period games in industrial economics relies heavily on the principle of (subgame) perfection . Roughly, in a (subgame) perfect Nash equilibrium each player's strategy is a best response to those of its rivals, subject to the constraint that no player's strategy can involve threats that it would not be the player's interest to carry out if his bluff were called. For example, a strategy involving reducing price to zero if any entry occurs contains a threat that is not generally credible (because it generally does not maximize post-entry profits) and is thus ruled out by the perfectness restriction. In a variety of economic settings the ability to make credible threats can induce other actors to behave "nicely" to avoid the threatened behavior. Much attention has accordingly been devoted to devices that firms can use to obtain credibility. These generally involve taking irreversible actions, which would often be undesirable except for their impact on the incentives and behavior of others, and go under the general heading of commitment. If an established monopolist could build a Doomsday Machine (as in the film Dr. Strangelove) that would somehow force it to drive market price to zero if entry occurred, and if it could make the existence of that device common knowledge, it could credibly deter entry. The ability to commit often (but not always) gives an advantage to the first player to move in economic 6 games . Much interest in recent years has attached to repeated games or supergames -- in which a relatively simple constituent or stage game (such as the one-period Cournot or Prisoners' Dilemma games) is played repeatedly by a fixed set of players. Strategies of simply playing Nash equilibrium strategies of the constituent game in each period form a Nash equilibrium of the repeated game. But strategies in the repeated game may involve taking actions conditional on past history, and there are usually many other equilibria when the horizon is infinite. In fact, many of the main results in the supergame literature are variants on the so-called Folk Theorem, which says that virtually any set of payoffs can arise in a perfect equilibrium if the horizon is long enough and the discount rate is low enough (Fudenberg and Maskin (1986)). Information Players are said to have incomplete information if they do not know their opponents' payoff functions and to have imperfect information if they do not observe the actions of all players. Most interest has attached to games in which information is incomplete and asymmetric. Each firm in a Cournot setup might know its own costs, for instance, with the probability distribution from which cost function parameters are drawn assumed to be common knowledge. In multi-period games with incomplete or imperfect information, it is natural to require players to optimize at each move using subjective probabilities that they update according to Bayes rule. This requirement yields Bayesian Nash equilibria. In such equilibria, actions in any period - 8 may affect other players' actions in future periods by altering their beliefs. These models thus often exhibit generalized signalling (Milgrom and Roberts (1987)): costly actions are taken for the purpose of altering other actors' beliefs. Since all parties are rational, such signalling cannot bias any player's beliefs on average but may nonetheless occur because all players interpret their observations in light of others' incentives to induce bias. Multiple equilibria are the norm in multi-period games of incomplete information, and Folk Theorems indicate that a small amount of incomplete information can produce almost any equilibrium payoffs when the discount rate is low and the horizon is long (Fudenberg and Maskin (1986)). A particularly interesting class of games of asymmetric information is explored in agency theory (Hart and Holmstrom (1987)). In the basic agency problem, one party (the principal) hires another (the agent) to act on his behalf in the first period. The principal can generally observe the consequences of the agent's action second-period , but his information about that action (e.g., the level of managerial effort) or about the relevant environment (e.g., the level of demand) is inferior to that of the agent. The principal's task is to design a compensation scheme based on observables in order to maximize his own utility or wealth, subject to the constraints that he be able to hire an agent and that the agent will then p act in his own self-interest, as defined by the compensation scheme. Approaches to Empirical Research The early years of industrial economics were marked by the production of a number of book- length industry studies, often relying heavily on information - 9 - made public during antitrust cases. These comprehensive works remain a rich source of examples of business behavior, such as the evolution of price leadership in the interwar U.S. cigarette industry (Nicholls (1951)), that seem to involve the exercise of monopoly power. This literature also provides a sense of business reality and a detailed understanding of particular markets not frequently encountered in more formal studies. But relatively few industry studies of this sort have been done in recent years . Inter- industry Studies In the early 1950' s, Joe Bain (1951, 1956) shifted the focus of empirical research in industrial economics away from industry studies by showing the apparent power of statistical analysis of industry- level cross -section data. This approach seemed to promise more rapid and objective development of general relations than the case study approach. Most of the cross-section studies that filled the journals during the 1960's and 1970' s used government- supplied data and ran regressions designed to "explain" differences in industry -average profitability. Critics of this research strategy have noted serious limitations of available data. Government statistics often ignore foreign competition and regional markets and define industries that do not correspond to economic product markets. Accounting profitability is at best a noisy measure of economic profitability; problems include the accounting treatments of 9 inflation, advertising, and depreciation (Fisher and McGowan (1983)). There is no fully satisfactory way to handle diversified firms that operate in multiple markets. In addition, it is difficult to construct defensible proxies for a number of variables, including expectations and fundamental attributes of products and technologies, that are important in theory. If unobservable variables are correlated with the independent variables used in 10 - cross-section regressions, as often appears likely, coefficient estimates will be biased. A second set of criticisms turns on the difficulty of using crosssection data to identify key structural parameters. Economists usually argue that cross-section studies can reveal differences among long-run equilibria as long as deviations from equilibrium are uncorrelated with independent variables. But in the long run almost all observable industrylevel variables are affected by firms' decisions and are thus logically endogenous. While lack of identification is not an absolute bar to inference (Breusch (1986)), its definitive symptom, the existence of more than one plausible structural interpretation of estimated parameters, is frequently encountered in the cross-section literature. Despite these problems, inter- industry studies have an important role to play. It is difficult to design broad public policies, such as antitrust and tariff policies, without a feel for the main economy-wide relations (structural or otherwise) among affected markets. A number of recent interindustry studies rely on specially constructed data sets to deal with some of the problems noted above. Comparisons between matched industries in different countries, for instance, hold constant a host of unobservable industry-specific quantities (Pryor (1972), Baldwin and Gorecki (1985)), while industry- level panel data can reveal the effects of macroeconomic conditions and permit analysis of stability and change (Domowitz, Hubbard, and Petersen (1986)). Interview and survey methods can provide information not otherwise available (Scherer, et al (1975), Levin, et al (1987)). A great deal of interest has recently attached to the use of firm-level panel data (Mueller (1986)). While there are significant differences in 11 industry-average profitability, there are often even greater differences within industries (Schmalensee (1985)), and variations in the performance of leading firms over time is often not well explained by changes in the industries in which they participate (Cubbin and Geroski (1987)). Econometric Industry Studies Many industrial economists have reacted to the limitations of the inter- industry approach by studying particular industries. Industry-specific studies cannot describe economy-wide patterns, but, like the earlier case-study literature, such research can provide reliable data points that can inform both theorizing and interindustry research. A number of studies involve comparisons of geographically isolated markets for a single product and thus hold constant unobservable industry-specific variables (Benham (1972)). Such variables are also held constant in before-and-after analyzes of the effects of exogenous industry- specif ic structural changes (Rose (1987)). In the last decade or so, changes in stock prices over short periods have been employed with increasing frequency in industrial economics (Schwert (1981)). On the widely-accepted assumption that the stock market makes full use of publicly available information, stock price changes over some period, corrected for movements in the market as a whole, give the expected present value of the change in profit associated with firm-specific or industry-specific events of that period. In recent years many authors have used firm- level panel data to estimate industry-specific structural models designed to reveal directly the intensity -- and sometimes the pattern --of rivalry. Research of this sort involves particularly heavy investment in data set construction and in developing modeling strategies tailored to available data. Accordingly, a - 12 - host of techniques for econometric industry analysis have been developed, but most have been employed only once or twice. Much of this literature has been concerned with estimation of variants of the following quasi-supply relations: P = MCj^q^Xj) + (l+A i )qiP'(Q,Z) i-1 N. (3) If the A^ are all zero, equations (3) are just the Cournot first-order conditions (2) with the addition of exogenous variables X- and Z that shift firm i's cost and market demand, respectively. The A- are conjectural variations that are best interpreted as reduced form parameters that summarize the intensity of rivalry that emerges from what may be complex patterns of behavior. If estimated A,s are all equal to minus one, sellers behave as if perfect competition prevailed; higher values of A correspond to larger (P-MC) gaps and thus to less intense rivalry (Iwata (1974)). Marginal cost is usually not treated as directly observable in this work; the X- usually include input prices. The quasi-supply relations are often estimated with the industry inverse demand relation, P - P(Q,Z), and sometimes also with total cost or input demand functions. Identification of the A's may rest on the availability of exogenous variables in Z that change the slope of the demand curve (Lau (1982)), or on information about marginal cost or its determinants (Iwata (1974), Panzar and Rosse (1987)). Some studies in this literature have test alternative models of conduct, such as competition and collusion (Bresnahan (1987)); others examine differences in conjectural variations over time or among firms (Geroski, Ulph, and Ulph (1986)). Still others develop alternative 13 approaches to the detection of non- competitive behavior in particular settings (Baker and Bresnahan (1985), Panzar and Rosse (1987)). Laboratory Experiments Given the difficulty of obtaining detailed data on an informative set of natural market experiments, a good deal of interest has recently attached to the use of laboratory experiments to test industrial economic hypotheses (Plott (1982), Smith (1982)). Many variables that are unobservable outside the laboratory (such as beliefs and marginal costs) can be fixed in experimental settings, and the sensitivity of behavior to environmental and institutional changes can be explored directly. While the experiments reported in the literature to date have frequently been criticized as artificially simple, they generally do involve actors with financial incentives to optimize and markets of at least the same order of complexity as those studied in theoretical analyses. Developments in computer software and experimental procedures will likely make more "realistic" experiments possible. Still, laboratory research seems best suited for testing the predictive power and robustness of particular theories; it is less useful for determining the class of real markets for which particular theories are useful. II. BUSINESS STRUCTURE AND BEHAVIOR There are three main points of tension between the textbook model of firm behavior and organization and reality. First, it is not obvious that the managers of real firms maximize profits. Second, few long-run average cost curves seem to be either U-shaped or everywhere declining, so that the textbook models of competition and natural monopoly do not explain how the 14 - scales of many real firms are determined. Third, the textbook model deals with single-product firms and ignores their internal structures, even though real firms produce multiple products and must decide the scopes of their activities and their internal organization. This section considers research bearing on these three points. Managerial Behavior There have historically been two main criticisms of the traditional assumption that firms maximize profits, properly generalized where appropriate to mean maximization of shareholders' wealth. While both have force, neither has yet produced a superior alternative assumption. Behavioral Theories The first criticism begins by noting that many decisions that managers make regularly are much harder than the simplified problems with which theorists struggle. Thus limits to human information processing capabilities -- bounded rationality -- make strict profit maximization fundamentally implausible. And nobody who observes real firms 12 closely can avoid noticing managerial blunders. But, while there is evidence that managers often follow simple rules of thumb in lieu of consciously maximizing, it has proven difficult to characterize such rules at the level of generality required for a tractable model implying systematic departures from profit-maximization. Additional difficulties arise because competition acts to weed out rules of thumb that do not at least approximately maximize profit (Nelson and Winter (1982)). This evolutionary process plainly does not work instantly or perfectly in real markets, but it has proven difficult to specify its imperfections in a useful, general way. 15 - Agency Relations The second major critique of profit-maximization, which has attracted considerably more attention in recent years, begins with the observation that most large corporations are not managed by their owners. Thus managers are likely to have objectives other than maximizing owners' wealth. Moreover, many boards of directors are dominated by managers, not owners. And, while owners can and do replace directors and managers whose performance is unsatisfactory, the mechanisms available for this purpose (takeovers and proxy fights) are hardly f rictionless . Thus managers are likely to have some freedom to pursue their own objectives at owners' expense. This critique led initially to the development of models in which managers maximized some specific personal objective, such as revenue, employees, growth, or managerial perquisites, subject to a profitability constraint imposed by product and capital markets. It is not clear what is the "correct" managerial objective, and the determinants of the crucial profitability constraint are typically left unspecified. Related empirical work has produced mixed results (Smirlock and Marshall (1983)). More recently, the tools of agency theory have been employed to model the implications of asymmetric information when ownership and control are separated (Jensen and Meckling (1976)). The firm is viewed through the lens of agency theory as a set of contracts (some provisions of which may be fixed by law or custom) among input suppliers. These contracts are generally incomplete; they do not fully specify the consequences for all parties of all possible actions in all possible states of nature. Incompleteness may occur because of asymmetric information (e.g., about managerial effort) or because some observable variables (e.g., the riskiness 16 - of investments) are too complex to be objectively verified by third parties, so that contract provisions involving them would be unenforceable. Optimal contract terms are then derived under relatively specific assumptions about information and strategy sets. These contracts are then often compared informally to actual laws, customs, and institutions. Departures from profit maximization are usually treated as managerial slack or failure to provide effort, not as pursuit of alternative objectives. This research has turned up a number of theoretical phenomena discussed at more length in the next section: actions may be rationally undertaken mainly to affect others' perceptions even when the others are aware of this possibility and are not fooled, for instance. In many situations, optimal contracts cannot induce the behavior that would occur under full or symmetric information, and systematic departures from profit-maximization are predicted in a variety of settings. But few of these predictions have been tested empirically, and no tractable, general alternative to the profit-maximization assumption has yet emerged from this research. Agency- theoretic work on the firm spills over into finance when it considers the operation of capital markets; it spills over into labor economics when it considers employment arrangements and superior/subordinate relations. The tools of agency theory have also been used to study the design of institutions or mechanisms for regulating natural monopolies or supervising public enterprises under asymmetric information. Many models have been analyzed, and prescriptions seem sensitive to details of the assumed information structures. Beyond the result that "cost plus" regulatory (or other) contracts are rarely optimal, little in the way of operational policy guidance has yet emerged from this work (Joskow and 17 Schmalensee (1986)). Similarly, no terribly strong arguments for privatization of government -owned natural monopolies have been developed (Vickers and Yarrow (1988)). Conglomerate Mergers An interesting set of issues broadly related to the separation of ownership and control has been raised by the wave of conglomerate mergers and acquisitions -- combinations of firms that are not participants in the same product markets -- in the U.S. in the 1980's. Shareholders of acquired firms seem generally to benefit from the announcement of these events, and acquiring firms' shareholders do not suffer visible losses (Jensen and Ruback (1983)). This favorable ex ante verdict from the stock market has been interpreted as implying that mergers on balance enhance efficiency, often by replacing inept management. But studies of actual post-merger performance paint a rather different picture. Numerous studies in the U.K. and the U.S. have found post-merger declines in productivity, profitability, market share, and even stock prices (Cowling, et al (1980), Ravenscraft and Scherer (1988)). It is unclear why ex ante and ex post evaluations of mergers point in such different directions, though the latter suggest the possibility that at least some mergers serve managers better than shareholders in the long run. Scale and Concentration Most work on the determinants of firms' scales has been motivated by a 13 desire to understand how seller concentration is determined. Common measures of seller concentration increase as the number of sellers declines and as their shares become less equal; each thus gives an industry's location on some line between competitive and monopoly structures. Measures of this sort include the four-firm concentration ratio, the share of output accounted for by the four largest sellers, and the Herfindahl -Hirschman H index, the sum of all sellers' squared market shares. These and other plausible concentration measures are highly (but not perfectly) correlated, and they tend to change slowly over time. It seems clear that firms' scales -- and thus market concentration -- reflect what Scherer (1980, ch. 1) has termed basic conditions of technology and demand as well as business decisions and historical accidents. Concentration is thus endogenous in the long run. Economies of Scale Rank correlations of manufacturing industries' concentration levels between industrialized nations are very high (Pryor (1972)), suggesting that some important common factor is at work. Technology is the most obvious candidate. Industrial economists have traditionally devoted considerable attention to the hypothesis that the more important are economies of scale in any particular industry, the higher will be seller concentration in that industry, all else equal. Empirically, long-run average cost (LAC) curves seem generally to be Lshaped: at small scales average cost declines with increases in output, but average cost is approximately constant for output rates above some minimum efficient scale (MES). The importance of scale economies is typically measured by the ratio of MES to the total capacity or output of the industry, sometimes augmented by a measure of the steepness of the LAC curve at scales below MES . Estimates of MES have been obtained by interviewing engineers and executives, by studying the variation of cost or profitability with scale, by seeing what sizes of plants or firms seem to prosper, and by assuming 19 that some fraction (usually half) of an industry's output is produced in efficient plants either in the country of interest or in some larger country (typically the U.S.)- Estimates based on real data inevitably reflect competitive conditions and historical investment patterns, along with the characteristics of best-practice technology that are of primary interest. Answers given in interviews may be speculative when questions go beyond the design decisions with which interviewees are familiar: many people design plants, but few design firms. Despite these measurement problems, a large number of studies have found significant positive relations between seller concentration and the market share of a MES plant or (in only a few studies) firm. But the leading firms in many U.S. markets are apparently much larger than MES, so that concentration is higher than is strictly required for production efficiency (Scherer (1980, ch. 4)). (In smaller national markets the opposite problem is often encountered, particularly where high tariff barriers are present.) A related finding is that the expected negative relation between market growth and changes in concentration tends to be weak. Similarly, among large industrialized nations, concentration levels do not decline much with increases in the size of the economy (Pryor (1972)). The sizes of leading firms tend to increase with the size of the national market, in part through increases in the extent of multi-plant operations (Scherer, Beckstein, Kaufer, and Murphy (1975)). Learning by Doing Since it was noticed during World War II that the labor required to build particular types of ships and aircraft declined with the cumulative volume of production, similar learning economies have been observed in a wide variety of settings. But only a few empirical studies 20 have analyzed variations in the importance of learning in particular processes (Lieberman (1984)). Even less work has been done on variations in the extent to which the benefits of one firm's learning spill over to other firms or lower the costs of other products produced by the same firm. Confining himself to the case, stressed in the business strategy literature, in which learning spillovers are completely absent, Spence (1981) explored the analogy between learning economies and economies of scale. Holding constant the ultimate total cost decline that learning can produce, Spence argued that learning would affect concentration most like economies of scale when learning economies are exhausted at "moderate" values of cumulative output. If exhaustion occurs at "low" cumulative output, a new firm needs essentially to incur only a small fixed cost to fully exploit economies of learning. At the other extreme, if full exploitation of learning economies requires "large" cumulative output, large differences in cumulative production imply only small cost differences. Unfortunately, I know of no empirical studies of the impact of learning economies on market structure. Other Forces If LAC curves are indeed approximately flat above MES , apparently "excessive" concentration is not surprising. Relative sizes of firms that have attained MES might well change because of apparently random innovations in production and marketing, with "better" firms growing at the expense of their rivals (Demsetz (1973)). Any particular innovation might tend to increase or decrease concentration, depending on whether it was made by a relatively large or relatively small firm. As this reasoning would suggest, U.S. manufacturing industries that experience large increases or decreases in concentration tend to show above-average increases in - 21 - productivity and below-average increases in price (Gisser (1984)). A variety of stochastic processes that might plausibly summarize this mechanism tend over time to produce skewed firm size distributions with considerable inequality in firm sizes, broadly consistent with the facts in 14 most U.S. and U.K. industries. Similarly, concentration can be maintained or increased by strategic behavior aimed at deterring entry or disadvantaging small rivals. Another process leading to "excessive" concentration is horizontal mergers -- combinations of competitors. The importance of this process has been vigorously debated, particularly in the U.K. (Curry and George (1983)). On balance, mergers seem to have been important sources of concentration in some E.E.C. nations, but not in the U.S., where policy toward horizontal mergers was quite strict from the early 1950's until the Reagan years. Shareholders of rival firms tend to gain from major horizontal mergers, as the frequently-hypothesized relation between concentration and monopolistic behavior implies, but the size of the gain appears unrelated to the level of concentration (Eckbo (1985)). On the other hand, some horizontal mergers do seem to raise prices (Barton and Sherman (1984)). Scope and Organization When firms produce multiple products, as virtually all real firms do, longrun cost functions cannot be described solely in terms of single-product economies of scale. Indeed in the multi-product context, product-specific average costs are not in general well-defined, and the definitions of economies of scale and natural monopoly must be significantly generalized (Baumol, Panzar, and Willig (1982)). Moreover, it seems clear that the 22 - boundaries between firms and markets and the internal organization of business firms is not determined only by the technology of production; the technology of transaction governance and supervision also matters. Economies of Scope One of the more useful concepts that emerges from recent work on multi -product cost and production functions is economies of scope, which are present when total cost can be reduced by consolidating production of multiple products within a single enterprise. Roughly, economies of scope arise if (but not only if) there are scale economies in the provision of services used to produce more than one output: the same switch can be used for both local and long-distance telephone service, for instance, or the same trucks can be used to deliver a wide array of products to grocery stores. One would expect firms to design product lines to exploit important scope economies, just as one would expect generally to observe firms large enough to exploit important economies of scale. But, while a number of authors have estimated multi-product cost functions, serious empirical use of multi-product cost concepts is not common. Transaction Governance The agency theory view of the firm is complementary to a line of research based on the argument that under competitive conditions, economic activity will be organized so as to economize on production costs plus transactions costs (Williamson (1985)). The many forms of transaction governance observed in practice can be thought of as forming a continuum, with classic spot markets and internal governance within firms at the extremes and contracts of varying duration and complexity in between. Work on transactions costs has concentrated on the identification of transaction attributes that generally affect the comparative performance of - 23 alternative governance structures in a world of selfish, boundedly rational actors, asymmetric information, and incomplete contracts. The transaction attribute most stressed in recent work has been asset specificity , the extent to which a particular transaction requires tangible or intangible assets that would be of substantially less value if redeployed to alternative transactions. Asset specificity is closely related to the notion of sunk costs, costs that could not be recovered if a particular activity were abandoned. The costs of digging a coal mine are sunk, for instance, since they would be lost if the coal business were abandoned. But no asset specificity is involved if a coal mine can easily sell on a spot market to many alternative customers. On the other hand, if an electric generating plant is built at the mouth of a coal mine, asset specificity is important, since the value of both the mine and the plant would decline if the mine had to sell its coal elsewhere and the generating plant had to ship in its coal from other mines. The general argument is that when asset specificity is important, contractual incompleteness inevitably puts at least one party to the transaction at considerable risk, since the value of his investment would decline substantially if the transaction broke down. Even if ex ante many firms compete to dig a coal mine next to some particular power plant, ex post, after one firm has dug a mine, there is bilateral monopoly for the remainder of the life of the transaction. High degrees of asset specificity are predicted to lead to complex long-term contracts or internal governance within firms. This and related predictions from this framework are difficult to test because it is not clear, for instance, how asset - 24 - specificity can be routinely quantified. Still, a fair number of empirical studies have produced supportive results (Joskow (1987)). Internal Organization Depending on the technology of supervising employees, individually and in groups, and on a particular firm's market environment (s) and long-run strategy, different internal structures may be optimal. Considerable research, much of it outside the usual boundaries of economics, has been done on the determinants and effects firms' internal structures (Caves (1980)). This work seems to have shown, among other things, that both the rise of middle management around the turn of the century (Chandler (1977)) and the more recent shift toward organizations based on multiple, relatively independent operating divisions (Williamson (1985)) reflected organizational innovations of considerable value under a fairly broad range of market and strategic conditions. Vertical Relations Two closely related lines of work have focused on vertical integration decisions and on contractual arrangements between manufacturers and firms providing wholesale and retail distribution services. This work has been motivated in large measure by the traditional hostility of antitrust authorities toward vertical mergers -- combinations of a buyer and a seller -- and toward a set of contractual provisions that are called vertical restraints in the U.S.. These provisions limit a distributor's freedom to compete -- for instance by specifying prices to be charged at retail. Much of the literature on vertical integration employs the agency theory or transactions costs framework and thus focuses on sources of efficiency gains. But vertical integration may also be a response to or source of competitive imperfections. A number of early authors argued that 25 industry-wide vertical integration that eliminated an intermediate product market could make entry more difficult by requiring an entrant to both produce and consume that product. Vertical integration may be profitable but have at best ambiguous welfare effects if it permits a monopoly manufacturer to price discriminate or to avoid downstream substitution away from its output in production, or if it arises as a response to rationing caused by price rigidities (Carlton (1979)). Very little empirical work has been devoted to integration related to market imperfections, however. Because vertical restraints typically limit retail competition, antitrust authorities have historically viewed them as signs of retailer cartels. But it is now clear that individual manufacturers can sometimes use vertical restraints to compete more effectively. For instance, a manufacturer might want to fix retail markups in order to induce all retailers to compete by providing demand- enhancing services from which all would benefit (Telser (I960)). Alternatively, placing floors on retail prices might allow high-cost "prestige" stores to stock the product, and thereby to provide a quality signal to buyers, by removing the threat of price competition from low-cost discounters. But when competition is imperfect at manufacturer or retailer levels, the net welfare effect of privately-profitable vertical restraints is often ambiguous (Rey and Tirole (1986)), in part because they change the nature and intensity of rivalry among manufacturers or retailers. The limited empirical work that has been done in this area (primarily case studies) suggests that vertical restraints serve a wide variety of purposes and that they rarely reflect retailer cartels. But generally ambiguous welfare analyses make it hard to make strong policy prescriptions. 26 - III. PRICE, OUTPUT, AND PROFITABILITY A central problem of industrial economics since its emergence as a distinct field has been to devise techniques for using observable variables {market structure , broadly defined) to predict conduct in and performance of markets that do not meet the strict structural conditions of perfect competition. This Section and the next review work on this problem, which is made difficult because, as we saw above in the case of concentration and will discuss further below, market structures are themselves endogenously determined. The focus here is on choices of price, output, and capacity; non-price rivalry is considered in Section IV. I begin with research on the exercise of monopoly power or, equivalently , market power. It is useful here and in what follows to distinguish between short-run and long-run market power. Short-run market power is the ability to raise price profitably above marginal cost; it arises whenever firms face downward- sloping demand curves. Long-run market power is the ability to earn persistently supra-normal profits by setting price above average cost. In the textbook tangency equilibrium of Chamberlinian monopolistic competition, for instance, firms have short-run market power but no long-run market power. Most work on the exercise of market power employs variants of the standard monopoly model. In an industry with more than one firm, sellers' profits depend on the intensity of rivalry and, in the long run, on the entry of new firms. The next two subsections review theoretical work on these dimensions of behavior. The final two subsections consider related empirical work on market conduct and performance. 27 The Exercise of Monopoly Power Price Discrimination A common symptom of monopoly power is price discrimination , which can be roughly defined as selling units of related goods at different percentage markups over marginal cost (Phlips (1983)). In order to discriminate profitably, a firm must be able to affect the prices it receives for its output, to sort units potentially demanded according to their optimal prices, and to avoid arbitrage. The first of these conditions is satisfied whenever firms have short-run monopoly power; price discrimination is consistent with free-entry, zero-profit equilibria involving no long-run power. The monopoly models that dominate this literature are thus potentially components of models of discrimination under other market structures. Following Pigou, it is useful to consider three basic types of price discrimination. A monopolist practicing first-degree or perfect discrimination leaves all her customers just indifferent between buying and not buying. Unlike a non-discriminating monopolist, she does not restrict output. Instead, she maximizes total surplus, as under competition, and then appropriates it all. It does not follow that real price discrimination is generally efficiency-enhancing, however, since first-degree discrimination is a limiting case found only in texts and journals. The simplest of Pigou' s other two types is third degree discrimination, which involves sorting customers into groups according to their demand elasticities and charging group-specific prices that vary inversely with elasticity. Case studies provide a rich array of sorting mechanisms: discounts for air travelers who reserve far in advance sort tourists from - 28 - business travelers, for instance, and supermarket coupons are used only by price-sensitive consumers. In these and other cases, transactions costs seem to be the main check on arbitrage. Since price discrimination makes marginal rates of substitution unequal, an increase in total output is a necessary condition for thirddegree discrimination to increase (Marshallian) social welfare (Varian (1985)). Output is unaffected by discrimination if all group demands are linear and all groups make purchases under both uniform and discriminatory pricing; total output is more likely to increase if sales to some groups are profitable only under discrimination. Ambiguous welfare results of this sort make it hard to prescribe general policies toward price discrimination. The final Pigouvian type, second-degree discrimination, involves nonlinear pricing, in which the buyer's average cost per unit depends on the quantity purchased (Maskin and Riley (1984)). The simplest case of nonlinear pricing is the so-called two-part tariff: buyers must pay a fixed charge, F, for the right to purchase any amount at a per-unit cost of P. If individual demand curves do not cross, profits are maximized with F positive and P between marginal cost and the ordinary monopoly price (Oi (1971)). Intuitively, it pays a monopoly to reduce P a bit if it can capture some of the increased consumer's surplus by raising F. If there are a finite number of consumer types with non-crossing demand curves, then under general nonlinear pricing regimes with a finite number of consumer types, all types but the one with the largest demand have marginal valuations for the good that exceed marginal cost, and all types but the one with the smallest demand enjoy positive consumer's surplus. Under some conditions the optimal - 29 nonlinear price schedule can be implemented by allowing buyers to select from a set of two-part tariffs. Actual pricing decisions and theoretical studies often involve variations on and combinations of the last two Pigouvian themes. An extensive literature has developed on spatial price discrimination. Random variations in price over time or space may profitably sort buyers according to their costs of search. Multi-product monopolies must consider crossprice elasticities and may find it profitable to sell bundles of two or more products or to use nonlinear pricing (Spence (1980)). Finally, one might think that a monopolist selling a durable good could discriminate intertemporally by lowering price over time, thus effecting first-degree discrimination by sweeping out the demand curve. But sophisticated buyers will anticipate price cuts under these conditions and will postpone their purchases until price falls to marginal cost, so that the monopolist would actually be better off if he could commit never to change price (Gul, Sonnenschein, and Wilson (1986)). A number of authors have studied the problem of pricing to maximize the welfare generated by a natural monopoly that is subject to a break-even constraint. This research, which has had a significant impact on public utility pricing, is closely related to work on optimal commodity taxation. Since price discrimination is generally profit-maximizing, profitconstrained welfare maximization generally involves departures from marginal-cost pricing in the direction of discriminating monopoly pricing. The use of nonlinear pricing is often particularly attractive in this setting, since it can Pareto-dominate linear pricing (Willig (1978)). 30 - Non-Price Consequences Spence (1975) showed that a single-product monopoly could choose a sub-optimal or supra-optimal quality level depending on the details of the demand structure. In order to maximize profit, a monopoly considers the effect of quality only on the reservation price of the marginal consumer, who is indifferent between purchasing and not, rather than on the value of the product to the average consumer. For basically the same reason, multi-product monopolies may offer too much or too little variety. If consumers differ in their willingness to pay for increments to quality, a multi-product monopolist will generally increase quality differences above socially optimal levels (under standard assumptions by lowering qualities at the low end of the product line) in order to facilitate price discrimination. Oligopoly Theory The Holy Grail of research in oligopoly theory has been the ability to use observable quantities to predict the intensity of rivalry in markets dominated by a small number of sellers. The literature now contains dozens of formal oligopoly models. These have provided insights that can be used to structure the analysis of particular markets, but they have given us a multitude of possibilities rather than the Holy Grail. Indeed, collectively they suggest that the Grail may not exist. Cournot vs. Bertrand This point is illustrated by two important oligopoly models that were introduced well before the emergence of industrial economics as a distinct field: the one-period Cournot (outputsetting) and Bertrand (price -setting) models. When products are homogeneous, the former predicts behavior intermediate between competition 31 - and monopoly with any finite number of sellers, with competition generally emerging in the limit as the number of sellers grows. Multiplying each of equations (2) by q • , adding, and rearranging, we obtain (P - MC)/P = H/E, (4) where E is the market price elasticity of demand, -P/P'(Q)Q, H is the sum of 2 — squared market shares, 2(q-/Q) , and MC is average marginal cost, Eq-MC-/Q. If all N firms have the same cost functions, H = 1/N, and the gap between price and marginal cost declines smoothly from the monopoly level to zero as N rises. The Bertrand model, in contrast, predicts essentially competitive outcomes with two or more sellers when products are homogeneous. If all sellers have the same cost functions and there are no capacity constraints, no pure strategy equilibrium with prices above marginal cost can exist, since any seller could increase profits by undercutting the lowest price slightly and capturing the entire market. Even when products are differentiated, outcomes tend to be noticeably more competitive when price is the strategic variable than when quantities are chosen. Intuitively, any single seller's demand curve has roughly the slope of the market curve when rivals' quantities are fixed; its demand curve is much flatter when it can steal sales from rivals by undercutting their fixed prices. Neither of these classic models is fully satisfactory. The mechanics of price determination are unclear in the Cournot model, while the Bertrand model depends on the absence of capacity constraints. A natural way to unify them is by observing that prices are generally more easily adjusted than capacities and considering a two-period game with capacities chosen in - 32 the first period and prices chosen in the second. Unfortunately, the equilibria of this game may either be Cournot or substantially more competitive, depending on how excess demand (which is never observed in equilibrium) is assumed to be rationed (Davidson and Deneckere (1986)). Two-period games of this sort, in which irreversible first-period commitments are made with a view to affecting second-period play, provide a useful approach to modelling a wide variety of situations. By signing contracts binding themselves to matching the lowest price offered by another firm, for instance, or engaging in a variety of related "facilitating practices," sellers may be able to support collusive outcomes (Salop (1986)). The observation that government policies (such as export subsidies) may serve as valuable commitments to firms in open economies has led to a fundamental reexamination of the case for free trade (Brander and Spencer (1985), Krugman (1986)). Fudenberg and Tirole (1984) have shown that the qualitative nature of first-period strategies in two-period games of this sort depends simply on the signs of two second-order partial derivatives of firms' payoff functions. (See also Bulow, Geanakoplos , and Klemperer (1985).) In particular, first-period strategies generally depend critically on whether the second-period game is of the Bertrand or the Cournot type. While in Cournot models a firm's best response to an aggressive increase in its rivals' output is generally to retreat by reducing its own output, in a Bertrand model (with differentiated products) the best response to an aggressive price reduction is usually to counterattack by cutting price. Supergames and Collusion In an influential paper, Stigler (1964) argued that oligopoly theory should be based on the theory of cartels. Any 33 - cartel has two tasks (Scherer (1980, ch. 5-7)). Its first task is to agree on a course of action -- a set of firm-specific outputs, for instance. Agreement is likely to be more difficult the more sellers that must be involved and the greater the differences among their costs and products. Stigler and most subsequent authors have placed more stress on the second task: to deter violations of the cartel agreement. When prices are raised to monopoly levels, each seller stands to gain by making undetected price cuts or output increases. Such cheating is less attractive the more quickly it can be detected and the more severe the punishment that can be credibly threatened. Cartel members may facilitate detection by dividing customers among themselves or adopting a number of related practices. Stigler noted that these same two problems must be solved by firms that attempt tacit collusion , on which the subsequent literature has concentrated, and try to mimic the explicit or overt collusion of a cartel without a formal agreement . Because detection and punishment take time, the supergame framework has often been employed to study the stability of collusive agreements, most often with variants of the Cournot model as the stage game. But with full information, collusion cannot emerge as a perfect equilibrium when the number of periods is finite and known in advance. To see why, suppose the stage game is Cournot. In the last period, firms face a one-period Cournot game, and the outcome must be the Cournot solution. Threats to behave otherwise are not credible. But since nothing done in the second- last period can affect what happens in the last period (beliefs are fixed with full information) , the firms face a one-period Cournot game at the start of the second-last period as well. By backward induction, rational players - 34 will simply repeat their one-period Cournot strategies whenever there is a known, finite last period. Thus collusive equilibria can only appear when the horizon is infinite. But in this case the Folk Theorem comes into play: in a large class of models there seem almost always to exist many collusive perfect equilibria, in which average payoffs exceed those in the stage game equilibrium. Suppose, for instance, that the basic Cournot game discussed in Section I is to be played an infinite number of times and that MC-(q-) = M, a constant for all firms. Then the single-period monopoly output, Q , is well-defined. Let q c be the single-firm Cournot output obtained by solving equations (2). Suppose each firm's strategy is to produce q = Q /N in period 1 and in every later period in which the previous period's total output has been Q , and to produce q c otherwise. If all other firms play this strategy, firm i could increase its profits in any single period by increasing its output. But in all subsequent periods it would then earn only Cournot, not monopoly profits. (If all other firms will produce q c , firm i's best response, by definition, is to produce q c also.) If the discount rate is low enough, the present value of these future losses will exceed the single-period gains from cheating, and the strategies discussed above will form a Nash equilibrium in which the monopoly output is produced in all periods. Of course, as Stigler stressed, firms may not be able to observe each others' outputs directly. But even if players can only observe market price, which depends on industry output and a random variable, there generally exists a continuum of collusive equilibria supported by trigger price strategies for any finite number of sellers (Green and Porter (1984)). These involve producing a low (collusive or monopoly) output unless market 35 price drops below some level, and then (assuming symmetry) producing q for some punishment period. (This threat is credible, as above, since if everyone expects everyone else to produce q , the best response is to follow suit.) In these equilibria cheating never occurs, but punishments are nonetheless sometimes carried out. This literature shows clearly that the more damaging the threats that can be credibly made and the smaller the gains from cheating, the greater the scope for profitable collusion (Abreu (1986)). Thus, somewhat paradoxically, the best sustainable collusive outcomes may be more monopolistic when the stage game is Bertrand than when it is Cournot, since the single-period Bertrand equilibrium involves zero profits, and excess capacity that can be used to fight price wars may instead sustain monopoly prices. Collusion may be more effective at business cycle troughs than at peaks if cheating is more profitable when demand is strong (Rotemberg and Saloner (1986)). The supergame literature raises some serious questions that have not yet been completely answered. In the absence of explicit collusion, how can firms select a single equilibrium from a continuum -- particularly if (as in reality) the firms are not identical? What are we to make of the fact that collusive equilibria generally exist for reasonable numbers of firms -- is collusion really almost universal? Is it plausible to think that cheating on collusive understandings never occurs? What happens if firms can renegotiate collusive agreements during a punishment period (Farrell and Maskin (1987))? The supergame literature seems so far mainly to have added to the long list of possible behavior patterns developed in 36 other branches of oligopoly theory, not to have placed strong restrictions on observable conduct. Entry and Exclusion Bain (1956) defined barriers to entry as factors that make it possible for established firms in an industry to enjoy supra-normal profits without attracting new entry. Without entry barriers, there can be no long-run market power; collusive behavior cannot succeed in raising profits in the long run. Thus preventing the entry of new firms is roughly as important in the long run as restraining rivalry among established sellers. Bain listed four sources of entry barriers: economies of scale, cost advantages of established firms, product differentiation advantages of established firms, and absolute capital costs. This list has generated both controversy and research on the possible exploitation of these factors to deter entry or induce exit. Recent work here, as in other areas, has paid particular attention to the implications of asymmetric information. Scale Economies In the presence of economies of scale, a viable entrant would add a non-negligible amount to total industry output. Bain (1956) argued a monopolist would engage in limit pricing to deter entry in this case by setting pre-entry output high enough (generally above the ordinary monopoly level) so that the addition of an entrant's output would force price below cost. But this argument has a serious game- theoretic problem: the implicit threat to maintain output in response to entry is not credible, since the incumbent (quantity-setting) firm would generally do better to reduce production. 37 - Spence (1977) observed that an incumbent's irreversible pre-entry investment in capacity might make threats of this sort credible by lowering its post-entry marginal cost, thus enhancing its incentives to maintain high output. (See also Dixit (1979).) Similarly, learning economies may induce an established firm to increase its pre-entry output in order to lower its post-entry marginal cost. In a variety of two-period models, an established monopoly over- invests in the first period to deter entry. Pre-entry output generally exceeds the monopoly level, as in limit-pricing, and profit may be much lower unless scale economies are very important (Schmalensee (1981)). The welfare implications of this behavior are generally ambiguous, since entry tends to be socially excessive in the presence of economies of scale (Mankiw and Whinston (1986)). The effects of scale economies also depend critically on timing assumptions and on the importance of sunk costs. In the limiting case of a perfectly contestable market there are no sunk costs, so that firms can enter or exit an industry costlessly, and entrants can enter, undercut incumbents' prices, and exit before incumbents can react. Under these strong assumptions about costs and differential reaction lags, and with other sources of entry barriers assumed away, potential entrants can enforce essentially competitive outcomes even in natural monopolies (Baumol, Panzar, and Willig (1982)). More generally, the higher are sunk costs, the greater the risk assumed by entrants, and thus the less attractive is entry. Thus barriers to exit, tangible and intangible sunk costs that make exit unattractive even when economic profits are negative, also serve to discourage entry. 38 In some markets scale economies imply that capacity is most economically added in large lumps, and investment costs are mostly sunk. Under these circumstances entry may be rationally prevented by preemption , the seizing of a discrete opportunity by an incumbent firm with market power before it can be used by an outsider to enter. The value of a new plant to an incumbent monopolist in a growing market is the difference between the monopoly profit it would enjoy with the plant and its share of the duopoly profit that it would receive if the potential entrant built the plant and entered. The value to the entrant is its share of duopoly profit in the latter case. As long as monopoly profit exceeds total duopoly profit, the plant will be worth more to the incumbent -- who will thus have an incentive to build it before the market has grown enough to attract an entrant. Other Bainian Barriers The effects of incumbents' cost advantages on entry incentives is sensitive to assumptions regarding post-entry rivalry. If the post-entry game would be Bertrand (with simultaneous moves) even a tiny cost advantage of an established monopoly serves to deter entry. But in the Cournot case, entry may be profitable despite higher costs. Indeed, with linear demand and constant costs, it is easy to show that high-cost but profitable entry may lower total surplus in the latter case. Switching costs may be important sources of product differentiation advantages of established firms in some markets. Switching costs may be objective, as in the case of computer systems, or subjective, deriving from a satisfied customer's rational reluctance to experiment with an untried entrant (Schmalensee (1982)). While it seems clear that these costs can advantage early entrants, the critical role of expectations in buyers' decisions makes multi-period modelling difficult outside steady states. 39 Bain's argument that an entrant's need to invest absolutely large sums of money might serve as a barrier to entry has been widely criticized because it seems to rest on capital market imperfections; incumbents also had to invest large sums. Bain might have been groping toward the sunk cost issues discussed above. Or, he might have anticipated the point that even perfectly competitive capital markets may be seriously affected by asymmetric information regarding a potential entrant's prospects. Information and Reputation Information asymmetries can rationalize a variety of policies to deter entry or induce exit (Roberts (1987)). If its costs are unobservable , an established monopolist may set price below the monopoly level, as in earlier limit -pricing models, in order to signal to potential entrants that its costs are lower than theirs would likely be. On the other hand, if potential entrants know only that their post-entry costs would be similar to those of an incumbent monopolist, the latter may set price above the monopoly level in order to signal that its costs are high and the market is thus relatively unattractive (Harrington (1986). Of course, since rational actors understand opponents' incentives perfectly and probability distributions of cost levels are common knowledge, nobody is fooled on average in equilibrium in either case. Imperfect information can also provide a rationale for predatory pricing, a legal term of art generally taken to mean charging unprofitably 18 low prices in order to eliminate an established rival. Until relatively recently, the following points were taken as a proof that predatory pricing is rarely if ever rational. The predator's losses generally exceed the prey's, since the prey can shut down temporarily, while the predator must make substantial sales to keep price low. Even if the prey is driven into 40 - bankruptcy, the predator may need acquire the prey's assets in order to avoid their being operated by a new rival. But then it will surely be cheaper simply to merge with the prey at the outset than to incur losses driving it from the market. A number of recent studies have argued that potential entrants might well attach some positive probability (assumed of course to be common knowledge) to the possibility that an established monopoly is irrational -- that it will always prey on entrants regardless of the costs. Then a rational established firm facing a finite set of potential entrants will often find it optimal to prey on the first few entrants in order to build (or, more precisely, to avoid destroying) a useful reputation for irrationality (Kreps and Wilson (1982)). With incomplete information, predation may also serve to lower the cost to the predator of acquiring the prey (Saloner (1987)). Unfortunately, since unobservable beliefs play a critical role in reputation models, these models place relatively weak restrictions on observed behavior; they imply the potential rationality of predation under almost any observable conditions. Does Market Structure Matter? 19 Let us now turn to empirical research. Many of the industry case studies discussed in Section seemed to detect tacitly collusive patterns of behavior in a variety of concentrated markets. But later, more quantitative studies have produced less clear-cut evidence of market power. Profitability Differences Oligopoly theory suggests at least the plausibility of the hypothesis that there is a negative relation between seller concentration and the average intensity of rivalry. Bain (1951) - 41 - argued that this implies that concentration should be positively correlated with industry-average profitability, and he found some support for such a correlation. Literally hundreds of subsequent studies have examined the relation between concentration and profitability in cross-section data. Through the early 1970' s, most such studies found a weak, positive correlation between concentration and industry-average profitability. The weakness of this relation was generally attributed to problems of defining markets and measuring profitability, and these results were generally interpreted as confirming the hypothesis that concentration tends to facilitate collusion and otherwise limit rivalry. Then Demsetz (1973) provided a plausible and disturbing alternative interpretation. To illustrate his argument, suppose that the single-period Cournot model developed above describes price formation in all markets, regardless of the level of concentration. Since cross-section regressions aim to reveal differences in long-run equilibria, suppose further that all production takes place under constant returns to scale (all firms are above MES) but that costs may differ within individual markets. Then equation (4) implies that for any individual industry, n/(PQ) = H/E, (5) 20 where II/(PQ) is the industry's rate of return on sales. For any value of N, H will be larger the greater the variance in firms' costs. This model thus predicts that in industries in which all firms are roughly equally efficient, concentration and industry-average profitability will be low. In industries in which some firms are noticeably more efficient than others, the more efficient firms will tend to capture large - 42 market shares, so that concentration will be high. And more efficient firms will earn rents, so that industry- average profits will also be high. Thus concentration and industry-average profitability will be positively correlated even though there is no collusion anywhere. This formal model probably overstates the dependence of concentration on idiosyncratic cost differences in light of the high correlations between concentration levels in different nations. But Demsetz's basic argument has received some empirical support. Bain's (1951, p. 320) did note that in his data, "Smaller firms tended to fare about the same regardless of industry concentration; the dominant firms in general had earnings rates that were positively influenced by concentration," and other U.S. studies have confirmed this finding. Similarly, at the firm or business unit level, market share is strongly correlated with profitability in samples that include many industries, and the coefficient of concentration is negative or insignificant in profitability regressions including market share (Ravenscraft (1983)). On the other hand profitability is not strongly related to market share in a sizeable fraction of manufacturing industries, (Porter (1979)). A variety of attempts to discriminate between the Bain and Demsetz interpretations have produced mixed results -- suggesting at least that both mechanisms may be at work in the economy. The 1970' s also saw the publication of a host of industry- level studies in which the concentration-profitability correlation was zero or negative. In U.S. data this correlation weakened dramatically in that decade (Domowitz, Hubbard, and Petersen (1986)); U.K. data seem to yield a monotonic relation between concentration and profitability very reluctantly (Geroski (1981)). Thus not only is it now hard to interpret a significant 43 - positive correlation between concentration and profitability, it is hard to 21 find such a correlation in many data sets. Bain (1956) noted that collusion could not sustain high profits in the long run in the absence of barriers to entry. This calls for an interactive (concentration x barriers) specification, but such specifications have not fared well empirically, perhaps in part because it is difficult to measure barriers to entry empirically. The most robust interactive result of this general sort is that the impact of imports on domestic profitability seems to be higher when domestic concentration is high (Caves (1985)). A sizeable number of authors have simply added proxies for various sources of entry barriers to regressions of profitability on concentration. In these linear (concentration + barriers) specifications, measures of scale economies or capital requirements of entry tend to be positively related to profitability, as do measures of advertising intensity. (The interpretation of the advertising results is discussed in Section IV.) In Bain's (1951) data, if one takes average profitability in the unconcentrated subsample to be the competitive rate of return, it follows that monopoly profits in the concentrated subsample average less than 5% of 22 sales. Indeed, even ignoring concentration, observed profitability differences, which are magnified by short-run disequilibria, are generally small relative to those implied by theoretical comparisons between 23 competition and monopoly. For this reason, studies of the total social cost of market power based on observed profitability differences tend to produce tiny deadweight loss estimates. Indus try -Specific and Behavioral Evidence Inter- industry profitability studies suffer from the limitations of accounting data and the inability to 44 measure a host of industry-specific variables. A number of authors have dealt with these problems by examining the correlation between seller concentration and the level of price across markets (often geographically separated) within individual industries. Most find a significant positive relation (Branman, Klein, and Weiss (1987)), tending to support an association between concentration and restrained rivalry. And there is some evidence that prices, like profits, are raised by tariff protection of concentrated industries. But few price studies attempt to investigate systematically the effects of conditions of entry. The wave of econometric industry studies that have appeared in recent years generally conclude that firms set price above marginal cost (Cubbin (1975), Bresnahan (1987)). Estimates of X- in equations like (3) always exceed (-1) and seem to be positive more often than negative. The data necessary for these studies are most readily available for concentrated industries, particularly those that have been subjected to antitrust prosecution, many of which sell differentiated products, and many different techniques that have been employed in this work. Thus very little has been learned from econometric industry studies about general relations between market conduct and observable elements of market structure. But this work does suggest strongly that short-run market power is exercised in at least some concentrated industries. The experimental literature mirrors recent theoretical findings: behavior in laboratory markets seems sensitive to small changes in information and institutional structures (Plott (1982)). In both largenumbers and monopoly situations, the cases that have received the most attention in this literature, performance seems to vary considerably 45 depending on whether prices are posted, negotiated, or called out. A wide variety of outcomes have been observed in broadly similar experimental oligopoly markets. Some practices that have been alleged to facilitate collusion (Salop (1986)) have been observed to have this effect in the laboratory (Grether and Plott (1984)). Many market settings and hypotheses about strategic behavior have been investigated experimentally only once or twice; some parts of the theoretical literature have remained untouched by experiments. Like econometric industry studies, laboratory experiments have not yet yielded a set of robust empirical findings that can serve to replace or underlie a general formal theory of imperfect competition. But they do seem generally to support the hypothesis that (exogenous) market structure affects behavior . Two additional bits of evidence deserve mention here. Hay and Kelley (1974) found that price-fixing conspiracies, at least those that were detected by U.S. antitrust authorities, tended to occur in concentrated industries. And Hall (1987) has argued that the assumption of short-run monopoly power provides the best explanation for the observation that productivity varies pro-cyclically in many industries. Rent Dissipation and Rent Sharing The preceding discussion suggests that short-run market power is not uncommon, but the high profits that would be predicted by long-run market power are rare. It would seem that either the rents produced by pricing above marginal cost are dissipated, perhaps by entry or non-price rivalry, - 46 or they are shared by firms' owners with suppliers of other inputs. I deal here with entry and rent sharing and treat non-price rivalry in Section IV. Entry If entry is easy, we have known since Chamberlin (and been recently reminded by contestability theory) that prices above marginal cost are consistent with zero economic profits. Official data usually show large numbers of small entrants in most industries, though most obtain tiny market shares and small new entrants have particularly high failure rates (Dunne, Roberts, and Samuelson (1987)). One might explain away the lack of a robust positive correlation between concentration and profitability by arguing that collusive behavior tends to attract small inefficient entrants, whose performance depresses industry averages. But the fraction of output produced in inefficiently small plants seems if anything to be negatively related to concentration. On the other hand, the inefficient entry hypothesis is consistent with the finding that tariff protection increases this fraction, particularly in concentrated industries (Baldwin and Gorecki (1985)). Alternatively, one might argue that monopoly rents are largely dissipated in the process of obtaining market power and deterring the entry of effective rivals. High estimates of the welfare cost of market power are implied by this argument (Cowling and Mueller (1978)), but the theoretical and empirical case for substantial rent dissipation of this sort is somewhat weak (Fudenberg and Tirole (1987)). In particular, little direct evidence of strategic behavior to deter entry has been detected in industry studies (Lieberman (1987)), though unfortunately few attempts have been made to detect it. A few studies have examined correlates of measures of the importance of 47 entry. Estimates of the market share of a plant of minimum efficient scale and of the capital cost of such a plant tend to be negatively related to observed entry, as does advertising intensity. Profitability is not generally strongly correlated with subsequent entry, but it is unclear whether this reflects expectations that significant entry would lower profits or the difficulty of measuring profitability. Labor Costs One might expect managerial behavior that is not in shareholders' interests to be more prevalent, ceteris paribus , when rivalry, and thus market discipline, is weak. And one might expect costs to rise as a consequence of such behavior, either because managers treat themselves to high salaries, plush offices, and large staffs or because they simply fail to perform the unpleasant task of cost control. There appears to be little empirical support for this view of the world (Smirlock and Marshall (1983)), but measurement problems are obviously severe. A good deal of work has recently been done on inter- industry wage differences that cannot be explained by differences in worker characteristics (Krueger and Summers (1987)). Like market concentration, these differences seem stable over time and highly correlated internationally. A number of authors have found that after controlling for worker characteristics, wage rates tend to be high in industries with high profitability (Dickens and Katz (1987)), suggesting that monopoly rents may be largely captured by workers. Rose's (1987) before-and-after study of trucking deregulation in the U.S. indicates that unionized workers captured over 2/3 of the rents produced by regulation. On the other hand, the pattern of wage differences in the Eastern Bloc seems to be highly correlated with that in the West. For this and other reasons, the exact 48 roles of technology, market power, and unobservable worker characteristics in the determination of wage differences remain controversial, though the view that monopoly profits tend to be shared with workers (particularly unionized workers) is coming to be widely held. IV. NON- PRICE RIVALRY Despite the picture painted by most microeconomics texts, business managers do not devote all their waking hours to setting price, output, and capacity. Major changes in capacity are infrequent, and prices tend to be rigid, especially in concentrated industries (Carlton (1986)). Important decisions regarding product quality and variety, advertising, and research and development are more frequent in many market settings. These decisions in turn often have important effects on the evolution of market structures. There is relatively little in the academic empirical literature -- or even in the folklore of antitrust --to suggest that non-price rivalry is often much muted by collusive behavior. Perhaps this is because it is more difficult to monitor rivals' research, advertising, and design activities than their prices and because it takes longer to retaliate along these dimensions than to change prices. At any rate, the literature on non-price rivalry has been more concerned with the social efficiency of noncollusive behavior than with the possibility of collusion. Little support has been found for the notion that non-price rivalry generally dissipates rents in a socially optimal manner. - 49 Product Selection Competing sellers rarely choose to offer exactly identical products, since product differentiation generally makes firm demand curves less elastic and thus tends to enhance short -run market power. And product- specif ic fixed costs (related to design, tooling, and introductory advertising, among other things) imply that production of all possible products is rarely an optimal or equilibrium outcome in any market. Equilibria and Optima Three types of models dominate the theoretical literature on product selection. In representative consumer models (Dixit and Stiglitz (1977)), there is a single buyer who consumes all products on the market and whose utility increases in the number of products available; variety is valued for its own sake. These models are consistent with Chamberlinian large-group monopolistic competition, since a change in any one product's price affects all others symmetrically. They have been used to show that intra- industry international trade, which effectively enlarges markets, can increase welfare by increasing equilibrium variety (Helpman and Krugman (1985)) . The other two types of models involve heterogenous buyers who purchase only one product and products that are described as points in a space with dimensions corresponding to product characteristics. In these models rivalry tends to be localized because a change in any product's price mainly affects its nearest neighbors. Thus even in markets with many firms or brands, all sellers may be effectively in small-numbers situations. In models of horizontal differentiation (Salop (1979)), buyers would make different choices if all possible products were available for free. The analysis of these models generally takes an explicitly spatial form; greater 50 variety gives buyers on average products closer to their ideal points. In contrast, vertical differentiation arises if buyers agree on quality comparisons among all possible products but differ in their willingness to 24 pay for increments to quality (Shaked and Sutton (1983)). Models of all three types indicate that market equilibria rarely involve optimal arrays of products. The optimal (second-best) array would maximize consumer's plus producer's surplus (conditional on firms' pricing rules) . Variety tends to be under-supplied because (absent perfect price discrimination) the profit produced by a new product is less than its contribution to total surplus. But variety tends to be over-supplied because (with multiple sellers) the profit earned on a new product generally exceeds its contribution to total industry profit, since the profit earned by rivals' existing products fall. whether variety is under- or oversupplied on balance depends on the details of the model studied. Entry Deterrence It has been argued that established firms may find it profitable to bar entry by preempting locations in the space of potential products (Schmalensee (1978)). The argument rests on product-specific economies of scale and basically parallels the discussion of preemption in Section III: any given product opportunity is more valuable to an established monopolist than to an entrant because entry would increase rivalry and reduce total profits from all products. But without productspecific barriers to exit the game -theoretic problem of the original limitpricing model reappears: the threat to leave "nearby" products in place after entry may not be credible (Judd (1985)). Empirical Studies The marketing literature contains a large set of techniques for estimation of the demand sides of markets with horizontal - 51 - and/or vertical differentiation (Shocker and Srinivasan (1979)). But these techniques have not been employed by economists. The relevant empirical literature in industrial economics is thin and concentrates on methods of econometric industry analysis when products are differentiated. For instance, Bresnahan (1987) uses a complex econometric model to test conduct hypotheses in a market with vertical differentiation. Baker and Bresnahan (1985) present a reduced form technique, which avoids the need for structural estimation of demand parameters, for estimating product-specific net demand elasticities that capture the effects of rivals' reactions to price changes. Little work has been done on testing for localized rivalry or distinguishing among alternative forms of differentiation. Advertising Advertising has long polarized industrial economists. Some view it as a device for differentiating products, and thus increasing market power, and for building barriers to entry. Others view advertising as a source of consumer information, which thus reduces market power, and as a means of effecting entry by informing consumers of new products. Since advertising ranges from uninformative televised skits about well-known products to newspaper ads that provide detailed price and availability information, it would not be a great surprise if both groups were sometimes right. Theoretical Analyses A number of authors have constructed models of advertising rivalry, treating advertising outlays simply as demand shifters (Friedman (1983)). Such rivalry dissipates profits most effectively when advertising has strong effects on market shares, since then firms' advertising elasticities of demand exceed the corresponding industry - 52 elasticity. (Recall the comparison of Bertrand and Cournot models in Section III.) Related models show how economies of scale in advertising interact with those in production to determine the net advantages of size. But to analyze the effects of advertising rivalry on welfare and conditions of entry, one must know exactly how advertising shifts demand. If advertising alters tastes, for instance, welfare conclusions depend on which set of tastes is used to evaluate advertising- induced output changes (Dixit and Norman (1978)). And, while imperfect information is a potentially important source of market power even when there are no barriers to entry, equilibrium levels of even purely informative advertising are not generally socially optimal (Grossman and Shapiro (1984)). Under some circumstances, equilibrium advertising outlays may provide quality signals to alert consumers, since high-quality producers have the greatest incentive to have buyers sample their wares, but such signalling inevitably involves waste. Overall, the existing theoretical literature indicates that advertising equilibria are generally not welfare optima and that the nature and extent of the differences depend on the details of the model. If advertising has long-lived effects on demand, it may be rational in a two-period model for an established firm to over-advertise in the first period to deter potential second-period entry. But optimal first-period strategies depend on exactly how advertising is assumed to affect demand and, as in any two-period model, on the type of second-period game assumed. Evidence A number of studies have found that advertising/sales ratios in consumer goods industries first rise and then fall as concentration increases (Buxton, Davies , and Martin (1983)). While this suggests the possibility that the intensity of advertising rivalry diminishes as 53 - concentration reaches high levels, bivariate relations between endogenous variables are inevitably difficult to interpret. In many cross -section studies, manufacturing industry advertising intensity, typically measured by the advertising/sales ratio, is strongly 25 correlated with accounting profitability. This correlation was initially interpreted as revealing the ability of advertising outlays to differentiate products and create entry barriers. But because advertising is expensed, rather than treated as an investment, accounting profitability is generally over-stated when advertising has long-lived effects on demand. The over-statement is greater the higher is the advertising/sales ratio and the more slowly advertising effects decay (Demsetz (1979)). Similarly, if (partial) collusion produces high price-cost margins, both optimal advertising spending and profits will generally be high. The existing evidence does not definitively rule out any of these structural interpretations, in part because it is difficult to estimate the rate at which advertising effects decay or to observe exogenous determinants of advertising outlays. A number of empirical studies suggest that the effects of advertising on market performance depend critically on the nature of the advertising involved and on the roles played by retailers and other information sources (Porter (1976)). In particular, it appears that restrictions on retailer advertising tend to raise prices (Benham (1972)). Research and Development It is a commonplace that technical progress, the development and use of new products and processes, is the most important source of increases in 54 consumer welfare in modern economies. Slight reductions in the rate of progress outweigh any plausible estimates of the static welfare costs of monopoly power after only a few years. It is also frequently noted that this subject has received much less study than its importance warrants. But, perhaps because of productivity slowdowns during the 1970' s, studies of the sources of technical change have multiplied in the last few years. Models of Technological Rivalry Theoretical work in this area has generally assumed a known, possibly stochastic relation between research and development (R&D) spending and the advance of knowledge. Most studies have considered noncooperative equilibria in which firms incur R&D costs in the hope of securing a single possible patent. It is sometimes assumed (particularly when firms are asymmetric in some respect) that the firm that spends the most gets the patent. More recent work tends to assume that spending levels instead affect success probabilities; some studies assume that several successive successes are necessary to win the patent. It has long been accepted that the market system is unlikely to yield the socially optimal rate of technical progress. The traditional view has been that there is generally too little technical progress. Unless patent protection is permanent and patent-holders can practice perfect price discrimination, the private returns to innovation will fall short of the social returns. Monopolies not threatened by entry have particularly weak incentives to innovate, since innovation in effect ends the profit flow produced by their initial monopoly position (Arrow (1962)). But recent work that explicitly models multi-firm R&D rivalry makes it clear that there can be both too much R&D spending and too much technical progress in equilibrium (Dasgupta and Stiglitz (1980)). Duplication of R&D 55 efforts is a source of social waste, and intense competition for a valuable patent can lead to innovation occurring sooner than would be socially optimal. The efficiency problem here parallels that in the product selection literature; increased R&D spending by a single firm involves an externality because it lowers the expected profits of rivals. A new wrinkle is that because patents are awarded to the first firm to innovate, there is an incentive to adopt excessively risky research strategies, since it doesn't matter if one loses by a day or a decade. If patent protection is imperfect, so that rivals benefit from each others' R&D, waste is reduced, but so are incentives to invest in research (Spence (1984)). Market structure is clearly endogenous in the presence of R&D rivalry, since success brings with it some (generally temporary) market power. A number of authors have explored the possibility of preemptive patenting to 27 deter entry. In the simplest case, an incumbent monopoly will always outbid a potential entrant for a patent on a new production process that either could use, exactly as a new plant or a new product is more attractive to an established monopoly than to a new entrant (Gilbert and Newbery (1982)). But preemption is less likely to be rational if the patent does not yet exist (since the monopoly generally has less to gain from invention) , if there is uncertainty in the research process (since a potential entrant may have a positive probability of winning the patent race with even a very small-scale research effort), or if there are multiple patents that can be used to effect entry (Reinganum (1983)). In general, whether R&D rivalry tends to perpetuate concentrated market structures depends on the details of the model studied (Vickers (1986)). 56 Empirical Studies A number of studies make clear some of the limits of O Q the theoretical literature. Most research is devoted to the development of new products, not new processes, and development (post- invention) spending far outweighs research spending in most industries. In some industries (e.g., chemicals) patents are effective and important instruments for preventing imitation, but they can often be invented around, and in many industries (e.g., electronics) patents are neither effective nor important (Levin, Klevorick, Nelson, and Winter (1987)). In some cases the time required to copy an innovation is the main source of an innovator's rewards, even though copying is usually cheaper than innovating. It would seem that corporate R&D efforts can only rarely be well described as patent races with a single prize. Many authors have attempted to test Joseph Schumpeter's assertions that large firms and concentrated industries are disproportionately important sources of technical progress. But, aside from very small firms, which pose particular measurement problems, R&D spending as a percentage of sales does not seem to rise with firm size in most industries (Cohen, Levin, and Mowery (1987)). Moreover, the largest firms are not disproportionately important producers of major innovations, nor are they quickest in all cases to adopt innovations originating elsewhere. And, adjusting for differences in technological opportunity, increases in seller concentration do not appear to spur R&D effort. Schumpeter also stressed that R&D rivalry shapes market structures, a theme that runs through the theoretical literature and is broadly consistent with the arguments of Demsetz (1973) discussed above. But this mechanism has received little explicit empirical attention (Temin (1979)). 57 - V. STATUS AND IMPLICATIONS In this final section I offer a brief overall assessment of the state of industrial economics and discuss some implications for research priorities and policy design. Status of the Field Industrial economists have adopted a common theoretical language in recent years and have produced a host of formal models. This work has uncovered a number of general principles, such as the importance of credibility and the consequent value of commitment, that have proven useful in a wide variety of contexts. And our understanding of a number of classic problems, including entry deterrence and cartel stability, has been considerably advanced. But we have also learned two unpleasant features of the gametheoretic approach to the analysis of imperfect competition. First, even apparently simple multi-period games of incomplete information often have multiple (perfect Bayesian Nash) equilibria that can be uncovered only by very sophisticated analysis. The assumption that boundedly rational humans can solve the much more complex games they face in real life seems to push the rationality assumption very far indeed. (Chess is soluble in theory, for instance, but not in practice.) But it is not 29 clear how to replace that assumption. Nor is it clear, despite a great deal of effort devoted to refining the equilibrium concepts discussed in Section I, how to deal in general with models possessing multiple perfect Bayesian Nash equilibria. - 58 Second, the predictions of game- theoretic models seem delicate and are often difficult to test. Important qualitative features of equilibria often depend critically on whether prices or quantities are choice variables, on whether discrete or continuous time is assumed, on whether moves are sequential or simultaneous, and, perhaps most disturbing of all, on how players with incomplete information are assumed to alter their beliefs in response to events that cannot occur in equilibrium. When information is incomplete, strategies depend on unobservable beliefs, and the often empirically questionable assumption that key parameters and probability distributions are common knowledge is central to the analysis. I do not mean at all to suggest that the game- theoretic approach should be scrapped. It can't be wrong in principle to spell out explicitly the details of the situation analyzed and to derive their implications rigorously. And there is simply no attractive alternative approach available. Still, recent theoretical research has taught us much more about what might happen in a variety of market situations than about what must happen conditional on observables. Advances have also been made on the empirical front, particularly in the analysis of individual industries. But, while the empirical research discussed in the preceding sections has uncovered a number of interesting regularities, it has not yet managed substantially to erase the impression that "anything is possible" left by the theoretical literature. Empirical studies in most areas are still concerned with the existence of hypothesized effects rather than with precise estimation of their magnitudes. Debates still rage, for instance, on whether there is any structural relation at all between market concentration and the intensity of rivalry. Industrial - 59 economists can thus speak the same theoretical language and yet disagree sharply as to the empirical relevance of particular theoretical results. Research Strategies Most central questions in industrial organization have by now received considerable game- theoretic attention; the problem is not too little theory but too many different theories. It would appear that research on the theoretical front should be aimed, at least in part, at unification of diverse models and identification of particularly non-robust predictions. Until game- theoretic analysis either begins to yield robust general predictions or is replaced by a mode of theorizing that does so, it seems a fair bet that most major substantive advances in industrial economics will come from empirical research. Only empirical studies can reveal which theoretical models are "empty boxes" and which have wide domains of applicability. And without the discipline provided by a solid base of facts, theorists cannot be expected to concentrate on deducing the implications of empirically interesting assumptions. Much of the most valuable and persuasive empirical research in industrial economics employs carefully-constructed data sets. In many cases these are industry-specific; most industrial economists are more confident about the workings of a few well-studied markets than about markets in general. Still other data sets use interviews or surveys to supplement government statistics or exhibit both time-series and cross-section variation. Since data collection is usually neither intellectually exciting nor highly valued by the economics profession as a whole, progress in 60 industrial organization may depend critically on the availability of financial support for this important activity. Policy Design As I noted at the start of this essay, industrial economists have long been concerned with public policies toward business, and the set of such policies has expanded in recent years . In some domains we are much better able to meet the demand for policy prescriptions than in the past; in others we have mainly learned how little we can confidently assert. On the positive side, the conceptual and empirical tools available for 30 the analysis of individual markets have been considerably improved. The procedures now used by U.S. antitrust authorities to evaluate proposed mergers, for instance, are radically different from and, on balance, much sounder than those used in earlier decades. The quality of economic analysis in individual antitrust cases and in debates about regulatory policies affecting particular industries has risen sharply. On the negative side, recent research has cast doubt on many positive and normative relations that were once widely believed to be generally valid. This makes it harder to speak confidently about policies that apply across the economy. In particular, it now seems clear that the level of seller concentration is at best a poor predictor of the intensity of rivalry, so that simple concentration-based rules that once seemed attractive now have little appeal. Recent theoretical research suggests that market conduct depends in complex ways on a host of factors, and the empirical literature offers few simple, robust structural relations on which general policies can be 61 confidently based. Moreover, formal models of imperfect competition rarely generate unambiguous welfare conclusions. In such models, feasible policy options usually involve movements toward but not Co perfect competition, so that welfare analysis involves second-best comparisons among distorted equilibria. In particular, there is no guarantee that making markets "more competitive" will generally enhance welfare, particularly if non-price rivalry is intensified. Even though it is sometimes painful to recognize the limitations of existing knowledge, it can also be quite exciting. Industrial economics today is an intellectually lively field. And the practical importance of understanding the supply side of the economy is certainly not diminishing. 62 - References Abreu, D. (1986). 'Extremal Equilibria of Oligopolistic Supergames . ' Journal of Economic Theory, vol. 39, pp. 191-225. Arrow, K.J. (1962). 'Economic Welfare and the Allocation of Resources for Inventions.' In The Rate and Direction of Inventive Activity (ed. R.R. Nelson). Princeton: Princeton University Press. Axelrod, R. (1984). The Evolution of Cooperation. New York: Basic Books. Bain, J.S. (1951). 'Relation of Profit Rate to Industry Concentration: American Manufacturing, 1936-1940.' Quarterly Journal of Economics , vol. 65, pp. 293-324. (1956). Barriers to New Competition . Cambridge: Harvard University Press. Baker, J.B. and Bresnahan, T.F. (1985). 'The Gains from Merger or Collusion in Product-Differentiated Industries.' Journal of Industrial Economics , vol. 33, pp. 427-444. Baldwin, J.R. and Gorecki , P.K. (1985). 'The Determinants of Small Plant Market Share in Canadian Manufacturing Industries in the 1970 's.' Review of Economics and Statistics , vol. 67, pp. 156-161. Barton, D.M., and Sherman, R. (1984). 'The Price and Profit Effects of Horizontal Merger: A Case Study.' Journal of Industrial Economics, vol. 33, pp. 165-177. Baumol, W.J., Panzar, J.C., andWillig, R.D. (1982). Contestable Markets and the Theory of Industry Structure . New York: Harcourt Brace Jovanovich. Benham, L. (1972). 'The Effects of Advertising on the Price of Eyeglasses.' Journal of Law and Economics , vol. 15, pp. 337-352. 63 Brander, J. A. and Spencer, B.J. (1985). 'Export Subsidies and International Market Share Rivalry.' Journal of International Economics, vol. 18, pp. 83-100. Branman, L. , Klein, J.D., and Weiss, L.W. (1987). 'The Price Effects of Increased Competition in Auction Markets.' Review of Economics and Statistics , vol. 69, pp. 24-32. Bresnahan, T.F. (1987). 'Competition and Collusion in the American Automobile Industry: The 1955 Price War.' Journal of Industrial Economics , vol. 35, pp. 457-482. Breusch, T.S. (1986). 'Hypothesis Testing in Unidentified Models.' Review of Economic Studies , vol. 53, pp. 635-651. Bulow, J., Geanakoplos, J. and Klemperer, P. (1985). 'Multimarket Oligopoly: Strategic Substitutes and Complements.' Journal of Political Economy, vol. 93, pp. 488-511. Buxton, A.J., Davies, S.W. and Lyons, B.R. (1984). 'Concentration and Advertising in Consumer and Producer Markets.' Journal of Industrial Economics , vol. 32, pp. 451-464. Carlton, D.W. (1979). 'Vertical Integration in Competitive Markets Under Uncertainty.' Journal of Industrial Economics, vol. 27, pp. 189-209. (1986). 'The Rigidity of Prices.' American Economic Review, vol. 76, pp. 637-658. Caves, R.E. (1980). 'Corporate Strategy and Structure.' Journal of Economic Literature , vol. 28, pp. 64-92. (1985). 'International Trade and Industrial Organization: Problems, Solved and Unsolved.' European Economic Review, vol. 28, pp. 377-395. - 64 and Porter, M.E. (1977). 'From Entry Barriers to Mobility Barriers.' Quarterly Journal of Economics, vol. 91, pp. 241-267. Chandler, A.D. (1977). The Visible Hand: The Managerial Revolution in American Business . Cambridge: Harvard University Press. Cohen, W.M. , Levin, R.C., and Mowery, D.C. (1987). 'Firm Size and R&D Intensity: A Re - Examination. ' Journal of Industrial Economics, vol. 35, pp. 543-565. Comanor, W.S. and Wilson, T.A. (1979). 'The Effect of Advertising on Competition: A Survey.' Journal of Economic Literature, vol. 17, pp. 453-476. Cowling, K. and Mueller, D. (1978). 'The Social Costs of Monopoly Power.' Economic Journal , vol. 88, pp. 727-748. Cowling, K. and Waterson, M. (1976). 'Price-Cost Margins and Market Structure.' Economica , vol. 43, pp. 267-274. Cowling, K. , et al (1980). Mergers and Economic Performance . Cambridge: Cambridge University Press. Cubbin, J. (1975). 'Quality Change and Pricing Behaviour in the U.K. Car Industry.' Economica, vol. 42, pp. 45-58. and Geroski , P. A. (1987). 'The Convergence of Profits in the Lonj Run: Inter-firm and Inter- industry Comparisons.' Journal of Industrial Economics , vol. 35, pp. 427-442. Curry, B. and George, K.D. (1983). 'Industrial Concentration: A Survey.' Journal of Industrial Economics , vol. 31, pp. 203-255. Dasgupta, P. and Stiglitz, J.E. (1980). 'Uncertainty, Industrial Structure, and the Speed of R&D.' Bell Journal of Economics, vol. 11, pp. 1-28. - 65 - Davidson, C. and Deneckere , R. (1986). 'Long-Run Competition in Capacity, Short-Run Competition in Price, and the Cournot Model.' Rand Journal of Economics , vol. 17, pp. 404-415. Demsetz, H. (1973). 'Industry Structure, Market Rivalry, and Public Policy.' Journal of Law and Economics , vol. 16, pp. 1-10. (1979). 'Accounting for Advertising as a Barrier to Entry.' Journal of Business , vol. 52, pp. 345-360. Dickens, W.T. and Katz , L.F. (1986). 'Interindustry Wage Differences and Industry Characteristics.' In Unemployment and the Structure of Labour Markets (ed. K. Lang and J. Leonard). Oxford: Basil Blackwell. Dixit, A.K. (1979). 'A Model of Entry Suggesting a Theory of Entry Barriers.' Bell Journal of Economics , vol. 10, pp. 20-32. and Stiglitz, J.E. (1977). 'Monopolistic Competition and Optimum Product Diversity.' American Economic Review, vol. 67, pp. 297-308. and Norman, V. (1978). 'Advertising and Welfare.' Bell Journal of Economics , vol. 9, pp. 1-18. Domowitz, I., Hubbard, R.G., and Petersen, B.C. (1986). 'Business Cycles and the Relationship between Concentration and Price-Cost Margins.' Rand Journal of Economics , vol. 17, pp. 1-17. Dreze, J.H. (1985). 'Uncertainty and the Firm in General Equilibrium Theory.' Economic Journal (Supplement) , vol. 95: 1-20. Dunne, T. , Roberts, M.J., and Samuelson, L. (1987). 'Patterns of Firm Entry and Exit in U.S. Manufacturing Industries.' Mimeographed, Pennsylvania State University. Eckbo , B.E. (1985). 'Mergers and the Market Concentration Doctrine: Evidence from the Capital Market.' Journal of Business , vol. 58, pp. 325-349. - 66 - Evans, D.S. (1987). 'The Relationship Between Firm Growth, Size, and Age: Estimates for 100 Manufacturing Industries.' Journal of Industrial Economics , vol. 85, pp. 657-674. Farrell, J. and Maskin, E. (1987). 'Renegotiation in Repeated Games.' Mimeographed, University of California, Berkeley. Fisher, F.M. and McGowan, J.J. (1983). 'On the Misuse of Accounting Rates of Return to Infer Monopoly Profits.' American Economic Review, vol. 73, pp. 82-97. Friedman, J.W. (1983). 'Advertising and Oligopolistic Equilibrium.' Bell Journal of Economics , vol. 14, pp. 464-473. (1986). Game Theory with Applications to Economics. Oxford: Oxford University Press. Fudenberg, D. and Maskin, E. (1986). 'The Folk Theorem in Repeated Games with Discounting and with Incomplete Information.' Econometrica , vol. 54, pp. 533-554. and Tirole, J. (1984). 'The Fat-Cat Effect, the Puppy-Dog Ploy, and the Lean and Hungry Look.' American Economic Review (Papers and Proceedings) , vol. 74, pp. 361-366. and (1987). 'Understanding Rent Dissipation: On the Use of Game Theory in Industrial Organization. ' American Economic Review (Papers and Proceedings) , vol. 77, pp. 176-183. Geroski, P. A. (1981). 'Specification and Testing the Profits-Concentration Relationship: Some Experiments for the U.K.' Economica , vol. 48, pp. 279-288. (1988). 'In Pursuit of Monopoly Power: Recent Quantitative Work in Industrial Economics.' Journal of Applied Econometrics , forthcoming. - 67 , Ulph, A. and Ulph, D. (1986). 'A Model of the Crude Oil Market in Which Conduct Varies Over Time.' Economic Journal (Supplement) , vol. 97, pp. 77-86. Gilbert, R.J. and Newbery, D. (1982). 'Preemptive Patenting and the Persistence of Monopoly.' American Economic Review, vol. 72, pp. 514- 526. Gisser, M. (1984). 'Price Leadership and Dynamic Aspects of Oligopoly in U.S. Manufacturing.' Journal of Political Economy, vol. 92, pp. 1035- 1048. Green, E.J. and Porter, R.H. (1984). 'Noncooperative Collusion under Imperfect Price Information.' Econometrica , vol. 52, pp. 87-100. Grether, D.M. and Plott, C.R. (1984). 'The Effects of Market Practices in Oligopolistic Markets: An Experimental Examination of the Ethyl Case.' Economic Inquiry, vol. 22, pp. 479-507. Grossman, G. and Shapiro, C. (1984). 'Informative Advertising with Differentiated Products.' Review of Economic Studies, vol. 51, pp. 63- 82. Gul , F. , Sonnenschein, H. and Wilson, R. (1986). 'Foundations of Dynamic Monopoly and the Coase Conjecture.' Journal of Economic Theory, vol. 39, pp. 155-190. Hall, R.E. (1987). 'The Relation between Price and Marginal Cost in U.S. Industry. ' Mimeographed, Hoover Institution. Harrington, J.E., Jr. (1986). 'Limit Pricing when the Potential Entrant is Uncertain of its Cost Function.' Econometrica, vol. 54, pp. 429-437. Harris, M. and Townsend, R. (1981). 'Resource Allocation under Asymmetric Information.' Econometrica, vol 49, pp. 33-64. 68 - Hart, 0. and Holmstrom, B. (1987). 'The Theory of Contracts.' In Advances in Economic Theory, Fifth World Congress (ed. T. Bewley) . Cambridge: Cambridge University Press. Hay, D.A. and Morris, D.J. (1979). Industrial Economics . Oxford: Oxford University Press. Hay, G.A., and Kelley, D. (1974). 'An Empirical Survey of Price-Fixing Conspiracies.' Journal of Law and Economics, vol. 17, pp. 13-38. Helpman, E. and Krugman, P. (1985). Market Structure and Foreign Trade. Cambridge: MIT Press. Iwata, G. (1974). 'Measurement of Conjectural Variations in Oligopoly.' Econometrica , vol. 42, pp. 947-966. Jensen, M.C. and Meckling, W. (1976). 'Theory of the Firm: Managerial Behavior, Agency Costs, and Capital Structure.' Journal of Financial Economics , vol. 3, pp. 305-360. and Ruback, R. (1983). 'The Market for Corporate Control.' Journal of Financial Economics , vol. 11, pp. 5-50. de Jong, H.W. (1986). 'European Industrial Organization: Entrepreneurial Economics in an Organizational Setting. ' In Mainstreams in Industrial Organization , Book I (ed. H.W. de Jong and W.G. Shepherd). Dordrecht: Kluwer . Joskow, P.L. (1987). 'Contract Duration and Relationship-Specific Investments.' American Economic Review, vol. 77, pp. 168-185. and Klevorick, A.K. (1979). 'A Framework for Analyzing Predatory Pricing Policy.' Yale Law Journal, vol. 89, pp. 213-270. and Schmalensee, R. (1986). 'Incentive Regulation for Electric Utilities.' Yale Journal on Regulation, vol. 4, pp. 1-49. 69 Judd, K.L. (1985). 'Credible Spatial Preemption.' Rand Journal of Economics , vol. 16, pp. 153-166. Kamien, M. and Schwartz, N. (1982). Market Structure and Innovation. Cambridge: Cambridge University Press. Krattenmaker , T.G. and Salop, S. (1986). 'Anticompetitive Exclusion: Raising Rivals' Cost to Achieve Power Over Price.' Yale Law Journal, vol. 96, pp. 209-295. Kreps , D.M. and Wilson, R. (1982). 'Reputation and Imperfect Information.' Journal of Economic Theory, vol. 27, pp. 253-279. Krueger, A.B. and Summers, L.H. (1987). 'Reflections on the Inter-Industry Wage Structure.' In Unemployment and the Structure of Labour Markets (ed. K. Lang and J. Leonard). Oxford: Basil Blackwell. Krugman, P.R., ed. (1986). Strategic Trade Policy and the New International Economics . Cambridge: MIT Press. Lau , L. (1982). 'On Identifying the Degree of Competitiveness From Industry Price and Output Data.' Economics Letters, vol. 10, pp. 93-99. Levin, R.C., Klevorick, A.K., Nelson, R.R. , and Winter, S.G. (1987). 'Appropriating the Returns from Industrial R&D.' Mimeographed, Yale University. Lieberman, M.B. (1984). 'The Learning Curve and Pricing in the Chemical Processing Industries.' Rand Journal of Economics, vol. 15, pp. 213- 228. (1987). 'Excess Capacity as a Barrier to Entry: An Empirical Appraisal.' Journal of Industrial Economics, vol. 35, pp. 607-627. Mankiw, N.G. and Whinston, M.D. (1986). 'Free Entry and Social Inefficiency.' Rand Journal of Economics , vol. 17, pp. 48-58. 70 - Maskin, E. and Riley, J. 'Monopoly with Incomplete Information.' Rand Journal of Economics , vol. 15, pp. 171-196. Milgrom, P. and Roberts, J. (1987) 'Informational Asymmetries, Strategic Behavior, and Industrial Organization.' American Economic Review (Papers and Proceedings) , vol. 77, pp. 184-93. Mueller, D. (1986). Profits in the Long Run. Cambridge: Cambridge University Press . Nelson, R.R. and Winter, S.G. (1982). An Evolutionary Theory of Economic Change. Cambridge: Harvard University Press. Nicholls, W. (1951). Price Policies in the Cigarette Industry. Nashville: Vanderbilt University Press. Novshek, W. (1985). 'On the Existence of Cournot Equilibrium.' Review of Economic Studies, vol. 52, pp. 85-98. Oi, W.Y. (1971). 'A Disneyland Dilemma: Two-Part Tariffs for a Mickey-Mouse Monopoly.' Quarterly Journal of Economics, vol. 85, pp. 71-96. Pakes, A. (1987). 'Mueller's Profits in the Long Run.' Rand Journal of Economics , vol. 18, pp. 319-332. Panzar, J.C. and Rosse, J.N. (1987). 'Testing for "Monopoly" Equilibrium.' Journal of Industrial Economics , vol. 35, pp. 443-456. Phlips, L. (1983). The Economics of Price Discrimination. Cambridge: Cambridge University Press. Plott, C.R. (1982). 'Industrial Organization Theory and Experimental Economics.' Journal of Economic Literature , vol. 20, pp. 1485-1527. Porter, M.E. (1976). Interbrand Choice, Strategy, and Bilateral Market Power. Cambridge: Harvard University Press. 71 (1979). 'The Structure Within Industries and Companies' Performance.' Review of Economics and Statistics , vol. 61, pp. 214-227 (1980). Competitive Strategy . New York: Free Press. Pryor, F.L. (1972). 'An International Comparison of Concentration Ratios', Review of Economics and Statistics , vol. 54, pp. 130-140. Ravenscraft, D.J. (1983). 'Structure-Profit Relationships at the Line of Business and Industry Level.' Review of Economics and Statistics , vol. 65, pp. 22-31. and Scherer, F.M. (1988). Mergers, Sell-Offs, and Economic Efficiency . Washington: Brookings Institution. Reinganum, J. (1983). 'Uncertain Innovation and the Persistence of Monopoly.' American Economic Review, vol. 73, pp. 741-748. Rey, P. and Tirole , J. (1986). 'The Logic of Vertical Restraints.' American Economic Review, vol. 76, pp. 921-939. Roberts, D.J. (1987). 'Battles for Market Share.' In Advances in Economic Theory, Fifth World Congress (ed. T. Bewley) . Cambridge: Cambridge University Press. Rose, N.L. (1987). 'Labor Rent-Sharing and Regulation: Evidence from the Trucking Industry.' Journal of Political Economy, vol. 95, pp. 1146- 1178. Rotemberg, J.J. and Saloner, G. (1986). 'A Supergame -Theoretic Model of Business Cycles and Price Wars During Booms.' American Economic Review, vol. 76, pp. 390-407. Saloner, G. (1987). 'Predation, Mergers, and Incomplete Information.' Rand Journal of Economics , vol. 18, pp. 165-186. 72 Salop, S.C. (1979). 'Monopolistic Competition with Outside Goods.' Bell Journal of Economics , vol. 10, pp. 141-156. (1986). 'Practices that (Credibly) Facilitate Oligopoly Coordination.' In New Developments in Che Analysis of Market Structure (ed. G.F. Mathewson and J.E. Stiglitz) . Cambridge: MIT Press. Scherer, F.M. (1980). Industrial Market Structure and Economic Performance , 2nd Ed. Chicago: Rand McNally. , Beckstein, A., Kaufer, E., and Murphy, R.D. (1975). The Economics of Multi-Plant Operation . Cambridge: Harvard University Press . Schmalensee, R. (1978). 'Entry Deterrence in the Ready-to-Eat Breakfast Cereal Industry.' Bell Journal of Economics , vol. 9, pp. 305-327. (1981). 'Economies of Scale and Barriers to Entry.' Journal of Political Economy, vol. 89, pp. 122-138. (1982). 'Product Differentiation Advantages of Pioneering Brands.' American Economic Review, vol. 72, pp. 349-365. (Errata, vol. 73, p. 250.) (1985). 'Do Markets Differ Much?' American Economic Review, vol. 75, pp. 341-351. and Willig, R.D., eds . (1989). Handbook of Industrial Organization . Amsterdam: North-Holland. Schwert, G.W. (1981). 'Using Financial Data to Measure the Effects of Regulation.' Journal of Law and Economics , vol. 24, pp. 121-158. Shaked, A. and Sutton, J. (1983). 'Natural Oligopolies.' Econometrica , vol, 51, pp. 1469-1483. 73 - Shocker, A.D. and Srinivasan, V. (1979). 'Multiattribute Approaches for Product Concept Evaluation and Generation: A Critical Review.' Journal of Marketing Research , vol. 16, pp. 159-180. Smirlock, M. and Marshall, W. (1983). 'Monopoly Power and Expense-Preference Behavior: Theory and Evidence to the Contrary.' Bell Journal of Economics , vol. 14, pp. 166-178. Smith, V.L. (1982). 'Microeconomic Systems as an Experimental Science.' American Economic Review, vol. 72, pp. 923-955. Spence , A.M. (1975). 'Monopoly, Quality, and Welfare.' Bell Journal of Economics , vol. 6, pp. 417-429. (1977). 'Entry, Capacity, Investment, and Oligopolistic Pricing.' Bell Journal of Economics , vol. 8, pp. 534-544. (1980). 'Multi-Product Quantity-Dependent Prices and Profitability Constraints.' Review of Economic Studies, vol. 47, pp. 821-842. (1981). 'The Learning Curve and Competition.' Bell Journal of Economics , vol. 12, pp. 49-70. (1984). 'Cost Reduction, Competition, and Industry Performance.' Econometrica , vol. 52, pp. 101-122. Stigler, G.J. (1964). 'A Theory of Oligopoly.' Journal of Political Economy, vol. 72, pp. 44-61. (1968). The Organization of Industry . Homewood: Irwin. (1971). 'The Theory of Economic Regulation.' Bell Journal of Economics and Management Science, vol. 2, pp. 3-21. Stoneman, P. (1983). The Economic Analysis of Technological Change. Oxford: Oxford University Press. 74 - Telser, L.G. (1960). 'Why Should Manufacturers Want Fair Trade?" Journal of Law and Economics , vol. 3, pp. 86-105. Temin, P. (1979). 'Technology, Regulation, and Market Structure in the Modern Pharmaceutical Industry.' Bell Journal of Economics, vol. 10, pp. 429-446. Tirole, J. (1988). The Theory of Industrial Organization. Cambridge: MIT Press . Varian, H.R. (1985). 'Price Discrimination and Social Welfare.' American Economic Review, vol. 75, pp. 870-875. Vickers , J. (1985). 'Delegation and the Theory of the Firm.' Economic Journal (Supplement) , vol. 95, pp. 138-147. (1986). 'The Evolution of Market Structure When There is a Sequence of Innovations.' Journal of Industrial Economics, vol. 35, pp. 1-12. and Yarrow, G. (1988). Privatization: An Economic Analysis . Cambridge: MIT Press. Williamson, O.E. (1985). The Economic Institutions of Capitalism. New York: Free Press . Willig, R.D. (1978). ' Pareto-Superior Nonlinear Outlay Schedules.' Bell Journal of Economics , vol. 9, pp. 56-69. Yarrow, G. (1985). 'Strategic Issues in Industrial Policy.' Oxford Review of Economic Policy, vol. 1, pp. 95-109. - 75 Footnotes \* I am pleased to acknowledge a great debt to my teachers, colleagues, and students at MIT, who have taught me much of what I know about industrial economics. I am especially indebted for helpful comments on an earlier version of this essay to Timothy Bresnahan, Joseph Farrell, Paul Geroski, Donald Hay, Andrew Oswald, Nancy Rose, Julio Rotemberg, Jean Tirole , John Vickers , and two anonymous referees. As usual, creditors cannot be held liable for defects in the final product. 1. It is conventional in industrial economics to use "rivalry" instead of "competition" when markets may be only imperfectly competitive. 2. Two other limitations of this essay should be noted. First, I confine my attention to the English-language literature. Continental European research tends to have a more Austrian flavor and to stress the importance of disequilibrium behavior and the effects of institutions (de Jong (1986)). Second, because comprehensive bibliographies are available in Schmalensee and Willig (1989) and Tirole (1988) , I have felt free to cite an idiosyncratic selection of classic, neglected, recent, illustrative, and atypical studies, along with some surveys, tilting toward recent writings that discuss earlier contributions. My apologies to those whose important works were thereby omitted from this essay's long but seriously incomplete bibliography. 3. For a detailed exposition of game theory, see Friedman (1986). The discussion of game- theoretic work in this essay has been heavily influenced by Fudenberg and Tirole (1987) and Milgrom and Roberts (1987). 76 6. The alternative normal form or strategic form condenses all this and simply gives payoffs as functions of the players' strategies. The extensive form seems to be more convenient and informative in most applications in industrial economics. 5. Some properties of such equilibria are discussed in Section III. It is worth noting that there exist plausible cost and demand functions that do not yield a unique pure-strategy Cournot equilibrium (Novshek (1985)) 6. In a Bertrand (price-setting) duopoly with differentiated products, for instance, the second mover has an advantage because he can undercut the first mover. Issues of commitment can also arise in the specification of dynamic models. The open- loop equilibrium concept, for instance, assumes that players decide once and for all what moves they will make at each date. Open-loop equilibria thus rest on the assumption that players can commit to ignore to their rivals' subsequent moves. (In closed- loop equilibria, which are generally more plausible, players' strategies consist of functions that map histories into actions (or probability distributions) at each date.) Similarly, discrete- time models involve the assumption that players cannot move within periods, so that period length (as measured by the discount factor) often affects the nature of the equilibrium. 7. If potential agents have hidden knowledge about their differences before the principal makes his hiring decision, the situation is said to involve adverse selection: less able potential agents, with poorer alternatives, may try harder to be hired. Information asymmetries that arise after hiring give rise to moral hazard: if the agent's effort cannot be directly observed, he may have little or no incentive to work 77 - hard. 8. The analysis of agency- theoretic problems is often simplified considerably by invoking the revelation principle (Harris and Townsend (1981)). Suppose that possible agents differ according to the value of some parameter, 6, that the principal cannot observe directly and that affects performance. Then the revelation principle says, roughly, that any optimal scheme in which equilibrium compensation depends on 8 is generally equivalent to a scheme in which the agent is asked to report his 6 to the principal and is given incentives that make it optimal to tell the truth. One can thus limit attention to compensation arrangements of the latter form. 9. Since oligopoly theory deals with the relation between price and marginal cost, not with the rate of return on investment, it can be argued that the best performance measure would be the Lerner index, (price - marginal cost)/price. This argument has led to the use of the so-called price-cost margin, (revenue - labor and materials cost)/revenue , in some studies, but there is no reason to think that marginal cost is accurately measured by unit labor and materials cost. Moreover, rates of return on investment, not price-cost margins, should be equalized under competition. 10. Some authors have used the ratio of a firm's market value to the estimated replacement cost of its assets -- Tobin's q -- to measure profitability. This approach does not avoid accounting problems, of course, since replacement cost estimates must be based on accounting data. - 7! 11. This is the most natural and common generalization, but it ignores two problems. When capital markets are incomplete, shareholders will not necessarily vote unanimously for wealth maximization --or for anything else (Dreze (1985)). And in strategic settings, owners' interests may be best served by managers who do not aim to maximize profit (Vickers (1985)). If the owners of a monopoly want to deter entry, for instance, they might want to hand control of the firm over to irrational managers who would be willing to incur any losses necessary to drive any entrants out of business -- as long as these managers can convince all potential entrants of their irrationality and job security. 12. There is an obvious tension between these observations and the extreme rationality assumed in many game- theoretic models. 13. Curry and George (1983) provide a useful survey of the literature on measures and determinants of seller concentration. 14 . In the U.S. data, which have been most intensively studied in this respect, it appears that the variance in year-to-year firm growth rates declines with firm size, while the mean growth rate declines somewhat with both size and firm age (Evans (1987)). 15. Stigler (1968) offered an alternative definition: costs that must be borne by an entrant that were not borne by an incumbent. The main difference is that scale economies cannot constitute an entry barrier according to Stigler. The related concept of mobility barriers, obstacles to mimicking other firms' strategies, is often useful in industry analysis (Caves and Porter (1977)). 79 - 16. Most models of entry deterrence assume a single established firm. In oligopolies, incentives to over- invest to deter entry are reduced because deterrence is a public good but increased because investment tends to raise pre-entry profits if price is above cost. In some models the second effect dominates, and oligopolies facing potential entrants invest more (and deter entry more effectively) than a monopoly would. 17. Some recent work has treated cost advantages as endogenous, stressing the ability of firms under some conditions to advantage themselves, and possibly induce exit, by actions in input markets that differentially raise rivals' costs (Krattenmaker and Salop (1986)). 18. Most proposed policy rules for evaluating charges of predation employ tests based on market structure and cost-price relations (Joskow and Klevorick (1979)). Such rules lack formal welfare- theoretic rationales and are not well-suited for handling the sort of predation discussed in the next paragraph, though they do serve in practice to dispose of many groundless cases brought by high-cost producers. 19. Pakes (1987) and Geroski (1988) provide useful discussions of recent empirical work on the topics considered here. 20. Generalized versions of this equation appear in Cowling and Waterson (1976) and a number of later studies. 21. It is plausible to suppose that high buyer concentration would tend to reduce the effect of seller concentration on profitability. The few empirical tests of this countervailing power hypothesis have produced rather mixed results, however. - 80 - 22. The argument underlying this assertion is spelled out in my chapter in Schmalensee and Willig (1989). 23. On the other hand, it is worth noting that accounting profitability differences among large firms, as well as large firms' market shares, seem to persist for long periods (Mueller (1986)). 24. A related literature studies situations in which buyers can verify quality only by use, and sellers have reputations for quality. High quality products will then be priced above cost in equilibrium and yield a flow of excess profits. If not, firms with reputations for high quality will prefer to exploit them by secretly lowering quality (and thus cost) and selling at the same price as high quality products until buyers catch on. 25. Comanor and Wilson (1979) survey much of the literature on the arguments discussed in this paragraph. These arguments are also relevant to the positive correlation between profitability and research and development intensity reported in several studies. 26. There are also interesting recent studies of technology adoption, particularly in the presence of network externalities (which imply that the value of a technology to any one user increases with the number of users) , and on the strategic uses of patent licensing. 27 . The issues discussed in this paragraph also arise in connection with the acquisition of natural resource deposits. 28. Surveys of the empirical literature are provided by Kamien and Schwartz (1982) and Stoneman (1983). 51 - 29. Note that learning arguments have very little appeal here, since allowing for the possibility of rational learning requires formulating a new and more complex game. For an interesting alternative (evolutionary) approach to this class of problems, see Axelrod (1984). 30. These developments have also made industrial economists better able to provide useful strategic advice to business decision-makers, and the academic and commercial markets have generally reacted rationally. J / V - U 6 k Date Due tt« •68 r\ 'Q; BASEMENT B MY06189 NOV 1 6,gg, jUU. j 5 1994\* Lib-26-67 MIT LIBRARIES 3 TDfiD DD5 37b b2b

NBER WORKING PAPER SERIES PERSONNEL ECONOMICS Edward P. Lazear Paul Oyer Working Paper 13480 http://www.nber.org/papers/w13480 NATIONAL BUREAU OF ECONOMIC RESEARCH 1050 Massachusetts Avenue Cambridge, MA 02138 October 2007 We thank the editors, James Malcomson, Bentley MacLeod, Scott Schaefer, Jan Zabojnik, and participants in the Handbook conference at Stanford for comments. This is a draft of a chapter to appear in the forthcoming Handbook of Organizational Economics, edited by Robert Gibbons and D. John Roberts and published by Princeton University Press. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research. © 2007 by Edward P. Lazear and Paul Oyer. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source. Personnel Economics Edward P. Lazear and Paul Oyer NBER Working Paper No. 13480 October 2007 JEL No. J3,J41,M5 ABSTRACT In this review of the personnel economics literature, we introduce key topics of personnel economics, focus on some relatively new findings that have emerged since prior reviews of some or all of the personnel economics literature, and suggest open questions in personnel economics where future research can make valuable contributions to the literature. We explore five aspects of the employment relationship - incentives, matching firms with workers, compensation, skill development, and the organization of work - reviewing the main theories, empirical tests of those theories, and the open questions in each area. Edward P. Lazear Graduate School of Business Stanford University Stanford, CA 94305-5015 and NBER lazear@gsb.stanford.edu Paul Oyer Graduate School of Business Stanford University 518 Memorial Way Stanford, CA 94305-5015 and NBER pauloyer@stanford.edu 1 1. Introduction Personnel economics is the application of economic and mathematical approaches to traditional topics in the study of human resource management. This includes topics such as compensation, turnover, and incentives that are inherently economic, as well as those that do not at first appear to be economic topics (such as norms, teamwork, worker empowerment, and peer relationships.) Using the tools from advances in game theory, information economics, econometrics, and other areas of economics, personnel economics has come a long way over the last few decades. It now produces a large share of the labor economics literature, has earned its own code in the Journal of Economic Literature classification system (M5), and has its own working group within the National Bureau of Economic Research. In this paper, we review how this literature has developed and discuss some avenues for fruitful future research. Personnel economics provides both positive and normative analysis of human resources practices and methods. On the positive/descriptive side, we study a range of questions. When do firms choose to use one form of compensation over another? When are teams important? How do firms find the right workers? When are certain benefits or stock grants given to workers? The list extends to any decision an employer has to make with respect to its dealings with employees. That is, personnel economics attempts to describe how human resource practices can best address an employer’s goals, subject to the constraint that employee actions will be affected by those practices. Perhaps because the subject was taken up by business school economists whose job it is to teach managers what to do, personnel economics has not shied away from being somewhat prescriptive. Several MBA textbooks and numerous MBA classes that are either focused on personnel economics or cover it as part of a broader subject deliver research findings to current and future managers.1 In many ways, the basic outlines and topics of this chapter are similar to the syllabi for the MBA classes we teach at Stanford (though the details are surely very different.) Given the importance of human resources to many organizations’ success and the fact that labor accounts for the majority of business costs, bringing research findings to managers can have an important effect. 1 A recent paper by Bloom and Van Reenan (2007) suggests that managers can benefit from using “modern” managerial practices, including some specific practices studied in personnel economics. 2 In this review, we will introduce the key topics of personnel economics. We will also focus on some relatively new findings that have emerged since prior reviews of some or all of the personnel economics literature, such as Gibbons and Waldman (1999), Malcomson (1999), Prendergast (1999), Murphy (1999), and Lazear (1999, 2000b). Throughout the discussion and in the conclusion, we will suggest open questions where future research can make valuable contributions to the literature. Personnel economics research has focused on five aspects of employment relationships – incentives, matching firms with workers, compensation, skill development, and the organization of work. Though some research focuses on only one of these, much of the literature touches on multiple parts of jobs. For example, incentives have an effect on which workers accept positions and the organization of the workplace depends on the skills a firm develops (or selects for). Nonetheless, these five areas of employment relationships allow for a broad categorization of research in personnel economics. So we explore each of these in turn, reviewing the main theories, empirical tests of those theories, and the open questions in each area. Before going on, it is worth briefly noting what personal economics is not. Personnel economics is a branch of labor economics, but there is a great deal of labor economics that is not personnel economics. Studies that do not consider interactions between a firm and workers or do not explicitly or implicitly consider a firm trying to maximize some objective (usually profits) are not personnel economics. Many findings in labor economics are important conceptual or empirical inputs to, but not part of, personnel economics. For example, labor economists estimate Mincer-ian earnings functions of the returns to experience and the return to schooling and they estimate labor supply functions. While these equilibrium outcomes are very important issues to firms designing personnel policies, they are not strategic choices of the firm. Similarly, most personnel economics is not policy-oriented. Personnel economics models typically focus on welfare within a given employment relationship rather than on overall social welfare function. In the absence of a market failure, firms’ best interests will be the same as those of the economy as a whole. While there are some exceptions (for example, policy and personnel economics influence one another in the study of workplace discrimination), the main contributions of personnel economics to social welfare are in research advances and helping managers to run their businesses more efficiently rather than in influencing policymakers’ decisions. 3 2. Incentives Firms and employees naturally have opposing interests in that employee effort typically leads to benefits to the firm and costs to the employee. However, there can be gains from trade if the value to employers is enough to compensate workers for the cost of their effort. Encouraging employee effort is a central issue in personnel economics and, as with much of economics, has led to a large literature on incentives. When effort is contractible, providing efficient incentives is trivial. But, in the more realistic case where there is hidden action (moral hazard), one of the keys to successful personnel practices is to design incentives for employees. When effort is not contractible, one potential way (though not the only way) to get the socially efficient level of worker effort is to pay the worker the full value of output – that is, sell the firm to the worker. The firm could set a base salary to meet the worker’s participation constraint and let the worker keep all the marginal fruits of his effort. The base salary will need to be negative – that is, the worker will pay the firm for the right to work there – in order for the firm to find the employment arrangement profitable. In other words, moral hazard problems can be solved by making employees the residual claimants on their effort. If individuals own all their output, they will efficiently match the marginal benefits of their effort to the marginal cost. There are employment arrangements that involve selling the firm to the worker (such as taxi drivers who rent cabs for a shift and keep all fares), but these jobs are exceptions. Why does this simple model fail for the vast majority of workers who have positive base wages and piece rates of less than one (and usually zero)? There are many reasons, but they generally relate to the facts that worker output (that is, measures of the worker’s productivity) is usually imperfectly related to inputs (that is, the worker’s effort), that firms cannot always credibly commit to reward effort ex post, and that individuals are typically more risk averse than employers. 2.1.1 The Trade-Off Between Risk and Incentives 4 Most theoretical principal-agent models assume that output is an increasing function of effort and it is also influenced by some unobservable random shock.2 The employee’s utility is decreasing and concave in effort (that is, his disutility of effort is positive and convex.) Because he is risk averse, utility is increasing and concave in income. The complication, which makes it impossible to simply pay the agent for his inputs, is that the principal can only observe the output measure, so this is the only variable upon which compensation contracts can be based. If the two parties could write enforceable contracts based on effort, the problem would be simple. The risk neutral firm would simply pay for effort and the agent would be completely insured. The problem for the principal is to set the compensation scheme to maximize profit subject to incentive compatibility and participation constraints. A key implication is that there is now a tradeoff between insurance and incentives. Because workers do not like risk, the firm must dampen the relation of wages to output. If compensation is based largely on output, the firm may have to provide substantial risk compensation. But if compensation is not based on output, workers will put forth little effort. The second-best solution, though the “best” compromise, forces the worker to bear risk and provides the firm with an inefficient amount of effort. This model has two key predictions. First, the stronger the incentives, the harder the employee will work. As we will discuss below, there is ample empirical evidence to suggest that this relationship holds in most employment relationships. Second, there is a trade-off between risk and incentives. Everything else equal, in settings where factors beyond the employee’s control have a relatively large effect on output, incentives will be weaker. As we also discuss below, probably at least partially due to an inability to make “everything else equal”, the empirical evidence on this is decidedly mixed. 2.1.2 Distortions in Performance Measures Imperfections with performance measurement can lead firms to limit the use of incentives even when employees are risk neutral. In addition to exogenous shocks to output measures, the usefulness of a performance measure may also be limited by employees having private 2 See Prendergast (1999) for a somewhat more detailed description of moral hazard theory. See Johnson (1950) and Cheung (1969) for important early compensation analyses from the agricultural economics literature and Ross (1973), Mirrlees (1974), Stiglitz (1975), Bergson (1978), and Holmstrom (1979) for early modern treatments of the moral hazard problem. 5 information or their ability to influence the measure without actually advancing the principal’s goals. Gibbons (1987) considers the case where only the worker knows the difficulty of the job and only the worker knows his true action. Under these circumstances, Gibbons shows that workers will restrict output (that is, there will be a “ratchet effect”) when the firm cannot commit not to use information it learns about the difficulty of the task. He argues that this contributed to the decline of piece rate systems which were extensively used at the end of the 19th century. Another class of models considers the implications of employees that can affect the output measure differently from their effect on actual output. Lazear (1986a) briefly discussed the implications for incentives and performance measurement when workers can increase quantity at the expense of quality. Baker (1992) provides a more general model of a firm that wants to maximize some non-contractible objective (call it V). The firm can write compensation contracts based on P, which is imperfectly correlated with V. The employee has private information about how his actions affect both P and V. The firm wants to provide incentives so that the employee will exert more effort, but has to trade this off against the fact that the employee may exert effort when that effort has a large effect on P but not V. That is, the firm has to worry about the fact that incentives drive efficient effort and “gaming” at the same time. Holmstrom and Milgrom (1991) explore the limits of measurement from a slightly different perspective. They derive a “multi-tasking” model where an employee engages in two tasks that affect actual output and measured output differently. Incentive contracts may drive the employee to under-invest or over-invest in one or both tasks. They discuss how this may explain the limits of incentive pay for teachers. While incentives may lead teachers to work harder, they may also focus their effort on teaching rote skills that are rewarded on standardized tests at the expense of teaching logical reasoning and other skills where the payoff is not easily measured in the short-term.3 As we discuss below, these models are generally consistent with several empirical studies showing unintended responses to incentives.4 2.1.3 Subjective Measures of Performance 3 Lazear (2006) lays out how the performance measures should be designed and how much information about them should be given to employees when employees engage in multiple tasks. 4 One case study that shows direct evidence of the importance of multi-task issues is Brickley and Zimmerman (2001). They show that teacher ratings rose and research output fell at the University of Rochester’s Simon School of Business after the administration increased the relative importance of teaching in its reward structure. 6 Given the limits of objective performance measures, what alternatives do firms have for motivating workers? The answer for most employees is various incentives based on subjective measures of performance. That is, the performance of most employees is monitored by some supervisor(s) that gets a reasonably accurate, but non-verifiable, signal of how good a job that employee does. Good performance can then be rewarded through a variety of mechanisms, including increases in base salary, subjectively determined bonus payments, or promotions. Baker, Gibbons, and Murphy (1994), Bull (1987), and MacLeod and Malcomson (1989), model the use of subjective performance evaluation. They analyze the optimal mix of incentives based on imperfect objective measures and perfect (but unverifiable) subjective measures. These models highlight the trade-off between strong incentives from well-designed subjective measures versus the fact that employees have to trust firms to follow through on the inherently implicit contracts involved in any non-verifiable performance measure. For example, many Wall Street firms distribute year-end bonuses to employees based on largely subjective measures of the individuals’ contributions. The shareholders of those firms would gain in the short-term by announcing that nobody had earned any bonus in a given year, but the firm would find it hard to motivate employees in the future. As the pioneering career concerns models of Fama (1980) and Holmstrom (1999) discuss, immediate pay-for-performance is not necessary if current performance affects an employee’s reputation and future compensation.5 Gibbons and Murphy (1992), building on these models, analyze employees that exert effort both in response to explicit incentives and in the hopes of improving employers’ beliefs about their ability. Firms do not need to provide strong current-period incentives for young workers, who have more reason to be motivated by future promotions and opportunities. They find some support for the model in the way Chief Executive Officer compensation varies with age and experience. Chevalier and Ellison (1999), who study mutual fund managers, find further support for age-based variation in the need for objective incentives. In addition to the limits that come with using implicit contracts for enforcement, subjective performance evaluation also is limited by differences between actual and measured performance. There are interesting models of at least three ways that subjective assessments can 5 See Prendergast (1999) and Gibbons and Waldman (1999) for more detail on career concerns models. 7 differ from actual productivity. First, employees can take actions that affect their supervisors’ assessments. This issue has been studied in a series of papers on “influence activity” starting with Milgrom and Roberts (1988). The key idea is that employees will try to influence their bosses’ decisions in their favor (and not necessarily in the interests of the firm) if their pay does not properly reflect their productivity. Firms face a trade-off between taking advantage of the manager’s information and letting the manager manipulate decisions in her own best interests. For decisions that do not affect the distribution of benefits within the firm, this presents no problem. But when decisions have distributional implications, the firm may want to commit not to use information provided by agents. Meyer, Milgrom, and Roberts (1992) and Schaefer (1998) show how the firm can benefit from a crisis (that is, very bad alternatives for agents) in order to reduce influence activity and align the interests of employees and the firm.6 Second, Prendergast and Topel (1996) consider a manager that favors some workers over others. That is, the manager affects the payoffs of the employees and the distribution of employee payoffs, in turn, affects the utility of the manager. The firm has to trade off the informational advantage of the manager regarding the employees’ true performance, the taste for the manager to have more discretion, and the distortion to incentives caused by favoritism. Favoritism leads the firm to reduce incentives and to rely on imperfect (but non-corruptible) objective performance measures. Finally, MacLeod (2003) considers the optimal contract when the manager and employee have different opinions about the employee’s performance. He shows how the optimal implicit contract will vary with the degree of correlation between the manager’s assessment of the employee’s performance and the employee’s self-assessment. The model predicts the tendency of managers to bunch or compress employee’s ratings. MacLeod (2003) also studies the effects of favoritism, allowing for a more general set of contracts than Prendergast and Topel (1996) study. He shows that favoritism reduces the incentives of employees that are discriminated against, leading pay to suffer both due to discrimination and lower effort. This has the interesting implication that, holding performance constant, measured differences in wages will understate the importance of manager bias because it will ignore the effect on the employees’ incentives. 6 Also see Fairburn and Malcomson (2001), who develop an integrated model of incentives and promotions when managers can engage in influence activity. They argue that managers with sufficient monetary incentives will be resistant to influence activity when making promotion decisions, which both makes promotion-based incentives stronger and more efficiently matches workers to appropriate jobs. 8 These models of subjective performance evaluation and implicit contracts have made important contributions to personnel economics and our understanding of compensation structure. However, while these models are consistent with anecdotal examples and case studies, they have not been backed up by a great deal of rigorous empirical corroboration. Hayes and Schaefer (2000) provide compelling (though indirect) evidence on the importance of implicit contracts and subjective performance evaluation of executives. They argue that a board of directors that uses subjective evaluation will reward executives in the current period for actions that will affect objective measures of performance in future periods. They then show that there is an important empirical relationship between current pay and future performance and that this relationship is stronger at firms where they would expect subjective performance measures to be relatively advantageous for providing efficient incentives. 2.1.4 Relative Performance Evaluation Some performance measures, whether objective or subjective, can be improved by making them depend on an employee’s relative performance rather than some absolute standard. Relative performance evaluation (RPE) can be used in two different ways. First, firms can use the performance of a peer group of competitors to filter out shocks that are common to the whole peer group. This helps the firm to lower the risk (and the associated compensation premium) imposed on individual employees. As Gibbons and Murphy (1991) and Antle and Smith (1986) discuss, this idea has the empirical implication that pay will be increasing in the employee’s own performance but decreasing in the average performance of the reference group. As we discuss below in Section 4.3, at least in the case of Chief Executive Officers (where their performance and that of a peer group are easily observed), there is surprisingly little support for filtering out common shocks of a peer group. A second form of RPE is to use fixed rewards for a fixed group of people where the distribution of rewards is based on the order of the participants’ performance. This form of reward system, generally known as a tournament, can vary from a group of people competing for a single promotion (such as to CEO) to a large set of prizes that diminish in value as a person’s relative performance gets lower (such as at a golf tournament or in the memorable real estate sales tournament in the drama Glengarry Glen Ross.) 9 Tournament theory, as modeled by Lazear and Rosen (1981), models promotions as a relative game. The compensation at one level of the firm, in addition to motivating individuals at that level, motivates those at lower levels. There are three basic principles of tournament theory. First, prizes are fixed in advance and depend on relative rather than absolute performance. Second, larger spreads in wages at different levels of the hierarchy motivate those at lower levels to put forth more effort. Third, there is an optimal spread. Although a greater spread increases effort, at some point the additional wages necessary to compensate workers for the increased effort is larger than the additional output generated. An important variable in the Lazear and Rosen (1981) model is the amount of noise – that is, to what degree luck affects the probability of winning. When there is more noise (so that luck becomes relatively more important and effort relatively less important), workers will try less hard to win because effort has a smaller effect on whether or not they win. In production environments that are very uncertain, large raises must be given in order to offset the tendency by workers to reduce effort. Similarly, the number of slots available affects effort. If a firm has 100 vice-presidents and only one slot for president, most VPs will give up trying to become president thinking that the chance of promotion is too slim. At the other extreme, if there were 100 VP slots and 100 president slots so that every VP knew that she would be promoted to president, VPs would become complacent. The firm can use the size of the spread between pay at various levels to manage effort. When luck is unimportant, the wage spread will be small. But when luck is important, the firm needs to increase the spread to increase effort. This may help explain differences in salary structure across countries or across industries. Lazear and Rosen’s (1981) model implies that riskier industries (where this is interpreted as risk that affects individual output) should have larger wage spreads than less risky industries to induce workers to put forth the appropriate amount of effort.7 These basic ideas of tournament theory have been extended in numerous ways. For example, while Lazear and Rosen (1981) focus on relative performance strictly as a means of generating incentives, Green and Stokey (1983) highlight the potential usefulness of tournaments in removing common shocks from risky performance measures. They show how individual incentive contracts compare to tournaments based on variation in how common the shocks to 7 This prediction also comes from incentive models where agents are risk averse. 10 agent output are, how easily the firm can observe these shocks, and how many employees participate in the tournament. Nalebuff and Stiglitz (1983) allow for risk-averse employees, allow the number of participants in the contest to vary, and allow for both rewards and penalties. They derive a large set of results regarding when tournaments will be more efficient than incentive schemes that are not based on relative performance and how the optimal incentive system will vary with the number of participants, environmental uncertainty, and other factors. Dye (1984) and Lazear (1989) consider how the potential for collusion, sabotage, or other forms of non-cooperative behavior in tournaments counter the incentive value of a large spread in rewards. Chan (1995) considers how firms should balance the incentive value of internal promotion tournaments against the value of retaining the option to hire an outsider when a senior opening occurs. In the basic Lazear and Rosen (1981) framework, firms can consider outsiders and keep tournament incentives strong by increasing the rewards for being promoted. However, this can cause problems such as influence activity or sabotage. Chan (1995) shows that a potentially useful alternative is to give insiders a handicap in the tournament. Patterns of pay and promotions within firms and in other contest contexts fit the predictions of tournament theory. Devaro (2006) uses a sample of recent hires and their initial promotions to fit a structural model of tournaments. He shows the importance of relative, rather than absolute performance in determining promotions. Knoeber and Thurman (1994) provide empirical evidence consistent with several specific predictions of tournament theory in the context of rewards for producers of broiler chickens. They show, for example, that spreads between prize values affect output. Drago and Garvey (1998), using a survey of Australian firms, show that individuals are less helpful and work harder when promotion incentives are strong. Ehrenberg and Bognanno (1990) show that golfers are affected by prize spreads in tournaments. Finally, Bull, Schotter, and Weigelt (1987) do laboratory experiments on how people respond to contests with piece rates and tournaments, so “effort” is chosen and stated. The results of their experiments are generally supportive of tournament theory, though they find that less able people expend more effort than the theory would predict. 2.1.5 Alternative to Incentives I – Monitoring 11 An alternative to financial incentives is to simply monitor workers. If a supervisor can keep close watch over employees, she can insure that the employee takes the best action. However, monitoring is typically imperfect, which led to the idea of efficiency wages. Shapiro and Stiglitz (1984) suggest that a firm will pay workers a wage premium when it cannot perfectly monitor effort. Workers, facing a choice between earning this wage premium if they work hard and facing a probabilistic chance of being caught and fired if they shirk, choose to work hard. Efficiency wage theory implies a negative relationship between monitoring intensity and wages.8 The empirical evidence in favor of this prediction is mixed, at best.9 This may be because the conditions that make efficiency wages optimal for individual employers may not exist at a large subset of firms in the economy. That is, there could well be many firms that can monitor very effectively or use output-based incentives, while other firms (such as the one studied by Cappelli and Chauvin, 1991) use efficiency wages. A recent study by Nagin, Rebitzer, Sanders, and Taylor (2002), while not directly testing the relevance of efficiency wage theory, finds empirical support for the underlying relationship between employee productivity and the probability of catching employees that do not act in the firm’s best interests. They run a field experiment at a telephone solicitation company where employees spend their workdays calling people to ask for donations to charitable organizations. They find that the rate at which workers “cheat” by trying to get paid for donations people did not actually make varies in the opposite direction of the rate at which managers audit donations. When the probability of getting caught goes up, workers cheat less.10 However, based on a survey of the telephone solicitors, Nagin et al (2002) conclude there is little evidence to support another key idea in efficiency wage models – that those workers who value a job the most will react the most to monitoring. 2.1.6 Alternative to Incentives II – Intrinsic Rewards 8 The monitoring/pay relationship is the primary implication for the personnel economics literature. But Shapiro and Stiglitz (1984) actually focus on the fact that efficiency wages also lead to equilibrium unemployment. Bulow and Summers (1986) also model equilibrium efficiency wages and focus on equilibrium unemployment. They discuss how the reaction to Ford Motor Company’s $5 day for assembly line workers, introduced in 1914, is consistent with efficiency wage theory. 9 See Malcomson (1999) and Gibbons and Waldman (1999) for further discussion of the empirical relevance of efficiency wage theory. 10 Duflo and Hanna (2005) show that monitoring works in increasing teacher attendance in India. In their case, the employees do not have to infer the monitoring rate. 12 One way to save on both the costs of incentives and the costs of monitoring is to rely on the intrinsic value workers get from doing their jobs well. If firms pick workers carefully and create the right work environment, won’t their employees be productive? In other words, standard moral hazard models assume that all productive effort is distasteful to the employee. But perhaps this is not a reasonable assumption. All else equal, people who are intrinsically motivated to do certain types of jobs (or, put another way, those that simply enjoy certain jobs) will go work for firms that let them do those jobs.11 If the labor market is reasonably competitive, then no matter how much workers enjoy the job, the firm will still have to pay them to do it because employers will bid wages up to the person’s marginal product of labor. But, if people like their job enough, why would firms ever need to provide incentives or monitor workers? The answer to this, and the limit to intrinsic motivation, is not that people are not intrinsically motivated or that they don’t like their jobs (though this is surely true in many cases.) Rather, it’s that the efficient outcome in most employment relationships is for the employee to dislike her job on the margin. Put another way, the key assumption in moral hazard models is not that all effort is distasteful. The key assumption is that the marginal disutility of effort is convex. Many employees are intrinsically motivated to expend at least some effort at their jobs. However, at some point between zero and twenty-four hours of work in a given day, the marginal cost of an additional hour of work becomes positive. If a somewhat intrinsically motivated worker had no incentives or monitoring, she would do some work but would stop at some point. Assuming the person is still productive at the effort level where marginal work becomes costly, it is efficient to induce her to work more. It makes sense for her to work until the marginal disutility of her effort hits the marginal value to the firm. This in no way implies that intrinsic motivation is not important. In fact, as our discussion of selection and matching below will make clear, firms should work hard to find workers who are motivated so as to decrease their own costs of compensation. But it does suggest that, on the margin, agency theory and other economic principles about motivation and incentives apply even when workers “like” their jobs. 11 See Prendergast (2007) for a model of how different forms of intrinsic motivation should affect the matching of people to employers. 13 While this very simple discussion of intrinsic motivation gets at the key economic issue regarding intrinsic motivation, it misses some potentially important and subtle effects having to do with the context of intrinsic motivation. Firms may affect employees’ intrinsic rewards with the signals they send. Benabou and Tirole (2003) formalize this idea and argue that a traditional economic model can be reconciled with many of the findings by psychologists about the demotivating effects of incentives.12 They set up a model where both the employee and the firm have private information. The employee has better information about his abilities and his own interests, while the firm has better information about the attractiveness of the task. The contract offered by the firm affects the employee’s beliefs about the task or the firm’s beliefs regarding the employee. Under certain conditions, incentives can have negative effects on the employee. 2.2 Responses to Incentives – Empirical Studies In a variety of situations where economists have been able to study the effects of incentives in relatively controlled settings, employees respond to incentives. However, while incentives “work”, several studies have shown some of the unintended consequences that are created by employee incentives. That is, the “distortions” predicted by some of the theoretical work appear to be important empirically. Some of these studies, because they focus on individual firms, could be capturing mistakes. Others, which show systematic and long-term costs of incentive schemes, suggest that incentives cannot be implemented perfectly but that the costs are worth bearing. In this section, we review the empirical work on incentives. We first review several studies that establish that employees respond to incentives. We then discuss studies that show the importance of distortions in incentive schemes. There are several important empirical papers that consider how incentives operate in group settings. We defer the discussion of these studies to Section 6, where we discuss the organization of work. Lazear’s (2000a) study of the Safelite Glass Company provides a case study of the effect of incentives on both effort and worker selection. Safelite switched from paying windshield glass installers an hourly wage to paying them a piece rate per windshield installed. Because Safelite implemented the new pay scheme at different times at different locations, Lazear (2000a) could 12 For a review of many of these papers, see Deci, Koestner, and Ryan (1999). 14 isolate the effect of the pay scheme while controlling for other company-wide changes. Also, because he was able to observe many new workers under both pay schemes and could control for on-the-job learning (that is, tenure effects), he could separate the degree to which any change in productivity was due to changes in individual worker behavior and to changes in the types of workers that Safelite was able to attract. He finds that the piece rate system increased productivity at Safelite by about 44%. About half of this was the pure productivity effect. That is, the typical windshield installer that worked at Safelite under both pay schemes increased his productivity by about 22% upon the implementation of the piece rate scheme. The rest of the increase in productivity was due to the fact that the piece rate scheme led to self-selection of more productive workers into Safelite (and the self-selection of less productive employees out.) Shearer (2004) studies how piece rates affect the productivity of tree planters in British Columbia. Working in cooperation with the company that employs the tree planters, Shearer was able to implement a true experiment where a treatment group was randomly assigned to be paid a piece rate while a control group was paid an hourly wage. This means he cannot estimate a selection effect of piece rates, but insures the validity of a causal interpretation of the incentive effect that he measures. Shearer (2004) finds that workers paid through a piece rate are approximately 20% more productive than those paid by the hour, which is nearly identical to the effect in Lazear (2000a). Bandiera, Barankay, and Rasul (2007) consider how the implementation of an incentive pay system affects the productivity of managers of fruit pickers on a UK farm. The pickers themselves are on a piece rate system at all times, but the farm made an unannounced switch to incentive pay for managers in the middle of one season. Bandiera, Barankay, and Rasul (2007) show that the average picker’s productivity increased by 21% when his manager’s pay was related to his productivity. They go on to show that much of this increase is due to the managers focusing their effort more carefully on those pickers where it will have the largest marginal effect on productivity, but that at least half is due to the managers being more discriminating in which workers they select.13 Freeman and Kleiner (2005) provide a useful reminder that, while incentive pay generally increases productivity, it does not necessarily increase profits. They study a footwear 13 Other studies that show evidence that incentives affect employee behavior, but where the exact “incentive effect” is not as easily interpreted, include Gaynor, Rebitzer, and Taylor (2004), Groves, Hong, McMillan, and Naughton (1994), and Fernie and Metcalf (1999). 15 manufacturer that switched from piece rate pay to an hourly wage as part of an attempt to cut costs and change its manufacturing processes. The firm implemented numerous other changes at the time of the change in compensation system, so it is difficult to isolate the exact effect of the change in pay scheme on productivity. But Freeman and Kleiner (2005) do find a significant reduction in productivity, and an increase in profits, when piece rates were removed.14 These studies all show that incentives can be a powerful managerial tool for affecting individuals’ behavior. Asch (1990) shows the same thing, but highlights the more problematic side of incentives. She studies Unites States Navy recruiters who were measured, and in some cases paid, based on their ability to enlist sailors. To the extent that productivity affected compensation for the recruiters, it was based on whether or not they met annual quotas. Therefore, recruiter incentives were low if they either had already reached their quotas or if it became apparent that they would not reach them. Also, due to discounting, the return to recruiter effort grew as the end of the measurement period approached. Asch (1990) shows that recruiting success was higher near the end of the year as the end of the measurement cycle approached. So, while incentives “worked”, they did not work consistently. It is difficult to determine whether this inconsistency had any negative ramifications for the Navy. Courty and Marschke (2004) examine the response of another set of government workers – managers of job training centers – to non-linear incentives. However, in this case the incentives are indirect in that performance affects the budget of the job training office and not the employees’ compensation. Courty and Marschke (2004) show that managers act so as to increase the expected amount of incentive payouts near the end of each measurement period, but that their actions generate real costs in terms of lowering the quality of overall training. The incentives may lead to benefits in terms of overall quality, but it seems likely that these benefits could be captured without the costs imposed by the focus on year-end results. The analyses by Asch (1990) and Courty and Marschke (2004) suggest that incentives can have a costly and inefficient side to them. However, because these studies are each based on a single institution (and one that is not subject to competition), it is feasible that they are simply studies of mistakes. Over time and facing competition, perhaps the Navy and the job training incentive systems would get changed or driven out of existence. However, there are broader 14 Note, however, that this profit effect could reverse in the long-term if employees become more productive in the future due to current effort. See Friedman and Kelman (2007), who show that short-term financial incentives in British hospitals had long-run effects on productivity. 16 (though less direct) studies of unintended consequences of incentives which suggest that firms choose to live with some of these consequences for the sake of capturing some other benefits. Oyer (1998), for example, analyzes the effects of the fact that firms tend to plan their budgets and their incentive systems around a fiscal year. Executives and salespeople typically have contracts with a non-linear relationship between pay and performance. Salespeople often have annual quotas, for example, and many executives receive certain bonus payments only if they surpass some target. This provides these employees with incentive to try to exert some discretion over when certain results are achieved. For example, a salesperson who is rushing to meet a quota near the end of the year may offer a customer a big (but unnecessary) price break if the customer orders immediately. He shows that, controlling for the calendar seasonality of a firm’s industry, firms tend to sell more (and at lower margins) near the end of fiscal years than they do in the middle of the year. They also tend to sell less early in the year, suggesting that salespeople and/or executives “borrow” from the next year to meet the current year’s quota.15 Larkin (2007), using data from a large software company, finds that salespeople’s reaction to their incentive contracts cost the firm 6-8% of potential revenue. The specific example studied by Larkin (2007), as well as the broader pattern identified by Oyer (1998), raise the interesting question of why these non-linear contracts are so prevalent given that they create inconsistency in the production process and cut profit margins at the end of the year. However, given that these fiscal-year effects have been so common for so long, it appears there must be some benefit of these contracts that outweighs these apparent costs. Chevalier and Ellison (1997) show that non-linearities also lead to persistent distortions for mutual fund managers, though these non-linearities are not chosen by firms.16 Specifically, they show that there is a non-linear relationship between mutual fund investment returns over a calendar year and investment inflows into those funds. As a result, near the end of the year, some mutual fund managers have an incentive to change the level of risk in their funds. They show that such changes in risk do occur, suggesting that mutual fund managers choose their portfolios’ risk with more than just maximizing return in mind. Again, given that this distortion has been 15 See Healy (1985) and Murphy (2000) for further evidence that there is a relationship between executive contracts and timing of performance within fiscal years. 16 Also see Brown, Harlow, and Starks (1996), who reach similar conclusions about mutual fund managers. 17 taking place in a competitive market for a long period, it appears that the costs imposed by this behavior of mutual fund managers are outweighed by some benefits.17 2.3 Empirical Relevance of the Risk/Incentive Trade-off In a series of papers over the last decade, Canice Prendergast has explored the trade-off between risk and incentives that we discussed above. First, in a review of the incentives literature at that time (Prendergast, 1999), he argued that the empirical evidence was mixed on the risk/incentives trade-off.18 Then, in a series of papers, Prendergast (2000, 2002a, 2002b) more carefully lays out the evidence on the risk/incentive trade-off and offers several explanations for why the risk/incentive trade-off may be difficult to find in the data even if the underlying idea is correct. In Prendergast (2002a), risky environments are likely to be those where a manager’s private information is more valuable. A firm will be willing to pay the additional costs to compensate a manager for additional risk in order to make sure the manager uses his private information profitably. In other words, the marginal value of the manager’s effort is increasing in environmental risk such that the natural trade-off between risk and the cost of incentives may be overwhelmed by a positive relationship between risk and the benefits of incentives. 19 Initial attempts to test Prendergast’s model have been generally supportive (see Adams, 2005, and DeVaro and Kurtulus, 2006), though they are limited by the difficulty of finding exogenous variation in decision rights delegation. Prendergast (2002b) presents a model of subjective assessments and shows that either favoritism or costs to the manager of making assessments can lead firms to find incentives more valuable when risk is higher. Prendergast (2000) argues that, if the costs of monitoring effort directly are correlated with risk, then firms may use more incentives in riskier environments because the requisite risk premium is smaller than the cost of the additional resources that would be required to directly monitor the agent’s actions. 17 For other examples of problematic responses to incentives and measurement, see Dranove, Kessler, McClellan, and Satterthwaite (2003) and Jacob and Levitt (2003). 18 See Aggarwal and Samwick (1999a) and Jin (2002) for recent evidence that the trade-off holds for American CEOs. 19 Zabojnik (1996) draws similar implications in a model where a portion of market risk is related to the manager’s marginal product and is revealed to the manager before she chooses her effort, though he does not model the decision rights of the manager. 18 Oyer (2004) offers an alternative reason risk and the value of incentive contracts can be positively correlated, even when the pay scheme is not meant to affect the agent’s actions. He develops a model where the employee’s reservation wage and the firm’s value or profits are correlated with one another because they are both affected by macroeconomic shocks. The firm ties the employee’s pay to the firm’s success in order to lower contracting and renegotiation costs if a shock affects the employee’s outside opportunities. In environments where shocks (and, therefore, risk) are greater, the firm will tie more of pay to firm performance in order to more closely match compensation to the employee’s outside opportunities. However, if the environment gets too risky, the firm will abandon performance-based compensation and renegotiate wages when conditions change. This suggests that the incidence of “incentives” will be negatively related to risk but the amount of incentive will be increasing in risk. Prendergast (1999) summarized the research on the risk/incentive trade-off as of the time he wrote by saying, “there is some evidence that contracts are designed to optimally trade off risk against incentives” and “it would not appear that on the margin, the risk measures that have been considered are the true constraining factors on the provision of incentives.” These conclusions are still appropriate. But, in the last several years, some of the extensions to the basic moral hazard model reviewed above have uncovered possible reasons the risk/incentive trade-off has been so elusive to empirical researchers. The challenge in the years ahead is to design empirical strategies and measures that can confirm or refute these newer models, as well as the basic moral hazard model. 3. Matching Firms and Workers Matching firms with workers would be an easy process if labor were a commodity like some other inputs. However, labor is probably the most heterogeneous of all inputs in production functions. This is true on both sides of the market – the value of a given worker is likely to vary dramatically across potential employers and the disutility of effort associated with work will vary for a typical worker across the firms she might work for. Matching the right firms to the right workers creates economic value of a magnitude that few other economic processes can. Given the importance of job market matching, it is not surprising that economists have been studying the selection process since well before there was a field known as personnel 19 economics. Important theoretical contributions have come in two varieties – game theoretic models of asymmetric information and models of efficient matching with symmetric learning about worker productivity. 3.1 Learning Models Suppose individual i's output (y) in period t, if employed at firm j, can be written as ijt i ij ijt y = α + μ + ε where α is the innate ability of worker i, μ is the “match” productivity of the worker i/firm j combination, and ε is a productivity shock. Suppose all three variables on the right-hand side of the equation are random variables with mean zero and variance σα, σμ, and σε, respectively. Consider the case where σα = σμ = 0. In this case, each individual is identical and labor is a commodity. Efficiency is independent of how workers are matched to firms. If this were a reasonable representation of the world, then the field of personnel economics would not study selection of workers. Now consider the somewhat more interesting example where σμ = 0, but σα > 0. In this case, workers differ in their productivity but their productivity is independent of where they work. Again, selection is uninteresting in this environment as long as information about an individual’s value of σ is symmetric between firms and the worker.20 Now consider the more interesting example where σμ > 0. In this case, which is the core of the matching model in Jovanovic (1979), a worker’s expected productivity in any given period depends on where she works. In this case, it is important, in terms of economic efficiency, to maximize the firm/worker match quality (μ). Though exactly who captures the value of this match quality will differ depending on the competitive environments in the labor and production markets, the total size of the economy will increase in average match quality. In the absence of any cost of changing jobs, workers would switch jobs several times early in their careers in search of the best match. But, given some search or other transaction cost in job switching, workers will switch jobs only if the expected gains to doing so are large enough. 20 While selection issues are not important in models with symmetric learning about general ability (that is, ability that is equally useful to all employers), these models have interesting implications for compensation. See Section 4.1 below. 20 There are two key empirical implications of matching models. First, turnover rates will decrease in job tenure (that is, the longer a person stays in a job, the less likely she is to leave the job in any given period.) Second, wages will increase in job tenure.21 Both of these implications have proven consistent with empirical patterns in numerous studies.22 Unfortunately, these two implications are also consistent with models of firm-specific human capital. While the patterns in the data may appear consistent with both models, the managerial implications of these models are drastically different. If σμ is very high, so that match quality varies, employers should invest carefully in screening and selection. But if match quality does not vary and firm-specific human capital is an important driver of productivity, firms should focus on training and other human capital development. Note that the matching literature is vague about what underlies the value of a match (μ). Two recent models explore this. Lazear (2004a) models a world where all skill is completely general, but employers value mixes of skills differently. Hayes, Oyer, and Schaefer (2006) emphasize the value of co-worker specific match quality. Employees are more valuable at some firms than others because they work more productively with the employees at some firms. While both these ideas can help get to the root of matching, they do not help with distinguishing matching from firm-specific human capital. The mix of skills and/or relationships with coworkers can be developed after taking a job, which would make them a form of firm-specific human capital rather than ex ante match quality.23 Two recent papers that look more directly at the importance of matching are Andersson, Freedman, Haltiwanger, Lane, and Shaw (2006) and Woodcock (2006). Andersson, et al (2006) examine the matching of the most talented software engineers to the firms with the highest returns to talent, while Woodcock (2006) analyzes the importance of individual skill, firm effects, and match-specific productivity on wages. Empirical research along these lines, which has become more feasible as rich employer-employee datasets have become available, could 21 This implication will not be strictly true for a given worker over his job tenure if the match quality is immediately determined and incorporated into the wage. But if the quality of the match is learned over time, this implication will hold. 22 Miller (1984) and McCall (1990) extended the basic matching model by considering job-specific and occupationspecific matching. They generate implications regarding the optimal strategy for trying jobs of different types at the beginning of careers and find supportive empirical evidence. 23 Backes-Gellner and Mure (2005) test Lazear’s model using German data on employer-provided training. Their results suggest that the skills mix idea is enhanced by on-the-job training, though this does not exclude the possibility that employers select employees that already have (or show the ability to acquire) the range of skills the firms need. 21 shed some light on how the matching process takes place and how much value it creates.24 Also, research at single firms or small groups of firms that analyzes job applicants, those offered jobs who do not accept, and those who accept positions might also inform research and practice in selection and matching. 3.2 Asymmetric Information Models While matching models focus on firms and workers that are equally well informed, other models consider how employees match to firms when one party is better informed. An important example of how asymmetric information can be critical in the labor market is when a person knows her ability but a firm has only a noisy estimate. If all workers were honest, firms could simply ask job applicants about their ability and make hiring decisions based on what the applicants say. Less qualified applicants have considerable incentive to exaggerate their qualifications, so firms must find another way to extract this information. Two solutions to this problem have been suggested. In the pioneering work of Spence (1973), employers use costly signals to infer the ability of applicants. The cost to an individual of obtaining an education is inversely related to her ability (and, therefore, her productivity as an employee.) Signaling will only solve selection problems when acquiring the signal is sufficiently costly and when the cost of acquiring it is inversely related to the person’s ability. Also, signaling is an inefficient solution to the selection process because the cost of acquiring the signal is a deadweight loss. Another way to separate people of differing skill is to use selfselection (see Salop and Salop, 1976). If some portion of compensation or other parts of the employment relationship differ in their value to prospective employees, and if these differences in value are related to productivity, then more productive employees will self-select into an organization. While self-selection has the advantage of not wasting the resources involved in acquiring a signal, it relies on employers being able to find a condition of employment that will separate people based on ability. Lazear (1986a) models how incentive schemes can be useful for signaling and selfselection. Suppose that potential workers know their ability but employers do not. If productivity 24 Oyer and Schaefer (2007) also use a newly available data source – directories of employees from law firms’ web sites – to study the match between firms and workers. 22 differences cannot be measured on the job, then all employees will be paid the average workers’ productivity. But suppose an employer can measure an individual’s productivity at some cost. Lazear (1986a) assumes that all employers in the marketplace observe the person’s productivity when one employer measures it, so the firm will only undertake this measurement if the employee pays for it (presumably through lower compensation.) Once productivity is measured, employee wages are bid up to their individual productivity level. If measurement is costless, then the system fully unravels and everyone is paid exactly their productivity. But, if measurement is costly enough, then those with relatively low productivity are not willing to pay to be separated from those with the lowest productivity. As a result, some firms pay a fixed salary for all workers and attract those of relatively low ability. Other firms set up compensation systems related to output and measure workers. These firms attract workers above some ability cutoff. This model implies that, in industries and occupations where it is relatively costly to measure workers' output, workers will be more likely to be paid straight salaries. This is one reason salespeople are paid commissions, which is output-based pay, while most high level service workers are paid salaries. In the latter case, it is more costly to obtain a decent measure of output. More pooling occurs and proxies for input are used as the pay basis instead. Now consider a different type of asymmetric information model that can lead to inefficiencies in the labor market. Suppose each employee has ability α as in the prior discussion and firms cannot observe an individual’s ability before hiring her. However, suppose the firm gets a precise estimate of the person’s ability shortly after hiring her. Then, as Greenwald (1986) shows in an extension of the early work by Akerlof (1970), under certain conditions, there can be an inefficient level of employee mobility.25 Firms use their informational advantage to retain high ability workers when they receive outside offers while not responding to outside offers made to mediocre performers (known as “lemons” in the paper.) Fearing this information disadvantage will lead them to the winner’s curse and overpaid lemons, firms are reluctant to make offers. Thus, in some cases where a worker’s talents would be better put to use at a new firm, this efficient movement will not take place. In some labor markets, this may not be an issue. For example, the productivity of professional athletes (and academic economists) is readily observable to other teams 25 Lazear (1986c) also formalizes the adverse selection issue in the employment setting by putting the problem in an efficient turnover and auction context. 23 (universities) and the markets for these professionals’ services are quite liquid with significant volume of trade. However, in other markets, ability is not as easily observed by outsiders and this observability need not be completely exogenous. Several models have advanced Greenwald (1986) and Akerlof (1970) by analyzing how employers can best take advantage of their inside knowledge of their workers’ abilities. For example, Waldman (1984) considers strategic assignment of employees to jobs within a firm when outside firms take these assignments as signals of employees’ ability. He shows that employees may not be assigned to the jobs where they would be most efficient and that pay will depend on job assignment rather than strictly on ability. Milgrom and Oster (1987) extend this idea by considering two classes of workers – a regular group and a “disadvantaged” group. If it is easier to hide the disadvantaged group than the regular group, then firms will not promote or pay the skilled members of the disadvantaged group as much as they pay other skilled workers. This leads to persistent discrimination (that is, lower wages and under-representation in senior positions) against the disadvantaged workers and less investment in human capital by members of this group. Bernhardt (1995) also assumes that incumbent firms can exploit an informational advantage regarding employee ability and that this advantage may vary across certain classes of workers. He develops a model that is consistent with several stylized facts of some labor markets. The model also predicts the so-called “Peter Principle” where some managers are promoted to positions that are no longer appropriate for their skill levels.26 While layoffs, or even voluntary turnover, may reveal a “lemons” problem in some cases, Gibbons and Katz (1991) point out that this will not be the case when a firm fails.27 They predict that, because of the inferences drawn by employers, the reemployment wages of a worker at a plant that closes will be higher than the wages of a worker who is laid off from a continuing operation. They find some empirical evidence to support this prediction, though follow-up studies have argued that their results are fragile (see Song, 2007, and other work cited there.) Nonetheless, the insight that there might be more stigma attached to losing a job at a downsizing 26 For alternative explanations of the “Peter Principle”, see Lazear (2004b) where, on average, early success is at least partially good luck and Fairburn and Malcomson (2001) where risk averse managers may inefficiently promote underlings to provide them with more incentive. 27 This “stigma” of certain layoffs is similar to the signal sent by not getting outside offers in Lazear (1986c). 24 firm than at one that is closing an operation has become widely used in the labor economics literature. While the asymmetric information literature has been quite successful at developing realistic models that match some basic stylized facts, there has been relatively little direct empirical testing or estimation of these models.28 There are several reasons for this and, while empirical work would certainly be informative, it will likely continue to be difficult. Probably the biggest challenge to empirical researchers is the lack of measures or proxies for some of the key parameters in these models. For example, it is hard to find a convincing measure of the degree to which an employee’s ability is observable to other employers or a proxy for the costs of measuring an individual’s output. 3.3 Firing and Displacement In addition to selecting workers when they hire them, firms have the opportunity to change their selection of workers at any time by layoffs. If firms could freely adjust wages to an employee’s marginal product, then worker displacement would not be an issue because workers would make efficient voluntary decisions about when to leave. However, due to wage compression within firms, norms against nominal wage reductions, and other factors, workers often do not voluntarily quit when firms would like them to. Most worker displacement is simply the result of some significant negative shock to a firm or its industry or is the result of an individual proving to be significantly less productive than the firm anticipated. This type of turnover does not generate particularly interesting insights into how firms manage their human resource systems, so we do not cover it in detail here. It is worth noting, however, that displacement typically has negative and economically significant consequences for workers due to loss of firm-specific human capital (or match-specific productivity), stigma of losing a job, or other factors. See Farber (2005a) and Kletzer (1998) for reviews of the frequency and effects of job loss in the United States and Hallock (2006) for a discussion of the layoff process. One way that displacement affects the way firms run their personnel systems is through limits that are placed on firm’s ability to dismiss workers. A firm making decisions about who to 28 See DeVaro and Waldman (2006) for one recent empirical study on this issue, focusing on a single large firm. 25 hire and whether to hire may act differently if it loses the option of correcting bad hiring decisions through dismissal. Consider a firm weighing the possibility of taking on a particular candidate. The firm trades off the expected costs and benefits of hiring that person versus its next best option (which may be an alternative candidate or it may be to not hire at all.) The firm will typically have an imperfect estimate of the benefits an employee will bring the firm because it does not know the person’s ability (or match) exactly and the firm’s business environment could change in a way that affects the potential employee’s marginal product. Consider a firm that is choosing between candidate A and candidate B. Normalize the person’s expected marginal product of labor in a given period to (and, therefore, their reservation wage) to 0. But suppose that candidate A produces 0 for sure, candidate B produces either z or -z (each with 50% probability), and output-contingent contracts are not feasible. Assuming everyone is risk neutral and the person is hired for only one period, a firm would be indifferent between the two candidates. However, if the employment match can last multiple periods, then, under certain conditions and all else equal, the firm will prefer type-B workers for their option value. In equilibrium, type-B workers will earn a wage premium and only certain types of firms will want these workers. Lazear (1998) develops a model along these lines, where he explores the market equilibrium when workers have option value.29 The option value of risky workers will be highest in contexts where workers are somehow tied to their firms and where it is relatively easy for employers to dispose of workers that turn out to be unproductive. In many cases, firms face barriers to dismissing workers due to legal or other institutional (for example, union) reasons. This makes the cost of employing workers higher, lowering labor demand, and can lead to some combination of lower wages and lower employment.30 In addition to lowering overall labor demand, firing costs can reverse the logic of preferring risky workers for their option value because substantial negative realizations relative to expected productivity can be very costly. That is, a firm that cannot lower wages substantially will want to fire a type-B worker that turns out to have productivity of -z. If firing costs are substantial enough, then the costs of hiring an unproductive type-B worker can outweigh the option value of having a 29 See Bollinger and Hotchkiss (2003) for an analysis of risky hires in professional baseball. Lazear (1998) may also apply to academic labor markets because moving costs tie professors to their institutions while the tenure system allows schools to displace workers relatively easily. 30 As Lazear (1990) points out, perfectly efficient labor markets could undo the negative consequences of firing costs. However, as Lazear (1990), Deterzous and Karoly (1992), DeLeire (2000), Acemoglou and Angrist (2001), and Autor, Donohue, and Schwab (2006) show, there is substantial evidence that firing costs increase unemployment. 26 productive type-B worker. Oyer and Schaefer (2002) model how equilibrium wages and employment will change for different types of workers when these workers differ in the observability of their ability or productivity. They argue that increases in firing costs that are constant across workers should increase the wage premium to seniority because employers have more precise estimates of the ability of workers with more experience. They find support for this model empirically, using labor market and civil rights litigation data around the time of an increase in employer liability for discrimination in displacement. In addition to affecting the “riskiness” of workers a firm would want to hire, firing costs will also affect the types of workers that a firm will displace. If the costs of firing a worker increase, a firm will be willing to live with a worse worker because the benefits of firing the previously marginal worker will no longer outweigh the costs. This implies that those workers that a firm fires when firing costs are high will be, on average, of lower ability than the ones it fires when firing costs are low. Empirical support for this idea has been found using variation in firing costs generated by three different institutional settings – American Civil Rights legislation, institutional barriers to worker displacement in Europe, and firing restrictions imposed by state courts and legislatures. Oyer and Schaefer (2000) find evidence consistent with this by using variation in firing costs by race. Extending this logic one step, the “stigma” that Gibbons and Katz (1991) argue will go along with displacement should be greater when firing costs are greater. Canziani and Petrongolo (2001) formalize this idea and find empirical support using data from Spain around the time of easing of firing restrictions in the mid-1980’s. Kugler and SaintPaul (2004) develop a similar model, focusing on the choice of firms to hire workers that are unemployed or working at other firms. Using variation in firing restrictions across American states, they find that the “lemons” effect increases in firing costs. 4. Compensation 4.1 The Level of Compensation How much will firms pay their employees? In the absence of incentive issues and any firm-specific productivity, compensation will be just enough to keep employees from leaving the firm (as long as the firm does not lose money). In equilibrium, this will be equal to the expected 27 marginal product of the employee’s labor. But how do firms and employees determine an employee’s marginal product? How do they know what is a reasonable amount to pay? For some employees with a lot of experience, firms and workers have a pretty good idea of what a reasonable wage is. But, when an employee enters the labor market, firms have to rely on observable features such as the person’s education, performance in interviews, and performance on tests. Over time, firms and employees learn more about the employees’ skill and tailor jobs and his pay accordingly.31 Farber and Gibbons (1996) model this learning process. Consider a person entering the labor market. The firm that hires him expects his productivity (yit) to be ( , ) it F i Xit y = α where αi is the innate ability of worker i and X is a set of characteristics that are observable to the firm (such as education). Because neither the firm nor the individual know the exact value of α, initial wages are based largely on X. However, over time, the firm observes a set of noisy signals about yit and updates its estimate of α. Farber and Gibbons (1996) derive empirical implications of employer learning on the measured correlation between pay and X, as well as how pay is related to characteristics firms cannot observe at the time of hiring. Using data from the National Longitudinal Survey of Youth, they find empirical results that are broadly consistent with employer learning playing a large role in the development of wages over careers.32 Altonji and Pierret (2001) develop this idea further and point out that, until the learning process is fairly developed, employers are statistically discriminating on the basis of the observable X characteristics. A key point in these models is that the level of pay will more closely approximate the worker’s marginal product as the firm and worker learn about the worker’s true ability. Competitive labor markets do not require that workers earn their expected marginal product in any given period. Firms may set implicit contracts across multiple periods as a means of facilitating long-term relationships, providing incentives, or more efficiently sharing risk. Lazear (1979) focuses on the long-term relationship and incentives ideas. He notes that a firm can provide incentives through fixed wages if the firm pays the worker less than his marginal product early in his career and more than his marginal product later. After a person has been at 31 See Waldman (2007) for a more detailed survey of careers in organizations. 32 Gibbons, Katz, Lemieux, and Parent (2005) and Lluis (2005) extend this learning model and the empirical analysis in Farber and Gibbons (1996) by noting that employees may have differing productivity in different sectors. Firms and workers learn about workers’ innate skill, as well as which sector best fits their skills, over time. 28 the firm for a while, he has incentives to perform well and avoid being fired in order to enjoy the rents he has been promised later. Because this ties the worker to the firm, it may also encourage development of firm-specific human capital. Older workers are “overpaid” in this model and need to be induced to leave the firm either through a mandatory retirement policy or an appropriate pension scheme. Harris and Holmstrom (1982) model wage dynamics over careers when employees are risk averse and when firms and workers learn symmetrically about workers’ ability. In their model, workers are insured against negative shocks to their realizations on ability. In equilibrium, individuals never have wage decreases and wages increase faster than productivity. Frank (1984) offers an alternative explanation for why workers may not be paid their marginal product and why wages are generally thought to be more compressed than productivity. He argues that, if workers care about their pay relative to their peers within the firm, high ability workers will be willing to take a wage discount for the value they get from being near the top of the pay scale. A look inside firms shows that actual wage dynamics are driven largely by the jobs people hold. That is, individual jobs have a fairly narrow band of possible wages. The learning process that leads to people settling into the appropriate compensation scheme over time appears to be largely about finding the right job for the person, rather than finding the right pay for different people doing similar jobs. This can be seen in the firms studied by Lazear (1992), Baker, Gibbs, and Holmstrom (1994), and several more recent studies (see Waldman, 2007, for more details.) 4.2 The Mix of Compensation Money isn’t everything, but everything can be expressed in terms of its monetary equivalent. That is, in most markets, consumers pay some amount of cash in return for a good or a service. In some labor markets, such as temporary help or “under-the-table” work, transactions take this form. But most people sell labor services in exchange for a range of cash and other compensation. Interpreted broadly, the “other compensation” can include standard non-cash benefits such as health care, retirement benefits, and employer-sponsored child care as well as “amenities” such as a large or nicely decorated office, low risk of injury or death, or a job that 29 consists of largely interesting tasks. If all firms had the same cost structure, all employees had the same preferences, and employees could buy any amenities or benefits from other sources for the same prices the employer would pay to procure them, then labor would be traded for money and individuals would buy their own basket of benefits to best suit their needs. However, either due to institutional features such as tax incentives or simply due to economies of scale, firms often have a comparative advantage in providing benefits or amenities relative to workers. In addition, some firms are better at providing benefits than other firms and individuals vary in terms of how they value benefits. For example, a restaurant can more costeffectively provide meals to its employees than a firm that manufactures ball bearings and some people place a higher value on food provided by their employer than other people.33 These forms of differentiation lead to two important conclusions. First, it is often efficient for firms to provide a compensation package that mixes cash with other things.34 Second, the total economic value of employment relationships can be enhanced by matching workers who value a given benefit or amenity with firms that have a comparative advantage in providing that benefit. That is, one of the determinants of the match-specific surplus discussed in Section 3.1 is the degree to which a firm efficiently provides amenities valued by its employees. The classic work by Rosen (1974) laid out the theoretical foundations for this job differentiation and Rosen (1986) provides an intuitive discussion of how this model leads to the theory of compensating differentials in labor markets. Rosen (1974) shows how market equilibrium prices allow inferences about the monetary value of some characteristic. For example, suppose that we want to determine the monetary equivalent value of interesting work in a firm. Define X as the proportion of tasks on a given job that are new to the employee. Suppose that data were available on wages and the proportion of tasks on a job that were new per day. Consider the following regression: Wage on job j = a + b (Proportion of new tasks per day on job j) 33 For alternative justifications of certain benefits, see Rajan and Wulf (2006), Marino and Zabojnik (2007), and Oyer (2008). They consider the relationship between benefits, employees’ cost of effort, and productivity. 34 Employer-provided health care and pensions account for a large share of the cost of employee benefits, as well as of economics research on benefits. Gruber (2000) reviews the huge literature on employer health insurance. The large literature on pensions focuses on the various types of pensions offered and how they affect retirement behavior. See surveys by Lumsdaine and Mitchell (1999) and Lazear (1986b). 30 The coefficient b reveals the value that the market places on having flexibility on the job. The coefficient reflects the market value and not necessarily the tastes of any one individual. Employees that value new tasks relatively highly can earn rents by taking jobs at firms that offer many new tasks and firms that can provide new tasks to workers at relatively low cost can earn rents by providing a relatively high proportion of new tasks. Rosen (1974, 1986) discusses the conditions under which the market “price” reflects individual tastes, firm technologies, or neither. There is an empirical literature that attempts to examine the relation of wages to nonmonetary benefits and to determine the “price” of various benefits and amenities.35 These analyses are very challenging, however, due to the effects of unobserved ability. Consider two people who are observationally equivalent (same age, education, etc.), but one has more skill that is observable to employers and unobservable to an econometrician. The higher skill employee will be able to command higher total compensation. However, due to regressive taxes on cash compensation and the fact that income effects will generally make the marginal utility of workplace amenities (relative to cash compensation) higher for higher paid workers, higher ability workers will take some of their additional compensation in the form of workplace amenities. As a result, Ehrenberg (1980) found that pensions and wages were positively correlated (when there should be a tradeoff between the two) and numerous studies have shown that wages are positively related to benefits.36 This does not mean that hedonic wage theory or the concept of compensating wage differentials is wrong. It just means that measuring the market price of these differentials is difficult. There are at least three potential ways that future research can get around the challenges presented by unobserved ability. First, if there are variables that are correlated with whether a person gets a benefit at work but not otherwise related to that person’s earning potential, then instrumental variables can be used to identify the salary/benefit trade-off. Olson (2002) uses this approach to estimate that married women will accept a 20% salary reduction in return for health insurance, using their husbands’ union status and firm size as an instrument for whether the women have employer-provided health insurance. While this approach can be useful when 35 See Antos and Rosen (1975), Thaler and Rosen (1976), Brown (1980), and Woodbury (1983) for early examples of this literature. 36 Hamermesh (1999) and Pierce (2002) highlight an interesting implication of the income effects of benefits. They show that changes in wage inequality in the U.S. in recent decades understate the total increase in compensation inequality as benefits have become relatively more generous for high-paid workers. 31 looking at health insurance, the value of other benefits is unlikely to be high enough to allow for precise estimation of the salary/benefit trade-off. Second, detailed information on multiple job offers to the same person can provide estimates of trade-offs, assuming the value of the employee is roughly equal to the various employers. Stern (2004) takes this approach by surveying scientists about the compensation and research components of their job offers. His estimates suggest scientists are willing to accept substantial wage decreases to engage in on-thejob research. However, highlighting a potential weakness of this approach, his estimates are imprecise because his sample size is small and respondents’ recall of job offer details is likely to include substantial measurement error. Finally, given that benefits are such an important part of human resource policies, a large firm might be willing to work with economists to design experiments that randomize across workers by location. Large firms that have many locations and are known for providing generous benefits (such as Starbucks or Whole Foods Markets) are promising candidates. Laboratory experiments may also be helpful for learning about how people view benefits relative to cash, but the short-term and relatively low stakes nature of that environment is likely to be an important limitation. 4.3 Equity Ownership Over the last few decades, stock and stock options have grown as compensation tools. This probably reflects several factors, including attempts to increase worker incentives and the fact that, as economic growth has made workers wealthier, employee risk aversion may have gotten lower. Think about a middle manager at a large firm. If this person were truly risk neutral, he would be indifferent between taking his compensation in the form of cash or in terms of company stock. The stock would have the benefit of aligning his incentives with those of other shareholders. However, given that this manager’s efforts have only a trivial value on stock holdings, it would be difficult to generate incentives unless the firm issued him huge amounts of stock (as in Holmstrom, 1982.) In this case, risk costs would certainly become important. So why do firms grant stock to employees? One possible answer is to retain employees, as in the Oyer (2004) model we discuss elsewhere. Another possible answer is to attract the right employees. There may be several “sorting” benefits of making equity compensation part of the employment agreement, even in the 32 absence of incentives. The manager who has many stock options and a low base salary does not earn much unless the company does well. Thus, a manager who is willing to take a job under these circumstances reveals that he believes in the company. This information may be valuable to investors, whose information about the true value of the company is not as good as that of managers. See Salop and Salop (1976) and Lazear (2005). Similarly, equity compensation, like any other non-cash benefits, may be used when a firm can provide it at relatively low expense and the employees value it. That is, firms can use equity-based pay to lower compensation costs by attracting employees that are particularly optimistic about the firm’s prospects. A slightly different, but closely related, idea is that, if workers that are optimistic about a firm are also the most productive (perhaps because of enthusiasm or an understanding of the firm’s environment), then equity-based pay would just lead to the type of self-selection in Salop and Salop (1976) that we discussed above. See Oyer and Schaefer (2005) and Bergman and Jenter (2007) for theoretical and empirical analyses of this idea in the context of broad-based stock option plans. While attraction and retention may be important reasons to grant equity to employees, perhaps the most important justification for equity-based pay is to generate incentives. This explanation is likely to apply in small firms or among very high-level managers at large firms. These employees can have an important impact on the firm’s value and the incentive effects of ownership can outweigh the inefficiency in asking these employees to bear the risk of factors beyond their control that affect firm value. 4.4 Executive Compensation One set of workers that has been widely studied, with much of this work falling in the arena of Personnel Economists, is top executives. There are several reasons why executives have received more than their share of attention in the personnel economics literature.37 First, executives (and especially Chief Executive Officers, or CEOs) receive a great deal of compensation. The average CEO of a large American company now makes several million dollars per year and the growth rate over the last few decades has been much faster than for compensation more generally. Second, due to disclosure regulations, publicly traded firms have 37 This is certainly true on a per-worker basis and is probably still true on a per-compensation-dollar basis. 33 to provide information about both executive compensation and company performance. This allows empirical economists to create large datasets of individuals’ pay and “performance” (where performance is a firm’s stock return or measures based on a firm’s or division’s accounting statements.) For these reasons, CEOs and other executives have been widely studied. Murphy (1999) documents the rise in executive-related academic work, as well as providing a summary of this literature. We refer readers interested in details on the institutional features, the level and structure, and the politics of executive pay to Murphy’s review. We will, however, highlight a few stylized facts from the executive compensation literature that relate directly to personnel economics. First of all, as one might expect given the large effect CEOs can have on firms’ outcomes and the relatively easy availability of performance measures, CEOs have high levels of incentive-based pay. Much of this is based on explicit incentives such as performance-contingent bonuses and equity ownership. A smaller, but important, amount of incentives comes in the form of subjective or implicit incentives including adjustments to salary and bonuses not tied to objective metrics. These incentives vary in ways that most theories predict, with stronger incentives at firms where the marginal product of the CEO is likely to be higher (such as larger firms, firms with more capital, and unregulated firms.) Murphy (1999) also shows that executive turnover is somewhat based on firm performance, but that the relationship is not as strong as one might expect and, at least as of the time of his writing, had weakened over time. He also shows that there is a surprising lack of relative performance evaluation (RPE) in executive pay. Executives are rewarded (penalized) for good (bad) macroeconomic conditions, though it would seem easy for firms to filter this uncontrollable risk out of pay/performance contracts. Murphy’s (1999) survey ends with suggestions for future research. He asked why RPE is scarce?38 Numerous explanations have been suggested since Murphy posed the question. First, in trying to justify the lack of RPE for executives, Garvey and Milbourn (2003, 2006) have provided both “supply” (that is, executive-driven) and “demand” (firm-driven) explanations. They argue that there is RPE for some executives, but that the average effect is small because there are many cases where RPE is not used. In Garvey and Milbourn (2003), they argue that 38 See Bertrand and Mullainathan (2001) for details on the lack of RPE in executive compensation. Note that Murphy (1999) also posed the question of why has executive pay increased so much in recent decades. Several explanations have been put forward. However, because these explanations are largely a debate about corporate governance rather than personnel policies, we do not address that issue here. 34 RPE is unnecessary once an executive attains sufficient wealth. Once his assets that are not part of current compensation are large enough, he can completely undo any position in the firm’s stock in his private portfolio such that the use of RPE is irrelevant. Consistent with this idea, they find that younger and less wealthy CEOs, who are less likely to have sufficient assets to unwind compensation contracts imposed by the firm, have higher levels of RPE. In Garvey and Milbourn (2006), they suggest that CEOs will try to influence boards of directors to limit RPE when the market is performing well. They then show empirically that CEOs are more likely to have their pay indexed during bad times than during favorable markets. Aggarwal and Samwick (1999b), influenced by an idea proposed in Fershtman and Judd (1987), suggest that RPE will be counter-productive when products are strategic substitutes (as in differentiated Bertrand markets) and when market competition is imperfect. Pay based on relative output will lead managers to be too aggressive in product market competition. They find support for this empirically by showing that the typical level of RPE is higher when product market competition is more intense. Oyer (2004) models a case where agents’ reservation wages are driven by systematic (that is, market-wide) shocks. In order to reduce contracting and turnover costs, firms use equity ownership knowing that compensation will fluctuate with market conditions. Options or shares in the firm’s stock can therefore serve two purposes. While providing shares for incentive purposes would be most efficiently done with RPE, the secondary purpose of efficient contracting when reservation wages vary is most efficiently done with pay that is indexed to the market. Yet another possible explanation for the lack of RPE is that the marginal return to effort may be correlated with the state of the market. Suppose the firm’s value (П) is an additively separable function of the CEO’s effort (e), macroeconomic conditions (θ), and an idiosyncratic shock (ε), so we can write П = v1e + v2θ + ε. Then it is clearly optimal to filter out the macroeconomic effects. But if the marginal product of effort (v1) is correlated with the macroeconomic shock (θ), then the optimal contract will be an increasing function of θ. Just as we might expect CEO pay to use RPE, we might also expect relative outcomes to affect CEO dismissals. Barro and Barro (1990) who focused on bank CEOs, found that pay is not related to relative outcomes but turnover is. Two more recent and broader studies, however, show that CEO turnover is affected by both overall market conditions and relative performance 35 of individual firms. Jenter and Kanaan (2006) and Kaplan and Minton (2006) both show that CEO turnover is higher when the firm’s industry performs poorly and when the firm underperforms the industry. 5. Skill Development The analysis and discussion to this point has been naïve in that it took the person’s ability and match-specific productivity as given. In reality, people develop skill over time both before and after joining a given firm and this skill development affects productivity. As Becker (1964) laid out in his classic study of human capital, people can develop general skill that is equally useful to multiple employers (and, therefore, becomes part of what we called ability, or α, in Section 3.1) or that is more useful to one employer than to any other (and, therefore, becomes part of the match-specific component, or μ, in Section 3.1.)39 Becker (1964) argued that the efficient way to divide the investment costs of skill is for firms to pay for firm-specific human capital acquisition of workers and for workers to pay for their own general human capital. Any other means of splitting human capital investments costs will lead to the potential for the firm or the worker to “hold up” the other party after sinking the cost of a relationship-specific investment. This allocation of human capital investment clearly applies in most situations and, in particular, when the investment is large. For example, almost all people pay the cost of their own secondary, college, and graduate education. There are exceptions such as firms that sponsor their employees getting MBAs, but these arrangements often come with a contractual commitment not to leave the firm for some set period. Similarly, firms generally pay employees the going wage during training periods and early in job tenure while the employee is largely acquiring the skill and knowledge necessary to be productive in a given firm.40 39 One important idea related to human capital is the extent to which incentives to acquire general and firm-specific human capital are related to promotions, careers within organizations, and even optimal organizational form. See Kahn and Huberman (1988) on up-or-out contracts, Prendergast (1993) on the relationship between promotion and firm-specific human capital, and Levin and Tadelis (2005) on partnerships in high-skill industries. 40 Note that it is often difficult to determine whether a firm or a worker have paid for the acquisition of skill because we do not generally have complete information on both the compensation arrangements in a given job and what those arrangements would be in the absence of skill acquisition. 36 Recently, a number of papers have argued that Becker’s optimal allocation of human capital investments does not appear to be applied in certain situations and tried to explain why. There are situations that are fairly clear examples of firms paying for the acquisition of workers’ general human capital. For example, Stanford University offers free computer training in a variety of programs to all employees, sometimes including programs not related to a person’s current job. Employees who take advantage of this suffer no compensation ramifications and are allowed to attend the training while being paid. A series of recent papers have attempted to explain why situations may arise where firms find it in their best interest to pay for employees’ acquisition of general human capital. Most of these papers focus on the same informational advantage of incumbent firms modeled by Greenwald (1986) and discussed in the prior section. An example is Autor (2001), who focuses on a specific example that clearly violates Becker’s model. Manpower Associates is a large U.S.-based temporary help services firm. This firm offers potential “temps” free computer training. The training provides skill in common computer programs, is completely free to the person, and is not coupled with any obligation to work for Manpower. A person can take this class and then go down the street to another temp agency or any other employer and use her new computer skills. Why would Manpower offer such a program? Autor (2001) argues that the process of training gives the firm asymmetric information about the ability of the temporary worker. The firm then can use that information to match temps to assignments in a way that creates rents, at least until the person has been out on enough assignments that another temporary agency can infer her ability and where to assign her. He argues that this information advantage may be made even more valuable when free training helps a temporary help agency attract higher quality workers. He develops a model of the agency’s acquisition and exploitation of private information, arguing that it will lead to particular relationships between market wages, training, and the degree of local temporary labor market competition. He then uses data from a Bureau of Labor Statistics survey of Temporary Service suppliers and finds evidence that is consistent with the model’s predictions. For example, firms that offer free training offer slightly lower wages, but this gap gets smaller in more competitive markets. Acemoglu and Pischke (1998) offer a related explanation for firms making investments in general human capital, based on the information structure in Greenwald (1986). In their model, firms have some level of monopsony power over workers and can capture some of the rents 37 associated with acquisition of general human capital. Because workers do not capture the full value of general human capital investments, they will underinvest in these skills. Some firms may therefore find it profitable to sponsor this training. They find evidence consistent with the predictions of their model using a test, in the spirit of Gibbons and Katz (1991), that looks at the relative wages of German apprentices that stay at the firm that trains them, those that quit and move to another firm, and those that quit for an exogenous reason (random selection into the military.) In another study, Acemoglu and Pischke (1999) show that, even in the absence of information asymmetry, firms may sponsor general training if wages in the labor market are compressed relative to workers’ marginal products. This compression, which they argue can arise from search costs or other transaction costs or from institutional factors such as unions and minimum wages, may differ across labor market settings leading firms in some markets to be relatively likely to sponsor human capital investments. Another justification for firms making investments in worker general human capital is that it may provide incentives for the employee to stay at the firm while in training. Flaherty (2006) analyzes data from a single non-profit institution’s training programs and from a crosssection of firms. She finds that training is related to increased employee retention. The work by Autor, by Acemoglu and Pischke and, to a lesser extent, by Flaherty focuses on general training of relatively low-skilled workers. There are interesting examples of firmprovided general human capital for higher skilled workers, often including cases where the training is highly visible to outside employers. For example, both of us (the authors) previously held positions at business schools that offered part-time MBA programs for people with full-time jobs. Most of these students were sponsored by their employers and were not paid less than people doing similar jobs (though some firms required workers to partially repay these investments if they leave before a specified time.) Future research focused on advanced training and degree programs might provide useful insights into optimal human capital investment for high-skill workers. 6. Organization of Work 38 To this point, we have acted as though a firm picks a single worker to do a well-defined set of tasks. However, in reality, firms are made up of a group of diverse people doing different sets of tasks. In this section, we briefly look at how firms decide who does what, who works with whom, and who works for whom. 6.1 Job Design Most firms beyond some minimum size set up organization charts that show individual jobs designed to do some set of tasks. Several studies of individual companies, including Lazear (1992), Baker, Gibbs, and Holmstrom (1994), and Gibbs and Hendricks (2004), show how firms divide jobs into well-defined pay grades. Given this importance of jobs, how do firms divide tasks into jobs? Rosen (1978) develops a model of firms and workers matching based on the tasks firms need done and the comparative advantage of workers in performing each of these tasks. (See Sattinger, 1993, for a review of assignment models.) In many ways, this model matches firms to workers based on tasks the same way that Rosen (1986) models firms and workers matching based on employee preferences. A similar model underlies the analysis in Gibbons, Katz, Lemieux, and Parent (2005), which highlights the importance of assigning workers to their most productive sector. Gibbons and Waldman (2006), on the other hand, develop a model where workers develop task-specific human capital so that a person’s job assignment has important effects on how her career develops.41 6.2 Teams, Worker Interaction, and Human Resource Practices Successful human resource management is not simply a matter of giving each individual proper incentive to engage in work in isolation. If it were, there would be no need for the firm. Perhaps the greatest value of the firm is that it provides a mechanism for people to work together and take advantage of complementarities in their skills and interests. The evidence that there can be gains from assembling workers into teams or groups has to be done at a firm-by-firm level 41 For other models of how tasks get divided into jobs, see Cremer (1986) on how career concerns affect job design, Garicano and Rossi-Hansberg (2006) on dividing tasks based on skill level, and models discussed in Section 6.3 below on the division of tasks between managers and other employees. 39 because of the need for data on productivity and team structure. As a result, much of this work falls into the category of “insider econometrics.” Ichniowski and Shaw (2007) review this method elsewhere in this volume. We will briefly discuss a few papers related to the issue of group incentives. Hamilton, Nickerson, and Owan (2003) study a garment manufacturer that switched from an individual piece rate system to group-based piece rates. Workers were given some discretion over when they made the switch. Though the most productive workers had the most to lose financially from the switch, they tended to be early adopters. Productivity went up (and stayed higher) overall, suggesting that collaboration was valuable at this company. The gains were greater for more heterogeneous teams and more able team members had a bigger influence on others’ productivity than less able team members. Not all manufacturers get such positive results from the switch to teams, however. King (1998) provides anecdotal evidence of a much different (and much less successful) transition to teams at Levi’s. Boning, Ichniowski, and Shaw (2001) look at the effects of team production in the steel industry. Most firms in their sample pay some sort of group-based incentive, but only some organize their workers into cooperative teams. They find that productivity is higher at firms that use teams and that this effect is especially large where the product and production process are more complicated. Incentive pay is also associated with higher productivity at these firms.42 What are the sources of these group-based productivity benefits? In some cases, groups allow people to solve problems better and to take advantage of different skills. Gant, Ichniowski, and Shaw (2002, 2003), again focusing on the steel industry, show that interactions among a larger set of workers are associated with higher levels of production. Mas and Moretti (2006) show how supermarket clerks are affected by the productivity of other clerks working at the same time. They find that high-productivity clerks increase the productivity of peers. Like Hamilton, Nickerson, and Owan (2003), they find larger productivity gains when groups of workers are more diverse in skill. Mas and Moretti (2006) also find evidence suggesting social pressure is important, because productive checkers have larger effects 42 Incentives are group-based in these steel mills. However, individual incentive pay can be used to undo some of the potentially problematic aspects of production in groups of people that may try to capture some of the rents of production (either from other workers or from the firms.) Bandeira, Barankay, and Rasul (2006) show that managers on a fruit picking farm allocate workers more efficiently when the managers are given incentive pay. Some of this productivity improvement comes at the expense of the managers’ friends, who receive worse assignments under the incentive pay plan. 40 on peers that can see them while they work. Similarly, Ichino and Maggi (2000) show that absenteeism and misconduct has an effect on peers in an Italian bank. In addition to these field studies, Falk and Ichino (2006) find noteworthy peer effects in an experimental environment. Subjects were given a mundane task (stuffing envelopes) and no financial incentive. Average worker productivity was higher when each person worked in the same room as another person than when people worked in isolation. The least productive workers were the most affected by working in a group, so the increase in their output more than made up for the decrease in output of the most productive workers. The results of all these studies taken together suggest that, at least in many environments, people prefer working in groups to working in isolation, people working in groups feel some pressure to keep up with the efforts of those around them, and/or the most productive workers pressure others into working harder. As the theoretical work of Milgrom and Roberts (1990) and Kandel and Lazear (1992) suggests, the effectiveness of team-based systems is likely to vary from firm-to-firm due to complementarities across human resources practices. That is, team-based productivity may be more important when firms invest in selecting employees carefully, training people in the system and other practices. Testing this idea empirically is challenging, however, because measuring productivity is difficult and human resource practices are adopted endogenously. Several studies (see, for example, Black and Lynch, 2001, Cappelli and Neumark, 2001, and Bloom and Van Reenan, 2007) have shown that adoption of progressive workplace practices are correlated with one another and with productivity. However, it is difficult to make any causal statements about these relationships. Ichniowski, Shaw, and Prennushi (1997) allay this concern to some degree by looking at productivity of individual steel production lines around the time of adoption of modern human resource practices. They show that these practices increase productivity in these factories and that the marginal effect of each practice is increasing in the adoption of other practices. 6.3 Hierarchies Most people work in a hierarchy with several levels of reporting responsibility. There are many contributing factors to this standard institutional arrangement and we will not investigate 41 this in any great detail because it is the focus of another chapter in this volume (see Garicano and Van Zandt, 2006.) As we noted in some detail above, one benefit of hierarchies is that they provide incentives. Hierarchies can also be a substitute for incentives, however, in that they can provide a mechanism for monitoring employees directly. Lucas (1978), formalizing ideas in Manne (1965) and elsewhere, develops a model where some people are simply endowed with superior management skills. These skills can be interpreted as monitoring, decision making, motivating, or any other skill that would be more valuable for a manager than for other employees. Rosen (1982) distinguishes between the managerial tasks of making choices about what to produce, strategy, etc. (“management”) and insuring that employees carry out these decisions (“supervision”). To best utilize superior management skills, it would be efficient to grow firms ever larger. However, there are diseconomies of scale in supervision, so optimal firm size trades off these two factors. Rosen’s (1982) model can explain skewed firm size within an industry and skewed compensation within a firm. Garicano (2000) models hierarchies as a means of matching problems to those in an organization that can most efficiently solve them. There would be no need to use a hierarchical structure for this issue if everyone knew who was capable of solving which problems. But, when there are costs of communicating this knowledge, expertise organized by levels efficiently gets the right person to the right problem. Hierarchies are arranged with the highest skill person at the top, the next highest skilled people at the next level, and so on to the lowest skill workers at the bottom. Each new problem enters the firm at the lowest level and gets bumped up the hierarchy until it reaches the person that can solve it. Given the increasing importance of services and “knowledge workers”, this idea of knowledge-based hierarchies has become more important. See Garicano and Van Zandt (2006) for discussion of recent related work. 7. Conclusion We have described a great deal of progress in the personnel economics literature. This research has helped explain how firms operate and has helped shape business education curricula in recent years. So it seems sensible to ask “where do we go from here?” There is plenty of room for more work, especially empirical work, in all the areas that we have discussed. But it seems 42 likely that the highest returns will come from careful analyses of the selection, sorting, and matching processes. How do firms find the right workers? How important is it that firms find the right workers for their particular context? Business school curricula and the business press are filled with many case studies of firms that highlight finding people that match their specific needs (Southwest Airlines, SAS Institute, and Whole Foods Market are prominent examples). How do these and other companies generate economic value by finding the right people? Another question worth further pursuit is how employees value different fringe benefits. As employees become wealthier and as technology breaks down the work/leisure divide, firms are likely to find non-cash compensation more strategically useful. Some of the types of on-thejob skill development we discussed can also be thought of as non-cash benefits. Why do firms train employees in computer skills and send them to executive education training? Is it to increase their productivity, to reward them, or to signal something to or about them? In addition to these broad areas, we have also noted some specific questions presented by recent work that are worth pursuing. What explains the lack of evidence to support the risk/incentive trade-off predicted by moral hazard models? Are American executives overpaid? How would we ever know? Other questions will emerge as external and institutional features evolve. For example, if the returns to skill continue to increase, how will firms manage the reorganization process as they pay for greater skill while trying to economize on low-skill jobs? What problems, if any, will result if internal pay differentials grow too large? Two empirical methods have become useful in answering these and other questions and we expect them to continue to grow in influence in personnel economics. First, studies of single firms, often involving a field experiment, can provide precise analysis of the effects of various human resource policies. This has proven effective in studying incentives, team-based production, and the role of jobs, despite the lack of generalizability of these studies. Well-crafted future studies (and perhaps experiments) can broaden the value of this method to selection, the wage/benefit trade-off, the use of subjective performance evaluation, training, and other areas. The other method is the exploitation of employer-employee datasets, which have become much more prevalent in recent years.43 These data can provide rich time series detail about firms’ pay structure, growth, employee tenure, etc. The lack of detail on firms’ human resources 43 See Abowd, Kramarz, and Margolis (1999) for a recent survey of some of the data and some of the papers that have used them. Also, see Lazear and Shaw (2007) for a series of studies using datasets of this sort from a variety of countries to analyze pay and turnover patterns within and across firms. 43 practices makes it challenging to frame an economic question in a manner that can be addressed across the whole dataset. But careful modeling and/or focusing on certain industries or types of workers often helps make the exercise manageable. Personnel economics has come a long way in the last few decades. We have learned a great deal about the underlying economics of human resource policies. But, through the methods and ideas we have just mentioned, as well as others we have overlooked or that have not yet arisen, we believe there will be much more learned in the next few decades. 44 References Abowd, John M., and Francis Kramarz, 1999. “The Analysis of Labor Markets Using Matched Employer-Employee Data,” in Handbook of Labor Economics, Vol. III, eds. Orley C. Ashenfelter and David Card. Amsterdam: North Holland. Acemoglu, Daron and Joshua D. Angrist, 2001. “Consequences of Employment Protection? The Case of the Americans with Disabilities Act,” Journal of Political Economy, 109(5), 915- 957. Acemoglu, Daron and Jorn-Steffen Pischke, 1998. “Why Do Firms Train? Theory and Evidence,” Quarterly Journal of Economics, 113(1), 79-119. Acemoglu, Daron and Jorn-Steffen Pischke, 1999. “The Structure of Wages and Investment in General Training,” Journal of Political Economy, 107(3), 539-572. Adams, Christopher, 2005, “Agent Discretion, Adverse Selection and the Risk-Incentive TradeOff,” Working Paper, Federal Trade Commission. Aggarwal, Rajesh K. and Andrew A. Samwick, 1999. "The Other Side of the Trade-off: The Impact of Risk on Executive Compensation," Journal of Political Economy, 107(1), 65-105. Aggarwal, Rajesh K. and Andrew A. Samwick, 1999. "Executive Compensation, Strategic Competition, and Relative Performance Evaluation: Theory and Evidence," Journal of Finance, 54(6), 1999-2043. Akerlof, George A., 1970. “The Market for ‘Lemons’: Quality Uncertainty and the Market Mechanism,” Quarterly Journal of Economics, 84(3), 488-500. Altonji, Joseph G. and Charles R. Pierret, 2001. “Employer learning and statistical discrimination,” Quarterly Journal of Economics, 116(1), 313-350. Andersson, Fredrik, Matthew Freedman, John C. Haltiwanger, Julia Lane, and Kathryn Shaw, 2006, “Reaching for the Stars: Who Pays for Talent in Innovative Industries?” NBER Working Paper #12,435. Antle, Rick and Abbie Smith, 1986. “An Empirical Investigation of the Relative Performance Evaluation of Corporate Executives,” Journal of Accounting Research, 24(1), 1-39. Antos, Joseph R., and Sherwin Rosen. 1975. "Discrimination in the Market for Teachers," Journal of Econometrics 2 (May), 123-150. Asch, Beth J., 1990. “Do incentives matter? The case of Navy recruiters,” Industrial and Labor Relations Review, 43(3), 89S-106S. Autor, David H., 2001. “Why Do Temporary Help Firms Provide Free General Skills Training?” Quarterly Journal of Economics, 116(4), 1409-1448. 45 Autor, David H., John J. Donohue, and Stewart J. Schwab, 2006. “The Costs of WrongfulDischarge Laws,” Review of Economics and Statistics, 88(2), 211-231. Backes-Gellner, Uschi and Johannes Mure, 2005. “The Skill-Weights Approach on Firm Specific Human Capital: Empirical Results for Germany,” working paper, University of Zurich. Baker, George, 1992. “Incentive Measures and Performance Measurement,” Journal of Political Economy, 100, 598-614. Baker, George P., Robert Gibbons, and Kevin J. Murphy, 1994. “Subjective Performance Measures in Optimal Incentive Contracts,” Quarterly Journal of Economics, 109, 1125-1156. Baker, George P., Michael Gibbs, and Bengt Holmstrom, 1994. “The Wage Policy of a Firm,” Quarterly Journal of Economics, 109(4), 921-955. Bandiera, Oriana, Iwan Barankay, and Imran Rasul, 2006, “Social Connections and Incentives in the Workplace: Evidence from Personnel Data,” working paper, London School of Economics. Bandiera, Oriana, Iwan Barankay, and Imran Rasul, 2007, “Incentives for Managers and Inequality Among Workers: Evidence from a Firm-Level Experiment,” Quarterly Journal of Economics, forthcoming. Barro, Jason R. and Robert J. Barro, 1990. “Pay, Performance, and Turnover of Bank CEOs,” Journal of Labor Economics, 8(4), 448-481. Becker, Gary S., 1964. Human Capital. New York: Columbia University Press. Bénabou, Roland and Jean Tirole, 2003. “Intrinsic and Extrinsic Motivation,” Review of Economic Studies, 70(244), 489-520. Bergman, Nittai K. and Dirk Jenter, 2007, “Employee Sentiment and Stock Option Compensation,” Journal of Financial Economics, 84(3), 667-712. Bergson, Abram. 1978. “Managerial Risks and Rewards in Public Enterprises,” Journal of Comparative Economics, 2, 211-25. Bernhardt, Dan, 1995, “Strategic Promotion and Compensation,” Review of Economic Studies, 62:2, 315-339. Bertrand, M. and S. Mullainathan, 2001. “Are CEO’s Rewarded for Luck? The Ones without Principals Are,” Quartery Journal of Economics, 116(3), 901-932. Black, Sandra E. and Lisa M. Lynch, 2001. “How to Compete: The Impact of Workplace Practices and Information Technology on Productivity,” Review of Economics and Statistics, 83(3), 434-445. 46 Bloom, Nick and John Van Reenan, 2007, “Measuring and Explaining Management Practices Across Firms and Countries,” Quarterly Journal of Economics, forthcoming. Bollinger, Christopher R. and Julie Hotchkiss, 2003. “The Upside Potential of Hiring Risky Workers: Evidence from the Baseball Industry,” Journal of Labor Economics, 21(4), 923- 944. Boning, Brent, Casey Ichniowski, and Kathryn Shaw, 2007. “Opportunity Counts: Teams and the Effectiveness of Production Incentives,” Journal of Labor Economics, 25(4), 613-650. Brickley, James A., and Jerold L. Zimmerman, 2001, “Changing Incentives in a Multitask Environment: Evidence from a Top-tier Business School,” Journal of Corporate Finance, 7(4), 367-396. Brown, Charles, 1980. “Equalizing Differences in the Labor Market,” Quarterly Journal of Economics, 94(1), 113-134. Bull, Clive, 1987. “The Existence of Self-Enforcing Implicit Contracts,” Quarterly Journal of Economics, 102(1), 147-159. Bull, Clive, Andrew Schotter; and Keith Weigelt, 1987. “Tournaments and Piece Rates: An Experimental Study,” The Journal of Political Economy, 95(1), 1-33. Bulow, Jeremy I. and Lawrence H. Summers, 1986. "A Theory of Dual Labor Markets with Application to Industrial Policy, Discrimination, and Keynesian Unemployment," Journal of Labor Economics, 4(3), 376-414 Canziani, Patrizia and Barbara Petrongolo, 2001. “Firing Costs and Stigma: A Theoretical Analysis and Evidence From Microdata,” European Economic Review, 45(10), 1877-1906. Capelli, Peter and David Neumark, 2001. “Do ‘High-Performance’ Work Practices Improve Establishment-Level Outcomes?” Industrial and Labor Relations Review, 54(4), 737-775. Capelli, Peter and Kevin Chauvin, 1991. "An Interplant Test of the Efficiency Wage Hypothesis," Quarterly Journal of Economics, 106, 769− 87. Chan, William, 1996. “External Recruitment versus Internal Promotion,” Journal of Labor Economics, 14(4), 555-570. Cheung, Steven N. S. 1969. The Theory of Share Tenancy: With Special Application to Asian Agriculture and the First Phase of Taiwan Land Reform. Chicago: University of Chicago Press. Chevalier, Judith & Ellison, Glenn, 1997. "Risk Taking by Mutual Funds as a Response to Incentives," Journal of Political Economy, 105(6), 1167-1200. Chevalier, Judith and Glenn Ellison, “Career Concerns of Mutual Fund Managers,” The Quarterly Journal of Economics, 114(2), 389-432. 47 Courty, Pascal; and Gerald Marschke, 2004. ”An Empirical Investigation of Gaming Responses to Explicit Performance Incentives,” Journal of Labor Economics, 22(1), 23-56. Cremer, Jacques, 1986. “Cooperation in Ongoing Organizations,” The Quarterly Journal of Economics, 101(1), 33-50. Deci, Edward L., Richard Koestner, and Richard M. Ryan. 1999. “A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation,” Psychological Bulletin, 125(6), 627-668. Deleire, Thomas, 2000. “The Wage and Employment Effects of the Americans with Disabilities Act,” The Journal of Human Resources, 35(4), 693-715. Dertouzos, James N. and Lynn A. Karoly, Labor-Market Responses to Employer Liability, Santa Monica: RAND Corporation, 1940. DeVaro, Jed, 2006, “Internal Promotion Competitions in Firms,” RAND Journal of Economics, 37(3), 521-542. DeVaro, Jed and Fidan A. Kurtulus. 2006. “An Empirical Analysis of Risk, Incentives, and the Delegation of Worker Authority,” Working Paper, Cornell University. DeVaro, Jed and Michael Waldman. 2006. “The Signaling Role of Promotions: Further Theory and Empirical Evidence,” Working Paper, Cornell University. Drago, Robert and Gerald T. Garvey, 1998. “Incentives for Helping on the Job: Theory and Evidence,” Journal of Labor Economics, 16(1), 1-25. Dranove, David Daniel Kessler, Mark McClellan, and Mark Satterthwaite, 2003. “Is More Information Better? The Effects of ‘Report Cards’ on Health Care Providers”, Journal of Political Economy, 111(3), 555-588. Duflo, Esther, and Rema Hanna, 2005. "Monitoring Works: Getting Teachers to Come to School," NBER Working Paper, no. 11880. Dye, Ronald A., 1984. “The Trouble with Tournaments,” Economic Inquiry, 22(1), 147-150. Ehrenberg, Ronald G. 1980. “Retirement System Characteristics and Compensating Wage Differentials in the Public Sector,” Industrial and Labor Relations Review, 33(4), 470-483. Ehrenberg, Ronald G. and Michael L. Bognanno, 1990. “Do Tournaments Have Incentive Effects?” The Journal of Political Economy, 98(6), 1307-1324. Fairburn, James A. and James M. Malcomson, 2001. “Performance, Promotion, and the Peter Principle,” Review of Economic Studies, 68(1), 45-66. Falk, Armin and Andrea Ichino, 2006, “Clean Evidence on Peer Effects,” Journal of Labor Economics, 24(1), 39-57. 48 Fama, E., 1980. “Agency Problems and the Theory of the Firm,” Journal of Political Economy, 88(2), 288-307. Farber, Henry, 2005a. "What Do We Know About Job Loss in the United States?" Working Paper #498, Industrial Relations Section, Princeton University. Farber, Henry and Robert Gibbons, 1996. “Learning and Wage Dynamics,” Quarterly Journal of Economics, 111(4), 1007-1047. Fernie, Sue and David Metcalf, 1999. “It's Not What You Pay it's the Way that You Pay it and that's What Gets Results: Jockeys' Pay and Performance,” LABOUR: Review of Labour Economics & Industrial Relations, 13(2), 385-411. Fershtman, Chaim and Kenneth L. Judd, 1987. “Equilibrium Incentives in Oligopoly,” The American Economic Review, 77(5), 927-940. Flaherty, Colleen N., 2006, “The Effect of Employer-Provided General Training on Turnover: Examination of Tuition Reimbursement Programs,” working paper, Stanford University. Frank, Robert H. 1984. “Are Workers Paid Their Marginal Products?” American Economic Review, 74(4), 549-571. Freeman, Richard B. and Morris M. Kleiner, 2005. "The Last American Shoe Manufacturers: Changing the Method of Pay to Survive Foreign Competition," Industrial Relations, 44(2), 307-330. Gant, Jon, Casey Ichniowski and Kathryn Shaw, 2002. “Social Capital and Organizational Change in High-Involvement and Traditional Work Organizations,” Journal of Economics & Management Strategy, 11(2), 289-328. Gant, Jon, Casey Ichniowski and Kathryn Shaw, 2003. "Working Smarter by Working Together: Connective Capital in the Workplace," Working Paper. Garicano, Luis, 2000, “Hierarchies and the Organization of Knowledge in Production,” Journal of Political Economy, 108(5), 874-904. Garicano, Luis, and Esteban Rossi-Hansberg. 2006, “Organization and Inequality in a Knowledge Economy,” Quarterly Journal Economics, 121(4), 1383-1435. Garicano, Luis, and Timothy Van Zandt. “Hierarchy,” forthcoming in Handbook of Organizational Economics, eds. Robert Gibbons and D. John Roberts. Princeton: Princeton University Press, 2006. Garvey, Gerald and Todd Milbourn, 2003. “Incentive Compensation When Executives Can Hedge the Market: Evidence of Relative Performance Evaluation in the Cross Section,” The Journal of Finance, 58(4), 1557-1582. 49 Garvey, Gerald and Todd Milbourn, 2006. “Asymmetric Benchmarking in Compensation: Executives are Rewarded for Good Luck But Not Penalized for Bad,” Journal of Financial Economics 82(1), 197-226. Gaynor, Martin, James Rebitzer and Lowell Taylor, 2004. “Physician Incentives in Health Maintenance Organizations,” Journal of Political Economy, 112, 915-931. Gibbons, Robert. 1987. “Piece-Rate Incentive Schemes,” Journal of Labor Economics, 5(4), 413-29. Gibbons, Robert and Lawrence Katz, 1991. “Layoffs and Lemons,” Journal of Labor Economics, 9(4), 351-380. Gibbons, Robert, Lawrence Katz, Thomas Lemieux and Daniel Parent, 2005. “Comparative Advantage, Learning, and Sectoral Wage Determination,” Journal of Labor Economics, 23(4), 681-724. Gibbons, Robert, and Kevin J. Murphy, 1990, “Relative Performance Evaluation for Chief Executive Officers,” Industrial and Labor Relations Review, 43, 30s-51s. Gibbons, Robert, and Kevin J. Murphy, 1992. “Optimal Incentive Contracts in the Presence of Career Concerns: Theory and Evidence,” Journal of Political Economy, 100(3), 468-505. Gibbons, Robert, and Michael Waldman, 1999. “Careers in Organizations: Theory and Evidence,” in Handbook of Labor Economics, Vol. III, eds. Orley C. Ashenfelter and David Card. Amsterdam: North Holland. Gibbons, Robert, and Michael Waldman. 2006. “Enriching a Theory of Wage and Promotion Dynamics inside Firms,” Journal of Labor Economics, 24(1), 59-107. Gibbs, Michael, and Wallace Hendricks. 2004. “Do Formal Salary Systems Really Matter?” Industrial and Labor Relations Review, 58(1), 71-93. Green, Jerry R. and Nancy L. Stokey, 1983. “A Comparison of Tournaments and Contracts,” Journal of Political Economy, 91(3), 349-364. Greenwald, Bruce, 1986. “Adverse Selection in the Labor Market,” Review of Economic Studies, 53(3), 325-347. Groves, Theodore, Yongmiao Hong, John McMillan and Barry Naughton, 1994. "Autonomy and Incentives in Chinese State Enterprises," Quarterly Journal of Economics, 109, 183-209. Gruber, Jonathan, 2000. “Health Insurance and the Labor Market,” in Handbook of Health Economics, eds. A.J. Culyer and J.P. Newhouse. Amsterdam: North Holland. Hallock, Kevin F., 2006. “Layoffs in Large U.S. Firms from the Perspective of Senior Managers,” Research in Personnel and Human Resources Management, 25, 137-179. 50 Hamermesh, Daniel S., 1999. "Changing Inequality In Markets For Workplace Amenities," Quarterly Journal of Economics, 114(4), 1085-1123. Hamilton, Barton H., Jack A. Nickerson, and Hideo Owan, 2003. “Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation,” Journal of Political Economy, 111(3), 465-497. Hayes, Rachel M., Paul Oyer, and Scott Schaefer, 2006. “Co-Worker Complementarity and the Stability of Top Management Teams.” Journal of Law, Economics, and Organizations, 22, 184-212. Hayes, Rachel M. and Scott Schaefer, 2000. “Implicit Contracts and the Explanatory Power of Top Executive Compensation for Future Performance.” RAND Journal of Economics, 31, 273-293. Harris, Milton and Bengt Holmstrom, 1982. “A Theory of Wage Dynamics,” Review of Economic Studies, 49(3), 315-333. Healy, Paul M., 1985. "The Effect of Bonus Schemes on Accounting Decisions," Journal of Accounting and Economics, 7, 85-107. Holmstrom, Bengt, 1979. “Moral Hazard and Observability.” Bell Journal of Economics, 10, 74- 91. Holmstrom, Bengt, 1982. “Moral Hazard in Teams,” Bell Journal of Economics, 13(2), 324-340. Holmstrom, Bengt, 1999. “Managerial Incentive Problems: A Dynamic Perspective”, Review of Economic Studies, 66(1), 169-182. Holmstrom, Bengt and Paul Milgrom, 1991. “Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design.” Journal of Law, Economics, and Organization, 7, 24-52. Ichino, Andrea, and Giovanni Maggi, 2000. “Work Environment and Individual Background: Explaining Regional Differentials in a Large Italian Firm,” Quarterly Journal of Economics, 115(3), 1057-1090. Ichniowski, Casey, Kathryn Shaw, and Giovanna Prennushi, 1997. “The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines,” American Economic Review, 87(3), 291-313. Ichniowski, Casey, and Kathryn Shaw, 2007. “Insider Econometrics,” forthcoming in Handbook of Organizational Economics, eds. Robert Gibbons and D. John Roberts. Princeton: Princeton University Press. Jacob, Brian A., and Steven D. Levitt, “Rotten Apples: An Investigation of The Prevalence and Predictors of Teacher Cheating”, Quarterly Journal of Economics, 118(3), 843-877. 51 Jenter, Dirk and Fadi Kanaan, 2006. “CEO Turnover and Relative Performance Evaluation,” NBER Working Paper No. 12068. Jin, Li, 2002. "CEO compensation, Diversification, and Incentives," Journal of Financial Economics, 66(1), 29-63. Johnson, D. Gale. 1950. “Resource Allocation under Share Contracts,” Journal of Political Economy, 58, 111-23. Jovanovic, Boyan, 1979. “Job Matching and the Theory of Turnover,” Journal of Political Economy, 87(5), 972-990. Kahn, Charles and Gur Huberman, 1988. "Two-Sided Uncertainty and Up-or-out Contracts," Journal of Labor Economics, 6, 423-444. Kandel, Eugene and Edward P. Lazear, 1992. “Peer Pressure and Partnerships,” Journal of Political Economy, 100(4), 801-817. Kaplan, Steven N. and Bernadette A. Minton, 2006. “How has CEO Turnover Changed? Increasingly Performance Sensitive Boards and Increasingly Uneasy CEOs,” working paper, University of Chicago. King Jr., Ralph T., 1998, “Levi’s Factory Workers are Assigned to Teams, and Morale Takes a Hit,” Wall Street Journal, May 20. Kletzer, Lori G., 1998. “Job Displacement,” Journal of Economic Perspectives, 12(1), 115-136. Knoeber, Charles R. and Walter N. Thurman, 1994. “Testing the Theory of Tournaments: An Empirical Analysis of Broiler Production,” Journal of Labor Economics, 12(2), 155-179. Kugler, Adriana D. and Gilles Saint-Paul, 2004. “How Do Firing Costs Affect Worker Flows in a World with Adverse Selection?” Journal of Labor Economics, 22(3), 553-584. Larkin, Ian, 2007. “The Cost of High-Powered Incentives: Employee Gaming in Enterprise Software Sales,” working paper, University of California at Berkeley. Lazear, Edward P., 1979. “Why is There Mandatory Retirement?” Journal of Political Economy, 87(6), 1261-1284. Lazear, Edward P., 1986a. “Salaries and Piece Rates,” Journal of Business, 59(3), 405-431. Lazear, Edward P., 1986b. “Retirement from the Labor Force,” in Handbook of Labor Economics, eds. Orley C. Ashenfelter and Richard Layard. Amsterdam: North Holland. Lazear, Edward P., 1986c. “Raids and Offer Matching,” Research in Labor Economics, 8, 141- 165. Lazear, Edward P., 1989. “Pay Equality and Industrial Politics,” Journal of Political Economy, 97(3), 561-580. 52 Lazear, Edward P., 1990. “Job Security Provisions and Employment,” Quarterly Journal of Economics, 105(3), 699-726. Lazear, Edward P. 1992. “The Job as a Concept,” in Performance Measurement and Incentive Compensation, ed. William J. Bruns, Jr. Cambridge: Harvard Business School Press. Lazear, Edward P. 1998. “Hiring Risky Workers,” in Internal Labour Market, Incentives, and Employment, eds. Isao Ohashi and Toshiaki Tachibanki. New York: St. Martin’s Press. Lazear, Edward P. 1999. “Personnel Economics: Past Lessons and Future Directions,” Journal of Labor Economics, 17(2), 199-236. Lazear, Edward P., 2000a, “Performance Pay and Productivity,” American Economic Review, 90(5), 1346-1361. Lazear, Edward P., 2000b, “The Future of Personnel Economics,” 110, Economic Journal, F611- 39. Lazear, Edward P., 2004a. “Firm-Specific Human Capital: A Skill-Weights Approach,” NBER Working Paper 9679. Lazear, Edward P., 2004b. “The Peter Principle: A Theory of Decline,” Journal of Political Economy, 112(1), S141-S163. Lazear, Edward P. 2005. “Output-Based Pay: Incentives, Retention, or Sorting?” Research in Labor Economics, 23, 1-25. Lazear, Edward P., 2006. “Speeding, Terrorism, and Teaching to the Test,” Quarterly Journal of Economics, 121(3), 1029-1061. Lazear, Edward P. and Sherwin Rosen. 1981. “Rank Order Tournaments as Optimum Contracts,” Journal of Political Economy, 89(5), 841-864. Lazear, Edward P. and Kathryn Shaw, 2007. An International Comparison of the Structure of Wages, Boston: NBER. Levin, Jonathan, and Steven Tadelis, 2005, “Profit Sharing and the Role of Professional Partnerships,” Quarterly Journal of Economics, 120(1), 131-171. Lluis, Stephanie, 2005. “The Role of Comparative Advantage and Learning in Wage Dynamics and Intrafirm Mobility: Evidence from Germany,” Journal of Labor Economics, 23(4), 725- 768. Lucas, Robert E., Jr., 1978, “On the Size Distribution of Business Firms,” Bell Journal of Economics, 9(2), 508-523. 53 Lumsdaine Robin and Olivia Mitchell, 1999. “New Developments in the Economic Analysis of Retirement,” in Handbook of Labor Economics, eds. Orley C. Ashenfelter and David Card. Amsterdam: North Holland. MacLeod, W. Bentley and James M. Malcomson. 1989. “Implicit Contracts, Inventive Compatibility, and Involuntary Unemployment,” Econometrica, 57(2), 447-480. MacLeod, W. Bentley, 2003. “On Optimal Contracting with Subjective Evaluation,” American Economic Review, 93(1), 216-240. Malcomson, James M., 1999. “Individual Employment Contracts,” in Handbook of Labor Economics, Vol. III, eds. Orley C. Ashenfelter and David Card. Amsterdam: North Holland. Manne, Henry G., 1965. “Mergers and the Market for Corporate Control,” Journal of Political Economy, 73(2), 110-120. Marino, Anthony M. and Jan Zabojnik, 2007. “Work-Related Perks, Agency Problems, and Optimal Incentive Contracts,” working paper, USC Marshall School. Marshall, Alfred, The Principles of Economics, London: MacMillan, 1890. Mas, Alexandre and Enrico Moretti, 2006, “Peers at Work,” NBER Working Paper #12,508. McCall, Brian, 1990. “Occupational Matching: A Test of Sorts,” Journal of Political Economy, 98(1), 45-69. Meyer, Margaret, Paul R. Milgrom, and D. John Roberts, 1992. “Organizational Prospects, Influence Costs, and Ownership Changes,” Journal of Economics and Management Strategy, 1(1), 9-35. Milgrom, Paul and Sharon Oster, 1987. “Job Discrimination, Market Forces, and the Invisibility Hypothesis,” The Quarterly Journal of Economics, 102(3), 453-476. Milgrom, Paul R. and D. John Roberts, 1988. “An Economic Approach to Influence Activities in Organizations,” American Journal of Sociology, 94(supplement), S154-S179. Milgrom, Paul and John Roberts, 1990. “The Economics of Modern Manufacturing: Technology, Strategy, and Organization,” American Economic Review, 80(3), 511-528. Miller, Robert A., 1984. “Job Matching and Occupational Choice,” Journal of Political Economy, 92(6), 1086-1120. Mirrlees, James, 1974. “Notes on Welfare Economics, Information and Uncertainty,” in Essays in Equilibrium Behavior under Uncertainty, eds. M. Balch, D. McFadden, and S. Wu, Amsterdam: North Holland. Murphy, Kevin J., 1999. “Executive Compensation,” in Handbook of Labor Economics, Vol. III, eds. Orley C. Ashenfelter and David Card. Amsterdam: North Holland. 54 Murphy, Kevin J., 2000. “Performance Standards in Incentive Contracts,” Journal of Accounting and Economics, 30, 245–278. Nagin, Daniel S., James B. Rebitzer, Seth Sanders, and Lowell J. Taylor, 2002. “Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment,” American Economic Review, 92(4), 850-873. Nalebuff, Barry J. and Joseph E. Stiglitz, 1983. “Prizes and Incentives Towards a General Theory of Compensation and Competition,” Bell Journal of Economy, 14(1), 21-43. Olson, Craig A., 2002. “Do Workers Accept Lower Wages in Exchange for Health Benefits?” Journal of Labor Economics, 20(2), S91-S115. Oyer, Paul, 1998. “Fiscal Year Ends and Non-Linear Incentive Contracts: The Effect on Business Seasonality,” Quarterly Journal of Economics, 113, 149-185. Oyer, Paul, 2004. “Why Do Firms Use Incentives That Have No Incentive Effects?” Journal of Finance, 59, 1619-1649. Oyer, Paul, 2008. "Salary or Benefits?" Research in Labor Economics, forthcoming. Oyer, Paul, and Scott Schaefer, 2000. “Layoffs and Litigation,” Rand Journal of Economics, 31(2), 345-358. Oyer, Paul, and Scott Schaefer, 2002. “Litigation Costs and Returns to Experience,” American Economic Review, 92(3), 683-705. Oyer, Paul, and Scott Schaefer, 2005. “Why Do Some Firms Give Stock Options To All Employees?: An Empirical Examination of Alternative Theories,” Journal of Financial Economics, 76, 99-133. Oyer, Paul, and Scott Schaefer, 2007. “Personnel-Economic Geography: Evidence from Large US Law Firms,” working paper, Stanford University Graduate School of Business. Prendergast, Canice, 1993. “The Role of Promotion in Inducing Specific Human Capital Acquistion,” The Quarterly Journal of Economics, 108(2), 523-534. Prendergast, Canice, 1999, “The Provision of Incentives in Firms,” Journal of Economic Literature, 37(1), 7-63. Prendergast, Canice, 2000. “What Trade-Off of Risk and Incentives?” American Economic Review, 90(2), 421-425. Prendergast, Canice, 2002a. “The Tenuous Trade-Off between Risk and Incentives,” Journal of Political Economy, 110(5), 1071-1102. Prendergast, Canice, 2002b. “Uncertainty and Incentives,” Journal of Labor Economics, 20(2), S115-S137. 55 Prendergast, Canice, 2007. “The Motivation and Bias of Bureaucrats”, American Economic Review, forthcoming. Prendergast, Canice, and Robert Topel, 1996. “Favoritism in Organizations”, Journal of Political Economy, 104: 958-978. Rajan, Raghuram G. and Julie Wulf, 2006, “Are Perks Purely Managerial Excess?”, Journal of Financial Economics, 79(1), 1-33. Rosen, Sherwin, 1974. “Hedonic Methods and Implicit Markets: Product Differentiation in Pure Competition,” Journal of Political Economy, 82(1), 34-55. Rosen, Sherwin, 1978. “Substitution and Division of Labour,” Economica, 45(179), 235-250. Rosen, Sherwin, 1982. “Authority, Control, and the Distribution of Earnings,” Bell Journal of Economics, 13(2), 311-323. Rosen, Sherwin, 1986. “The Theory of Equalizing Differences,” in Handbook of Labor Economics, eds. Orley C. Ashenfelter and Richard Layard. Amsterdam: North Holland. Ross, Stephen A. 1973. “The Economic Theory of Agency: The Principal's Problem,” American Economic Review, 63(2), 134-39. Salop, Joanne and Steven Salop, 1976. “Self-Selection and Turnover in the Labor Market,” Quarterly Journal of Economics, 90(4), 619-627. Sattinger, Michael, 1993. “Assignment Models of the Distribution of Earnings,” Journal of Economic Literature, 31(2), 831-880. Schaefer, Scott, 1998. “Influence Costs, Structural Inertia, and Organizational Change,” Journal of Economics and Management Strategy, 7(2), 237-263. Shapiro, Carl and Joseph E. Stiglitz, 1984. “Equilibrium Unemployment as a Worker Discipline Device,” American Economic Review, 74(3), 433-444. Shearer, Bruce, 2004. “Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment,” Review of Economic Studies, 71(2), 513-534. Song, Younghwan, 2007. “Recall Bias in the Displaced Workers Survey: Are Layoffs Really Lemons?” Labour Economics, 14(3), 325-345. Spence, Michael, 1973. “Job Market Signaling,” Quarterly Journal of Economics, 87(3), 355- 374. Stern, Scott, 2004. “Do Scientists Pay to Be Scientists?” Management Science, 50(6), 835-853. Stiglitz, Joseph E. 1975. “Incentives, Risk, and Information: Notes Toward a Theory of Hierarchy,” Bell Journal of Economics and Management Science, 6, 552-79. 56 Thaler, Richard and Sherwin Rosen, 1976. “The Value of Saving a Life: Evidence from the Labor Market,” in Household Production and Consumption, ed. Nester Terleckyj. Studies in Income and Wealth no. 40. New York: Columbia University Press (for NBER). Waldman, Michael, 1984. “Job Assignments, Signalling, and Efficiency,” The RAND Journal of Economics, 15(2), 255-267. Waldman, Michael, 2007. “Connecting Theory and Patterns of Evidence in Internal Labor Markets,” forthcoming in Handbook of Organizational Economics, eds. Robert Gibbons and D. John Roberts. Princeton: Princeton University Press. Woodbury, Stephen A., 1983. “Substitution Between Wage and Nonwage Benefits,” American Economic Review, 73(1), 166-182. Woodcock, Simon D., 2006, “Match Effects,” working paper, Simon Fraser University. Zabojnik, Jan, 1996. “Pay-Performance Sensitivity and Production Uncertainty,” Economic Letters, 53, 291-296.  
Institute of Social Studies Graduate School of Development Studies COPING WITH CAPITAL INFLOWS SURGE THE MACROECONOMIC IMPACT OF THE CAPITAL INFLOWS AND THEIR MANAGEMENT WITH RESPECT TO CHILE, THAILAND AND KENYA A Research Paper presented by: AHMED ARAB ADEN (ETHIOPIA) In Partial Fulfillment of the Requirements for Obtaining the Degree of: Master of Arts in Development Studies Specialization: ECONOMICS OF DEVELOPMENT Members of the Examining Committee: Dr P. de Valk Dr K. Jansen The Hague, December 2002 This document represents part of the author's study programme while at the Institute of Social Studies; the views stated therein are those of the author and not necessarily those of the Institute. Research papers and theses are not made available for outside circulation by the Institute. Enquiries: Postal Address: Institute of Social Studies P.O. Box 29776 2502 L T The Hague The Netherlands Telephone: +31-70-4260 460 Telefax: +31-70-4260 799 email: zubli@iss.nl Location: Kortenaerkade 12 2518 AX The Hague The Netherlands Coping with Capital Inflows Surge The Macroeconomic Impact of the Capital Inflows and their Management with Respect to Chile, Thailand and Kenya Table of contents Acknowledgement List of tables and figures Chapter one: Introduction 1.1) Obj ectives 1.2) Background 1.3) Problem Statement 1.4) Hypothesis 1.5) Data and Methodology Page IV V 1 1 4 5 6 Chapter Two: The capital Inflows: Theory and Literature Review 2.1) Conceptual Issues 2.2) Theoretical Framework 2.3 Capital inflows Literature Review Chapter Three: Capital Inflows and their Macroeconomic Impact 3.1 Capital Inflows: cross-country comparisons 3.2 Macroeconomic Impacts 3.2.1 Impact on aggregate demand 3.2.2 Impact on exchange rate 3.2.3 Impact on current account 3.2.4 Impact on inflation 3.2.5 Impact on economic growth Chapter Four: The Countries Macroeconomic Policy Response 4.1 Macroeconomic policy Options 4.2 Variations in Policy Responses 4.3 MacroeconOlnic outcomes 4.4 Policy Lesson 4.5 Policy Implication Chapter Five: Conclusions II 7 9 15 19 22 22 24 27 29 31 34 44 46 47 48 50 For my beloved wife, Yasmin And Newly born baby girl, Siham III Acknowledgements I am very grateful to my first supervisor, Dr. Peter de Valk, for his Painstaking, detailed comments and encouragement, which improved my exposition of the paper considerably. I would also like to pass my gratitude to my second reader, Dr. Karel Jansen, for his helpful comments and providing me with material, which facilitated my work. I would like also to pass my sincere appreciation to ECD convenor, Dr. Howard Nicolas and all my lecturers. Last but not the least, lowe special thanks to ECD Programme Administrator, Marija Zubli and the ISS staffs. IV Ahmed Arab Aden The Hague, 2002 ;,.., List of tables and figures Tables 3.1 Composition of Capital Flows since 1990-1998 3.2 Use of capital inflows during surge periods 4.1: Summary for the policy measures AI: Selected Economic Indicators since 1985-1998 A2: Other Selected Economic Indicators since 1990-1998 Figures 3.1 Net Private Capital Inflows 3.2 Real Effective Exchange Rate 3.3 Current Account 3.4 Inflation 3.5 Real GDP v Chapter one: Introduction The main objective of this paper is to conduct a descriptive comparative study on macroeconomic impacts of capital inflows and their management in Chile, Thailand, and Kenya. The surge of Capital inflows in these countries took place in different time period and had different compositions (recently mainly in the fonn of volatile short-tenn capital inflows). The capital inflows had different macroeconomic impacts because of difference in policy environment and the nature of capital composition in these countries. The inflows boost investment in Thailand and Chile (more importantly in Thailand) whereas in Kenya stimulated more consumption and surge in imports. The short-tenn capital inflows to these countries brought about appreciation of the real exchange rates, deterioration of the current account, capital flight, and uncertainty in the domestic markets. To manage these massive short-tenn capital inflows governments undertook different policy measures as prudential measures and capital controls to counteract the speculative nature of the inflows. In general, the authorities managed to attain a short-lived relief from these capital controls. 1.1 Objectives of the Research The objectives of my research paper are to examine the nature and composition of the capital inflow and assess their macroeconomic impact on: - Aggregate demand - Real exchange rate (RER), Current Account of Balance of payment (BOP), Inflation - Economic growth. 1.2 Background The international community has long recongnized that developing countries need a substantial capital inflow of external sources in order to fill the savings and foreign exchange gaps to overcome widespread poverty and to lift living standards to acceptable levels. However, Diaz-Alejandro (1985) was among the first to warn of the danger of financial liberalization. Describing the Chilean experience of the late 1970s and early 1980s. Chile's first full liberalization of the capital account in 1979 was linked to the subsequent debt and currency crisis in 1982. This was largely the result of liberalizing capital flows when the domestic financial sector had not been reformed, and in the presence of an unsustainable fixed exchange rate, and backward-looking indexation of wages (see, Edwards and Cox, 1997). These weakness, combined with lack of prudential supervisory regulations and a deep recessions beginning at the end of 1981, generated a sharp reduction in capital inflows (capital flight), a deterioration in terms of trade, and resulted in a financial crisis that spread throughout the financial system by the beginning of 1983. To correct this the authorities under took massive structural reforms in the late 80s such as banking reforms. Hence, capital inflows again started to come in since 1989. Whereas, in Thailand, the resurgence of capital inflows stared since in 1988 due to full integration of the country's financial system with international financial markets. And since the early 1990s, offshore banking (at the Bangkok International Banking Facility, BIBF) has provided direct and relatively cheap links to the international credit markets. Capital inflows in recent years were dominated by private sector short-term capital (Ariyoshi, 2000). These developments bring about significant changes in the nature of capital inflows, and in their impact on the economy. The growing size and volatility of these inflows, pruiicularly in early 1995 threatened the economy. In May 1997, the baht (Thai domestic currency) came under sever speculative attack and the capital stated to move out of the country. In Kenya, liberalization of the financial sector began in 1989 with measures intended to harmonize interest rate regulations for banks and nonbanks financial institutions. The measures include easing rigidities in financial sectors by freeing prices, liberalizing foreign trade and foreign currency transactions, and relaxing credit ceilings and interest rate controls. Liberalisation of the capital account was made in 1991 with out restrictions. However, capital do not seem to come in. Instead, the capital was going out from the country following the liberalisation. The main lesson from Kenya's experience seems to be that rapid and wideranging liberalisation may have increased the country's vulnerability to capital flows by providing legal channel for capital flight (Ariyoshi, 2000, 66). The accumulation of external debt that eventually 'resulted in the international debt crisis (in 1970s, 1980s and 1990s) was the result of an episode of capital inflows to developing countries in the form of syndicated baIUC loans directed almost exclusively to public sector 2 borrowers. The recent history of the international finance market liberalisation, both domestic and international, is characterised by costly financial crises. The recent series of financial crises include the Asian crises, Mexico (in 1976, 1980s, 1995) and Argentina (in1980s), Brazil, Peru, Thailand and the crises in Chile. The international capital movements were also having important macroeconomic effects. They were helping finance higher investment and growth, but there was also a tendency for the real exchange rate to appreciate and current account to worsen. Monetary control became difficult as the inflows persisted and an increasing share of inflows came in the form of shortterm capital. The new capital inflows, however, were very different in character from those of the previous episode. The previous capital movements were in the form of multilateral official transfers (grants) and concisional loans. But the new capital inflows mainly are composed of direct foreign investments (FDI- long-tenn investment) and private portfolio investments (shortterm in nature). Perhaps even more surprisingly, in view of the enormous economic costs associated with the debt crisis, the resurgence of capital inflows was not viewed as an unmitigated blessing by the recipient countries. Indeed, the arrival of large amounts of foreign capital was perceived as posing serious challenge to domestic macroeconomic management. These crises shed light on the danger international short-tenn capital and importance of its control. In this regard, the authorities of capital recipient countries should play an essential role in the financial markets to overcome market failure and boost economic development. The important role of the country may include capital controls. D~velopmental sates like Chile and Thailand played this essential role of controlling and allocating capital directly into specific sectors in order to promote investment. This paper examines and reviews both the macroeconomic challenges posed by the arrival of the capital inflows as well as the policy responses undertaken by the recipient countries with respect to Thailand, Chile and Kenya. It is organised into five chapters. The first chapter provides introduction, an overview of the paper. Second chapter looks at some of conceptual issues and theoretical explanations offered for the macroeconomic impact of capital inflows by reviewing the literature on this issue. In chapter three I will analysis the macroeconomic impacts of resurgence of capital inflows in countries under discussion. In Chapter Four I will 3 tum into an exploration of the policy challenge posed by the inflows, macroeconomic outcomes, policy lessons and implications. The last chapter will be based on conclusions drawn from the analysis. An important issue that rose to the fore in association with th~ capital inflows of the early nineties was vulnerability of the recipient countries to sudden capital flows, which resulted in currency crisis (in 1997 the Thai baht was in crisis). 1.3) Problem Statement After capital account liberalisation, the foreign capital started moving to most developing countries during the 1980s and the 1990s. As reported by Eichengreen (1999), that net capital flows to developing countries tripled from USD 50 billion in 1987-89 to more than USD 150 billion in 1995-1997. Capital inflows or capital account transactions often are categorised into foreign direct investment (FDI) and pOlifolio investment. Direct investment usually involves long term projects and bring not only real assets such as fixed assets and working capital, but also financial flows (equity). Meanwhile, portfolio investment is closely classified as shorttenn flows and encompasses trade in securities like stocks, bonds, bank loans, and various forms of credits. Unlike FDI, portfolio investment has special characteristics and more volatile in nature. It is evident that the surge has primarily been an East Asian and Latin American phenomenon. In both cases, the pace of inflows accelerated after 1991. This phenomenon may recently have become more widespread, reaching South Asia as well as Sub-Saharan Africa (particularly Kenya and Uganda) in 1993 (Eduardo, 1995, 10). In economies that are trying to stabilise, macroeconomic problems associated with volatile capital inflows include loss of domestic monetary controls, real appreciation, and increased instability. ' The transition was led by East Asian countries that had been much less affected by the debt crisis than had the major countries in Latin America, but many of the latter also soon began to participate in the new capital inflow episode. To date, however, a large number of small lowincome countries, particularly in sub-Saharan Africa, remain much less integrated with international financial markets. 4 1.4) Hypotheses: a) Impact on exchange rate My first hypothesis is that when capital inflows such as foreign direct investment (FDI) and portfolio investments (short-tenn capital inflows) increase in a country will stimulate the aggregate demand. In addition, the demand for the local currency may increase which in tum lead to appreciation of local currency. This appreciation will lead to loss of the export competitiveness in the international market. From macroeconomic point of view the episode of capital inflows may lead to exchange rate appreciation pressure. All the countries under discussion followed real appreciation in the face of the capital inflows. b) Impact on Current account Large proportions of the capital inflows can be used to finance domestic absorption namely investment and consumption. This finance of domestic absorption may tend to increase the imports from abroad. So it is likely to run huge deficit in BOP and hence on current account. For instance, Both in Chile and Thailand increase in absorption during the surge of capital inflows were dominated by investment. But the changes in the composition of absorption were heavily biased in favour of investment in Thailand leading to huge current account deficits. However, the capital inflows in Kenya increased more consumption and imports. c) Impact on inflation When a capital inflow is associated with an upward shift in the demand for money (induced, say, by financial deregulation), no policy action is required because, 1n this case, the expansion of the monetary base will not be inflationary or threaten external viability. However, if authorities heavily sterilise the capital inflows the increase in foreign reserves may lead to an increase in reserve money (Thailand and Kenya), which in tum, increase inflationary expectations and a deterioration of external position leading to escalation of domestic prises through expenditure rise. It may be necessary, however, for the central bank to intervene in the (relatively thin) money and foreign exchange markets to smooth· fluctuations in the exchange rate and interest rates. One possible cause for conce:rn is that banking credit is likely to expand as money balances increase. With a poorly supervised and weak banking system, the increase in commercial banks' reserves could lead to excessive risk taking in lending activities, and measures may be needed to restrict bank intennediation. 5 d) Impact on economic growth in host countries Free Capital mobility may have important benefits such as it lead to economic efficiency and development. Particularly, it creates valuable opportunities for portfolio diversification, risk sharing, and trade. Capital mobility can enable investors to achieve higher risk-adjusted rates of returns. Capital mobility not only bring resources, but also financial market can contribute to enhancing efficiency in resource allocation, easing external constraint, technology transfer, and better management and business practices. In addition, international capital movement increases the availability of foreign savings to supplement domestic resources that deliver faster rates of growth. 1.5) Data and Methodology I will use data on capital inflows, real exchange rate, balance of payment, and economic indicators from these countries in general. I will use descriptive qualitative research method to make comparative study between the countries such as graphs and tables to explain the impacts and trends, and explore country policy response to manage the capital inflows. In Sum, the reasons that I opt for these countries are that Thailand is one of the largest capital importers in the South East Asia and also the same for Chile in Latin America. The episode of the inflows in these countries resulted in financial and economic crisis (1982 in Chile, 1997 in Thailand and in 1994 in Kenya). However, by 1992 there were signs that a few African countries, notably Kenya had begun to attract private capital inflows. Many of the less desirable side-effects of capital inflows also became evident. All in these countries real exchange rate pressures and current account deterioration emerged. Attempts to sterilize the foreign exchange transaction through either open market operations or increase reserve requirements often drove domestic real interest rates higher, acting as a further stimulus to inflows and increasing debt-servicing costs for the governnients. What is the appropriate policy response, then? The purpose of this paper is to answer some of these questions. The paper presents a theoretical framework to analyse the macroeconomic effects of and the policy responses to a surge in capital inflows. 6 Chapter 2: The capital Inflows: Theory and Literature Review This chapter basically focuses on the literature available on capital inflows such as providing definitions for basic concepts and presents a theoretical framework to analyse the macroeconomic effects of the inflows. Finally, it looks into different arguments about the role of capital inflows and its liberalisation. 2.1) Conceptual issues Conventionally, it is defined the current account deficit (of the balance of payment), CAD, as follows: CAD = Import - Exports - Net Factor Transfers from the Rest of the World (1) The current account is the balance of export earnings (X), net current transfers received (NTR), payment for imports (M) and net factor payments (NFP)-payment of investment income and interest, receipts of workers' remittance, and transfer of profits. All are expressed as percentage of GDP (Jansen, 1997). The current account balance would be identifying the flow of goods and services across the boarder. The current account is also the net result of savings and investment, private and public (Reisen, 1997). The CAD measures the rate at which the country is becoming indebted to the rest of the world. Thus, if measured without en'or, it represents the growth of the country's overall (i.e., including private and public sectors) net international indebtedness. International Reserves, R, are defined as official holdings of international short-term liquid assets issued by foreign governments (e.g., US Treasury Bills). Normally, international reserves are held in the form of obligations incurred by hard-currency governments, e.g., U.S. dollars, Euro, etc, and are held at the domestic central bank. Capital Inflows, Kl, are defined as follows: Kl=CAD+~R, (2) 7 Where, il.R denotes accumulation of international reserves. Notice that KI - il.R = growth of net international indebtedness. Thus, capital inflows are a gross concept because it does not net out accumulation of international reserves. Capital flows episodes are defined as situations in which there are a sudden and persistent increase in KI. In other words, net cap~tal inflows can be defined as the difference between the capital inflows and the capital outflows. It is the increase in net international indebtedness of the private and the public sectors and are measured by the surplus in the capital account of the balance of payments. Therefore, except for errors and omissions, the capital account surplus equals the excess of expenditure over income (which, in turn, is equal to the gap between national investment and saving) plus the change in official holdings of international reserves. Thus, increase in capital inflows can be identified with larger current account deficits and/or an accumulation of reserves. The expression "capital inflows" gives the impression that when KI rises more capital gets accumulated. To illustrate this, we recall that the following is an identity in national accounting: CAD = Total Consumption + Total Investment - GDP, (3) Therefore by equation (2) and (3): KI = Total Consumption + Total Investment - GDP +il.R. (4) Consequently, gIven GNP and il.R, an mcrease in capital. inflows· must be reflected m consumption and/or an investment boom. Capital inflows have many macroeconomic effects on the receiving economy, both in the short run and in long run. Different types of capital inflows have different impacts. The longterm flows: ODA (public sector external borrowing), FDI (foreign direct investment), longterm loans, bonds are more stable, whereas the short-term flows: PFI (portfolio inflows), bank deposits, short-term loans are more volatile and can reverse on short notice, creating serious adjustment problems to the economy. Much of ODA and FDI enter the country as commodities and capital goods rather than finance. But inflows of loans and of PFI may have significant effects on monetary balances and on the stability of financial markets. It is also 8 important to note that long-telm loans may also suddenly change: DFI inflows may suddenly halt, but will not reverse; access to long-term loans or to the bond market can suddenly top or become prohibitively expensive, but the impact on the economy will come more gradually in this case (Jansen, 2002). The central concern is that capital inflow, although definitely more desirable than capital out flows could be harmful to the recipient country if adequately not managed. The main concern is, thus, not so much with the inflows as such but with the potential outflows. If the 'pull factor' for the capital inflows are not due to domestic fundamentals, because of 'herding' behaviour of the investors may consequence potential danger of outflows. The initial policy reaction such as capital control to the rising capital inflows has been to intervene in the market for foreign exchange and sterilise the effects of the intervention through either open market operations, increase in reserve requirement, or both. Capital account control or liberalization is an old issue. 'Capital control' means vanous measures to restrain international movement of capital. As such, they would represent a firstbest policy intervention. If the distortion causing the problem that cannot be removed, a second-best option may be to limit foreign bOlTowing (Montiel, 1998). Thus, such measures have generally targeted only short-ternl capital flows. Long term flows, such as FDI (foreign direct investment), are not subject to them. Another approach has been to tax such inflows, either by an explicit tax or by making them subject to higher reserve requirements. Besides, capital controls are neither good nor effective because in most cases control over capital flows failed and private capital can evade the controls almost always (Edward, 1999). 2.2 Theoretical framework Financial markets have traditionally been inherently short-termists and volatile (see, for example, Keynes (1936), Kindleberger (1978) and Minsky (1982). However, the evid~nce gathered in these book seems to indicate that these markets both seem to have become more volatile and that this volatility has the potential to be transmitted in greater and more harmful ways on macroeconomic trends in developing countries. 9 An up surge in capital inflows requires an increase in the current account deficit by increasing aggregate demand, which push up inflation so that real exchange appreciates. These effects also depend on the exchange rate regime. Under a flexible exchange rate, the nomina.! exchange rate will appreciate with capital inflows, which affects demand for domestically produced goods and leads to the current account deficit. Under a fixed exchange rate regime the capital inflows are reflected in an increase in foreign reserves and an increase in the money supply and a fall in the domestic interest rates. This stimulates aggregate demand and is likely to result in inflation which leads to an appreciation of the real exchange rate (IBID). Although the conventional view IS that developing countries fundamentals determine behaviour of international financial markets, there is increasing evidence that in many cases it is the endogenous behaviour of international markets that conditions or strongly influences fundamentals in developing countries (see Fitzgerald, 2002). Thus demand and supply curve for emerging market assets are not independent; a supply-led large capital inflows affects the domestic economic situations (generating for example an asset price bubble) in a way that can increase the demand for assets. This makes regulation and other public interventions 111 international financial markets complicate effective regulation (Griffith-Jones, 2002). This section basically highlights theoretical understanding of the impact of the capital inflows into different macroeconomic variables in the analyses part of the paper. 1) total expenditure and fiscal implications From the balance of payments identity, changes 111 capital inflows (i.e. capital account inclusive of official transfers and errors and omissions) can be decomposed into changes in the trade balance, net factor payments, and international reserves. In other words, capital inflows can be used to finance either current account deficit and/or to accumulate reserves (see equation 2). Typically, a capital episode is associated with a rise in total expenditure (i.e., Total Consumption + Total Investment) through an increase in aggregate demand and thus a decline in saving so that the investment saving gap widens. 10 2) Debt maturity By definition, a capital inflows episode cOlTesponds to a situation in which' the government and/or the private sector are increasing the rate at which they fall into debt or lower their net wealth. Issuing debt and selling stocks to foreign residents are two forms of "borrowing." When a foreign resident purchases stocks of a local firm, he is entitled to share in distributed profits and to vote in the company's shareholders meetings. The fund accruing to the investor, therefore, depends on the firm's ability to generate, and willingness to distribute profits. Investors could actually lose the entire value of their investments if the firm goes bankrupt. In contrast, bondholders are the first in line for repayment, which in most cases independent of the firm's perfonnance. A high volume of short-term debt relatives to the stock of international reserves can be a major problem if the country entered into a balance of payments crisis. Short-term debt usually gives rise to the "bunching effect"-instability, where large chunks of debt mature in a short period of time. 3) Monetary consequences (bank sector, inflation) Banking sector Capital inflows episode is associated with increase in total expenditure, which in turn, increases the demand for "money," e.g., currency, and demandltime deposits. Thus, even if no international funds are channelled through the banking system, the higher transaction associated with the expenditure will lead firms and individuals to increase their bank deposits. If reserve requirements are less than 100 per cent, banks will probably increase loans. Trouble starts as capital inflows slow down. Suppose, for example, that capital inflows, KI, go from a positive to zero or become negative. KI is a flow, it measures new borrowing. Thus, when KI goes to zero, it simply means that suddenly no new loans are flowing to the. economy. According to equation [4], and taking the change in reserves, R, and GNP as given, a fall in capital inflows, KI, would induce a fall in total expenditure. Therefore, given the positive link between expenditure and deposit, the decline in expenditure would induce a fall in the demand for deposit. Thus, unless banks quickly find an alternatively source of finance, 11 they are likely to stop some of their sh011-maturity loans. This is a loan reversal shock! Hence, if the government refrains from intervening (through devaluation etc,) this might lead to bankruptcies. However, also predominant banking regulation during capital inflows surge may aggravate prevailing problems if banking supervision is weak and there are ineffeciencies in pricing risk, there may be additional reasons to limit banks' role in intermediating the capital inflows such as shifting deposits to the central bank as a reserve requirement measure. A major concern about the intermediation of international capital flows through the domestic banking system is that individual banks are subject to free or subsidised deposit (i.e., there is an implicit commitment by the authorities that banks especially those of large size will not be allowed to fail. It is well known that this deposits induces banks to increase their risk exposure. In several countries, there has been a sharp expansion of banle loans to finance private consumption. There is evidence that in some of these countries the percentage of nonperforming loans has recently increased over time. All these factors increase the vulnerability of the financial 'System to reversals in capital inflows (reversal that have the potential to end in financial crisis. It is the role of bank regulation and supervision to effectively diminish some of these risks. Regulations that limits the exposure of the banks to the volatility in equity and real estate markets could help insulate the banldng system from the potential bubbles associated with sizeable capital inflows. Inflation The previous section showed how banks may help to channel some of the credits to sectors that have no access to international capital markets but may simultaneously contribute to the magnification of financial difficulties stemming from a slowdown of capital inflows. Central bank's typical reaction to a bout of capital inflows is, however, fear of inflation. This would be fully justified by the logic that capital inflows episode could end up in higher inflation. However, central banks worry because they see monetary aggregates sharply rising, especially when the exchange rate is not allowed to appreciate. Surge in capital inflows particularly sh011-term and portfolio flows are likely associated with credit booms and asset price bubbles. In other words, to invest in bonds and stocks you need to have a stock and bond markets in the first place. 12 The concern is that the rise in money supply will fuel inflation. Monetarists like Milton Friedman have said that inflation is always and everywhere a monetary phenomenon. Higher money balances would induce people to spend more. By this he means that there can be only persistent increase in prices if there is a persistent growth of money. The centr~l bank could increase money supply at the beginning of the cycle to prevent the exchange rate not to appreciate. Financial trouble can be avoided but a sequence of capital inflows cycles-where, first, capital flows in, and then those flows subside (and vise versa)-is likely to put the economy on an inflationary path if intervention result in increase in money supply. This need not occur, however, if the assistance to the potentially affected financial institutions comes from the government finances rather than through a monetary expansion by the central bank (Calvo, 1998). African examples of this fear of monetary expansion associated with the rise in capital inflows and increased purchase of foreign exchange were evident in the capital inflows episode of Kenya in late 1993-1994. These fears led monetary authorities to engage in largescale sterilisation policies through the open market sale of government securities (i.e., through absorption of domestic money in exchange for domestic public debt), increases in reserve requirement, or both. 4) The role ofthe exchange rate Countries that experience a capital inflow may opt to let the nominal exchange rate appreciate. The capital inflows' potential inflationary can be completely avoided by refraining from intervention in the foreign exchange market. But at the cost of appreciating the excqange rate. If the authorities allow the nominal exchange rate to appreciate in response to capital inflows, the profitability of the traded goods sector will obviously be affected adversely. Export become more expensive compared to the import. The sectoral profitability between the traded and non-traded goods sectors would be reversed, which in tum, may damage strategic sectors of the economy. Beside, allowing the exchange rate to fluctuate introduces uncertainty . Given the change in reserve [R] and GNP level, the increase in capital inflows results in a ,. current account deficit and higher expenditure. To counteract this, authorities allow some appreciation. It follows that the same increase in expenditure will now call for a sharper appreciation rate. Under these circumstances, producers of nontradable goods are secured 13 from currency appreciation, but those in the tradable sector will suffer more due to the sharper currency appreciation. An appreciation of the nominal exchange rate in response to increased demand for domestic assets can take place without the need for any policy action in more flexible exchange rate regimes. However, if the prevailing arrangement is one where the rate is set by the authorities (i.e peg, crawling peg, narrow band), then, at some point, a decision has to be made whether a realignment will be undertaken (Reinhart et aI, 1996). There are several advantages allowing the nominal exchange rate to appreciate during period of heavy capital inflows (see Calvo, Leiderman, and Reinhart, 1994). First, it insulates the money supply, domestic credits, and the banking system from the inflows; this is particularly desirable if the inflows' are perceived to be of a highly reversible nature. Second, if the economic fundamentals warrant a real exchange rate appreciation, the adjustment comes via the exchange rate and not via higher inflation. Third, and related to the previous point, because of the pass-through from the exchange rate to domestic prices, an appreciation may help reduce inflation. Beside, the reasons given above for allowing the exchange rate to adjust in response to as shift in capital flows, there are other motives for allowing the exchange rate to fluctuate more freely in the presence of the large capital inflows. First, it introduces some uncertainty that may well discourage some of the purely speculative (and highly reversible) inflows. Bacchetta and van Wincoop (1994) argue, in the context of a two-country model, that an increase in exchange rate uncertainty creates a bias toward the domestic assets (since the rate of return on the foreign assets is now more uncertain), dampens the sensitivity of .the current account to most types of shocks, and reduces net capital flows. Indeed, the higher uncertainty acts like a Tobin tax. In the event of capital outflows, the greater flexibility takes some of the pressure off foreign exchange reserves. Second, it grants the monetary authorities a greater degree of interdependence and permits them to exercise more control over the monetary aggregates. The main disadvantage of a pure float is that massive capital inflows may induce steep and abrupt movements in the real exchange rate, which in tum, may impose a substantial adjustment burden on the economy. In particular, the concern in many countries has been that real appreciation will harm strategic sectors of the economy, like the nontraditional export sector. This result may be due to the existence of incomplete markets, to the extent that financial markets do not provide enough instruments to hedge against such uncertainty. 14 5) Current account In general, an upsurge in capital inflows requires an increase in the current account deficit. The inflow leads to an increase in aggregate demand, which pushes up infl,~.tion so that the' exchange rate appreciates and current account deficit rises as imports rise and exports fali. The rise in aggregate demand is reflected in an increase in investment and consumption and thus a decline in savings so that the investment-saving gap widens. Currents account deficit measure the extent of total external financing and signify a strong willingness on the part of non-residents to lend· to the country. In the 1990s, equity and portfolio investments began to oveliake direct investment, loans, and trade credit as the main fonns of external financing. Reisen (1997) suggests that offers of financing by non-residents should be resisted when they cause unsustainable currency appreciation, excessive risk-taking in the banking system, and a sharp drop in private savings, all of which raise sore points in tenns of the fundamentals-sentimentals dichotomy. In principle, large current account deficit are a cause for concern and even more when aidfinanced. Large deficits could be problematic if there is a sudden and unexpected slowdown of capital inflows and the government cannot offset it by running down international reserves. A sudden cut in the CAD will lead to a reduction in domestic expenditure (or absorption). Thus, the relative price of nontradable goods is likely to fall. A CAD usually reflects the existence of new loans from the rest of the world (except when the deficit is fully financed by running down international reserves, recall equation [2]). All the above mentioned mUltiple concerns have led policymakers in countries experiencing a surge in capital inflows to react by actively implementing a spectrum of policies (see Calvo, Leidennan, and Rehinhart, 1993 and 1994; Schadler et aI, 1993; and Montiel, 1995). 2.3) Capital Inflows Literature Review Controversy persists on the role of capital flows in boosting development and inducing macroeconomic instability. Major neoclassical arguments emphasise gains from financial liberalisation and international capital mobility based on the belief in efficient market. Under the efficient markets hypothesis, it would be pointless to discuss capital account controls. 15 Liberalisation is always perceived as beneficial to investors. They argue that financial market can contribute to enhancing efficiency in resource allocation, easing external constraint, technology transfer, and better management and business practices. They strongly support capital account liberalisation and thus against any of capital controls. They argue international capital movement increases the availability of foreign savings to supplement domestic resources and enable investors to diversify risks around the world (Edward, 1995). They argue capital controls limit international market opportunities and restrict domestic financial market competition that induces distortion and inefficiency in the financial system and economy as a whole. However these arguments are valid only with the assumption of 'efficient' financial markets. But tins is not true due to lack of better information system, herding and other market distortions (such as financial repression). The rationale for restricting international capital flows, by contrast, is grounded in the believe that market failures and distortions pervade capital markets around the world. One of the most cited distortions is that of information asymmetries. Information asymmetries are presented in goods markets, but it is in asset markets (money and capital markets) that they become pronounced. Although a finn producing a good is more knowledgeable about the quality of the product than is the buyer, it is not so difficult for a buyer to monitor the quality of, say, the computer clnps produced in Taiwan or in Thailand. Banks, for example, exist because of their superior knowledge about the value of the firms to which they lend. Problems of asymmetric information are more extensive in international capital markets, where geographical and cultural differences make harder the task of obtaining information. Traditionally, most developing countries have kept strong capital controls due to several reasons including balance of payment problem, macroeconomic stability, and national development (Johnston and Tamirisa, 1998). Among others, it is interesting that capital controls can be used for economic development in some cases. Several countries like Japan and Korea that achieved rapid economic development intentionally adopt capital controls in line with broad national development strategy and planning (Collier and Mayer, 1989). If capital outflows are strongly controlled then it would obviously help to increase savings and thus investment. And they are more important because capital inflows are not that much in the early period of development. The government as well can earn revenue from capital controls, which enables it to implement expansionary policy. Besides, capital controls as a form of 16 exchange rate control are likely to manage exchange rate in such a way that it would maintain foreign reserves, manipulate terms of trade for trade growth and stabilise economy. Capital account liberalisation can just lead to more instability. Investors may overreact to shocks, withdrawing en masse from countries at economic problems. Despite their benefits, surges in capital inflows also create new macroeconomic problems for the recipients, more pressure on current account deficit due to increasing domestic demand, inflationary effects, weaker monetary control, real exchange rate appreciation and the most problematic is the vulnerability to reversals. Most of all, investors' 'herd' behaviour brings out serious volatility, not related to the real economic fundamentals (Kim and Wei, 1999). These theories can support capital controls to address the economic instability due to rapid movement of shortterm capital. This is the message of several theoretical papers emphasising imperfect information. Latin America and East Asian countries experienced a strong real appreciation of their currencies followed by balance of payment crisis after a surge of massive capital inflows. With open capital market, national government can not have the autonomy of macroeconomic policy. In particular, it is difficult to adopt expansionary monetary policy with freedom of capital movement due to the possibility of capital outflow and attack on the currency, which leads government to restore to restrictions. This is why Paul Krugman, for example, has argued that emergency controls on capital outflows may be a good choice at times of severe speculative attacks fi:om domestic and foreign speculators. Although it is often argued that controls on capital outflows are also likely to reduce capital inflows, many of those who oppose restrictions on outflows favor controls on inflows emphasizing the "precautionary" role of these controls, in contrast to the destabilizing effect of controls on outflows. The list of those supporting restrictions on capital inflows has grown larger in the last decade, with the most ardent supporters including Stiglitz and Barry Eichengreen. Besides, Sl11ce all of the markets are not perfect and efficient in reality, the positive relationship between the capital account liberalization and economic efficiency is hard to be . justified. When there is a trade barrier free movement of international capital can result in misallocation of capital and difference in tax rate on capital generates capital movement to evade tax, not enhancing efficiency at all (Cooper, 1999). That is, according to 'theory of the second best' there is no reason that free international movement will enhance the efficiency. 17 Empirical studies also show it is hard to justify the neoclassical liberalization and economic growth (Rodrik, 1998). As we have seen, there are enough arguments for capital controls with the concern about instability and macroeconomic management. The important role of the government in financial market including capital control (such as taxes and restrictive monetary policy in the form of sterilisation) has been already well acknowledged by many theorists (Saxena, 1999). This follows from the past experience that in Chile and Thailand where weak banking system and assets bubble played a major role in bringing crisis in these countries. In conclusion, it is argued that capital account liberalization. should be preceded by the trade reform (McKinnon, 1991; Edward, 1992a) as well as financial sector reform. The scope of sequencing broadened to the issue of policy responses to the macroeconomic instability (Edward, 1992b), emphasizing the appropriate domestic policy and prudent macroeconomic management. 18 Chapter Three: Capital Inflows and their Macroeconomic Impact TIns chapter looks into the magnitude and the nature of capital inflows and· analyses the macroeconomic impacts of the surge of the capital inflows with regard to the countries under discussion. It also provides summary conclusion for each macroeconomic impact discussed. 3.1 Capital Inflows: cross-country comparisons The description of capital flow experience in this section is based on three countries, namely Chile, Thailand and Kenya. The capital inflow data is from the IMF's World Economic Outlook data set. Capital flows in this data set are classified into five categories: portfolio flows (bonds and equity), shod-term flows, FDI (foreign direct investment), other long-term flows, and errors and omissions. The capital inflows experience for each of the countries since 1990 -1998 is described as follows in Table 3.1 below. Table 3.1: Composition of Capital Flows since 1990-1998 (in percentage ofGDP) 1990 1991 1992 1993 1994 1995 1996 1997 1998 Chile Net Private Capital Flows 9.9 5.5 6.9 7.2 11.2 6.7 10.4 9.9 3.6 ShOit-term Capital Flows 4.8 1.4 4.6 2.4 2.6 1.4 0.1 Direct investment (FDI) 0.7 1.7 1.7 2.0 3.7 2.9 6.0 5.8 6.2 Net portfolio flows ( with errors and omissions) 1.0 1.7 1.9 1.6 0.8 0.3 0.6 2.5 -2.7 Thailand Net Private Capital Flows 12.8 10.7 8.7 8.3 8.6 12.9 5.7 -7.6 -16.9 ShOit-tenTI Capital Flows 7.6 7.5 5.0 3.0 5.5 7.7 5.6 Direct investment (FDI) 1.9 1.7 1.5 1.3 1.0 1.3 1.4 2.5 6.3 Net portfolio flows ( with errors and omissions) 2.3 0.3 0.0 4.2 1.6 1.2 1.8 2.2 3.0 Kenya Net Private Capital Flows 2.4 4.3 1.0 0.7 -3.2 -1.1 -1.2 0.1 -0.9 Short-tenTI Capital Flows 2.2 -0.7 4.1 2.5 -2.8 3.7 2.3 Direct investment (FDI) 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 Net portfolio flows ( with errors and omissions) 2.2 -4.2 0.2 1.6 4.0 3.7 6.5 4.0 3.9 Source: International Monetary Fund, World Economic Outlook Represents that data is not available 19 As can be seen from the table, on average since 1990 to 1995 Thailand attracted more private capital inflows compered to Chile. In 1984, the baht was devalued by 14.8 percent, government undertook export-led growth policies and gave priority to promoting capital flows through tax and institutional reforms while developing its financial market. This policy together with large positive interest differentials and a fixed exchange rate promoted large net capital inflows. However, after 1995 onward Thailand suffered more capital out flows, which put Thai baht in crisis in 1997. In the early 1990s, Chile experienced a surge in capital inflows (see figure 3.1 below). Large capital inflows began in 1990, attracted by two domestic factors (and several external ones); a) the successful transition to democratic rule, which arguably reduced 'the country risk premium, and b) a policy of tight money and high interest rates launched by the newly independent Central Bank. On average Chile received net capital inflows of7.9 percent of the GDP since 1990 to 1998, raising demand for the peso and putting up ward pressure on the real exchange rate (Edward, 1999). These led policy makers to introduce specific capital controls. Whereas Thailand due to the outflows, it received only net capital inflows of 4.8 percent of the GDP during this period. However, Kenya received only 0.2 percent, even less than a half percent of her GDP. Even if Kenyan authorities tried to eliminate barriers to entry in early 1990s, it doesn't seem to attract more capital inflows instead the country experienced capital flight. The trends ofthe inflows are shown in figure 3.1 below. Thailand actually registered larger share of short-term flows over the period as a whole compared to Chile and Kenya. This volatile short-term inflows conseqp.ent speculative attack on local currency. We can also note that the short-term inflows in Chile decreased since 1995 after strengthening of the capital control whereas the long-term FDI has increased dramatically. The use of capital controls in Chile has been part of a broad program of economic reforms involving a coherent set macroeconomic and structural policies implemented consistently throughout the period. The Chilean authority goals when they passed capital controls were to slow down the volume of capital flow and to tile the shortterm composition of the flows into long-tenn investment, so as to reduce real exchange rate appreciation. And also to allow Central Bank to maintain a high differential between domestic and international interest rates, and thus to conduct an independent monetary policy. 20 In Thailand private capital flows started to diminish in early 1997. The Thai government intervened heavily to SUppOlt the peg when the baht came under serious attack for the first time in February 1997 by pumping a large volume of liquidity. This lead to further capital outflows and a decline in foreign exchange reserves. In mid-May, Thailand announced capital controls. Finally, on July 2, 1997, the government allowed baht devaluation. Figure 3.1 Net R-ivate Capital Row s 15~ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ~ \_\_\_\_\_\_\_\_\_\_\_\_ ~ ~ 10-1-------------, o \ 2 5 -1----- ---~-- - ~ ="'""';;>""Il!.,~=~~." .. ,=",",=,=w"''''''''''~'''''''''"'''=",,~.,. \ C) 0 - I------r--:; -r--1----r---r-·T--·"..:,~~--T~!O.,.:.~~,Jrw~r.';'l.1:' .. t;;;;:--· .£9 L(') 0 ..- N ("i) ·"';t.~""ih -~- o ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ Q; -10 -I-------------------------~-- \_\_\_ Chile • Thailand ~~=~.,,~"w'u Kenya 0- ~ -15- \ \ -20\_L-\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ~ year '-----\_.---------\_.\_---------------- Source: All the figures in the paper are graphed the data in the Appendix Following a collapse of tea and coffee prices in 1987, Kenya was left with a huge budget deficit, a rapidly deteriorating current account positions, and a severe shortage of foreign exchange. Also Real GDP growth slowed and inflation had increased,' despite extensive price controls. By 1989, it became evident that without foreign currencies and structural refonns, Kenya would experience a severe economic downturn. To avoid a severe recession, the government embarked on a wide-ranging capital liberalisation program aiming at attracting foreign savings in 1991. The program intended to remove rigidities in financial sectors by freeing foreign currency transactions, and relaxing and then dismantling credit ceilings and interest rate controls. The same year, some enterprises were pennitted to hold foreign cU11'encies. 21 Despite the introductions of these liberalisation measures, in effect, it rather led to further capital outflows till 1994. During the same time, the economy exp~rienced \_ a sharp downturn from late 1991 onward with sharp dramatic increase in inflation. The misappropriation of public funds led to a further deterioration of economic conditions and by the early 1993 the . . economy was in crisis. However, after some corrections, it showed some reversal in capital outflows. Conclusion The aim of this section was to show the evidence that there were capital inflows indeed in these countries before speaking of their macroeconomic impacts. As we have witnessed there was surge of capital inflows, which posed macroeconomic challenge to these countries such as real appreciation of exchange rate and some other impacts that I will discuss below. 3.2 Macroeconomic Impact 3.2.1 Impact on aggregate demand National accounting identities equate the trade deficit to the excess of domestic absorption over production. The additional capital inflows can be used to accumulate foreign exchange reserves, increase domestic investment, or increase domestic consumption. Here the CAD and reserve accumulation is calculated as share of capital account (inclusive of grants & errors and omissions). Table 3.2: use of capital inflows during surge periods a! Allocation of capital account (%) Use of Domestic Marginal Investment Absorption (%) Impact of Capital flows Reserve Net Factor Net Resource Accumulation Payments balance deficit Consumption Inv't Chile 57 71 -28 21 79 34 Thailand 41 10 48 -196 296 174. a/ Surge periods are country-specific based on observed capital inflows profiles. For Chile the surge period is 1989 to 1993 while in Thailand the surge period is 1988 to 1992. Due to data limitation Kenya is not included in this part. Source: The World Bank and IMF 22 Table 3.2 provides evidence on this issue for Chile and Thailand experiencing the large surges of private inflows relative to their economies. The first three columns of this table express reserve accumulation, net factor payments, and the trade deficit as share-s of the capital inflows for each countries during their respective surge period. Column 4 and 5 refer to the domestic absorption use of the trade deficit financed with capital inflows by examining how increase in absorption between "pre-surge" and "surge" periods were allocated to increase in consumption (that is decreases in domestic saving) and investment. Column 6 estimates the fraction of capital inflows used for investment purposes once reserve accumulation and consumption leakages are deducted. Large proportions of the capital inflows have been used for reserve accumulation rather than CAD finances. The accumulation of reserve accounted on average 49 % of the inflows for both Chile and Thailand. That is to say, through sterilisation large proportions (49%) of the capital inflows were deposited as a reserve in the Central Banle In Chile, 57 per cent of the capital inflows were used for reserve accumulation whereas only 41 per cent is used in Thailand for reserve accumulation. They also differ in regard to change in the composition of absorption (column 4 and 5). Both in Chile and Thailand increase in absorption during this period were dominated by investment. Investment increment indicates that capital inflow that was channelled to current account was used to finance investment rather than consumption. But the changes in the composition of absorption were heavily biased in favour of investment in Thailand. In Thailand, during 1988- 91 spending cuts were introduced on nontradables, tight fiscal policy, so as to lower aggregate demand and curb the inflationary impact of the inflows whereas Chile appears to have experienced consumption booms led by private sector consumption. The last column shows the surge of capital inflows was favouring investment in Thailand than in Chile. In other wards, this last column on "marginal investment" measures the impact of additional unit of capital inflows, that was not used for reserve accumulation, and on investment. Conclusion The additional capital inflows can be used to accumulate foreign exchange reserves, increase domestic investment, or consumption. Therefore, from the above we can conclude'that both for Chile and Thailand used 49 % of the inflows as reserves accumulation and rest is used for CAD finances. The inflows in both cases favour investment but more importantly Thailand. 23 3.2.2 Impact on exchange rate My hypothesis is that when capital inflows such as foreign direct investment (FDI) and portfolio investments (short-term capital inflows) increase in a country, to under take investments necessitate buying the local currency and hence the demand for the local currency increase which, in tum, lead to appreciation of local currency. An import relatively become cheap and exports expensive. The competitiveness of the export in the international market falls. In Chile, the financial liberalisation in the mid 1970s resulted in speculative attach on domestic currencies and financial crisis in 1982 and 83. III December 1983, a peg regime replaced the fixed exchange rate aiming at maintaining a constant level of real exchange rate against the U.S. dollar. In December 1989, after 17 years of a military regime, a new government was democratically elected and also the domestic interest rates in Chile meant high compared to international levels. Thus, Chile regained access to international markets and capital begun to flow in to the country in the early 1990s, raising demand for the peso and putting up ward pressure on the real exchange rate (Edward, 1999) see also Figure 3.2. Chile RER rate followed an appreciation trend of 4% a year during the 1991-94 (Laurens, B., and Cardoso, J., 1998). By late 1990, exporters had begun to complain that the rapid strengthening of the peso in real terms (it had appreciated by more than 20% since 1985) was negatively affecting their ability to compete in international markets (Cowan and De Gregorio, 1997, p.3). In Thailand, capital inflows accelerated during 1988. The Thai economy started showing signs of overheating in mid-1993, despite the authorities' tight financial policies. Demand pressures were manifested in higher inflation and some widening of the current account deficit, prompting the authorities to tighten the monetary and nscal policies. The combinafio-n of a pegged exchange rate since 1984 and highly liberalised capital inflows, along with large interest rate differentials, created strong surge of volatile net capital inflows. The inflows were predominantly short-term (about 60% of the total in 1993), mainly in the form of short-tenn borrowing by banks (Ariyoshi, 2000). 24 Since late 80s the gradual appreciation continued because Thai inflation was higher than world market, leading to appreciation of the real exchange rate which was not sufficiently corrected by 1981 devaluation. This appreciation was in favour of importers, and they formed a strong lobby to maintain the increasingly overvalued exchange rate (Jansen, 1997). Figure 3.2 Real effective exchange rate 60 r--------------------------------------. 50~ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ -,~ \_\_\_\_\_\_\_\_\_\_\_ 1 "5 t ~~ 20 - a- --) -- ',~, i ro -:L~'~99"\_'992 -20 \_ \_\_~ it:~~~-'-ir.::~B ~ \_\_ ~'-;;-]'C \_\_\_\_ I -30 \_ ------------ \---l-' ----I iff:;. -40L-\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ~ \_\_\_\_ ~ year -o-Chile ='l.k.-'~ Thailand \_~Kenya '-----------------------------------------------' Remarkably little real exchange movement were observed, particularly given the size of the inflows. In fact, the real exchange appreciated slightly in the early 1990s, and began appreciating very slowly since (see figure 3.2 above). However, after the currency crisis RER fluctuated tremendously After more than a decade of exchange rate stability and impressive economic growth, the growing domestic and external imbalances and the emerging of banking problems since late 1996 the Thai baht came under severe speculative pressure in May 1997. In 1996, export growth began to slow down after growing 20 percent in 1995. Thai export contracted by 1 percent in 1996. This was mainly due to loss of competitiveness associated with Thai baht appreciation on real effective basis. The authorities imposed sever capital controls on May 15, 1997, to stabilise the foreign exchange market and stem speculative attacks on the baht. Financial institutions were asked to refrain from transactions with non-residents, such as baht lending through swaps and sales of baht against foreign currencies, that could facilitates a . build-up of baht positions in the offshore market. Later (in June), all such transactions were suspended. In July the central bank introduced a two-tier exchange rate system. Restrictions were eliminated on January 30, 1998. 25 In Kenya, the financial liberalisation was accompanied by a rise in real interest rates, often from very negative levels. During recent years, Kenya has liberalised remaining exch@g<; liberalisation of outflows, which should have reduced net inflows. However, outflows liberalisation may also lead to larger inflows of capital. A number of authors including Williamson (1991), and Laban and Larrain (1998), and Laurens and Cardoso (1998) in the case of Chile, have argued that liberalisation of outflows by reducing uncertainty of investors in the country and by lowering domestic asset prices can also increase capital inflows (Simone, et-al, 1999, p. 9). In Thailand, capital inflows were promoted at a relatively early stage while outflows were liberalised only gradually. Given the limited policy options, the authorities attempted to cope with capital inflows through a combination of monetalY, prudential, and market based capital control measures. Liberalisation of capital outflows was allowed in April 1991. Residents were permitted to export capital for investment purposes and so as to counteract the inflows. In effect the net capital inflows started to decline gradually. However, the net capital inflows increased sharply in 1994 reaching the same level as 90s. Despite this capital outflows liberalisation the net capital inflows has increased which may tell us that allowing capital repatriation has attracted more inflows busting the investors' confidence. In 1993, 60 per cent of the inflows were short-term borrowing by banks. Due to volatility behaviour of these inflows, the Thai baht also came under speculative pressures in early 1995 and sever speculative pressure in May 1997. Before 1997, the capital account had been almost fully liberalised on the inflow side, except for the reserve requirement on shortterm foreign borrowing, while outflows were liberalised only gradually. The authorities imposed capital controls on May 15, 1997. The control did not prevent outflows as the result of the sharp rise in the spread between the onshore and offshore interest rates created arbitrage opportunities, and thus incentives forcircumyenting the controls (Ariyoshi, 2000). In Kenya, contrarily all the restrictions in capital account were removed in order to attract more capital to the country to boost the economy which include measures that allow domestic investors to invest part of their capitals abroad. However, even if some capital outflows were permitted, capital didn't seem to come in. Instead, the wide-ranging liberalisation has increased the country's vulnerability by providing legal chalmels for capital flight. Since 1993, thus, the net private capital inflows remained negative showing nlOre capital outflows 39 than inflows. It is also important to note that capital inflows in this country are more related with the economic perfOlmance such as real GDP growth. Monetary Policy All of the recipient countries have implemented restrictive monetary policy in the form of sterilised intervention, orland increases in reserve requirement. Sterilisation Sterilisation-the exchange of domestic securities for foreign exchange-can help to insulate the domestic economy from the macroeconomic effects of capital inflows. Sterilisation would keep domestic interest rates higher than they would be in the absence of the sterilisation. At worst, this may provide incentives for further short-term inflows .. In addition, sterilisation results in an increase in the public debt and entails quasi-fiscal costs to the extent that the interest rate on domestic bonds is higher than that on foreign exchange reserves. The magnitude of these costs will be greater the higher the degree of capital mobility and the larger the gap between domestic and foreign rates of return. Thus the feasibility of this policy is also at issue. Even if sterilisation succeeds in limiting domestic monetary expansion, it may not insulate the economy from the effects of capital inflows. In Chile the initial policy response to the capital inflows include heavy sterilisation by means of open market operations (OMO), particularly in 1990-92. Index Central Bank bonds of several maturities were issued as part of large open-market operations. Sterilisation also attempted through the transfer of public sector deposits out of commercial banks to the Central Bank. While sterilisation of most of the inter/ention helped prevent a monetary expansion, this policy imposed sizeable costs on the central bank, reflecting the differential between the interest cost of sterilisation and the return on foreign assts (roughly 1 per cent of GDP annually during the 1990s) (Ariyoshi, 2000, 47). In Thailand, policy responses to the capital inflows include heavy sterilisation by means of open market operations, particularly in 1989-91. The discount rate was increased sharply and . commercial banks' access to refinancing facilities at the Central Bank curtailed. Sterilisation also attempt through transfer of public sector deposits out of commercial banks to Central Banlc. In 1987-92, a sharp increase in the share of government deposits held at the Central 40 Bank rather than at commercial bank were introduced. Sterilisation efforts were reduced in mid-1993. Overall, sterilisation together with regulatory controls imposed on capital inflows in 1995-96 seem to have reduced net capital inflows into Thailand particularly reducing the share of short-term net inflows from 62 per cent of the total capital inflows in 1995 to 32 per cent in 1996 and lengthened the maturity of the loans (the share of long-term loans rose from 14 per cent in 1995 to 34.3 per cent in 1996). It is difficult, however, to isolate the impact of the controls from those of the deterioration in investor confidence and other external factors (IBID,52). In Kenya, fear of monetary expansion associated with the rise in capital inflows and increased purchases of foreign exchange were evident in the capital inflow episodes in late 1993-94. These fears led monetary authorities to engage in large-scale sterilisation policies through the open market sale of government securities, increases in reserve requirement, or both (Calvo, 1998, p 17). That means absorption of domestic money in exchange for domestic public debt (e.g. Treasury Bills). Sterilisation is therefore equivalent to the central bank acquiring international reserves in exchange for domestic public debt. Investors may require an interest rate premium on domestic public debt. Therefore, the interest rate on domestic public debt will exceed that of international reserves, generating a larger fiscal deficit (since those deficits are usually borne by the central bank, they are called quasi-fiscal deficits) and stimulating additional inflows .. In the case of Kenya, the rising burden of servicing the public debt became evident, as in the 1993/94 budget domestic interest payments were 1.5 per cent of GDP above the level that was programmed. The interest differential led Kenya in 1994 to stop sterilisation (Calvo, 1998). Unremunerated Reserve Requirement (URR) The introduction of the reserve requirement was motivated by macroeconomic and prudential considerations. The URR is an indirect, price-based measure in the form of a one-year compulsory deposit at the central bank of a fraction of certain capital inflows. Its purpose is to reduce certain capital inflows by increasing their cost. 41 Regarding the prudential considerations, the measure was expected to discourage short-term external borrowing. By reducing the volume of external short-term debt the measure was expected to enhance fmancial stability and, thus, reduce external vulnerability (Le Fort, 1996). The reserve requirement increases the cost of foreign borrowing) or conversely reduces the rate of return for a foreign investor), thus filling all or part the gap between domestic and international interest rates (Cardoso, 1998). The main instrument for restricting capital flows in Chile has been the URR on capital flows. Chile introduced restrictions on capital inflows (URR) in June 1991. Initially, all portfolio inflows were subject to a 20% reserve deposit that earned no interest. In July 1992 the reserve requirement on portfolio inflows was raised to 30%, and the holding period was set at one year. During the same month, reserve requirements were extended to trade credits and loans related to direct foreign investment. In 1995 capital controls were extended to cover international issues of bonds. With markets in turmoil and the Chilean peso under attack, in June 1998 the reserve requirement was lowered to 10%, and in September of that year reserve requirement was eliminated. Since domestic interest rates in Chile meant high compared to international levels reserve requirement can be expected to help to minimise exchange rate appreciation pressures in the face of substantial capital inflows and to enhance the autonomy of monetary policy so that the effect of a tight monetary policy on the exchange rate would minimised. As such, the Chilean authorities have argued that, due to the non-remunerated reserve requirement, Chile's RER rate has appreciated less than in other countries of the region. However, the URR has received a lot of attentions and has been subject to an intensive debate. Since the URR was not universally applied to all foreign capital inflows, the regulations tended to lose their effectiveness over time, as ways circumventing them were developed channelling the inflows through exempted windows. Several studies have been done to test the effectiveness of the reserve requirement particularly by Valdes and Soto (1996), Soto (1997), and Edward (1998a, and 1998c). But, the literatures do not provide empirical evidence that would suppOli a prolonged effect of the reserve requirement on the real exchange rate. Cardoso and Laurens (1998) find that the introduction URR had only temporary effects on the composition of external financing, which 42 is consistent with the view that the private sector will attempt to circumvent any restrictions to capital movements. Quirk and Evans (1995) observe that net short-term private capital inflows recorded in the balance of payments decreased in 1991 with the introduction ofURR as part of capital control. However, they also observe that 'net errors and omissions' and the estimated trade misinvoicing also increased sharply in the same year. One possible i~terpretation of that evidence is that an increase in unrecorded short-term flows reflecting an attempt by the private sector to circumvent the capital restriction. Valdes and Soto (1996) use a capital control index in the context of a single equation model. The results suggest that there is no evidence of a positive long-run effect. Soto (1997) also finds a negative relationship between the reserve requirement and exchange tate volatility, 30 percent reserve requirement would reduce volatility by approximately 20 percent. It is measures implemented by Chile whose specific objective was to reduce market volatility. While Soto finds a small and short-lived positive effect on the RER, Edwards concludes that the behavior of the RER was not affected by the capital controls. In Thailand, in August 1995 the authorities started to introduce restrictions on capital inflows. The measures included a 7 per cent reserve requirement (held at the central bank) on nonresident baht accounts with less than one-year maturity and on finance companies' shorttenn foreign borrowing. Also reporting requirements were imposed for short foreign currency positions. Further tightening of the restrictions occurred from April to June 1996, when the 7 per cent reserve requirement was extended to new short-term offshore borrowing with maturity ofless than one year by commercial banks and BIBF banks. Tpward the end of 1996, all restrictions on foreign borrowing were eliminated (Ariyoshi, 2000). In Kenya, during the October 1993-March 1994 period the statutory cash ratio was raised in three steps from 12 per cent to 20 per cent. The increased sales of treasury bills and higher reserve requirements drove domestic interest rates higher. Since the Kenyan shilling was appreciating this period the rise in dollar returns was even more dramatic. By mid-1994 the high level of interest rates was increasing debt servicing costs, generating quasi-fiscal losses and stimulating additional inflows. At the same time, it was decided that intervention efforts would be scaled down considerably and the shilling was allowed to appreciate further (See Kimei, Mjerna, Tarimo, and Msutze, 1997). 43 Fiscal Policy Some countries have complemented monetary and exchange rate policies with fiscal measures, such as the taxation of capital inflows and/or a reduction in public expenditure. Taxes on short-term borrowing abroad were imposed in Chile in 1991. This policy conveys the powerful message that the authorities are concerned with short-term consequence speculative attack. Such policy can coexist with policies that encourage specially foreign direct investment. However, even if they can be effective in the short run, private sectors are quick to circumvent the taxation (Khan, 1995). In Kenya, to make short-term inflows costly authorities imposed distortionary taxation on short-term inflows (Calvo, 1998). This tax would increase the cost of borrowing and hence decrease returns from such inflows. As such, in effect reduce the amount of new shOli-term borrowing. Another policy reaction to capital inflows has been to tighten fiscal policy. In Thailand, during 1988-91 spending cuts were introduced on nontradables such as investment in infrastructure, so as to lower aggregate demand and curb the inflationary impact of the inflows (Montiel, 1998; Schadler, 1993; Reinhart, 1996). 4.2 Variations in Policy Responses This section examines policy measures with their respective policy objectives with which policies of each of these types were adopted by these countries during the capital inflows episode, as well as the modalities of their implementation. Measures to Impede Gross Inflows Controls, taxes, or other impediments to inflows were adopted in several countries, and they have taken many forms. The most straightforward of these consist of quantitative restrictions on foreign borrowing. For instance, Chile adopted URR in 1991. URR were supplemented with other measures to increase the cost of carrying foreign exchange liabilities. Chile' also adopted a stamp tax to foreign loans. Thailand has adopted URR in 1995 and in 1990 started taxing foreign borrowing. Chile sought to discourage inflows by increasing the risk associated with foreign borrowing. Chile adopted exchange rate band, and widened them during the 44 inflow episode. Chile pennitted extensive variation of the exchange rate within the band to stabilize the path of the exchange rate. In Kenya, in October 1993 to March 1994 12 % statutory cash ratio was increased to 20%. Encouragement for Gross Outflows Many countries relaxed controls on capital outflows when inflows were high. Thailand, Chile and Kenya removed a number of restrictions including measures such as explicitly pennitting residents to invest abroad, removing restrictions on repatriation of capital and interest by foreign direct investors (outflows were liberalized in Chile and Thailand in 1991 and 1992 in Kenya). Exchange Rate Policy In Chile and Kenya capital inflows have been associated with considerable real exchange rate appreciation than Thailand (Chile in 1990-1992 and again in 1994 whereas in Kenya in 1993 and 1994), possibly reflecting a greater weight given to inflation than a competitive target. In other words, allowing movement (appreciation) would absorb some of the pressure from capital inflows. Chile revalued its RER by 5 per cent in 1992 and revalued additional 10 % in 1994. Kenya in 1995 maintained flexible exchange rate system. Although various reasons account for the variations in the response of the real exchange rate, important differences in the composition of aggregate demand may play a key role in detennining whether the real exchange rate appreciate or not. In addition, differences in the domestic policy response are likely to play a key role in explaining the differences in real exchange rate behavior among the countries. Specifically, the behavior of public sector consumption influences the real exchange rate by affecting both the level and composition of aggregate demand. Other things being equal, the more restrained the fiscal stance at the time of capital inflows, the weaker the real exchange rate appreciation. There were fiscal spending contraction most markedly in Thailand during 1988-91, at the time of the inflows (see Schadler and others (1993)). In addition, comparatively effective sterilisation of capital inflows which was successful in limiting the expansion in credit and money aggregates and in aggregate demand may have contributed to the differences in real exchange rate behavior. 45 Measures to Restrict Base Money Growth By far all the countries examined here the most common response to the r~ceipt of capital inflows has been sterilized intervention. Thailand (1989-90) employed reduction of access to the discount window (repurchase market). Chile (1990-92) was particularly aggressive initially in their pursuit of sterilization when inflows accelerated, seeking to offset all effects of capital inflows on the monetary base, while Thailand was not so ambitious, seeking only to improve effects on the base (Reinhart and Dunaway, 1995). In Kenya aggregate money started increasing particularly in 1992-1993. Thus, To mitigate the effect the authority adopted large-scale sterilization. Since sterilization the aggregate money declined significantly (Table A2). In effect, the policy drove domestic interest rate even higher than before the inflows leading further inflows (Montiel, 1998). This suggests that whether desirable or not, sterilization remained a realistic option for these countries, at least in the short run. However, capital mobility is higher in the long run than short run as such sterilization may indeed represent only a temporary option in most cases. Restrictive Fiscal Policy Fiscal tightening was an important component of the policy response in Thailand particularly in 1988-93 than compared to other countries. For Thailand, it was one of these policies that helped to maintain moderate exchange rate appreciation compared to other countries under discussion before the crisis. Other countries followed taxing short-term borrowing from abroad as shown in Table A2. 4.3 Macroeconomic Outcomes How well did these measures succeed in preventing macroeconomic instability in the face of inflows? Significant real exchange rate appreciation was widespread except Thailand. Thailand also after the crisis experienced high volatility of exchange rate. Chile experienced a mild appreciation compered to volatility in Kenya. All of these countries examined avoided a' real appreciation, though the real exchange rate appreciated and fluctuated in Thailand after the baht crisis in 1997. Fiscal restraint appears to have played a role in avoiding stronger real appreciation as well as more rapid inflation (Montiel, 1998). 46 Increases in current account deficits have been common during inflows episode. This is because few governments will resist the temptation to let the real exchange rate appreciate as long as money is flowing in and financing their current account deficits. In the end, correcting overvaluation always leads to painful adjustment. Chile and Thailand registered larger deficits compared to Kenya. The current wave of capital inflows does seem to have been associated with an investment boom, particularly in Thailand during 1988-90. Thus, the increases in current account deficits have accommodated an increase in domestic saving, particularly sharp increase in Thailand (IBID). Chile appears to have experienced consumption booms led by private sector consumption. The Kenyan economy has been improving since 1993 to 1995. 4.4 Policy Lesson A systematic attempt to draw lessons about effectiveness based on these countries' experience would require linking individual policies to their outcomes in particular countries. The lessons to be reported in this section would be those that can be crudely inferred from cross-country patterns of policy choices and illacroeconomic outcomes. Administrative controls (capital controls and liberalisation of outflows) After the capital controls were imposed, inflows have slowed both in Chile and Thailand suggests that controls can work, at least in some cases and at least temporarily. Including Kenya, moreover, controls have been credited altering the composition of inflows in favour of those with longer maturities (Calvo; Montiel, 1998). With regard to liberalisation of outflows, it is clear from country experience that substantial inflows followed the removal of restrictions on outflows-as discussed by Bartolini and Drazen (1997). This is consistent with the view that the removal of restrictions on outflows simply attracts additional inflows. Flexible Exchange rate policy Concerning use of the exchange rate to achieve the same result, the lessons are several. Flexible exchange rate allows nominal exchange rate to movements to absorb some of the pressures exerted by capital inflows in the foreign exchange markets. Nominal exchange rate adjustment was essentially confined to two countries (Chile and Kenya). Thailand restrained itself from this policy because of fear of huge deficit it may consequence. 47 However, Countries that adopted managed to avoid real appreciation over the course of the surge episode. The linle between real appreciation and the emergence of current account deficits is not sound. Avoiding real appreciation has not necessarily implied avoiding current account deficits. The emergence of large current account deficits was not. restricted to countries that experienced real appreciation. For instance, Thailand had large adverse movement in the current account balance with stable real exchange rates before the baht CrISIS. Monetary Policy All the countries adopted URR and it has helped to reduce gross capital inflows by favouring medium and long-term capital inflows. However, lessons concerning sterilised intervention are mixed. Indeed, all the countries registered an increase in domestic interest rates over the period of sterilisation (see Frankel, 1994). Finally, sterilisation does not seem to have completely insulated recipient economies from the effect of capital inflows (Montiel, 1998). Fiscal Policy Even if all the three countries adopted this policy, experience suggest that this has not proven to be flexible instlUment in responding to inflows. However, real appreciation was avoided in countries such as Thailand that tightened fiscal policy in response to inflows. The frequency of real appreciation elsewhere supports the implication of theory that, in the presence of capital inflows, the avoidance of real appreciation requires a fiscal contraction (Montiel, ~ 998, p 38). 4.5 Policy Implication The policy implication that can be learned from the analysis is related both to macroeconomic management of capital inflows and to the style and pace of capital account liberalization. We have seen from the analysis, that capital flows to these countries pose two clearly related problems. The first one is that there was a stlUctural decline in capital inflows. The second, is . the nature of the inflows to be pro-cyclical and short-termist. I will therefore divide my policy suggestions from these two dimensions. Firstly, it is indispensable encouraging a recovery of private flows to these countries, especially long-term ones. Portfolio flows are generally the 48 most volatile private capital inflows and thus stabilising those already present is therefore as important as attracting additional flows. Secondly, authorities should use appropriate macroeconomic measures to diminish the pro-cyclical and short: termist of such flows. The objective of sustaining economic growth in the face of volatile capital inflows requires the use of a battery of policy instruments. There is a broader consensus that international capital mobility is a necessary component of the development process. However, this view mostly take no account of important real-world conditions such as informational bottlenecks, domestic market structure, the volume and timing of financing, and its costs and volatility. Therefore, with regard the style and pace of liberalization, slow and gradual liberalization of the capital account is desira.ble especially in the early stage of development, and should be done when the economy is properly stabilised, and has a healthy financial sector. There is a lesson that can be learned from Kenya's past experience that rapid capital account liberalisation does expose to a greater degree of volatility. Conclusion In sum, the above discussion has highlighted that the risks associated with capital inflows create policy dilemmas. There is no a single best policy for managing the inflows. The appropriate policy mix will depend on the nature of the inflows, their causes, and the macroeconomic and policy climate of the recipient country. Although no single policy exists for all countries, to limit some of the risks associated with short-tenn inflows, a reasonable sequencing of policies would consist of those flows through sterilised intervention, greater exchange rate flexibility, and/or increased marginal reserve requirements, accompanied by allowing appreciation of the currency. 49 Chapter Five: Conclusion As we have seen there were massive surge of capital inflows in these countries. These capital inflows posed macroeconomic challenge to these countries such as real appreciation of exchange rate, huge current account deficits and fluctuating economic growth. These macroeconomic impacts necessitated governments to undertake some policy measures to overcome these challenges. Some of these macroeconomic impacts and policies are summarised below. Macroeconomic impact Impact on aggregate demand, as we have seen, the additional capital inflows can be used to accumulate foreign exchange reserves, increase domestic investment, or increase domestic consumption. In other words, capital inflows can be used to finance either current account deficit and/or to accumulate reserves (see equation 2). Large proportions of the capital inflows have been used for reserve accumulation rather than CAD finances. The accumulation of reserve accounted on average 49 % of the inflows for both Chile and Thailand. The remaining 51 per cent were used to finance CUll'ent account deficit. The capital inflow that was channelled to current account was used to finance investment rather than consumption for both Chile and Thailand. The capital inflows in both cases favour investment but more importantly Thailand whereas in Kenya stimulated more consumption and surge in imports. Impact on real exchange rate, the capital inflows in Chile since ~arly 90s increased the demand for the peso and put up ward pressure RER. Chile RER rate followed an appreciation trend of 4 percent a year during the 1991-1994. By late 1990, exporters had begun to complain that the rapid strengthening of the peso in real terms was negatively affecting their ability to compete in international markets. Whereas in Thailand remarkably little real exchange movement was observed, given the size of the inflows. In fact, the real exchange appreciated slightly in the early 1990s, and began appreciating very slowly since. How~ver, after the currency crisis the stable RER fluctuated tremendously. Since 1991, the Kenya shilling was appreciating. Sterilisation and higher reserve requirement drove domestic interest rates higher, which stimulated additional inflows. Since then intervention efforts were scaled down and the shilling was allowed to appreciate further. 50 Impact on current account, In Chile the current account deteriorated since the inflows in 1990s. However, after the government introduced the capital control current account was improving till 1995. In 1995, after strengthening of the capital control on inflows the curren! account has improved but fluctuated immediately and showed some tendency to widen on average. In Thailand when capital inflows accelerated during 1988 lead to some widening of the current account deficit peaking at 8.5% of GDP in 1990 and declining since. In 1996, just before baht crisis, growth and investment deteriorated and exports declined sharply. Due to massive capital outflows the CAD was at surplus of on average 12.5 per cent of the GDP in 1998. Whereas the Kenyan current account had improved since 1989 to 1994 reaching surplus in year 1993 and 1994 due to cut in government expenditure but after 1994 it continued to fluctuate highly in deficit and deteriorated reaching worst in 1995. Impact on Inflation, Indeed, the arrival of large amounts of foreign capital was perceived as posing serious challenge to domestic macroeconomic management. The concern is that the rise in money supply will fuel inflation. Due to their policy measures countries have been successful in preserving macroeconomic stability with regard to inflation in the face of inflows, and all the countries managed to reduce the level of inflation. Impact on economic growth, since the surge of the capital inflows in Chile in 1989 economic growth was ranging between 5-7 per cent. However, after massive surge of volatile short-term capital inflows the RER had shown appreciation and current account deteriorated leading to deterioration of economic growth till 1994. Authorities strengthened its control on capital inflows in 1995. During the same year the growth showed improvement but fluctuated since then, following the same trend like the other countries under dis9ussion. In Thailand, capital inflows contributed to strong economIC performance. In 1988-90 witnessed an investment boom, very high growth and thereafter a slight deceleration took place. Growth fell to the 8-8.5% range. In 1995 growth and investment levels started to deteriorate in the face of an appreciating real exchange rate and capital inflows and exports declined sharply. In 1997, the country was in crisis. During this period like any other economic fundamentals the real GDP growth was in negative. For all the countri~s under review experienced economic turn down since 1996. But Thai economy was hit worst. 51 A significant step toward liberalization of CUlTent account and capital account transactions was made in Kenya in 1991. However, economic growth decelerated from 4.7 per cent in 1991 to -0.8 percent in 1992. To avoid recession govermnent embarked on wide-ranging liberalization program in 1992. The Kenyan economy has been improving since 1993 to 1995, however, continued to decline since. Macroeconomic policies The risks associated with capital inflows created policy dilemmas. There is no a single best policy for managing the inflows. The appropriate policy mix will depend on the nature of the inflows, their causes, and the macroeconomic and policy climate of the recipient country. Although no single policy exists for all countries, to limit some of the risks associated with short-tenn inflows, a reasonable sequencing of policies would consist of those flows through sterilised intervention, greater exchange rate flexibility, and/or increased marginal reserve requirements, accompanied by allowing appreciation of the CUlTency. 52 Statistical Appendix Table AI: Selected Economic Indicators 1985-1998 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1<;97 1998 Chile (In per cent of GOP) Net private capital flows -8.5 -7.7 -4.3 -0.7 6.6 9.9 5.5 6.9 7.2 11.2 6.7 10.4 9.9 3.6 Current account balance -8.6 -6.7 -3.6 -1.0 -2.5 -1.6 -0.3 -2.4 -5.8 -3.1 -2.0 -5.1 -4.9 -5.7 (Annual percentage change) Real GOP 2.5 5.6 6.6 7.3 9.9 3.7 8.0 12.3 7.0 5.7 10.6 7.4 7.6 3.4 Inflation 29.5 20.6 19.9 14.7 17.0 27.3 18.7 12.7 12.2 8.9 8.2 6.6 6 4.7 Thailand (In per cent of GOP) Net private capital flows 3.8 2.5 3.1 7 10.3 12.8 10.7 8.7 8.3 8.6 12.9 5.7 -7.6 -16.9 Current account balance -4.0 0.6 -0.7 -2.7 -3.5 -8.3 -7.5 -S.S -S.O -S.4 -7.9 -7.9 -1.9 12.4 (Annual percentage change) Real GOP 4.6 5.5 9.5 13.3 12.2 11.6 8.1 8.2 8.5 8.6 8.8 5.5 -1.3 -9.4 Inflation 2.4 1.8 2.5 3.8 5.4 6.6 4.7 3 4.6 4.6 7.5 4.8 7.6 4.3 Kenya (In per cent of GOP) Net private capital flows 3.7 3.1 1.9 1.5 2 2.4 4.3 1.0 0.7 -3.2 -1.1 -1.2 0.1 -0.9 Current account balance -1.9 -0.6 -6.3 -5.5 -7.1 -5.6 -1.1 -1.3 2.9 0.9 -4.5 -1.1 -3.5 -3.5 (Annual percentage change) Real GOP 4.3 7.2 5.9 6.2 4.7 4.7 1.4 -0.8 0.4 2.6 4.4 4.1 2.1 1.5 Inflation 13.0 4.8 7.6 11.2 12.9 20.6 14.6 33.6 54.6 6.6 6.9 10.8 8.3 2.5 Source: IMF Economic Outlook 2002 Data Set (for 1985- I 989) and Ariyoshi, 2000 (for 1990-I 998). S3 Table A2: Other Selected Economic Indicators since 1990-1998 (Annual Percentage change) 1990 1991 1992 1993 1994 1995 1996 1997 1998 Chile Reserve money 5404 23.7 21.7 13.6 20.7 13.9 15.9 16.0 -3.6 Broad money 23.5 28.1 23.3 2304 11.3 25.8 19.6 16.3 9.6 Nominal exchange rate! 13.6 11.3 2.0 12.7 -6.3 0.8 404 3.5 7.7 Real Effective exchange rate2 -3.8 6.5 lOA 004 5.8 1.7 3.9 9.6 -6.1 Thailand Reserve money 18.6 13.3 17.9 16.1 14.5 22.6 12.0 4.5 004 Broad money 26.7 19.8 15.6 18.4 12.9 17.0 12.7 2.0 6.1 Nominal exchange rate! -1.6 -0.0 0.9 ·0.1 -1.8 0.4 1.7 84.5 -22.3 Real Effective exchange rate2 -2.9 0.6 1.8 1.8 -2.6 3.0 504 -33.0 23.8 Kenya Reserve money 21.8 15.7 53.5 52.5 31.3 28.7 8.2 -1.5 -1.7 Broad money 20.1 19.6 39.0 28.0 27.4 12.5 15.9 9.8 3.1 Nominal exchange rate! 11.5 16.6 29.0 88.2 -34.2 24.8 -1.6 13.9 -1.2 Real Effective exchange rate2 -5.7 -2.1 8.6 -17.5 47.4 -18.3 12.3 1.9 0.1 Source: Extracted from Ariyoshi, 2000 Domestic currency units per u.S. dollar 2 Increase means an appreciation 54 References Agenor, P., and et-al, 1999, "The Asian Financial Crisis: Causes, Contagion and Consequences," Cambridge University Press, UK. Ariyoshi, A. and et-al, 2000, 'Capital controls: country experiences with their use and liberalisation ',occasional paper, IMF, Washington D.C. Bartolini, L., and A. H. Drazen, 1997, "Capital Account Liberalization as a signal," American Economic Review, 87, 138-54. Bhagwati, J., 1998b, "Yes to Free Trade, May be to Capital Controls', The Wall Street Journal, Page A-38, 11/16/1998. Calvo, G., 1993, "The Capital Inflows Problem: Concepts and Issues," IMF, paper on Policy analysis and assessment, PP AA/93/1 O. Calvo, G., A., Leiderman, L., and Reinhart, C.M., 1996, "Inflows of Capital to Developing Countries in the 1990s," Journal of Economic Perspectives, Vol. 10. Calvo, G.A., and Reinhart, C.M., 1998, "The Consequences and Management of Capital Inflows: Lessons for Sub-Saharan Africa," EGDI, Expert Group On Development Issues, University of Maryland. Cowan, K. and De Gregorio, J., 1997, "Exchange Rate Policy and Capital Account Management: Chile in the 1990s', Departamento de Ingeniera Industrial, Universidad de Chile. Eduardo, F, and et-al, 1995, 'The Surge in Capital Inflows to Developing Countries: Prospects and Policy Response', World Bank, policy working paper no., 1473. Edward, S., 1999, 'How effective are capital controls?' working paper 7413, National Bureau of Economic Research, Cambridge (http://www . nber.org/papers/w7 413 \_\_\_ , ed., 1995, 'Capital Controls, Exchange Rates, and Monetary policy in the World Economy. Cambridge University Press. \_\_\_ , 1998a, 'Capital Flows, Real Exchange Rates, and Capital Controls: Some Latin American Experiences,' Paper Presented at the Conference on Capital Inflows to Emerging Markets, NBER, Cambridge, Massachusetts, Feb 1998. \_\_\_ , 1998c, 'Interest Rate Volatility, Capital Controls and Contagion,' NBER Working Paper 6756, October 1998. Eichengreen, et-al, (1996b), 'Is there a Safe Passage to EMU? Evidence on capital Controls and a Proposal', edited by Jeffery A. Frankel and others. 55 Fernandez-Arias, E., and Montiel, PJ., 1996, "The Surge in Capital Inflows to Developing Countries: An analytical overview," World Bank Economic Review, Vol. 10 Ffrench-Davis, R. and Griffith-Jones, S. (1995) Coping witlt Capital Surges, Tlte Return Of Finance to Latin America, Boulder, CO: Lynne RienneI'. Griffith-Jones, S. Capital Flows to Developing Countries, Does the Emperor Have Clothes? Paper submitted at ISS 50th Atmiversary Conference Globalization, Conflict and Poverty 7-9 October 2002 Griffith-Jones, S., and et-al, 1999, "Private Capital Flows to Africa: Perception and Reality", Forum on Debt and Development (FONDAD), The Hague. Griffith-Jones, S., Montes, M. F., and Nasution, A. (2001) Short-Term Capital Flows and Economic Crises, Oxford University Press. Guerguil, M., and Kaufman, M., 1998, "Competitiveness and the evolution of the real Exchange rate in Chile," WP/98/58, IMF http:// econ. worldbank.org \_\_ , IMF, 2001, Chile: 2001 Atticle IV Consultation-Staff Report: Public Information Notice; IMF Country Report No. 011116. \_\_ , IMF, 2002, World Economic Outlook Data Set Jansen. K. (2002) Lecture notes on Macroeconomic Impact of Capital Flows, ISS. Jansen, K. (2001) Thailand, Financial Crisis and Monetary Policy, an article in Journal Of the Asian Pacific Economy 6(1), Taylor and Francis Ltd Jansen, K., 1997, "External Finance in Thailand's Development: An interpretation of Thailand's growth boom," Macmillan Press LTD, UK, in association with Institute of Social Studies. Jaspersen, F.Z., and lC. Ginarte, 1993, "Capital Flows to Latin America 1982-92: Trends and Prospects," (World Bank, mimeo). Khan, M.S., and Reinhart, C.M., 1995, "Capital Flows in the APEC Region," Occasional Paper 122, IMF, Washington Dc. Kim, W., and Wei, S., 1999, 'Foreign Portfolio Investors Before and During a Crisis.NBER Working Paper 6968. Laban, Raul and Felipe Larrain, 1998, "The Return of Private Capital to Chile in the 1990s: Causes, effects, and Policy Reactions," WP/R98-02, Harvard University. Laurens, B. and Cardoso, l, 1998, 'Managing Capital flows: lesson from the experience of Chile', working paper 98/168, IMF, December. 56 Lee, K., 1999, "Goodbye Capital Controls, Hello Capital's Controls: political Economy of capital decontrols and the crisis in Korea. Le Fort, G., 1999, "The Experience with Capital Controls and Capital Account Regulations in Chile," paper prepared for the Working Group on Foreign Capital Flows of the Financial Stability Forum, May 28. \_\_ , and C. L. Budnevich, 1996, "Capital Account Regulation and Macroeconomic Policy", Jerome Levy Economics Institute of Bard Collage, May. Loser, C.M. and Ebrill, L.P., 2001, 'Staff Report for the 2001 Article IV consultation: Chile', IMF, June 22. Maria, A. and Pontes, P.G., 1999, 'Capital Flow and Control', the Washington University, Washington D.C., IBI- Institute of Brazilian Business and Management Issues. Montiel, P. J., 1996, "Policy Response to Surge in Capital Flows: Issues and Lessons," In Private Capital Flows to Emerging Markets after the Mexican Crisis, ed. By Guillenno A. Calvo, Morris Goldstein, and Eduard Hochreiter, Institute for International Economics, Washington, D.C. \_\_\_ , 1998, "The Capital Inflows Problem," EDI Working Papers, Economic Development Institute of the World Bank. \_\_\_ , and C. M. Reinhart, 1998, "The Dynamics of Capital Movements to Emerging Economics during the 1990s," In Short-term Capital Movements and Balance of Payments Crisis, ed. by Stephanie Griffith-Jones and M. Montes, Oxford University Press. Nadal-De Simone, F., and Sorsa, P., 1999, "A review of capital account restrictions in Chile In the 1990s," WP/99/52, IMF. Ndikumana, L., and Boyce, J.K., 2002, "Public debt and private assets: Explaining Capital Flight from Sub-Saharan African Countries," University of Massachusetts, Amherst, Ma 01003. . Rajan, R.S., 1998, 'Restraints on Capital Flows: What Are They? IPS Working Paperno.3. Reinhart, Carmen M., and Steven Dunaway, 1996, "Dealing with Capital Inflows: Are There Any Lessons? WIDER Research for Action No. 28 Sawatsawang, N., 1998, "Thailand's financial crisis of 1997: A survey and critique of Dominant explanations," WP/278, Institute of Social Studies, The Hague Saxena, S. and Wong, K., 1999, 'Clmency crisis and capital controls: A Selective Survey', University of Washington. 57 Schadler, S., M. Carcovic, A. Bennett, and R. Kahn, 1993, "Recent Experiences with Surge In Capital Inflows, IMF, Occasional Paper No.1 08 Schulze, G., 2000, "The Political Economy of Capital Controls," Cambridge University Press, UK. \_\_ , UN, 2000, "Capital Flows and Growth in Africa," United Nations Conference on Trade and Development, Geneva. Valdes-Prieto, Salvador, and Marcelo Soto, 1998, "The effectiveness of Capital controls: Theory and Evidence from Chile," Empirica, Vo1.25, No.2, pp. 133-64. 58

Review of Political Economy, Volume 16, Number 4, 485–499, October 2004 The Changing Face of Mainstream Economics DAVID COLANDER\*, RICHARD P. F. HOLT† & J. BARKLEY ROSSER, JR‡ \*Middlebury College, Middlebury, Vermont, USA †Southern Oregon University, Ashland, Oregon, USA ‡James Madison University, Harrisonburg, VA 22807, USA ABSTRACT This article argues that economics is currently undergoing a fundamental shift in its method, away from neoclassical economics and into something new. Although that something new has not been fully developed, it is beginning to take form and is centered on dynamics, recursive methods and complexity theory. The foundation of this change is coming from economists who are doing cutting edge work and influencing mainstream economics. These economists are defining and laying the theoretical groundwork for the fundamental shift that is occurring in the economics profession. 1. Introduction If one reads the heterodox literature in economics these days, one gets the impression that modern mainstream economics is much like the economics of 50 years ago; it is called ‘neoclassical economics’ and is criticized in much the same way that earlier heterodox economists criticized the mainstream economics of the 1950s or 1960s. In this paper we argue that much of this criticism today is off the mark because mainstream economic thinking has changed. We argue that economics is moving away from a strict adherence to the holy trinity—rationality, selfishness, and equilibrium—to a more eclectic position of purposeful behavior, enlightened self-interest and sustainability. The paper develops our ideas by considering the nature of that change and the process and sociological dynamics by which the profession changes. 2. The Profession as a Complex System To understand our argument it is helpful to think of the profession as a complex system—a system that is too complicated to be fully understood by agents in the Correspondence address: David Colander, Middlebury College, Middlebury, Vermont, USA. Email: colander@middlebury.edu ISSN 0953-8259 print/ISSN 1465-3982 online/04/040485-15  2004 Taylor & Francis Ltd DOI: 10.1080/0953825042000256702 486 D. Colander et al. system or researchers studying the system.1 Complex systems cannot be understood from assumed first principles; they can only be understood through the process of change that underlies them. In the same way, researchers can best understand the economics profession by the process of change that characterizes it. Most considerations of the economics profession have tended to take a static view of the profession, which makes it seem as if it is an unchanging entity. That is the approach that most heterodox criticisms of the profession have taken. But that is not the way we see the economics profession; we see it as a dynamic entity, which generates a self-reproducing, evolving, complex system of interacting ideas. Getting a handle on such a dynamic entity and conveying its essence to others often requires giving it static classifications and organizing it into distinct periods. Historians of economic thought must do this to provide structure when considering past economists. But these classifications are crutches, not characterizations of reality. They are imposed by the observer and are not necessarily part of the essence of the profession at any point in time. Any static classification hides the dynamic change occurring underneath it. For this reason, the classifications used by historians of thought, such as ‘Classical’ or ‘Neoclassical’, while useful and perhaps even necessary, are nevertheless confining and miss important dimensions of the profession. 3. The Edge of Economics The changes in the profession are brought about by what we call ‘work at the edge of economics.’2 It is innovative and successful work at the edge of the profession that signals the future direction of change in economics and how the profession eventually comes to be viewed and understood by its elite. The very concept of an edge of the profession is designed to suggest a profession in which there are multiple views held within the profession, and goes against the standard classifications of economics. Those standard classifications convey a sense of the profession as a single set of ideas. In our view, that is wrong; it is much more useful to characterize the economics profession as a diverse evolving 1 For a general discussion of complex systems, see Auyang (2000). Unsurprisingly, defining ‘complexity’ is not simple. The physicist Seth Lloyd has documented over 45 definitions (see Horgan, 1997, p. 303, footnote 11). Many deal with computational or algorithmic meanings, with these applied in economics by Leijonhufvud (1993) and Albin with Foley (1998). Day (1994) provides a dynamic definition that a system does not converge to a fixed point, a limit cycle, or explode continuously due to endogenous factors. Pryor (1995) and Stodder (1995) use it to mean something like ‘complicated’ institutional and sectoral linkages. We note that any of these can lead to making it difficult for an agent or observer to understand the workings of an economic system. See Rosser (1999) for further discussion. 2 The term, edge of economics, refers to work challenging the previously considered ‘orthodox’ ideas. Initially, we described it as cutting edge work, but some of our colleagues have pointed out that cutting edge work can only be defined historically as work at the edge that has panned out. Comments by Larry Moss and Ken Koford were very helpful in redirecting us in our terminology. The Changing Face of Mainstream Economics 487 set of ideas, loosely held together by its modeling approach to economic problems.3 Standard classifications tend to miss the diversity that exists within the profession, and the many new ideas that are being tried out. They miss the important insight that one can be part of the mainstream and yet not necessarily hold ‘orthodox’ ideas. Standard classifications also emphasize a fairly narrow orthodox core of the profession and convey a picture of all conventional economists accepting this core. The reality is more complicated; conventional economists often hold a variety of views simultaneously. If the variance of views increases, while the core remains relatively unchanged, the static characterization of the profession will not change, but its dynamic characterization will. A large variance in acceptable views, such as has emerged in the profession over recent decades, signals that changes are likely in the future. In our view, the interesting story in economics over the past decades is the increasing variance of acceptable views, even though the center of economics has not changed much. For example, mainstream economists today such as William Baumol, George Akerlof, Thomas Schelling, Truman Bewley, and Paul Krugman, in important aspects of their thinking, are working outside of what is generally considered the orthodoxy of the profession. Yet, their ideas are widely accepted and discussed within the mainstream of economics. It is such work that has increased the variance of acceptable views in the profession. To capture that variance of acceptable views, static classifications must be seen for what they are—useful fictions that are meant for students and non-specialists. These classifications are backward looking, and, to be meaningful, they must be supplemented with a discussion of the variance of ideas acceptable to the mainstream. The reality is that at any point in time a successful discipline will have hundreds of new ideas being tried out, as new methods, new technology and new information become available. That is what happens at the edge of economics. This edge of economics has both intellectual and social elements. The intellectual aspect of economics at the edge fundamentally involves originality. This does not mean that all ideas at the edge are totally new. Ideas have origins, and grow better in some environments than in others. The history of economics is full of instances in which old ideas are rehabilitated or revived and found to be useful and advantageous within the new context that is emerging. In work at the edge, ideas that previously had been considered central to economics are being modified and broadened, and the process is changing the very nature of economics. What makes it possible for these ideas to take root now, but not in the past, are advances in analytic technology, such as non-linear dynamics, which has made it possible to study much more complex models than before, and developments in computing capabilities, which have made studies with simulations and agent-based models much more useful, allowing 3 Robert Solow (1996, p. 43) writes ‘Today if you ask a mainstream economist a question about almost any aspect of economic life, the response will be: suppose we model that situation and see what happens … modern mainstream economics consists of little else but examples of this process.’ 488 D. Colander et al. economists to study problems that do not have analytic solutions. Combined with advances in other disciplines relevant to economics, which makes the integration among disciplines easier, the combination of these advances has opened up completely new ways of integrating those ideas into the core beliefs of the field, and has changed the core beliefs in important ways. For example, developments in nonlinear dynamics now allow alternative models of processes that include sudden shifts from one equilibrium to another, and the development of agent-based modeling is allowing researchers to explore models with heterogeneous agents and to move away from a focus on unique equilibria.4 4. Change within the Profession Sociological issues impinge upon and constrain what is possible intellectually. The reproduction of ideas involves the social, political, and economic structures of the academic and policymaking establishments in which ideas are developed and transmitted. Ideas, however original and possibly wonderful, that do not become accepted by some of the elite of the profession, and which do not eventually get funded, will not be accepted and transmitted within the profession. To internally move the discipline to a new position, some of the profession’s elite must accept these ideas. In our view, what is occurring in economics today is a modification of the standard view of paradigm shifts proposed by Thomas Kuhn (1970). Kuhn argued that the driving forces of change in a discipline are ideas that challenge the very system of thought in a way that puts them outside the mainstream, and ultimately are introduced ‘funeral by funeral’ by a paradigm shift. It is easy to recognize that a paradigm shift has occurred as the shift is so dramatic and sudden. We see this view as not quite fitting the economics profession. From our dynamic perspective, an alternative channel exists that allows significant changes to occur within the mainstream of the profession in a way that is not apparent to the mainstream. These changes do not lead to sudden paradigm shifts, but instead lead to cumulative evolutionary changes that ultimately will be recognized as a revolutionary change. The changes leading to this ex-post revolution were initially accepted within the profession only gradually, more along the lines suggested by Imre Lakatos (1978). This alternative channel is the following: When certain members of the existing elite become open to new ideas, that openness allows new ideas to expand, develop, and integrate into the profession. In this case change within the profession can be accepted gradually, being introduced ‘data set by data set’ and ‘new technique by new technique’ as well as ‘funeral by funeral’. In some cases these new ideas will originate from outside the mainstream, from those who consider themselves heterodox, even if the acceptance of such ideas leads to their ‘normalization’ and removal from being identified as heterodox. These alternative channels allow the mainstream to expand, and to evolve to include a wider range of approaches and understandings. Eventually, sufficient 4 See Arthur et al. (1997), Judd & Tesfatsion (forthcoming), and Rosser (2004). The Changing Face of Mainstream Economics 489 change is made so that future historians of thought will consider the orthodoxy of the period changed. This, we believe, is already occurring in economics. Mark Blaug, one of the most distinguished current historians of economic thought, has pointed out that, beginning as early as the 1950s, the classification ‘neoclassical economics’ was no longer appropriate to characterize modern economics (Blaug, 1998, p. 2), an argument further developed by Colander (2000a). The difference between Kuhn’s view and ours concerns how changes generally come about in a profession. We suggest that changes, even ones that will eventually be considered revolutionary, often come from within and will not be noticed for years. Kuhn’s view suggests that they can only come from outside and are quite apparent when they occur. The dynamic approach of change within the profession that we are introducing here involves stealth changes, in which advocates of new ideas may gain acceptance among the elite of the profession, and even achieve positions of power and prominence within at least some leading academic institutions of economics. The change, however, is so gradual that the profession often does not notice that the change has occurred. The reason for the difference is the multiple dimensionalities that we see in the mainstream profession. Mainstream economics is a complex system of evolving ideas. Individuals in the profession see minute change upon minute change but do not have a perception of the aggregate of the changes. Only when historians of thought look back, after sufficient time has passed to gain some historical perspective, does the larger change become apparent. 5. The Process of Change Both the social and intellectual aspects of change must be taken into account in order to understand the evolution of ideas. The work at the edge is generally begun by younger researchers, and in some cases those who are doing heterodox work. However, their ability to do that work, and to have their work affect the profession, is dependent on the existence of crucial persons in the leading academic establishments, representing the mainstream of economics, who are open to seriously considering new ideas. These crucial people may be the ones who have developed what was considered the old orthodoxy, but their having developed it does not mean that they aren’t open to change and new ideas. There is nothing inconsistent with being one of the originators of a theory and simultaneously being a critic of that theory. Good economists simultaneously recognize the strengths and limitations of a theory, and are open to new approaches and ideas. A good example of a person that fits this category is Kenneth Arrow. Although he is associated with what is considered modern neoclassical orthodoxy, he was instrumental in introducing the complexity approach into economics.5 The consideration and ultimate acceptance of a new idea by a certain portion of the elite becomes a key to the process of how the conventional 5 Mirowski (2002, pp. 432–436) argues that an important influence on Arrow’s change of view was a former student, Alain Lewis (1985), whose work continues to be little known by most of the profession. 490 D. Colander et al. foundation of the discipline evolves. It is not crucial that those developing the ideas initially be at leading establishments. But they must be able to attract the attention of influential individuals at those institutions in order for their ideas to be published in venues that will receive attention, and for research along those lines to get funded. This allows students and advocates of those ideas to get hired at those institutions and thus to establish themselves within the mainstream of the discipline, even when the originators of these ideas remain somewhat outside the mainstream elite. 6. Orthodoxy, Heterodoxy, Mainstream It is helpful in making our argument to consider carefully the terms, ‘mainstream’, ‘orthodox’, ‘heterodox’, how they are used, and how they relate to our idea that the dynamics of change in a profession are at the edge of the profession. Let us start with the term, ‘mainstream economics’. In some sense, mainstream economics is the easiest of the above terms to define, although it may be the hardest to identify in practice. It is in large part a sociologically defined category. Mainstream consists of the ideas that are held by those individuals who are dominant in the leading academic institutions, organizations, and journals at any given time, especially the leading graduate research institutions. Mainstream economics consists of the ideas that the elite in the profession finds acceptable, where by ‘elite’ we mean the leading economists in the top graduate schools. It is not a term describing a historically determined school, but is instead a term describing the beliefs that are seen by the top schools and institutions in the profession as intellectually sound and worth working on. Because of this, mainstream economics usually represents a broader and more eclectic approach to economics than is characterized as the recent orthodoxy of the profession. In our view, the term ‘orthodox’ is primarily an intellectual category. It is a backward looking term that is best thought of as a static representation of a dynamic, constantly changing profession, and thus is never appropriately descriptive of the field of economics in its present state. Orthodoxy generally refers to what historians of economic thought have classified as the most recently dominant ‘school of thought,’ which today is ‘neoclassical economics.’ In our view, modern mainstream economics is quite different from this neoclassical concept of orthodox economics. Having the two terms is important for us because it allows us to make intertemporal comparisons between the most recently dominant school of thought, in this case neoclassical economics, and today’s evolving mainstream economics. To help us get a grasp of what we mean by neoclassical orthodoxy and how it relates to mainstream economics, it is important for us to first specify what we see as neoclassical economics. In our view, neoclassical economics is an analysis that focuses on the optimizing behavior of fully rational and well-informed individuals in a static context and the equilibria that result from that optimization. It is particularly associated with the marginalist revolution and its aftermath. Le´on Walras and Alfred Marshall can be viewed as its early and great developers, with John Hicks’s Value and Capital (1939) and Paul Samuelson’s The Changing Face of Mainstream Economics 491 Foundations of Economic Analysis (1947) as its culmination. When a dynamic context is assumed, individuals understand the probability distributions of possible outcomes over the infinite time horizon at the moment of decision. The neoclassical orthodoxy tests the results of that model by using conventional econometric techniques that are based upon a foundation of classical statistics. Perhaps the most important characteristic of the neoclassical orthodoxy is that axiomatic deduction is the preferred methodological approach. The difference between mainstream and orthodox becomes clearer when one recognizes two other aspects of the term ‘orthodox’. The first is that the name and specification of what is orthodox usually comes decades after that time when orthodoxy was supposed to exist; at the time it is a true orthodoxy, it generally has no name. Thus, orthodox specifications inevitably are backward looking, not current or forward-looking. Second, in economics at least, the name for the orthodox school usually comes from a dissenter, who opposed orthodox ideas, not from a supporter of the orthodox ideas. For example, Marx (1847) coined the term ‘classical economics’, even though the Classical school is seen as starting back in the late 1700s. Before Marx’s general classification there was no name for the classical orthodoxy. Similarly, the term ‘neoclassical economics’ was coined by Veblen (1900), referring to the economics of the last part of the 19th century as he tried to tie this period of economics to Classical economics, so as to make the argument that both are unscientific (Aspromourgos, 1986). In each case, the classification was made by an economist to create a better target for his criticism. Defining orthodoxy, and giving a name to it, gives a critic an easy target; it implies a static unchanging dimension of thought. But this static view is not characteristic of the economics field. At any point in time, and especially by the time that the term becomes generally used, a large part of the mainstream profession disagrees with important dimensions of what is then thought of as orthodoxy. Finally, let us consider the term ‘heterodox’. It is usually defined in reference to orthodox, meaning to be ‘against orthodox,’ and defines itself in terms of what it is not, rather than what it is. An economist who sees him or herself as heterodox does not subscribe to the current orthodox school of thought, as defined by the historian’s classifications. However, in our view, heterodoxy also has a sociological aspect. A self-identified heterodox economist has also defined his or her self outside the mainstream. Heterodox economists are highly unlikely to get funding through normal channels, such as the National Science Foundation, although they might receive alternative funding from a variety of sources. Thus, heterodoxy involves both sociological and intellectual aspects. Since many mainstream economists also do not accept important aspects of the orthodoxy, the additional feature that determines a heterodox economist is social; heterodox economists refuse to work within the framework of mainstream economics whether because of the nature of the modeling process used, or because of the assumptions made. This often causes a failure of communication between heterodox and mainstream economists, even when they may share similar views about the limitations of the ‘orthodox’ approach. In the economics profession, various schools—many of which have long histories—comprise heterodox economics. These schools have their own net 492 D. Colander et al. works and organizations and journals and academic institutions where they dominate. Often, the fundamental intellectual content of a heterodox school is its rejection of orthodoxy, or at least major elements of orthodoxy. In economics, at least, beyond this rejection of the orthodoxy there is no single unifying element that we can discern that characterizes heterodox economics. In fact, it is well known that many varieties of heterodoxy have more disagreement with each other than they do with orthodoxy. But it should also be said that different heterodox schools previously emphasized many of the ideas that are now on the edge of economics, and these schools can play an important role in developing new critiques of the orthodox. Among the most established of the heterodox schools with reasonably full systems of institutional support are Marxists, Post Keynesians, feminists, Old Institutionalists, and Austrians.6 If the field of economics were static and one-dimensional, these two classifications (orthodox/heterodox) would be sufficient, but it isn’t and they aren’t. The economics profession is dynamic and constantly changing. Since these classifications usually lag developments in the field by decades, the terms, ‘orthodox’ and ‘heterodox’, when used in a current setting, tend to be backwardlooking, describing beliefs that, while they still may show up in texts, are not strong convictions of many in the profession, and are being attacked by economists at the edge of the profession. To understand the dynamic aspect of the profession and the role of economists working at the edge, the distinction between mainstream and orthodox is central. The edge of economics is that part of mainstream economics that is critical of orthodoxy, and that part of heterodox economics that is taken seriously by the elite of the profession. Our argument is that modern mainstream economics is open to new approaches, as long as they are done with a careful understanding of the strengths of the recent orthodox approach and with a modeling methodology acceptable to the mainstream. For an economist working at the edge, attacking the profession is not sufficient; he or she must be developing new methods and ideas. In this approach the difference between mainstream and heterodox becomes far less important than whether they are doing work at the edge. In this case, both mainstream and heterodox economists are working on issues that challenge the neoclassical orthodoxy, but that orthodoxy is no longer descriptive of what the mainstream elite believes. The elite’s vision of economics is forward looking—these are the ideas that are exciting today, and here is where they may lead; the static classifications of economics are backward looking, emphasizing where economics has been. This concept ‘elite of the profession’ is elusive, but is understood by those in the profession. It is those mainstream economists who have made important contributions to thought in the past. It includes some (but not all) Nobel Prize winners, and most economists who have major chairs at top graduate programs. If one has standing offers from a number of top schools to come and teach there 6 We recognize that this characterization oversimplifies the state of heterodox economics. Not only are there many subcategories and schools within these main branches of heterodoxy, but there are many other schools or approaches as well. The Changing Face of Mainstream Economics 493 if one desires, and if one receives calls from the NSF about who to put on NSF panels, one is in the elite of the profession. Examples of well-known mainstream elite are Paul Samuelson, Kenneth Arrow, Robert Solow, Thomas Schelling, Amartya Sen, Joseph Stiglitz, Chris Sims, Michael Woodford, George Akerlof, Richard Thaler, Anne Krueger, and Jagdish Bhagwati. This is a very diffuse group. Recognizing that there is an elite element in the mainstream that plays a crucial role in what new ideas will prove to be part of the acceptable edge of economics raises two problems—one of how open the elite will be, and another of how these ideas then disseminate throughout the rest of the mainstream and the profession more generally. Our view is that the current elite are relatively open minded when it comes to new ideas, but quite closed minded when it comes to alternative methodologies. If it isn’t modeled, it isn’t economics, no matter how insightful. It is here that heterodox economics and the mainstream elite normally collide. Specifically, it is because of their method, not their ideas, that most heterodox find themselves defined outside the field by the elite.7 We are certainly not claiming that the mainstream is always pluralistic and open minded, willing to accept heterodox views with open arms. Far from it. They are human, and become fixed in their ways of looking at things and often reject alternative views without giving them serious consideration. That is part of human nature. This means that, in many unconscious ways, which we consider unfortunate, the mainstream elite can suppress the views of heterodox economists. Moreover, they often use their method as a tool to protect views that do not fit nicely into their way of thinking. What we are claiming is that their close-mindedness is generally unconscious, and representative of almost any group that has the power to be that way—including, in their own small spheres, many heterodox economists. We are also claiming that the worst types of heterodox suppression and narrow-mindedness are not carried out by the elite, but instead are carried out by economists whose professional credentials are mediocre for the very reason that they are not as imaginative and creative as the elite. Once an idea is expressed in an acceptable model the dissemination process is a long and drawn-out endeavor that works along the following lines. Work at the edge usually shows up first in working papers that are presented at graduate seminars and workshops. These are the incubators of new ideas in economics, although sometimes the ideas are initially generated by persons outside of those seminars. The ideas contained in these working papers will generate discussion among professors at graduate schools. Some will be panned; others will be tentatively accepted, and mentioned to professors at other schools. Some ideas will generate a buzz and, when they do, will attract intense interest. (This generally occurs before publication.) Eventually the idea will be published in a top journal, but that publication is often a tombstone, marking ownership of the idea more than it is a spreading of the idea. The diffusion of the idea throughout the elite of the profession will have already occurred, although sometimes an idea will be published and not get noticed until sometime later. Thus, Fo¨llmer 7 See Colander (2003) for a development of this point in reference to institutionalists. 494 D. Colander et al. (1974) initially proposed that ideas from statistical mechanics could be applied to analyzing heterogeneous agent models in economics, which was not followed up on until the 1990s, and Strotz (1956) first proposed the idea of hyperbolic discounting, also only taken seriously in the 1990s. As this process is occurring, the working paper or article will show up in core graduate program reading lists, and eventually make its way into graduate textbooks. The process from conception of an idea to its appearance in graduate textbooks can take up to ten years. Intermediate and upper level undergraduate textbooks usually take another five to ten years to include these ideas, although they may show up as a supplemental box, or an added paragraph earlier than this. Principles books take another five to ten years to actually incorporate the idea as a central element, although, like their undergraduate upper level counterparts, they may add them as addenda so that they look modern. There is a paradox in this diffusion process. The more central the idea, the less likely it is to be included in a central way in the texts. For example, complexity suggests the whole conception of equilibrium in an economy needs to be reconsidered, and experimental economics suggests that the entire approach to thinking about the appropriate mix of induction and deduction needs to be rethought. Such a reconsideration and rethinking would likely change the entire way textbook are structured, and the way the courses are taught. Such major changes are unlikely to show up even with the long lags that we discussed. Instead they will be simply added as an addendum to the existing core. (For a discussion of these issues, see Colander, 2000b.) Such changes resemble more the kind of changes that Kuhn discussed in his analysis of paradigm shifts, even if the shift has occurred in the more gradualist manner that we have been describing. Why the enormous lag? The reason is that the professors who actually teach the majority of the courses are most comfortable teaching what they have studied, and the publishing industry writes for that majority. Since the average undergraduate professor has been out of graduate school for a long period of time, the average professor (which the textbooks target as their audience) will generally be most comfortable teaching older material as the core of the course, with new material scattered throughout. The material shows up in higher level courses first because the higher the level of the course the more likely a specialist in the area is teaching the course, and that specialist is more likely to feel comfortable including new developments. This long lag should not be seen as a complete waste; it serves a useful function in that it provides a filtering process that eliminates those ideas that seemed wonderful, but turned out to be just fads. For example, in graduate work in macro, the focus is almost entirely on stochastic dynamic intertemporal general equilibrium models with infinitely bright individuals, that for many economists are of dubious value in understanding the macro economy. Undergraduate textbooks have taken only slight notice of this development, and the Keynesian IS/LM model has remained the core of many undergraduate macro texts even after it has all but been excluded from what is taught in graduate schools. New books reflecting the new graduate school approach have been published, but they have not been generally adopted at the undergraduate level. The Changing Face of Mainstream Economics 495 We suspect that the reason for this is that the new work in macro is simply a fad that will pass as researchers come to accept that the macro economy must be analyzed as a complex system. This lack of acceptance by the undergraduate texts reflects the uncertainty that many mainstream economists had with the rational expectations revolution in macro. While it was a logical extension of microeconomic reasoning, it did not seem reasonable to many, suggesting that something was wrong with the models that were based on it in its strong form. For this reason, the rational expectations revolution led to work in what might be called the complexity revolution, which is striving to provide stronger underpinnings for macro models generally. This work begins from the assumption of rationality, but seriously considers the problems of defining rationality in a complex environment and, when there are problems, accepts the complex environment as its reference point, rather than taking a simpler environment. The lags in this process can lead to a situation where an idea that has come to be viewed as somewhat old hat at the elite mainstream level may only finally be appearing in the principles textbooks. Consequently, textbooks, especially lower level texts, often do not reflect the diversity of views acceptable to the mainstream, but instead reflect an older orthodox position.8 Another important comparison between the mainstream and orthodoxy is that economists working within the mainstream can find their views evolving. For example, they might be working with a particular approach, but then change. Consider rational expectations and the New Classical revolution in macroeconomics. One of the early developers of rational expectations, Leonard Rapping, modified his views significantly and became a leading heterodox economist before his untimely death. Another example is Thomas Sargent, another of the leading figures in the application of rational expectations to macroeconomics. As a result of visiting the Santa Fe Institute he came to abjure a strict rational expectations view (Sargent, 1993). His more recent work with Lars Hansen and others (Hansen & Sargent, 2000) has attempted to provide quantitative approaches to dealing with Knightian uncertainty. Thus, he has moved out of orthodoxy, but has remained mainstream, and is on the edge of the edge of economics. As should be clear from the above discussion, in our view, the edge is where the action is in the profession. Whether what works at the edge is considered heterodox or mainstream is primarily a matter of the individual’s proclivity to fit within the existing mainstream, and the degree to which they directly attack, rather than softly criticize, the work of the elite. It should be pointed out that working at the edge does have its problems, especially for those whose proclivity is toward attacking, rather than working within the existing field, and hence finding themselves in heterodoxy. They face significant sociological problems of achieving acceptance from the established mainstream. 8 This lag of textbooks of mainstream thinking can be seen in earlier times as well. In his writings John Stuart Mill gave up the wage fund doctrine, but retained it in his principles book, stating that these new developments ‘are not yet ripe for incorporation in a general treatise on Political Economy’ (Mill, 1929, p. xxxi) 496 D. Colander et al. Economists considered heterodox often find it difficult to gain funding for their work, and they will be squeezed out of the decision making process at their universities. Those involved in working at the edge that are in the mainstream lack this sociological problem, but they also often find themselves at odds with those around them to some degree as they press against the boundaries of the mainstream. 7. Work at the Edge of Economics and the Complexity Vision We emphasize complexity as a defining factor of the new work at the edge of economics, because it appears to us to be the vision behind this work. But the actual work involves a number of fronts, and the people working on those fronts have varying degrees of connection to the broader complexity approach. Along with this, and interacting with it, is a new openness to ideas from other disciplines. Thus, modeling remains the central core of the mainstream approach, but the nature of the models and the assumptions underlying them are much more open, and transdisciplinary.9 More specifically: • Evolutionary game theory is redefining how institutions are integrated into the analysis. • Ecological economics is redefining how nature and the economy are viewed as interrelating. • Psychological economics is redefining how rationality is treated. • Econometric work dealing with the limitations of classical statistics is redefining how economists think of empirical proof. • Complexity theory is offering a way of redefining how we conceive of general equilibrium. • Computer simulations are offering a way of redefining models and how they are used. • Experimental economics is changing the way economists think about empirical work. These changes in turn have led to a broader set of changes in how mainstream economics sees itself. It is much more willing to accept that the formal part of economics has limited applicability, at least as currently developed. It is also far more willing to question economics’ special status of economics over the other 9 There is much discussion now regarding how one is to describe research that involves more than one discipline. The oldest term is probably multidisciplinary. However, this now usually is applied to situations where persons representing different disciplines get together and contribute ideas from their separate disciplines in ways that maintain the distinct identities of their disciplines, as in separate chapters within a book. A more recent term of use has been interdisciplinary that involves more integration of the ideas of different disciplines. However, this is often used in the sense of dealing with ideas that exist in the intersection of two disciplines, leading to particular specializations, e.g. ‘water economist,’ who knows about relevant aspects of both hydrology and economics. Following the lead of the ecological economists we favor the term transdisciplinary to describe the new developments at the edge, which implies a more thoroughgoing and profound interaction between the disciplines leading to some kind of new synthesis and transcendence. The Changing Face of Mainstream Economics 497 fields of inquiry and to integrate the methods of other disciplines into their economic analysis. The change that is occurring in economics is most clearly symbolized by two conferences held nearly a decade apart at the Santa Fe Institute. The first held in 1988 generated a book entitled The Economy as a Complex Evolving System (Anderson et al., 1988). Waldrop (1992) reported that this conference featured a set of largely mainstream economists and defenders of general equilibrium orthodoxy, assembled by Kenneth Arrow, and a set of physicists assembled by others. The economists mostly attempted to defend their mainstream approach, while they faced sharp challenges and ridicule from the physicists for holding relatively simplistic views. Although models using nonlinear dynamics and other complexity approaches have been developed for some time (Rosser, 1999), such approaches at that time remained outside the mainstream camp. The second conference saw a very different outcome and atmosphere than the first (Arthur et al., 1997). No longer were mainstream economists defensively adhering to general equilibrium orthodoxy. Now they were using methods adopted from biologists and physicists, many suggested at the earlier conference, in innovative ways. They were also much more open to complex economic analysis. These two Santa Fe conferences are representative of the change that occurred throughout the profession during this time. It was as if the ideas planted by earlier researchers in many areas, such as experimental economics, behavioral economics, and nonlinear dynamics, were taking root. Thus, by 1997, the mainstream accepted many of the methods and approaches that were associated with the complexity approach. What they had not accepted was the broader complexity vision. (For a discussion of that broader vision, see Colander, 2000c.) That broader vision is held by a much smaller group of economists, and it may or may not be held by the individuals working on the edge of economics. But as the work at the edge progresses and accumulates, it shifts the center of economist’s approach, and, in our view, eventually will create a new orthodoxy centered on a broader complexity vision. References Albin, P. S. with Foley, D. K. (1998) Barriers and Bounds to Rationality: Essays on Economic Complexity and Dynamics in Interactive Systems (Princeton: Princeton University Press). Anderson, P. W., Arrow, K. J. & Pines, D. (Eds) (1988) The Economy as an Evolving Complex System (Redwood City: Addison-Wesley). Arthur, W. B., Durlauf, S. N. & Lane, D. A. (Eds) (1997) The Economy as an Evolving Complex System II (Redwood City: Addison-Wesley). Aspromourgos, T. (1986) On the origin of the term ‘neoclassical’, Cambridge Journal of Economics, 10, pp. 265–270. Auyang, S. Y. (2000) Foundations of Complex-System Theories in Economics, Evolutionary Biology, and Statistical Physics (New York: Cambridge University Press). Blaug, M. (1998) The formalist revolution or what happened to orthodox economics after World War II, University of Exeter, 98/10 Discussion Paper in Economics. 498 D. Colander et al. Colander, D. (2000a) The death of neoclassical economics, Journal of the History of Economic Thought, 22, pp. 127–143. Colander, D. (2000b) A thumbnail sketch of the history of thought from a complexity perspective, in: D. Colander (Ed.) Complexity and the History of Economic Thought (New York: Routledge). Colander, D. (Ed.) (2000c) Complexity and the Teaching of Economics (Cheltenham: Edward Elgar). Colander, D. (2003) Are institutionalists an endangered species?, Journal of Economic Issues, 37, pp. 111–122. Day, R.H. (1994) Complex Economic Dynamics, Volume 1: An Introduction to Dynamical Systems and Market Mechanisms (Cambridge: MIT Press). Fo¨llmer, H. (1974) Random economies with many interacting agents, Journal of Mathematical Economics, 1, pp. 51–62. Hansen, L. P. & Sargent, T. J. (2000) Wanting robustness in macroeconomics, ftp:// zia.Stanford.edu/pub/sargent/webdocs/research/wanting.pdf. Hicks, J. R. (1939) Value and Capital (Oxford: Clarendon Press). Horgan, J. (1997) The End of Science: Facing the Limits of Knowledge in the Twilight of the Scientific Age, Paperback Edn (New York: Broadway Books). Judd, K. & Tesfatsion, L. (Forthcoming) Handbook of computational economics: Agent-Based Computational Economics (Amsterdam: North Holland). Kuhn, T. S. (1970) The Structure of Scientific Revolution, enlarged edn (Chicago: University of Chicago Press). Lakatos, I. (1978) The Methodology of Scientific Research Programmes: Philosophical Papers, Vol. 1 (Cambridge: Cambridge University Press). Leijonhufvud, A. (1993) Towards a not-too rational macroeconomics, Southern Economic Journal, 60, pp. 1–13. Lewis, A. (1985) Complex structures and composite models - an essay on methodology, Mathematical Social Sciences, 10, pp. 211–246. Marx, K. (1847) The Poverty of Philosophy, English translation in (1977) Karl Marx and Friedrich Engels: Collected Works (New York, Progress Publishers). Mill, J. S. (1929) Principles of Political Economy with Some of Their Applications to Social Philosophy, W. J. Ashley (Ed.) (London: Longmans, Green). Mirowski, P. (2002) Machine Dreams: Economics Becomes a Cyborg Science (Cambridge: Cambridge University Press). Pryor, F. L. (1995) Economic Evolution and Structure: The Impact of Complexity on the U.S. Economic System (New York: Cambridge University Press). Rosser, J. B., Jr. (1999) On the complexities of complex economic dynamics, Journal of Economic Perspectives, 13(4), pp. 169–192. Rosser, J. B., Jr. (Ed.) (2004) Complexity in Economics, Volumes I-III (Cheltenham: Edward Elgar). Samuelson, P. A. (1947) Foundations of Economic Analysis (enlarged edn, 1983) (Cambridge: Harvard University Press). Sargent, T. J. (1993) Bounded Rationality in Macroeconomics (Oxford: Clarendon Press). Solow, R. (1997) How did economics get that way and what way is it? Daedalus, 126, pp. 39–58. Stodder, J. P. (1995) The evolution of complexity in primitive economies: theory, Journal of Comparative Economics, 20, pp. 1–31. Strotz, R. H. (1956) Myopia and inconsistency in dynamics utility maximization, Review of Economic Studies, 23, pp. 165–180. Veblen, T. (1900) Preconceptions of economic science, part III, Quarterly Journal of Economics, 14, pp. 240–269. von Neumann, J., completed by Burks, A.W. (1966) Theory of Self Reproducing Automata (Urbana: University of Illinois Press). The Changing Face of Mainstream Economics 499 Waldrop, M. M. (1992) Complexity: The Emerging Science at the Edge of Order and Chaos (New York: Simon & Schuster). Woodford, M. (Ed.) (forthcoming) Festschrift in Honor of Edmund Phelps (Princeton: Princeton University Press).