Weiying Zhang

The Origin of the Capitalist Firm

An Entrepreneurial/Contractual Theory of the Firm

格致出版社
Truth & Wisdom Press

Springer

The Origin of the Capitalist Firm

Weiying Zhang

The Origin of the Capitalist Firm

An Entrepreneurial/Contractual Theory of the Firm

格致出版社
Truth & Wisdom Press

Springer

Weiying Zhang
The National School of Development
Peking University
Beijing
China

ISBN 978-981-10-0220-5 ISBN 978-981-10-0221-2 (eBook)
DOI 10.1007/978-981-10-0221-2

Jointly published with Truth and Wisdom Press

Library of Congress Control Number: 2017935841

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface to the English Edition

The subtitle of this book—*An Entrepreneurial/Contractual Theory of the Firm*—sounds a bit awkward at first glance, but it does in fact set out the theme of this book in a very precise way.

The theory of the firm in the Neoclassical setting is essentially a theory of production decision, namely the study of how to allocate input factors and choose production level in order to maximize profit by a given production unit (e.g. a firm). It emphasizes the determination of production decision by the price of input factor, production technology and the demand function. This well-established theory is meaningful in terms of understanding market equilibrium and the efficiency of resource allocation, but it overlooks the complex organizational relations and incentive problems within the firm, and therefore it is also called 'black box theory'. The earliest critique of the Neoclassical theory of the firm came from Ronald Coase, who argued in his classic the Nature of the Firm (1937) that the Neoclassical theory cannot even explain the existence of the firm, due to its assumption that transaction costs are zero. Under such assumption, all production can be completed by market exchanges between individuals; hence there is no need for any authority like the firm to facilitate exchange. The modern theory of the firm, which is originated by Coase and further developed by Oliver Williamson and others, is commonly referred to as 'contractual theory'. It treats firm as a form of substitute for market exchange and puts transaction cost at the heart of analyses.

The contractual theory of the firm theorizes the existence of the firm and explains the importance of ownership and incentive within the structure of the enterprise. It helps deepen our understanding of the firm as an institution. However, a big problem with contractual theory is its negligence of the role that entrepreneurs play in the enterprise, i.e. the firm in contractual theory is still the firm with no entrepreneurs. Although contractual theory is able to expound the firm's existence and the importance of ownership, the absence of entrepreneur deprives its ability to further explain why decision-makers of the firm should have the right to take the residual, and who should be entitled to choose the decision-maker. To put it simple,

contractual theory pays attention to the problem of incentive rather than the choosing of the runner of the firm. I believe the latter is more important than the former, as an enterprise without an entrepreneur is not a true enterprise.

I put entrepreneurs at the heart of this book. I will establish a theory of firm that is essentially entrepreneur-centric with the nature of contractual relationship, and therefore it is called "an entrepreneurial/contractual theory of the firm". I believe making decisions regarding a firm's operation (e.g. what and how to produce) is the single most important function of a market economy, and hence 'entrepreneurs' refer to people who are good at it. And within a given population, the number of people who are especially skilful at making decisions is generally small. The value of a firm is letting entrepreneurial people make decisions; so to best incentivize entrepreneurs making the correct decision, they have to shoulder the responsibility and be accountable for the consequences. Therefore they should take the residual as income rather than a contract-specified salary. Moreover, since there is the problem of asymmetric information about a particular entrepreneur's ability, it is imperative how to ensure that the decision-makers of the firm are entrepreneurial through an institutional arrangement. The observed institutions of the enterprise under capitalism are the natural product of the free market to solve this problem. It is not a deliberate choice by the legislatures, but a result of competition. 'Capital hires labour' is a system, which guarantees that only qualified people will be chosen as entrepreneurs (operators of the enterprise).

Since 1994, I have been trying to apply my theory of the firm to the reform of state-owned enterprises (SOEs), and have written a serious of articles, collectively published as*Theory of the Firm and Chinese Enterprises' Reform* (2015). My main conclusion is that, although SOEs can provide short-term incentives to their operators through means like profit sharing and bonus, there is no way to guarantee the long-term effect, let along setting up a system of selection for entrepreneurs. The reason behind is that bureaucrats acting ownership rights on behalf of the state are not the true owners of the capital. In spite of them having the right to choose the operator of the firm, they themselves are not accountable to the financial risks involved. What they care about is the private benefits from control of the enterprise, rather than the monetary profit. Therefore, it is impossible for them to have the incentive to select truly entrepreneurial candidates to be the operators of the SOEs. It is precisely for this reason that SOE should be called a political organization rather than an economic organization—political struggle for control is typical of SOEs. The problem of operator's selection and long-term incentive can only be resolved by the privatization of the SOEs, so that true ownership can be defined.

An Entrepreneurial/Contractual Theory of the Firm is also the basic framework that I use to analyse the corporate governance structure. During the past few decades, the corporate governance has been a very popular topic among both theorists and practitioners. But in my opinion, the main stream theories on corporate governance can be simply called "the manager-centred model". This manager-centred model neglects entrepreneurship and shows little trusts in the market system. If we

want to truly understand the market and corporate governance, and provide useful policy recommendations, we need to transform the manager-centred model to the entrepreneur-centred model. I tried to provide such theories in *Understanding Company* (2014).

My theories of the firm are further developed in the first part of *The Logic of the Market* (2015). In contrast to Coase's theory, I treat firms as a functional form of market rather than a substitute for market. The firm is an organization with joint liability. The efficiency of the market relies upon the trusts between market participants. It is through joint liability that the owner or entrepreneur has for the employees and suppliers in the supply chain, that consumers can hold producers responsible and accountable, and therefore put trust in them, which in turn lowers the transaction costs of the market, and improves the efficiency of the market. If there is no organization like firms, there will be no market exchange at massive scale, nor the emergence of global market. Hence, Coase was wrong in his interpretation of the firm as a substitute for market. Furthermore, for consumers, a big branded enterprise plays the role of a grand contractor, which acts on behalf of consumers to supervise many other medium and small enterprises along its value chain. It reduces the cost of supervision and makes consumers trust the numerous complex products on the market. In this sense Adam Smith's criticism of big enterprise has also drawbacks. An important implication is that we have to reconsider the economics foundation of anti-trust legislations. The economics of anti-trust laws assumes that the enterprise only performs the role of a production unit, and ignores its roles as a carrier of reputation and creativity. The so-called 'perfect competition' in the traditional economics is in fact of no competition, as pointed out by Hayek (1946). A 'perfectly competitive' market cannot be an efficient and orderly market.

Since its first (Chinese) publication in 1995, this book has become one of the most cited works in economics and management science in China. It has generated significant influence on debates and policy formation on China's state-owned enterprises reform. While more than two decades have passed, the ideas expressed in this book is still as relevant as when it was first published. I hope the publication of this English edition will make a contribution to literatures on both the theory of the firm in general and China's SOEs reform in particular.

July 2017 Weiying Zhang

References

Coase, Ronald, 1937, "The Nature of the Firm", *Economica*, IV, pp. 368–405.

Hayek, F. A., 1948, "The Meaning of Competition." in *Individualism and Economic Order*. 1976. London and Henley: Routledge and Kegan Paul.

Weiying Zhang, 2014, *Understanding Company: property rights, incentive and governance, the 2nd edition*. Shanghai People's Publishing House. (The first edition was published in 2007 by Economic Science Press.)

Weiying Zhang, 2015, *Theory of the Firm and Chinese Enterprises' Reform, the 3rd edition*, Shanghai People's Publishing House. (The first and second edition were published respectively in 1999 and 2006, by Peking University Press.)

Weiying Zhang, 2015, *The Logic of the Market: An Insider's View of Chinese Economic Reform*. Washington DC: the Cato Institute Press. (The first and second Chinese editions were published respectively in 2010 and 2012 by Shanghai People's Publishing House.)

Preface to the First Chinese Edition

This book is based on my D.Phil. dissertation at Oxford University. The thesis is about which factors determine the ownership of firms in a market economy: Why does capital hires labour rather than the other way round? Why do entrepreneurs supervise workers but not vice versa? Why do owners of capital choose managers rather than workers do? What factors determine what kind of people will become entrepreneurs in equilibrium? I attempt to find the fundamentals behind the above questions. Despite the fact that the thesis is very theoretical, the motivation behind the choice of this topic is very practical. In the late 1980s, I worked at the Institute of Economic System Reform of the State Commission for Restructuring the Economic System, and did research on economic theory and reform policy. During that period of bold economic reforms, many problems emerged and made entrepreneurship a hot topic in the field of economic theory in China. Economists almost unanimously agreed that emergence of entrepreneurs was essential to guarantee the success of the reform and the efficient operation of the market system. However, they disagreed on the way of building this team of entrepreneurs, especially regarding the relationship between the birth of entrepreneurs and ownership. The mainstream view at the time was that entrepreneurs are important, but ownership is irrelevant. Entrepreneurs are found through a fair and competitive market environment and independent decision-making power, not related to a particular ownership. Some even cited examples of the separation of control and ownership in market economies from Japan in arguing that it is only because owners have relatively little power so that entrepreneurs could emerge. At the time I held the opposite view. I treated entrepreneurship as a product of a specific arrangement in property rights—without well-defined ownership there can be no entrepreneurship. Therefore, the key to creating a team of entrepreneurs is through reforms in ownership rights (Zhang 1986a, 1986b). However, at the time I lacked forceful theoretical tools to counter the shallow, mainstream view. In October 1987, I went to Oxford as a visiting student and started learning the then recently developed theories of the firm. Although this newly acquired knowledge hugely enlightened me, they did not give me a ready-made theoretical weapon. For instance, people like Coase studied the reason behind the existence of the firm,

but did not answer the question of why capitalists rather than workers became the owner of the firms. Principal-agent theories developed in the late 1970s defined shareholders as principal and managers as agents, and studies how the principal designs contracts in order to best incentivize agents to achieve the targets of the principal. Under such setting, the principal-agent relationship is given, but for me, a more fundamental question is who should be the principal and who should be the agent in the first place. Why did the owners of capital become the principal? In particular, if, like the theories assume, the profit of the firm does not directly depend on the actions of the principal, then why could the incentive problem for the agent not be solved by making agent become principal? In September 1990, I went back to Oxford to start my doctoral studies, and chose the theme of my thesis to be *Why Capital Hires Labour*. I think that the key to answer this question is to further explore the trinity of operators, entrepreneurs and capitalists in the classical capitalist firm—why do decision-makers claim the residual to become entrepreneurs, why capitalists have priority in becoming entrepreneurs. Only with these problems being solved could we truly understand the institutions of modern enterprises. With the help of the fast-developing information economics, my research progressed relatively smoothly. At the end of 1991, the basic thinking and model-building work had already been finished. A draft of the thesis was submitted as a Master thesis and won the George Webb Medley Prize for the best thesis at Oxford. This made me believe that my research was a new insight into the above problems.

I graduated and came back to China in August 1994. I introduced my theory to my peers in China and applied it to the analyses of the reform of China's state-owned enterprises (SOEs). Hui Yu and Shijin Liu of China Academy of Social Sciences (CASS) translated part of the first chapter and published it on *Economic Research Journal* (1994). I gave a speech at a biweekly academic conference held by CASS's Institute of Economic Research at the invitation by Gang Fan. I systematically introduced my theoretical framework during the lectures on Industrial Organization for graduate students at Peking University and CASS. Besides, I wrote two articles on Chinese SOE reforms that were published on *Economic News, Reform Magazine and Chinese Industry and Business Times*. All of these efforts brought readers' interests and attention, as many of them wrote to or phoned me and said that they thought my theory was very path-breaking and wanted to read the full version. Some of my students also suggested that I publish a Chinese version of my doctoral thesis, which would later become this book.

Like most other theoretical works, this thesis almost inevitably used maths, although to be honest the maths involved here is basic, mostly within the range of calculus and probability theory, and is much easier than the maths used by many articles in academic journals. For those who have been trained in intermediate level microeconomics, reading this book should not be difficult. Or readers can use this book to check the level of one's microeconomics. If any reader finds reading this book hard, then I suggest that the reader refresh his microeconomics. I do not expect this book to be universally popular amongst economics students, but I believe that for those who are aiming at doing economic research, especially theory of the firm, it is worthwhile to thoroughly read this work.

Unlike the main body of the book, the four appendixes are about the Chinese economy. They are like the application of modern economic theory of the firm (including my own), and had already been published in academic journals. I collected these articles with the aim of providing some examples of the theories' application.[1]

I want to thank Hui Yu, Shijin Liu, Chunlin Zhang, Li Guiren, Ci-ao Zhou, Youchang Wu, Jie Ma, Zhonghua Wang, and Wei Yan for their contribution of translating the English version into the Chinese one. I also want to thank my students at Peking University and CASS and other readers for their interests in my research. Finally, I want to thank Chen Xin for his efforts in making the Chinese version published.

May 1995 Weiying Zhang

References

Weiying Zhang, 1986a, "Entrepreneurs and Ownership", *The Research Report on Economic System Reform*, No.30 (1986), Beijing: China Institute of Economic System Reform.

Weiying Zhang, 1986b, "Making True Entrepreneurs", *People's Daily*, September 19, 1986.

[1]The Appendices 3 and 4 are not included in this English version.

Acknowledgements

My first thanks go to my University Supervisor Donald Hay and College Supervisor Jim Mirrlees for their guidance. The implicit contract was that Donald was responsible for overall supervision of the thesis and Jim for technical problems of modelling. During 3 years, I met each of them virtually every 2 weeks in term time for discussion of various problems concerning both the ideas and formulation of the thesis. I am sure that I was one of the best supplied students at Oxford in terms of provision of supervisors' time. Their encouragement and academic advice were crucial for completion of this thesis. On several occasions, when I was blocked by some analytic problems, they found a way out for me. Donald has been particularly kind to me, and he has done for me much more than a supervisor is normally assumed to do, My gratitude to him is beyond power of expression in English.

My thanks also go to Meg Meyer for her supervision, when Donald was on Sabbatical leave during Michaelmas Term 1992. In particular, her constructive comments on my M.Phil thesis were very helpful to me in completing the present thesis, which is an extended version of my M. Phil thesis. Meg's help has been always available to me even after she gave up her official duty.

During my stay at Oxford both as a degree student and as a visiting student earlier, I have benefited from tutorials with several Oxford economists. Among them, I am particularly grateful to Christopher Bliss, John Vickers and E. Eshag. I would also like to thank Cyril Lin for his help in arranging my studies at Oxford.

Many of my friends at Oxford have helped me in one way and another. I would particularly like to thank Lizuo Jin, Tru-Gin Liu, Lina Song, Duo Qin, Duo Xie, Dahong Wang, Yuan Cheng, Gang Wei, Caigong Qin and Shaojia Liu for their help and friendship.

I am very grateful to the World Bank for granting me a Graduate Scholarship, to the London School of Economics for a Lionel Robbins Memorial Scholarship, and to the British Government for an Overseas Research Studentship. I am also grateful to Nuffield College for offering me a fully funded studentship in the third year (fund in the event was not taken up).

Finally, I would like to thank my wife Joyce Jinhong Ma for her companionship, understanding and support during my studies at Oxford. The birth of our son Luke Yasheng in the final stage of my studies has given me great joy. I dedicate this thesis to him with my love. I would also like to thank my mother-in-law for looking after Luke which released me from some housework at the crucial stage of the thesis, and my parents for their love.

Hilary Term 1994

Contents

Chapter 1
Introduction: Why Does Capital Hire Labour?

1.1 A Brief Description of the Thesis

The firm is the typical organizational form of the market economy. The most significant characteristics of the firm are the asymmetric contractual arrangements between different participants (factor-owners) in both distribution of returns and control rights. Within the firm, some participants are called "employers", while others are called "employees". Employers hold "authority" over employees and are entitled to claim the residual returns, while employees are obliged to obey the authority of employers within certain limits and are entitled to fixed wages. In the terminology of principal-agent theory, employers are principals and employees are agents. This "micro" asymmetry between employers and employees directly determines a "macro" asymmetry. In society, employers belong to an upper-class, while employees belong to a lower-class. For this reason, this topic about the firm attracts attention not just from economists but also from sociologists, political scientists, politicians and, in particular, social reformers.

The employment relationship takes place between capital and labour. An important question which has puzzled economists as well as others for long time is: why does capital hire labour rather than labour hire capital? This question is specially relevant today for two reasons. First, almost all socialist countries have experienced the failure of the socialist planned economy and have now begun a market-oriented reform program. Although Yugoslavia's experiment has shown that a labour-managed economy cannot be an efficient option, there is no guarantee that other socialist countries will not be attracted by the labour-hiring-capital system when they begin to deviate from the traditional planned economy. In particular, for ideological reasons, the labour-hiring-capital economy may be thought to be the only "acceptable" choice for some socialist countries. Secondly, in the joint-stock company, "ownership" is separated from management and the traditional conception of the employer is no longer as relevant as in the owner-managed firm. Instead, shareholders hire the management who in turn hire workers. That is, the traditional single agency relationship between a capitalist-entrepreneur and the workers has been

W. Zhang, *The Origin of the Capitalist Firm*, DOI 10.1007/978-981-10-0221-2_1

replaced by an agency-chain between capitalists and management, and management and workers. Many economists have focused their attentions on how capitalists as the principal make an optimal incentive scheme to induce the management (agents) to act in their best interests, or how the managerial behaviour deviates from shareholders' interests; but the most fundamental question is why the principalship should be assigned to capitalists rather than management in the first place. The logic behind this question, if the firm's output does not directly depend on the actions taken by capitalists, is why could not the incentive problem associated with the separation of ownership and management be solved by assigning the principalship to the management and let the management work for themselves? Or more generally, why do we need capitalists?

This thesis is intended to explore the elements determining the assignment of principalship within the firm: Why does capital hire labour rather than labour hire capital? Why does the entrepreneur monitor workers rather than workers monitor the entrepreneur? Why do capitalists rather than workers select the management of the firm? What factors determine who will be the entrepreneur in equilibrium? We are concerned with an economy in which all economic actions fall into two types: marketing and producing. By "marketing" we mean the activities of "discovering the relevant prices" (Coase 1937, p. 390) including speculating about profitable opportunities, forecasting market demands and making "judgmental decisions" (Casson 1982) of "what to do, and how to do it" (Knight 1921): in Schumpeter's words, setting up a production function. By "producing" we mean all the activities of transforming inputs into outputs "physically" under the given production function (technology) and according to marketing decisions.

Individuals in the economy are assumed to differ in (1) their marketing ability (entrepreneurial ability), denoted by θ; (2) personal assets, denoted by W_0; and (3) risk-attitudes, denoted by R. Because individuals differ in their marketing ability, it may be profitable for them to cooperate by setting up a "firm" through which individuals who have advantages in marketing specialize in making marketing decisions, while those who are not good at marketing specialize in producing (note that we assume that individuals are identical in their producing ability). Because of "uncertainty" (Knight 1921) and "team production" (Alchian and Demsetz 1972), the firm involves an agency problem — some member may take actions (e.g., shirking) which benefit himself but cost others. The key organizational issue is to design a contractual arrangement between different participants of the firm so as to make each member as responsible for his own actions as possible. We will argue that the member who does marketing should be assigned to be the principal to claim the residual return and to monitor others, not just because he is the major "risk-maker" but mainly because his actions are the most difficult to monitor. Thus he becomes the entrepreneur while those who do producing become the workers.

Under the assumption that personal assets W_0 are costlessly observable for all individuals while marketing ability θ is private information (or observable only at some cost), we will demonstrate that capitalists with high marketing ability will be the winners of the competition for being the entrepreneurs because their costlessly observable capital stocks can work as a device to signal information about marketing

ability, and the arrangement therefore saves transaction costs. In other words, when information of ability is asymmetric between the insider and outsiders, only those would-be entrepreneurs who possess enough personal assets can be trusted as qualified entrepreneurs. Capitalists are more likely to be honest, credible, responsible and industrious when they choose to be entrepreneurs. They have less incentive to overstate their entrepreneurial ability, or to overinvest. A capitalist can earn "pure" profit, because his capital economizes on transaction costs by signaling information.[1] In short, we show that capital-hiring-labour is a mechanism which guarantees that only qualified people will be chosen to be entrepreneurs (/managers); in contrast, if labour hires capital, the market for entrepreneurs (/managers) would be full of lemons (i.e., too many unqualified people would choose to do marketing).

Finally, we set up a general equilibrium entrepreneurial model of the firm, in which marketing ability θ, personal assets W_0 and risk-attitudes R are identified as the three key factors determining the choices of being an entrepreneur or a worker or a manager or a pure capitalist. We will show that there is a general equilibrium in which all individuals have their (constrained) optimal choices and different forms of the firm are chosen so that both the labour market and the capital market stay in equilibrium (goods market equilibrium can be understood as a by-product of labour market equilibrium and capital market equilibrium). In particular, we will show that the equilibrium relationships (both in pecuniary and non-pecuniary terms) between the firms' members depends on the joint distribution of marketing ability, personal wealth and risk attitudes in the population. Given asymmetry of distribution of entrepreneurial ability and distribution of personal wealth, the joint-stock company as a cooperation between capital and ability occurs if the costs of searching for high ability people are not prohibitively high. And it is also socially optimal to allow capitalists to select the management because the more personal assets a person holds, the more incentives he has in searching for high ability people. We argue that the major function of shareholders is to select a high ability manager rather than to monitor an incumbent manager.

One of the implications of our hypothesis is that because advantages of capital over labour are associated with information-cost saving, we may predict that these advantages will be diminishing as other signals become available. Education is one such signal, which may reveal some information on marketing ability and therefore help some MBA-holders to become managers. In the extreme case, if information is perfect, capital would become a pure production factor and would lose all its advantages over labour. In fact, in this case nobody has any advantages over others in marketing, and thereafter the firm becomes redundant in Coase's sense. However, if we believe that marketing is some kind of innate ability which is not entirely educable, capital will still enjoy advantages over labour in signaling information about a person's marketing ability.

[1] The reader who is familar with signalling models may find that the use of the term "signalling" here is somewhat misleading, since (initial) personal wealth is not a choice variable, unlike what is usually meant by a "signal" in the literature. What I really mean by "signalling" is that the entrepreneurial choice of the wealthy is more informative than that of the poor.

Some problems of terminology need to be emphasized. In this thesis, two kinds of terminology are used alternately, one borrowed from the traditional entrepreneurial theory, and the other from modern agency theory. Following Knight (1921), the "entrepreneur" is identified with two functions: making business decisions (marketing) and bearing business risks (residual-claiming). In our analysis, these two functions turn out to be only weakly separable. If a capitalist chooses to do marketing himself, he becomes an entrepreneur; if instead he selects another agent for marketing, the latter becomes a manager, while he becomes a security holder. In the latter case, they become the joint-entrepreneurs. For this reason, we say a joint-stock company is characterized by decomposition of entrepreneurship rather than separation of ownership and control. "Marketing ability" can be understood as conventionally used "entrepreneurial ability". We use "marketing ability" instead of "entrepreneurial ability" because in this thesis marketing comes first and entrepreneurship is derived rather than assumed. That is, when we deal with marketing, we do not know who is going to be the residual claimant. A "pure capitalist" is a security-holder who supplies the firm with capital but does not do marketing (in this thesis we do not distinguish between shareholders and bond-holders).

As in agency theory, a "principal" is defined as the party who bears some risk for the other party's actions and in return secures the right to monitor the latter; correspondingly, an "agent" is the party who may be not necessarily responsible for his own actions. In other words, in a principal-agent relationship, the "risk-taker" (principal) is not necessarily the "risk-maker" (agent). The agency problem exists as long as the contract can not be complete. However, unlike agency theory in which it is assumed that outcomes depend on only the agent's actions, the transactions with which we are concerned are such that outcomes may depend on actions taken by the agent as well as by the principal. The prime "incentive scheme" associated with the agency problem is to assign principalship: which party should be the principal. For this reason, the distinction between the agent and the principal should be treated with caution. In many cases the distinction makes sense only to some degree. For instance, in a classical firm, the entrepreneur bears all risks while workers receive fixed returns, and therefore we call the former the principal and the latter the agent. However, if we assume that workers' skill is "firm-specific" and their fixed return within the firm is higher than the market wage, they have to bear some risks for the entrepreneur's actions. In this sense, workers are the principal while the entrepreneur is the agent.[2] Another example is the partnership in which each partner has dual-identity: he is an agent as well as a principal.

In the literature, "authority" and "direction" are widely used interchangeably. In this thesis, they will be distinguished. The latter is associated with the marketing function, and the former with principalship. The marketing function requires the producing party to "obey" the marketing party's "direction" about what to do and how to do it. However this direction is not necessarily related to principalship which entitles the principal to monitor the agent's performance, to determine the

[2]For this reason, workers may demand to share some control over management in the case of "bankruptcy". See FitzRoy and Mueller (1984).

Table 1.1 A classification of types

		Marketing ability	
		θ_H	θ_L
Personal wealth	W_{0H}	**Type** E : (θ_H, W_{0H})	**Type** C : (θ_L, W_{0H})
		Entrepreneurs	**Capitalists**
	W_{0L}	**Type** M : (θ_H, W_{0L})	**Type** Z : (θ_L, W_{0L})
		Managers	**Workers**

reward-penalty and to hire/fire the agent. We identify that entitlement as "authority". The distinction can be best exemplified by the worker-managed firm in which workers as a whole hold authority over the appointed management while the manager still has the right to direct workers about what to do and how to do it. Another example is the relationship between the patient and the doctor. The patient has authority over the doctor, while the doctor directs the patient. The integration of "direction" and "authority" in a typical capitalist firm may be responsible for many economists ignoring this distinction.

The basic ideas of this thesis can be described by using a two-way classification of individuals (we omit the risk-attitude dimension).

In Table 1.1, for simplicity, we assume that both marketing ability and personal wealth have two-point distributions: for any individual, θ takes one of two values, θ_H (high) or θ_L (low); similarly, W_0 takes one of two values, W_{0H} (rich) or W_{0L} (poor). Thus there exist four types of individuals, denoted by E, C, M and Z respectively. Type E individuals are rich in both marketing ability and personal wealth: (θ_L, W_{0H}); Type C are rich in personal wealth but short of marketing ability: (θ_L, W_{0H}); Type M have high marketing ability but low personal wealth: (θ_H, W_{0L}); Type Z are poor both in marketing ability and personal assets: (θ_L, W_{0L}). If all transactions between individuals take place through spot-markets, each individual has to work as an individual businessman in dealing with both marketing and producing. Taking Type Z as the yard-stick, C has an advantage in capital factor, M's advantage is in marketing ability, and E has advantages in both capital and marketing ability. Obviously it may be profitable for different types of individuals to cooperate by forming a "firm" in which some individuals are specialized in marketing activities while others are specialized in producing activities. Two problems associated with the firm are: first, how to allocate the marketing function and producing function to different members; second, how to resolve the agency problem by assigning principalship. What we are going to demonstrate is that, as the firm substitutes for the spot-market, (i) Type E become entrepreneurs by doing marketing, monitoring producing-members, and claiming the residual; (ii) Type M become managers by doing marketing, monitoring producing-members but being monitored by Type C, and sharing some risk with Type C; (iii) Type C become capitalists by selecting and monitoring managers, and bearing risks; Type C and Type M together become joint-entrepreneurs; and (iv) Type Z become workers by specializing in producing, and receiving a fixed return.

We also show that bargaining power of each type depends on the joint distribution of θ and W_0 in population. For instance, an increase in the proportion of Type C in the population will disadvantage Type C but advantage Type M (managers) and probably Type Z (workers).

1.2 A Critical Review of the Theories of the Firm

More than fifty years ago, in his classic paper (1937), Coase pointed out that "economic theory has suffered in the past from a failure to state clearly its assumptions. Economists in building up a theory have often omitted to examine the foundations on which it was erected" (p. 386). It seems that even today economists are still suffering from a failure to examine some important assumptions on which their theories are based, although it should be recognized that the situation has been greatly improved since Coase (1937). One such failure is the assumption that capital hires labour. In neoclassical economics, the firm is treated as no more than a production function, and capital and labour as no more than production factors. As production factors, capital and labour are symmetric and their respective returns are determined by their respective contributions to production. In equilibrium, the wage is equal to the marginal productivity of labour and the interest rate is equal to the marginal productivity of capital. In Clark (1899) diagram, the triangle in the labour-marginal output space is equal to the rectangle in the capital-marginal output space; the converse is also true. Therefore, the observed profit in the long-run must be understood as some form of imputed factor returns (either as the wage of management or as the interest of capital). Although neoclassical economists based their theory on the assumption of capital hiring labour, they said nothing to guarantee that labour will not hire capital. In fact, as Paul Samuelson (1957) pointed out, it made no difference whether capital hired labour or labour hired capital. Perhaps the best we can infer from neoclassical economics is that "capital hires labour because capital is much more scarce than labour". Not surprisingly, using the neoclassical framework, Vanek (1970) could demonstrate that a labour-hiring-capital system can be as efficient as a capital-hiring-labour system in the long-run.[3]

Neoclassical economics of the firm has been challenged by many economists. All these challenges can been classified into three branches: (1) the contractual theory of the firm; (2) the entrepreneurial theory of the firm; and (3) the managerial theory of the firm. In the following, we survey each of them but with a focus on the first two.[4]

[3]The contributors to the literature of "neoclassical theory of labour-managed firms" include, among others, Ward (1967), Domar (1966), Vanek (1970), Meade (1972). This literature can be taken as a "mirror image" of neoclassical theory of the firm. It shows, in the neoclassical framework, what would happen if labour hires capital instead of capital-hiring-labour. We will not survey this literature.

[4]For a comprehensive survey of the theory of the firm, see Holmstrom and Tirole (1989).

1.2.1 The Contractual Approach to the Firm

The "mainstream" contractual theory of the firm was pioneered by Coase (1937), and further explored by Alchian and Demsetz (1972), Williamson (1975, 1979, 1980), Klein et al. (1978), Jensen and Meckling (1976, 1979), Leland and Pyle (1977), Ross (1977), Cheung (1983), Grossman and Hart (1986), Holmstrom and Tirole (1989), Hart and Moore (1990), and Aghion and Bolton (1992), among others. The most recent model is developed by Yang and Ng (1995). The common theme of this approach is that the firm is a "nexus of contracts" (written and unwritten, explicit and implicit). However, emphases are different from one author to another. The most influential theories are transaction costs theory and the agency theory. The former focuses almost exclusively on the relationship between markets and the firm (i.e., what are the boundaries of the firm? why does the firm exist?); the latter focuses on the relationship between internal structure of the firm and agency problems within the firm. In the following, we classify the transaction costs theory into two branches: the theory of indirect pricing and the theory of asset specificity; and we also classify the agency theory into two branches: the theory of agency costs and the principal-agent theory. [5] In addition, we will also discuss the fast growing literature on security design. It should be pointed out that the survey is selective rather than exhaustive.

1.2.1.1 Transaction Cost Economics (1): The Theory of "Indirect Pricing"

It is appropriate to summarize Coase (1937), Cheung (1983) and Yang and Ng (1995) under the title of "theory of indirect pricing". The key point of this theory is that the firm functions to save costs of direct pricing in markets (or market transaction costs).

Coase (1937) was the first to explore the rationale for the existence of the firm characterized by authority in terms of the transaction costs of using the market price mechanism. For him, the market and the firm are two alternative instruments for the allocation of resources, and can be substituted for one another; the difference between them is that in the market, the allocation of resources is directed by the impersonalised price, while within the firm, the same work is done by authority. The choice between them depends on the balance of market pricing costs and bureaucratic costs of the firm. The firm occurs because authority can greatly reduce the number of transactions which need separate pricing: the entrepreneur or the agent who contractually holds a limited set of use rights of inputs directs production activities without reference to the price of each activity.[6]

[5] The second classification is based on research methodology consideration.

[6] Coase writes: "It is true that contracts are not eliminated where is a firm, but they are greatly reduced. A factor of production (or the owner thereof) does not have to make a series of contracts with the factors with whom he is cooperating within the firm, as would be necessary, of course, if this cooperation were a direct result of the working of the price mechanism. For this series of contracts is substituted one" (p. 391).

No doubt, "authority" is a very important aspect characterizing the firm. But Coase failed to distinguish "authority" associated with principalship from "direction" associated with marketing function. Moreover, he failed to tell us why the authority of the firm is held by capitalists rather than workers. In fact, in a Coasian firm, as in a neoclassical firm, the relationship between capital and labour is still symmetric: it is irrelevant who holds the authority. Coase correctly pointed out that "the most obvious cost of 'organizing' production through the price mechanism is that of discovering what the relevant prices are" (p. 390), but he did not connect this cost with the internal structure within the firm. In the present thesis, we take this cost as a key to explore the asymmetric arrangement within the firm by focusing on differences of marketing-ability between individuals.

Cheung (1983) has refined and developed Coase's theory of the firm by giving a more insightful interpretation of the nature of the firm. For Cheung, the distinction between the market and the firm is a matter of degree, and they are just two different types of contractual arrangements. The firm occurs when a private factor owner surrenders use rights to the agent in exchange for income under a form of contract that binds the factor owner to follow directions instead of determining his own course of actions by continual reference to the market prices of a variety of activities he may perform. The firm is not designed to supersede "the market", rather, it replaces product markets with factors markets, or "one type of contract supersedes another type" (p. 10). Market transactions involve products or commodities, while "firm transactions" involve factors of production. Because of the costs of measuring and of obtaining information about a product, pricing by measuring some proxies for inputs often costs less than directly pricing output. However, pricing a proxy does not channel as full a set of information as pricing a product. Hence a choice between the two different types of contractual arrangements depends on whether the transaction costs saved in proxy pricing can more than offset the loss of certain information.

Cheung's argument that the firm replaces product markets with factor markets is penetrating. It implies that the argument that the firm can eliminate opportunism need not be decisive since the firm may bring opportunism from goods markets into factor markets.[7] A logical development is to investigate contracts of factor transactions (e.g., labour contract and capital contract), which would lead to the internal structure of the firm dominated by the incentive/monitoring problem of Alchian and Demsetz (1972) type. Unfortunately, Cheung does not go further. He disregards this problem simply by declaring that shirking behaviour would not occur if pricing costs are zero.[8]

Based on the original ideas of Coase and Cheung, (Yang and Ng 1993, 1995) developed a general equilibrium contractual model of the firm with consumer-producers, economies of specialization and transaction costs. What distinguishes their model is that they explicitly connect the internal structure of ownership of the firm with costs of pricing and identify the equilibrium organizational form of the firm with transaction efficiency. In their model, the choice exists not between markets or

[7] See later discussion on Williamson's vertical integration.

[8] Also see Cheung (1992).

the firm, but between autarky, markets and the firm. They show that the emergence of the firm that promotes the division of labour may increases transaction costs compared to autarky as long as the increase in economies of division of labour outweighs the increase in transaction costs; given the existence of the firm, the structure of ownership matters since different structures involve different transaction efficiencies. An asymmetric structure of residual claims can be used to improve transaction efficiency and to promote the division of labour by excluding the activity with the lowest transaction efficiency from direct pricing and trading. They show that the structure in which the manager claims the residual occurs because the transaction efficiency for labour hired to produce management services is much lower than for labour hired to produce a final good with management services as an intermediate input since it is prohibitively expensive to measure efforts exerted producing intangible management and to measure the output level of management services. The claim to the residual of the firm by the manager is the indirect price of managerial services.[9]

An important difference between Yang-Ng and Coase is that according to Coase, an increase in transaction costs will decrease the scope of markets and increase the size of firms, while, according to Yang and Ng, such an increase may decrease both market transactions and firm-type transactions, *if* transaction efficiency differs

[9]They refer to their theory as "theory of indirect pricing". Their story runs as follows. There are many ex ante identical consumer-producers in an economy, and each individual as a consumer must consume a final good, called cloth, the production of which requires an intermediate good, called management services, as an input. Each individual as a producer can choose to produce either one or both of cloth and management services. However, because of economies of specialization, each individual's optimal decision is a corner solution. He may choose autarky which implies that he self-provides cloth and self-manages the production, or chooses specialization of producing either cloth or management services. Autarky generates a low productivity but incurs no transaction costs, while specialization generates a high productivity but incurs transaction costs. Hence there is a trade-off between economies of specialization and transaction costs. If transaction efficiency is high, the division of labour will occur in equilibrium because economies of specialization outweigh the transaction costs for the division of labour; otherwise autarky will be chosen at the equilibrium. Suppose that transaction efficiency is sufficiently high such that individuals prefer the division of labour to autarky. Then there are three different structures of residual rights which can be used to organize transactions required by the division of labour. The first, called structure 1, is comprised of markets for cloth and management services; specialist producers of cloth exchange cloth for management services with specialist producers of management services. For this market structure, residual rights and authority are symmetrically distributed between trade partners and no firms and labour markets exist. The second, called structure 2, is comprised of the market for cloth and the market for labour hired to produce management services within the firm; the producer of cloth is the owner of the firm and specialist producers of management services are employees. Residual rights and authority are asymmetrically distributed between the employer and his employees. The third, called structure 3, is comprised of the market for cloth and the market for labour hired to produce cloth within the firm; The professional manager is the owner of the firm and specialist producers are employees. The choice between the three structures is determined by relative transaction efficiency of each structure. Assuming that transaction efficiency is much lower for management services than for labour, then the institution of the firm can be used to organizes the division of labour more efficiently because it avoids trade in management services. Suppose further that transaction efficiency for labour hired to produce management services is much lower than for labour hired to produce cloth (because of difficulty in measuring intangible management services). Then the division of labour can be more efficiently organized in structure 3 than in structure 2. The claim to the residual of the firm by the manager is the indirect price of management services.

among individuals.[10] From historical point of view, Yang-Ng's thesis is more robust. More 200 years ago, Adam Smith pointed out: "The division of labour is limited by the extent of the market." The reduction in transaction costs enlarges the extent of the market and motivates the division of labour and therefore expansion of the firm. My own view is that markets and the firm are substitutes at the *micro* level, but complementary at the *macro* level. What we have seen from history is that market transactions and firm-type transactions have been expanding simultaneously. This positive correlation cannot be an accident. The formation of the firm reduces transaction costs of using markets through variety of ways. It is worthwhile to explore this point further within Yang-Ng's framework.[11]

The major proposition of Yang and Ng that a residual claim on the firm by the manager is optimal because directly pricing management services is more costly is confirmed by our model,[12] although our conclusion is derived from a different approach.[13] The main differences between their model and our model are as follows. First, in their model, individuals are *ex ante* identical and the division of labour is chosen because of economies of specialization, while, in our model, individuals differ *ex ante* in marketing ability and the division of labour is chosen because different individuals have different advantages.[14] Although we agree that endogenous comparative advantage based on specialization is important for explaining the coexistence of markets and firms, we argue that, in the context of entrepreneurship, exogenous comparative advantage is more fundamental and realistic. Historically, the firm is set up by entrepreneurial people who are endowed with high ability to "price" activities and opportunities. It is because of this exogenous advantage that these people hold a prestigious status. In contrast, if individuals are identical *ex ante*, in equilibrium, all individuals should be indifferent between being a worker and being a manager, and no one can be superior to anyone else. This is clearly not true.[15] Secondly, Yang and Ng investigate only one dimension of the problem associated with equilibrium structure of ownership, i.e., measuring management services, while we consider two dimensions, that is, relative importance in production and effectiveness of monitoring. We will see that, without consideration of each member's relative importance, the effect of measuring (equivalent to monitoring) is indeterminate. Thirdly, in their paper, the internal structure of the firm is investigated from an external angle, that is, market transaction efficiency, while we analyze the internal structure from the internal perspective, that is, interaction between different members within the firm.

---

[10] This statement is my own inference. Yang and Ng do not exploit this point.

[11] Given that Coase developed his thesis when he was a young socialist, it is understandable that he took the firm as a substitute for markets. Not suprisingly, Coase and Williamson's arguments have been used by some Chinese economists against markets.

[12] see Chap. 2.

[13] I read Yang and Ng's paper after my thesis had been fully formulated.

[14] Yang and Ng claim that when Ricardo's concept of exogenous comparative advantage based on constant return to scale is used to generate gains to trade, the productivity implication of the institution of the firm cannot be explored. This is not correct.

[15] In a personal communication, Yang suggested that it might be promising to set up a model with combination of exogenous and endogenous comparative advantages.

In fact, they follow Cheung (1983) proposition that the firm is a special contractual form to supersede goods markets with factor markets, while we deal with the Alchian and Demsetz (1972) type problem. We believe that our approach gives more insight for understanding internal relations between different members of the firm. Fourthly, Yang and Ng examine only one scene of the whole drama of the capitalist firm (e.g., they have not sought to deal with the problem of why capitalists have priority in specializing in management services), while we try to explore the whole drama.

1.2.1.2 Transaction Cost Economics (2): The Theory of Asset Specificity, Incomplete Contracts and Vertical Integration

Another branch of the Coasian approach to the firm is explored by Williamson (1975, 1979, 1980) and Klein et al. (1978), and further developed by Tirole (1986), Grossman and Hart (1986) and Hart and Moore (1990), Riordan (1990), and Dow (1993a, b), among others. This theory takes the firm as a vertically integrated entity of incomplete contracts between successive production processes, and argues that the firm occurs because the vertical integration can eliminate or at least mitigate the asset specificity generated opportunism problem when contracts cannot be complete.

Williamson (1975, 1979, 1980) and Klein et al. (1978) take the same position as Coase (1937) that the firm as a transaction mode serves to economize on transaction costs. However, they are concerned with whether a firm should "buy" or "make" a specific input, or how large the firm should be, rather than why the firm exists in the first place. They focus on "asset specificity" and related opportunism as the main determinants of transaction costs. The arguments are as follows. If the transaction involves the relationship-specific investment (i.e., an asset is idiosyncratic), *ex ante* competition will be followed by *ex post* monopoly or monopsony, which invite "opportunist" behaviour to appropriate the quasi-rents of the specialized asset. This opportunism problem makes spot market transactions very costly in the sense that it makes the contractor's relationship-specific investments suboptimal, and negotiation and enforcement of contracts more difficult. As relationship-specific investments become more important, the transactions costs associated with mediating a vertical relationship using conventional spot markets increase. Therefore vertical integration is more likely to substitute for spot markets since under vertical integration, the opportunism is checked by authority. In his early work, Williamson emphasized the choice between the spot market on the one hand and vertical integration on the other hand. His later work and a great deal of Klein et al's work, however, consider the long-term contract alternative to vertical integration where the transaction costs within the firm are non-trivial. If vertical integration is not economic because of diseconomies associated with internal production, long-term contractual arrangements to govern transactions between independent agents will emerge to economize on transaction costs.[16]

It should be pointed out that when the firm decides to "make" by itself rather than "buy" an input in the spot market, it must buy factors of production (labour

[16]Tirole (1986) formalized Williamson's idea that "opportunism" leads to underinvestment in the relationship in the context of procurement. He shows that Williamson's presumption holds under

and capital). This is exactly what Cheung (1983) means when he says that the firm substitutes factor markets for goods markets. Thus, vertical integration may bring opportunism from the market into the firm. The choice between markets and vertical integration is actually the choice between the opportunism in the market and the opportunism within the firm.

Intra-firm opportunism involves both "idiosyncrasies" and "non-separability" (see Alchian and Demsetz 1972), and it has much to do with the internal structure of the firm. Williamson analyzes the internal structure of the firm in the same spirit as he analyzes the choice between the market and vertical integration. He identifies the (capitalist) firm with the "Authority Relation" between capitalist and worker characterized by the hierarchic structure,[17] and emphasizes "idiosyncrasies" rather than "non-separability" as the key factor accounting for the employment relation. Idiosyncrasies create a bilateral monopoly between workers who come to possess firm-specific skills and employers who can refuse to renew a contract for the corresponding jobs, which threatens to make investment in training unprofitable. The long-term employment relation attenuates this problem by tying pay to well-defined jobs, rather than to individuals, and by filling higher slots through internal promotion and making promotion itself contingent upon long-term performance rather than short-run assessment. In particular, in his 1980 paper which responds explicitly to radical critiques of the hierarchical organization, Williamson compares six modes in terms of efficiency considerations and concludes that the capitalist employment relation is the most efficient mode in aggregation.

While Williamson's positive contribution to the understanding of the firm's internal structure is significant, from the point of view of the present thesis he has said little about why capital employs labour rather than labour employs capital.[18] In fact, he does not even attempt to do this. In his analysis, capitalists have already become employers before a hierarchy is justified, and he is more concerned with why the employment relation is characterized by hierarchy, than why the role of the employer is performed by capitalists in the first place. But hierarchy is not necessarily inconsistent with labour-hiring-capital. For this reason, in this thesis, we focus on the

(Footnote 16 continued)
very general assumptions about bargaining and about ex post asymmetric information as long as the firm's (seller) investment is not observable by the sponsor (buyer). If investment is observable by the sponsor and thus may become a joint decision variable, the two parties may choose to under- or over-invest. He shows that the absence of commitment and asymmetry of information are crucial for Williamson's proposition. Also see Tirole (1986).

[17] The other capitalist mode listed by him is "inside contraction".

[18] Klein et al. (1978) extend their argument of vertical integration based on asset specificity to "why the owners of a firm (the residual claimants) are generally also the major capitalists of the firm. They wrote: "(O)wners may rent the more generalized capital, but will own the firm's specific capital.This observation has implications for recent discussions of 'industrial democracy', which fail to recognize that although employees may own and manage the firm (say, through their union), they will also have to be capitalists and own the specfic capital. It will generally be too costly, for example, for the worker-owners to rent a plant because such a specific investment could be rather easily appropriated from its owners after it is constructed." Also see Hansmann (1988) and Dow (1993a, b).

"horizontal" asymmetry of bilateral relations between employers and employees rather than vertical hierarchy. Our arguments do not depend on idiosyncrasies at all, though we agree with Williamson when he says that "to the extent the requisite information-processing and decision-making talents are not widely distributed, efficiency will be served by reserving the central information collection and decision-making position to the one or few individuals who have superior information processing capacities and exceptional oratorical and decision-making skills" (1975, p. 52).

Following Williamson and Klein et al. Grossman and Hart (1986) and Hart and Moore (1990) developed the model of ownership structure. They show ownership matters when contracts cannot be complete because of costs of specifying all the particular rights.[19] They make a distinction between the residual claim and the residual rights of assets (which are defined as those not specified in the contract), and identify ownership with the purchase of the residual rights. They argue that when two parties enter into a relationship in which assets will be used to generate income and it is too costly to list all specific rights over assets in the contract, it may be optimal for one party to purchase all the residual rights. In particular, when residual rights are purchased by one party, they are given up by a second party, and this inevitably creates incentive distortions; an efficient allocation of residual rights is such that gains in the purchaser's incentive can sufficiently offset loss in the seller's incentive. They argue that the ownership of the residual rights is more likely to be allocated to the party whose investment actions are the most important.

Grossman–Hart–Moore is a major contribution to the contractual theory of the firm. As they pointed out, until their models, the literature on transaction costs had emphasized that incomplete contracts could cause a non-integrated relationship (market transaction) to yield outcomes that are inferior to those that would be achieved with complete contracts; it was implicitly assumed that integration yields the outcomes that would arise under complete contracts. Their model goes beyond this point. According to them, the relevant comparison is not between the nonintegrated outcomes and the integrated outcomes but instead between one type of integration and another type; the problem is not only whether integration should occur, but also and more importantly, who should integrate with whom. Since when the residual rights are purchased by one party, they are lost by a second party, integration shifts the incentive for opportunistic and distortionary behaviour, but it does not remove these incentive problems. Optimal integration is that which assigns the control rights to the party whose investment decision is particularly important relative to the other party, whereas nonintegration is desirable when both investment decisions are "somewhat" important.[20]

Our major dissatisfaction with the Grossman–Hart–Moore model is their confusion of ownership of the firm with ownership of assets. Their definition of the firm

[19]Note that in their model, ownership matters, not because authority can eliminate the ex post bargaining problem as Williamson argued, rather, because control rights define the *status quo* which affects the equilibrium solution of bargaining which in turns affects the *ex ante* incentive to invest.

[20]At this point, it might be interesting to draw a parallel between Yang-Ng's model and the Grossman–Hart–Moore model. If it can be said that Coase (1937) theory of the firm has been

as being composed of the assets it owns is particularly problematic. In fact, one of the major advances made by the contractual theory is that the firm is a collection of contracts instead of physical assets. Ownership of assets does not characterize ownership of the firm.[21] They define ownership of the firm by the residual rights over assets rather than the residual claim, but fail to explain how residual rights are connected with residual claims. Their definition has led them to focus on authority over assets rather than authority over "actions", and relationship between "firm 1" with "firm 2" rather than relationship between capitalists and workers.[22,23] However, for a theory of the firm, the horizontal relationships between different members *within* the firm are more fundamental than the vertical relationships between different firms.[24] To understand fully the employment relation associated with the firm, both residual rights and residual returns are important. After all, historically they were fully integrated, and it is difficult to imagine that residual rights can be entirely separated from residual returns. Perhaps the most interesting problem is to

(Footnote 20 continued)
divided into two branches—one is explored by Cheung (1983) and the other by Williamson (1975, 1979), then Yang and Ng (1995) and Grossman–Hart–Moore model are respectively the ownership theories of each branch. Yang-Ng developed Cheung's theory of factor markets superseding goods markets by linking costs of pricing with residual claim structure of the firm, while Grossman–Hart–Moore developed Williamson's theory of asset specificity by linking opportunism with residual control structure of the firm. For Yang and Ng, the relevant comparison is not simply between goods markets and factor markets but between different structures of residual claim; similarly for Grossman, Hart and Moore, the relevant comparison is not simply between nonintegration and integration but between different structures of control rights. The two models differ in their definitions of ownership and their institutional focuses: Yang and Ng define ownership by the residual claim and focus on which member should be the owner of the firm (the residual claimant); Groosman, Hart and Moore define ownership by the residual control rights and focus on which firm should be the owner of the integration (the holder of residual rights). In their paper (1994), Yang and Ng refer to their model as a "theory of indirect pricing" and Grossman–Hart–Moore model as a "theory of asset specificity". They say that a complete story of the firm that occurs in reality may be then predicted by a blend of the theory of indirect pricing and the theory of asset specificity. The theory developed in this thesis is closer to Yang-Ng's than to Grossman–Hart–Moore's in the sense that we are also concerned with which member should be the residual claimant rather than which firm should hold control rights.

[21] Ownership of a stock company is attributed to its shareholders. But what a shareholder actually holds is the share of residual claim, rather than the share of assets since much of assets are held by debt-holders.

[22] I conjecture that the causality may go the other way: they adopted this definition for their analysis.

[23] Xiaokai Yang has pointed out to me that the Grossman–Hart–Moore model is a theory of optimum ownership rather than a theory of the firm since their propositions hold even without the firm. Yang's comments are coincident with ours. They identified ownership of the firm with ownership of assets and thus developed a theory of optimum ownership of assets in the context of the theory of the firm.

[24] In the fictional example of Hart and Moore (1990), the ownership of assets is held by a "big worker" rather than a "small worker". According to this logic, it might be argued that capital hires labour because capital is a "big worker" and labour is a "small worker". But their focus is still on the authority over assets rather the authority over people. What has to be demonstrated is why capital is a "big worker" while labour is a "small worker". In fact, what they try to explain is something like why an electricity generating company may own a coal mine, rather than why a capitalist may "own" a firm.

understand how their combination has evolved. In the present thesis, we use principalship instead of ownership to characterize the internal contractual arrangement between different members of the firm. Here principalship refers to both residual returns and residual rights. However, we adopt the right to monitor or to exercise authority rather than residual rights since, in the context of the firm, the residual rights are not a well-defined concept.[25]

Although Grossman–Hart–Moore's definition of ownership is problematic, their analytic framework is very powerful. By explicitly introducing wealth constraints into their framework, Aghion and Bolton (1992) developed a theory of capital structure based on transaction costs and contractual incompleteness.[26] In their paper, incomplete long-term contracts between an entrepreneur with no initial wealth and a wealthy investor is modeled as "vertical integration"; both agents have potentially conflicting objectives since the entrepreneur cares about both pecuniary and non-pecuniary returns from the project while the investor is only concerned about monetary returns. They consider (i) whether and how the initial contract can be structured in such a way as to bring about a perfect coincidence of objectives between agents and (ii) when the initial contract cannot achieve this coincidence how control rights should be allocated to achieve efficiency. They show that control rights matter whenever some important future variables have to be left out of the contract if they are difficult or impossible to describe initially, and that different control arrangements are efficient for different values of monetary returns and private benefits.[27] In particular, (i) the entrepreneur's unilateral control is efficient whenever the entrepreneur's private benefits are "comonotonic" with total revenues[28]; (ii) the investor's unilateral control is efficient whenever monetary benefits are comonotonic with total revenues; and (iii) the contingent control is efficient whenever neither monetary nor private returns are comonotonic with total revenues.[29] Aghion and Bolton interpret contingent control as a control allocation with debt financing. If the first-period signal represents a default-no default event, then the entrepreneur gets control as long

[25] Rights over the choice of business (in Coase's sense) or rights over the management itself?.

[26] Essentially, the Aghion–Bolton model is a security design model. We put it here for tracing development of Williamson's theory.

[27] Control rights refer to authority of deciding what action to choose when a special signal is realized.

[28] By "comonotonic" they mean that the action which generates the highest total revenues also brings about the highest private benefits (monetary benefits).

[29] Arguments (i) and (ii) are obvious since in these two cases the first-best can be implemented by the controller's self-motivation without renegotiation in equilibrium. The intuition for (iii) is as follows. When the entrepreneur's private benefits are not comonotonic with total benefits it may not be feasible to implement the first best action plan (with entrepreneur control) without renegotiation in equilibrium, while renegotiation with entrepreneur control may not be feasible since it may require an excessively large fraction of the project returns for the investor to bribe the entrepreneur such that too little is left for the investor's participation constraint to hold. When monetary returns are not comonotonic with total returns, renegotiation (with investor control) is necessary for the first-best action to be implemented, but renegotiation may not take place since the entrepreneur's wealth constraint prevents him from bribing the investor for choosing an action which yield a lower expected monetary return (and higher private return). Contingent control is some mixture of the entrepreneur control and the investor control: In the state where the private benefits are comonotonic with total benefits, the entrepreneur takes control, while in the state where monetary returns are comonotonic with total

as he does not default on his debt obligation; but the investor gets control in the event of default. The Aghion–Bolton model is very close to a model of why capital hires labour. In the present thesis, we take a different line to address this problem.[30]

The essence of asset specificity is the lock-in effect. FitzRoy and Mueller (1984) (also see Mueller 1976) developed a different version of Williamson's asset specificity theory to deal with the internal structure of the firm. In their model, "immobility" is the main determinant of the internal structure of the firm. The firm is a cooperative agreement. Even if all members are mobile at the time they join the firm, they may become immobile over time due to the transaction costs of exit and entry, or the accumulation of non-transferable human capital. The degree of a factor's immobility is measured by the difference between its return from the present employment and from the next best employment, net of transaction costs of changing employment. They argue that arrangements of authority within the firm are determined by the distribution of immobility among the members. When all members are equally immobile, authority should be equally shared and the agreement should be characterized by high trust and consensual decision making; when asymmetry in mobility exists between members, authority should be concentrated in the hands of the immobile members. The reason is that incentives to shirk are decreasing with immobility while incentives to monitor are increasing with immobility. The mobile member does not mind the behaviour of the other members so long as he continues to receive an income equal to his opportunity costs; the immobile member has to bear the full costs of the mobile member's opportunistic behaviour. Thus the immobile member can be expected to demand a more explicit statement of the duties to be performed by the mobile member to facilitate monitoring. The mobile member exercises control on other members via exit, while the immobile member must rely on voice. Therefore the fact that managerial-monitoring authority goes to capitalists can be explained by immobility of capital.[31] Although FitzRoy and Mueller claim that they follow Williamson's transaction costs approach, they are concerned with horizontal asymmetric distribution of authority within the firm much more than vertical hierarchy, as we are. We do think "immobility" is a quite useful dimension in understanding the firm's internal structure of authority. Particularly their arguments about the trade-off between monitoring rights and high mobility throw some light on the relationship between the role of "voice" and the role of "exit" in governing managerial behaviour. However, our criticism of Alchian–Demsetz's argument on the

(Footnote 29 continued)
benefits, the investor takes control. As a result, the first best action plans are implemented in both states.

[30]The entrepreneur' wealth constraint plays a similar role in both their model and our model.

[31]"The theory developed here allows us to explain the widely observed identification of 'ultimate' managerial-monitoring authority with the ownership of material capital in another way. As was noted, physical capital can easily erode or even be destroyed in the short-run by free-riding by short-term workers, who are not contractually liable for losses or bound to future service. The benefits to such a worker from continuing the employment relationship are of a very long-run nature and uncertain in any case, and in appropriate circumstances, may be plausibly outweighed by the immediate gains from free-riding. This, we suggest, is the basic economic justification for the capitalists' demand to monitor the employment contract,…" (p. 72).

costs of monitoring capital applies to their argument about immobility of capital.[32] Insofar as the financial form of capital is concerned, there is no reason to believe that capital is more immobile than labour. The arguments in the present thesis have nothing to do with immobility. We argue monitoring authority should be assigned to the marketing member not because the marketing member is more immobile but because he is more difficult to monitor; principalship should go with capitalists not because capital is more immobile but because it signals.

Riordan (1990) presents the first formal model to deal explicitly with the trade-off between market transactions suffering from the information problem and vertical integration suffering from the incentive problem.[33] In his model, a downstream firm (principal) faces a choice of either buying a non-standard component (or standard one) from a upstream firm in the market or making it within the firm. But when it decides to make rather than buy, it has to hire a manager responsible for producing (the vertical integration will transform the previous "owner-manager" into a "employee-manager").[34] In the case of buying, the principal has no information about costs of production, while the owner-manager has full incentive to make an effort in reducing costs; on the other hand, in the case of making, the principal can observe costs, while the employee-manager has no incentive to make an effort in reducing costs because of difficulties of monitoring.[35] *Ex post* the production decision is more efficient when the principal can observe cost than when she cannot. As a result, vertical integration conveys better information about costs and yields a more efficient quantity decision, but undermines managerial incentives for cost reduction. The trade-off between non-integration and integration is actually between a distorted production decision and a distorted managerial incentives. A preferred organizational mode will be chosen depending on which effect dominates, which in turn depends on the value of the component for the principal, and the sensitivity of the cost function to managerial incentives. Vertical integration is more likely to be preferred where the component is more valuable and cost function is less sensitive to effort.

Dow (1993b) developed a bargaining model of why capital hires labour. His model is also based on asset specificity and incompleteness of contracts. However, his model differs from the aforementioned models in that in his model, the viability of alternative organizational form of the firm in competitive markets may not depend on the total surplus it can produce but appropriability of quasi-rent by the supplier of specialized assets. He argues that whenever specific investments are noncontractible (perhaps it is too costly to write down all technologically relevant characteristics of the asset), authority within the firm can influence quasi-rent distribution of the sunk asset and

[32] See next subsection "The Theory of Team Production, Moral Hazard and Agency Costs".

[33] Also see Lewis and Sappington (1991) for a model of effects of technology changes on the trade-off between procurement and self-producing.

[34] Although Riordan does not cite Cheung (1983) paper, his model illustrates Cheung's idea that the firm replaces the product market with factor markets.

[35] In Grossman and Hart (1986), vertical integration does not change information structure.

thus the viability of alternative organizational forms.[36] An organizational form can persist in competitive markets if it can satisfy the specific asset owner's participation constraint, even if it produces less total surplus than an alternative form. He therefore argues that capital-managed firms will be the equilibrium organizational form in industries where capital is more specialized than labour, while labour-managed firms will be the equilibrium form in industries where labour is more specialized than capital.

Dow's model is close to the Aghion–Bolton model. They both deal with the conditions under which the capital-supplier should be allocated control rights. However, Dow's arguments relies too much on the *physical* form of capital. Much of his argument would be undermined if the physical form were replaced by a financial form, since in the latter case, capital-managed-firms could persist only if they can produce more total surplus. If labour-managed-firms can produce more total surplus, then there is no reason why worker-owners cannot bribe capitalists by making a debt contract.[37]

1.2.1.3 The Theory of Team Production, Moral Hazard and Agency Costs

While much of transaction cost economics is focused on the choice between markets and firms (vertical integration), the theory pioneered by Alchian and Demsetz (1972) is more concerned with the internal structure of the firm (horizontal integration). The literature of this theory is extensive. In this subsection, we pick a few representative contributions which are relevant to the present thesis.

Instead of focusing on the transaction costs of using markets, Alchian and Demsetz (1972) concentrated upon the incentive problem (monitoring costs) in explaining the firm's internal structure. For them, the essence of the firm is "team production". Team production occurs when an output is produced by the simultaneous cooperation of several team members and any one member's activity may affect the productivity of the other members. Because the final output is the joint result of the combined efforts of all the inputs working at the same time and the individual contribution of each member cannot be isolated and observed accurately, it is impossible to reward each member according to his own contribution. This generates a shirking problem: team members have no incentive to work hard. To reduce shirking, it is necessary for some team member to specialize in monitoring others. The monitor should be the residual claimant because otherwise he has no incentives to monitor. For monitoring to be effective, the monitor must hold rights to revise the contract terms and direct other

[36] In his paper, authority refers to rights to decide how much to produce. A crucial assumption of the model is that the authority-holder does not fully internalize the effect of his decision unless the other party's participation constraint becomes binding.

[37] In Dow's model, worker-owners can exploit capitalists because they can avoid responsibility for compensating for depreciation of a machine.

members because without these rights he cannot fulfill his functions effectively.[38] In addition, the monitor should be the owner of the team's fixed inputs because the cost of monitoring the use of inputs by a non-monitor owner is too high. Thus the classical capitalist firm comes into being.

Although the present thesis shares some propositions with Alchian–Demsetz, and even its motivation can be traced to the their argument, there are basic differences in both assumptions and propositions between them. First, in Alchian–Demsetz, all team members are originally homogeneous —at least in terms of monitoring costs, and therefore the monitor can be picked at random among members, and it is the only important to entitle the monitor to claim the residual returns so that he has incentives to monitor. In contrast, we assume that team members are originally heterogeneous in their marketing ability and functions within the firm, and it is this heterogeneity which dominates the choice of the monitor. Secondly, in Alchian–Demsetz, the monitor is specialized in monitoring, while in the present thesis monitoring is only a function of the entrepreneur or the joint entrepreneur who is specialized in marketing and bears business risk. Because the entrepreneur does not make a living from monitoring, he can delegate this function to someone else, leaving himself to concentrate on marketing activities. Finally, Alchian and Demsetz attribute the observed phenomenon of capital-monitoring-labour to the costs of monitoring the use of capital, while we explain this employment relation by emphasizing capitalists' responsibility for selecting qualified entrepreneurs and/or managers. Costs of monitoring seem a plausible explanation initially, but on reflection it is unconvincing. The fact that a driver should own rather than rent a truck because of monitoring costs does not necessarily imply that he must buy the truck with his own money.

If the monitor's residual claim results from the non-separability of each member's contribution within team production, a logical proposition is that relative difficulties in measuring each individual's contribution should have an effect in determining

[38] Despite letting the residual claimant hold so many rights, Alchian and Demsetz explicitly criticized the Coase's argument of authority of the firm. Coase (1937) argued that the key difference between an employer-employee relationship and one between independent contractors is that whereas an employer can tell an employee what to do, an independent contractor must persuade another independent contractor to do what he wants through the use of prices. Alchian and Demsetz pointed out that an employer typically cannot force an employee to do what he wants: he can only ask him or fire him if he refuses. However, this is not different from one independent contractor's firing another (quitting their relationship) if he is unhappy with the latter's performance. I offer two reasons that might have misled Alchian and Demsetz. The first is that Coase failed to make a distinction between"direction" associated with business decisions of what to do and how to do it and "authority" associated with principalship; the second is that Alchian and Demsetz failed to realize that obeying authority is part of the price which a wage-worker has to pay after contracting. Hart and Moore (1990) offer a reconciliation between Coase and Alchian and Demsetz: "While it follows Alchian and Demsetz in not distinguishing between the contractual form or nature of transactions in the two relationships, (our approach) captures the idea that one agent is more likely to do what another agent wants if they are in an employment relationship than if they are independent contractors. The reason the manager of Alchian and Demsetz's grocery store will be more likely to follow their wishes if they employ him than if they are customers is that in the former case his future livelihood (they control the assets the manager intends to work with), whereas in the latter case it does not").

who should be the monitor. In their survey paper on the theory of the firm, Holmstrom and Tirole (1989) indeed emphasize this point. They argue that ownership is important in solving incentive problems associated with the firm. In particular, ownership (principalship, in our words) should go with the input factor whose marginal contribution is hardest to assess, and capital hires labour because the contribution of capital is hardest to measure and because capital is easy to misappropriate.[39] It should be noted that the hypothesis implied here is heuristic, and it has not been elaborated. In fact, there is some superficial similarity between their argument and our arguments in the sense that both agree that the factors whose actions are the most difficult to monitor should hold principalship. But it is not meaningful to say the contribution of capital is the hardest to measure. The owner of capital becomes the principal not because capital's contribution is more difficult to measure than the manager's contribution, but because capital can be used to signal information about the would-be entrepreneur's ability which otherwise would be more costly to acquire. Of course, if "contribution" covers this signaling function, Holmstrom and Tirole's argument might coincide with ours. But such an explanation leads us to the question of cooperation between the risk-maker (manager) and the risk-taker (capitalist).

Jensen and Meckling (1976) can be viewed as the "managerial" analogue of Alchian and Demsetz (1972). They identify "agency costs" as the determinant of ownership structure of the firm; agency costs arise from the fact that the manager is not a full owner of the firm. Under partial ownership, on the one hand, when the manager makes an effort, he bears the entire costs but can only capture a fraction of benefits; on the other hand, when he consumes perks, he enjoys the whole benefits but bears only a fraction of costs. As a result, he has less incentive to work but more incentive to pursue perks, and the value of the firm is smaller than when he is a full owner. The difference in value is called "agency costs" which have to be borne by the manager himself under the rational expectation of outside owners.[40] Agency costs can be eliminated or at least mitigated by letting the manager become the full residual claimant. Nevertheless, the ability of the manager to be the full residual claimant is constrained by his personal wealth. Debt financing may also be helpful since, for a given investment and his personal assets, the manager's residual share increases with the fraction of the investment financed by debt. However, debt

[39] "Changes in ownership may imply inevitable transfers of return streams, because of incomplete contracting. Therefore ownership may be the only means by which proper financial incentives can be provided. Ownership should go with the input factor whose marginal contribution is hardest to assess. .....Reinterpreted in this way, Alchian–Demsetz Theory can be read as suggesting that the monitor is the owner because his product is important but diffuse" (p. 73).

"We believe it is more likely that the contribution of capital is hardest to measure, because capital is easy to misappropriate. Consequently, capital should hold title to the residual return stream. This idea deserves further elaboration. Our main point is that the allocation of return streams via ownership can be a significant component in understanding which factor becomes the owner" (Holmstrom and Tirole 1989, p. 73).

[40] Broadly, agency costs include the costs of structuring, monitoring, and bonding a set of contracts among agents with conflicting interests. Agency costs also include the value of output lost because the costs of full enforcement of contracts exceed the benefits. In equilibrium, the value of the firm is determined after deductions of all these agency costs.

financing may invite another kind of agency cost. Under debt financing, the manager as a residual claimant has more incentive to invest in riskier projects since he can enjoy the benefit of success but leave the cost of failure to the debt-holders because of the limited liability. These agency costs are also borne by the manager since the debt-holders also have rational expectation. The equilibrium ownership structure of the firm will be determined by balance between equity agency costs and debt agency costs.

We share many of Jensen–Meckling's arguments. In particular, the demonstration in Chap. 3 of the present thesis is inspired by their argument about bankruptcy costs. But, it is worth considering the differences. First, Jensen and Meckling do not explicitly pose the problem of why the residual claim is attached to capital. Secondly, they focus on the agency problem incurred by separation of ownership and management, while we are concerned with the more general agency problem associated with the firm. Thirdly, their argument about bankruptcy costs associated with debts is based on "limited liability", while our arguments rely only on an assumption of "non-negative consumption". Finally and most important in the present thesis, the choice of the qualified manager (i.e., consideration of a person's ability) is the major determinant of capital-hiring-labour, while in their paper, only the choice of projects (degree of risk) matters.[41]

Jensen and Meckling (1979) are more explicitly concerned with the efficiency of capital-hiring-labour by answering the challenge from the "neoclassical theory of self-management". For the present thesis, two arguments given by them need mentioning here. First, they point out that the "pure rental capital" assumption on which labour-management's theoretical optimality hinges is not relevant because that: (a) investment in such intangibles as R&D cannot take the form of renting physical assets only, and (b) given the monitoring costs associated with asset use, ownership rather than rental is frequently the most efficient decision even with respect to physical capital. Second, they argue that, without a stock market and marketable claims to firms, little incentive exists for professional and public evaluation of firms, and therefore there will be insufficient monitoring of managerial behaviour.[42] Our point is that although their arguments might be effective in attacking the worker-managed firm, they are not relevant in general.[43] First, as we have pointed out, in principle any capital investment can be financed through financial assets, which makes the monitoring costs of physical assets no longer a relevant argument for ownership of financial assets. Second, the worker-owned firm is not the only alternative to the

[41] In my M.Phil thesis, I formalized Jensen–Meckling's idea that the riskiness of the project to be chosen by the entrepresneur is increasing with his personal stake.

[42] "Employees of the pure-rental firm will also have an incentive in monitoring the performance of management, but no one in the pure-rental economy will have the same incentive to specialize in performance evaluation (monitoring) as exists in a corporate economy, because there is no way for any individual employee to capture more than a small fraction of the potential gains from such activities. It is therefore naive to believe that pure-rental managers will take the same pains as would corporate executives to seek out high-payoff new projects, to weed out projects which have negative payoffs, to control waste and shirking, etc" (1979, p. 485).

[43] However, some critics have pointed out that the underinvestment problem of the labour-managed firm can be cured by making membership rights marketable. (See Dow 1993a).

capitalist-owned firm; another alternative is the manager-owned firm in which the self-monitored manager hires both capital and other workers. The question to be addressed is why such a manager-owned firm is not common in the market economy.

Following Jensen and Meckling (1976), Leland and Pyle (1977) set up a formal model in which the capital stake held by the entrepreneur functions as a signaling device to solve the agency problem.[44] Under the assumption of asymmetric information about the mean of the project returns between the entrepreneur and outsiders, they show that the entrepreneur's own stake in the project can fully reveal his belief about the mean return of the project and a higher entrepreneur's share signals a higher project value. The influence of the Leland–Pyle signaling model upon the present thesis needs no elaboration. However, the major difference between them is that in the present thesis, shareholders hold residual claims because their shares are a signal, while in the Leland–Pyle model, shares are a signal because shareholders are residual claimants. In addition, in the present thesis, the stake held by the (would-be) entrepreneur signals ability, while in the Leland–Pyle model, it signals project quality.

Stiglitz and Weiss (1981) developed the first contractual model of credit-rationing. Although they analyze the problem in a rather different context, their arguments are relevant to the problem to be explored in the present thesis since credit-rationing is a phenomenon of capital-hiring-labour. Their model is based on asymmetric information between borrowers and lenders about the risk quality of investment projects. They argue that because of adverse selection and moral hazard problems, the investment projects become riskier as the interest rate charged by the lender increases[45]; and hence an increase in the interest rate may decrease rather than increase the total expected returns to lenders. This provides the incentive for lenders to ration credit rather than raise the interest rate when there is an excess demand for loanable funds.

We share Stiglitz and Weiss' arguments. However, their downplaying of the role of the collateral is problematic. The problem is their curious assumption that the individual's wealth is not observable to the lender. The function of collateral requirements is to exclude the poor from borrowing. But their assumption actually excludes any feasibility of imposing collateral requirements.[46]

---

[44] Another well-known signalling model of capital structure is Ross (1977).

[45] This is the case since the change in the interest rate may itself affect the riskiness of the pool of loans. The mechanism is as follows. Assume that all projects have the same means of returns, but differ in the probabilities of success. Thus, different borrowers have different probabilities of repaying their debts. The lender cannot separate "good borrowers' from "bad borrowers". But the interest may work as a screening device. Those who are willing to pay higher interest rates may, on average, be worse risks: they are willing to borrow at higher rates because they perceive their probability of repaying the loan to be low. As a result, as the interest rate rises, low risk borrowers drop out of the borrowing pool, and therefore the average riskiness of those who borrow increases, possibly lowering the lender's profits. Similarly changes in the interest rate may change the borrowers' behaviour. In particular, as the interest rate increases, the borrowers are more likely to undertake risky projects, because by doing so they can reduce the probability of repaying.

[46] Even disregarding this, I find that their theorem 9 cannot hold as long as the borrowing rate is greater than the safe investment return rate and the collateral requirement is greater than the investment (using their notations, i.e., $(1 + \hat{r}) \geq \rho^*$ and $C \geq 1$; the first is obviously true and

Eswaran and Kotwal (1989) developed an incentive model to answer explicitly why capital hires labour in the context of the classical capitalist firm. They show that because of limited liability, the moral hazard problem in capital markets may force the owners of capital to supervise the use of their own capital instead of lending it out in the market. Their arguments can be summarized as follows. Envisage a widget producing activity requiring two essential inputs: an entrepreneur's effort and a hired input, which could be considered as a composite of labour and physical capital. Capital is used to finance the hiring of inputs. The output from a given bundle of inputs is uncertain because of stochastic factors outside the entrepreneur's control. Consequently there is a finite probability that the borrower will default on the loan. Due to his limited liability, the entrepreneur effectively faces a lower price of capital than he would under full liability. Since his effort level is unobservable, he substitutes hired inputs for his effort, and consumes an amount of leisure which is in excess of what he would under full liability. This distortion in the input mix induced by limited liability results in a bankruptcy probability which, from the point of view of the creditor, is larger than it need be. This in turn provides the capitalist with the incentive to undertake production himself. The capitalist firm thus emerges as natural response of the capitalists to the moral hazard of borrowers.

Eswaran and Kotwal's arguments contain useful insights. Insofar as the problem of why capital hires labour is concerned, their arguments are complementary rather than competitive with ours. However, we believe that our argument of the informativeness of wealth in signaling marketing ability of the would-be entrepreneur is more fundamental in explaining capital-hiring-labour. What distinguishes entrepreneurs is their innate ability. Everyone can work hard, but only a small fraction of population can manage the firm well. Some capitalists lend out their capital instead of doing businesses themselves not because they trust that the borrower will work harder than they do, but because they trust that the borrowers are more competent than they are. The moral hazard problem of borrowers may explain why some "marginal" lenders take over supervision of production, but it cannot explain why there are pure lenders at all.[47] In addition, the Eswaran–Kotwal model cannot explain the organizational form of joint-stock companies, while our model can.

(Footnote 46 continued)
the second should be true in their context since they assume that the wealth $W_0 \geq 1$). In that case, investing with own wealth always brings about a greater expected utility than investing with borrowing (comparing equation (17) with (21) in their paper shows this). Using their notation, $\widehat{\widehat{W}}$ cannot be greater than $\hat{W}$.

[47] According to the Eswaran–Kotwal model, a capitalist will lend out only when his total capital exceeds the amount of his own investment; and he always invests more than borrower-entrepreneurs.

1.2.1.4 The Principal-Agent Theory

The principal-agent theory has been the most important development of contract economics in the past two decades.[48] What distinguishes it from the above surveyed agency theory is that all its results have been derived from formal models, and its major developments have been motivated by differences between theoretical predicted contracts and observed contracts. The theory has greatly improved economists' understanding of the internal relationships between capitalists, managers and workers and more generally of transaction relationships in markets. However, in this literature, the major contractual arrangement between capital and labour (i.e., assignment of principalship) is entirely predetermined: capitalists are principals and labourers are agents. The question is to explain how the principal (shareholder/manager) may control the agent (manager/worker) by designing an incentive contract, rather than to explain why the capitalist is the principal while the labourer is the agent.[49] In some sense it is this "imperfection" of the principal-agent theory that motivated our study of endogeneity of principalship.

The standard principal-agent theory is based on two basic assumptions: ($A1$) the principal has no (direct) contribution to the stochastic output (in a parameterized model, no effect on the distribution function of the output), and ($A2$) the agent's actions are not directly observable for the principal (although some indirect signals may be available). Under these two assumptions, the theory gives two basic propositions: ($P1$) for any incentive contract which maximizes the principal's expected utility subject to the agent's participation constraint and incentive compatibility constraint, the agent must bear some risk; ($P2$) if the agent is risk-neutral, the first-best result can be achieved by letting the agent bear full risk (i.e., become the only residual claimant).[50] These two propositions will break down once we relax the two assumptions. First, whenever the principal himself makes contributions to the output, the risk-neutrality of the agent no longer suffices to bring about the first-best result, since in that case, the full residual claim by the agent will inevitably distort the principal's incentive. Second, if the agent's actions are observable with some costs of observing, one's incentive loss from not sharing in the residual may be offset by the other's monitoring, and as a result, a residual-sharing contract may be dominated by one-sided residual claims. Then the most fundamental question is who should be the principal and who should be the agent. This is what we are concerned with in this thesis.

---

[48] The pioneering contributors to the principal-agent theory include Wilson (1969), Spence and Zeckhauser (1971), Ross (1973), Mirrlees (1974, 1975, 1976), Holmstrom (1979, 1982), Grossman and Hart (1983), among others. For an excellent survey, see Hart and Holmstrom (1987).

[49] In fact, all agency models predict that if the agent is risk-neutral while the principal is risk-averse, the optimal contract will be such one in which the agent bears all risk by promising the principal a fixed payment. This proposition can be explained as that the assignment of principalship is determined by risk-attitudes.

[50] Actually, according to the definition of principalship, such a first-best contract changes the agent into the principal.

1.2.1.5 The Theory of Security Design

The literature on security design takes a different direction from the traditional analysis (and the present paper), but it also sheds some light upon assignment of principalship. Building on the work of Grossman and Hart (1988) and Harris and Raviv (1988b), Harris and Raviv (1989) developed a model of packaging residual returns and voting rights of securities. They focus on securities serving as a control device to ensure that a superior candidate rather than an inferior candidate acquires control of the corporation. The main argument is that voting rights should be positively related to residual claims and risk-free "cheap votes" should never be issued. In other words, the right to select management by voting must be held by those who bear business risk. Reinterpreted in this way, our theory can be read as suggesting that, given that labour finds it easier to escape risk than capital (this can be ensured by non-negative consumption constraint and by costless observability of capital endowment), labour-hiring-capital is not optimal because it is a "cheap voting" system in which an inferior candidate with a big private benefit of control is more likely to win the contest for control of the firm. However, it is more appropriate to say that the Harris–Raviv model is concerned with which capitalist (security-holder) should have more voice in choosing management rather than with why capital should hire labour in the first place. In the present thesis, we do not deal with this problem in detail.

Security design models based on agency costs which address only the allocation of cash flows are given by, among others, Townsend (1979), Diamond (1984), Gale and Hellwig (1985), Chang (1987), Hart and Moore (1989), Williams (1989), and Bolton and Scharfstein (1990). In most of these models, it is assumed that there is an information asymmetry about a firm's returns between the insider (manager) and outsiders (investors) and the former can keep any income not paid out to the latter.[51] Under this assumption, it is concluded that debt is an optimal contract. In contrast, in the present paper, it is asymmetric information about ability rather than returns that exists. We predict that some capitalists are willing to buy debt only because there are other capitalists (maybe including the manager himself) —equity holders who might be less uninformed either because they have costless information advantages about the manager's ability (e.g., he is a relative), or because they have paid to acquire such information.[52]

1.2.2 The Entrepreneurial Approach to the Firm

Although the contractual approach to the firm is the most popular among economists today, the first challenge to neoclassical firm theory came from the entrepreneurial

[51] Chang (1987) and Williams (1989) considered a less severe problem by assuming that some returns or assets cannot be appropriated by the manager and therefore they can explain equity claim by outsiders.

[52] For a comprehensive survey on financial contracting theory, see Harris and Raviv (1992).

approach. For neoclassical economists, the firm is a production function; for contract theorists, the firm is a nexus of contracts; the entrepreneurial theory treats the firm as a personalized contrivance. If the contractual theory is concerned with the "demand" for the firm, the entrepreneurial approach penetrates the "supply" of the firm. From the point of view of the present thesis, the firm can not exist without entrepreneurship.[53]

Knight (1921) was the first economist to discuss the existence of the firm in terms of uncertainty and entrepreneurship. He pointed out that under uncertainty, "the actual execution of activity becomes in a real sense a secondary part of life; the primary problem or function is deciding what to do and how to do it" (p. 268). This "primary function" is the entrepreneurial function. Because uncertainty is uninsurable, the entrepreneur has to bear uncertainty. According to Knight, the firm is nothing but a contrivance through which, "[t]he confident and venturesome assume the risk or insure the doubtful and timid by guaranteeing the latter a special income in return for an assignment of the actual result" (pp. 269–270). He took the authority of the entrepreneur over workers within the firm as a compensation for the former for insuring the latter: "With human nature as we know it is would be impracticable or very unusual for one man to guarantee to another a definite result of the latter's actions without being given power to direct his work. And on the other hand the second party would not place himself under the direction of the first without such a guarantee" (p. 270). In short, for Knight, the entrepreneur is the employer (holds authority over workers) because he bears uncertainty.

It should be pointed out that one should not confuse Knight's uncertainty-bearing with the so-called "risk-sharing" view which says that "it ought to be an asymmetry of risk-attitudes between employers and employees that motivates them to agree to long-term employment contracts rather than using the spot-market" (Aoki 1984, p. 14),[54] although there is some superficial similarity between them. For Knight, the entrepreneur bears risks not necessarily because he is risk-neutral or less risk-averse, but because he is more confident and has better judgment and better knowledge, and because risk associated with his decisions is vulnerable to the moral hazard problem.[55] Of course, a risk-neutral person is more likely to be an entrepreneur than a risk-averse one. But his distinction between risk and uncertainty warns us that this point should not be overemphasized.[56]

Compared to Coase, Knight directly touched on the key feature of the firm — assignment of authority. In that sense, our theory is quite Knightian. The distinction between marketing and producing can be traced to Knight's distinction between the "primary function" of "deciding what to do and how to do it" and "the actual execution of activity"; marketing ability can be thought as the analogue of his

[53] The following review is restricted to the "traditional" entrepreneurial theories. Some mathematical entrepreneurial models of the firm will be reviewed in Chap. 4.

[54] This view is shared by Coase (1937) when he regards the "risk attitudes of the people concerned" as a reason for the emergence of the long-term contracts, particularly in employment relations.

[55] For more discussion about this point, see LeRoy and Singell (1987).

[56] Blanchflower and Oswald (1990) have correctly pointed out that for Knight attitude to risk is not the central characteristic which determines who becomes an entrepreneur.

entrepreneurial talent; and the main propositions are derived from our understanding of his "uninsurable uncertainty". However, Knight's entrepreneurial theory is flawed by two conceptual confusions. First, he did not separate "primary function" from bearing uninsurable risks; second, he failed to distinguish explicitly between the entrepreneur and the capitalist. He took it for granted that the entrepreneur who plays a "primary function" bears risks, and therefore he is also a capitalist.[57] It is these two conceptual confusions that made him vulnerable to criticism by economists such as Schumpeter (1934) who argued that uncertainty is borne by capitalists rather than by the entrepreneur so that the entrepreneur bears uncertainty only in so far as he is also a capitalist. This weakness is thoroughly exposed in the case of a corporate firm in which the decision-makers are not necessarily coincident with risk-takers. Knight attempted to remedy this weakness by arguing that the crucial decision in the corporate firm is to select the person who makes decisions, and any other decision-making or exercise of judgment is automatically reduced to a routine function. Therefore, in a corporate firm, ultimate entrepreneurship is located with the shareholders rather than managers unless the managers are also shareholders. This argument is only partially acceptable. In the present thesis, by making a conceptual distinction between marketing and risk-bearing, and between entrepreneur and capitalist, their associations have been explicitly justified. In particular, we characterize the corporate firm by decomposition of entrepreneurship rather than separation of ownership and control.

Other main contributors to the entrepreneurial theories include Kirzner, Schumpeter, Shackle, and Casson. For space reasons, their main ideas are reviewed under two headings. One is what is the function of the entrepreneur. The other is what is the relation between the entrepreneur and capitalist.

For the first, Kirzner (1973, 1979) takes the entrepreneur as a "middleman" who perceives the opportunities and makes profit by capturing these opportunities. He emphasizes that what distinguishes the entrepreneur from others are his "alertness" and special "knowledge". J. Schumpeter (1934, 1942) views the entrepreneur as an "innovator" who "reforms or revolutionizes the pattern of production". To be an innovator, one must be capable of ruthlessly smashing the opposition. Shackle's (1979) entrepreneur is endowed with a particularly creative imagination in making a choice. Casson (1982) attempts to synthesize and extend all these conceptions of the entrepreneur (including Knight's, of course). His definition is "an entrepreneur is someone who specializes in taking judgmental decisions about the coordination of scarce resources" (p. 23). He emphasizes that the entrepreneur is a "market maker". Like Knight, they all agree that the entrepreneur's reward is a residual return not

[57] But at one point Knight placed capital in a secondary role: "In actual society, freedom of choice between employer and employee status depends normally on the possession of a minimum amount of capital. However, demonstrated ability can always get funds for business operation. A propertyless employer can make the contractual payments secure by insurance even when they may involve loss…" (p. 274).

a contractual return. In the present thesis, the entrepreneur can be understood as a mixture of Knight's, Kirzner's, Schumpeter's, Shackle's, and Casson's ideas.[58]

On the second question, Kirzner denies capital as a necessity for someone to be an entrepreneur. He argues that entrepreneurial talent will find ways of securing control of resources, although lack of personal capital might present extra transactional difficulties.[59] However, at one point, Kirzner suggests that capitalists must inevitably exercise the quality of entrepreneurship. Schumpeter (1934) also downplays the importance of capital in entrepreneurship and argues that modern capital markets generally enable an entrepreneur to find a capitalist to bear the risks for him. But Casson takes the opposite view. He emphasizes that the entrepreneurs require command over resources if they are to back their judgment and that this is likely to imply personal wealth. He refers to people with entrepreneurial ability but no access to capital as "unqualified" (p. 333). This view is shared by us.

A word about the fate of entrepreneur. One of the most celebrated aspects of Schumpeter's work is his prediction of the obsolescence of the entrepreneur. He argued that the progress of capitalism would eventually reduce the importance of the entrepreneur because the entrepreneur was initially required to overcome resistance to change but now "innovation itself is being reduced to routine". We do not share this view. But our theory predicts that if marketing activities become routine, capitalists would be deprived of principalship.

1.2.3 The Managerial Theory of the Firm

The managerial theory of the firm was preceded by the development of an empirical thesis on the so-called "separation of control from ownership" in the seminal work of Berle and Means (1932). The Berle-Means hypothesis is that as share-ownership is widely dispersed in joint-stock corporations, authority over the firm has been transferred into the hands of management, and "owners" of the firm have been relegated to the position of means-suppliers.[60] Although the work of Berle and Means was very favourably, perhaps even uncritically, received at the time of publication, its influence on professional economists was not very great. It was not until the late 1950 s and 1960 s that theoretical managerial models of the firm became influential. The most celebrated models are those of Baumol (1959), Marris (1964) and Williamson (1964). All these three models maintained the Berle-Means hypothesis of manager-

[58] Superficially, the Knightian entrepreneur is quite different from the Schumpeterian entrepreneur. But, as FitzRoy and Mueller (1984) point out, in some ways the Knightian entrepreneur is a generalization of the Schumpeterian entrepreneur. In an uncertain environment, making decisions to explore profitable opportunities cannot be separated from innovations.

[59] "Entrepreneurial profits…are not captured by owners, in their capacity as owners, at all. They are captured, instead, by men who exercise pure entrepreneurship, for which ownership is never a condition" (1979, p. 94).

[60] "The concentration of economic power separate from ownership has, in fact, created economic empires, and has delivered these empires into the hands of a new form of absolutism, relegating "owners" to the position of those who supply the means whereby the new princes may exercise their power" (Berle and Means 1932, p. 116).

dominated firms. They are distinguished primarily by the assumed objectives of the managers and the assumed constraints imposed by shareholders. Baumol suggested that managers maximize revenue from sales, subject to a minimum profit constraint; Marris suggested that managers maximize growth, subject to a valuation ratio constraint; Williamson argued that managers maximize a managerial utility function including "staff" or "emoluments", subject to a minimum profit constraint.[61]

From the point of view of their hypotheses, the above three managerial models are anti-neoclassical. Yet methodologically they are quite neoclassical. In fact, if we suppose that the "owner-entrepreneur" also has preference for power, prestige and non-pecuniary consumption, none of which is perfectly substitutable for pecuniary income, the managerial models will lose their identities. This can be seen in Jensen and Meckling (1976) where even an "owner-entrepreneur" is not a "value-maximizer". In this sense, the observed conflicts between shareholders and management are nothing but the externalization of internal conflicts of preferences. That is, a shareholder's utility is a function of share value or profit only because he is not a manager; once he becomes an owner-manager, some other variables (such as growth and staff) will enter his utility function and he will no longer be a value-maximizer.

From the point of view of institutional economics, what the managerial models have provided is questions but not answers. The contractual models of the firm reviewed in the first subsection can be understood as a response to the challenges of the managerial models in the sense that the contractual theorists try to make managerial discretion endogenous by putting managers under a competitive but imperfect monitoring environment. However, neither the contractual models nor the managerial models have provided an appropriate explanation for the origin of "separation of control from ownership". One of our purposes in this thesis is to explore the origin of this. We shall make the equilibrium relationship dependent on the joint distribution of marketing ability and personal wealth (and risk-attitudes). In addition, the manager in the present model is much more entrepreneurial than in the managerial models.

1.3 The Plan of the Thesis

A complete theory of the firm must deal with at least the following three interrelated problems: (i) Why does the firm exist in the first place? (ii) How is principalship (residual claim and authority) assigned among the different members of the firm? (iii) What are the optimal contracts that the principal uses to control agents? Much of the literature on the theory of the firm has so far focused on the first and the third problems. As we have seen, although some economists have been concerned with

[61] In the 1970 s managerial theory was applied to other areas. One of such applications is Niskanen's (1968) model of bureaucracy, in which bureaucrats are assumed to maximize their budget in the ultimate interests of power, status or prestige and constrained only by the demand curve for the service they provide.

the second problem, so far there has been no convincing answer given to the question of why capital hires labour. The present thesis is intended to make a contribution to understanding this problem by combining the contractual approach with the traditional entrepreneurial approach. Our demonstration consists of three major steps. The first step is to show why principalship of the firm is assigned to the marketing member in the first place, by providing a rationale for the assignment of entrepreneurial role to the marketing member; the second step is to address directly the question of why capital hires labour by showing why priority in being an entrepreneur or authority in selecting management is given to capitalists; and the third step is to show how equilibrium relationships between different members of the firm change resulting from changes in distributions of personal wealth and marketing ability (and risk-attitudes) in the population. The thesis proceeds as follows.

In Chap. 2 (the first step), we demonstrate why principalship should be assigned to the marketing member rather than to the producing member to maximize total welfare (equivalently, to minimize agency costs). In so doing, we argue that differences of marketing ability between individuals are the original rationale for the occurrence of the firm; we identify marketing with Coase's "discovering the relevant prices" but focus on aspects ignored by Coase. We make a distinction between the self-monitored incentive and the being-monitored incentive. We argue that there is a trade-off between the self-monitored incentive and the being-monitored incentive associated with assignment of principalship, and it is optimal for the marketing member to be the principal because such a contractual arrangement can guarantee that total welfare are maximized. In Chap. 3 (the second step), a hidden information model is used to show why priority in being entrepreneurs is given to capitalists. In so doing, we focus on how the capital endowment of a would-be entrepreneur can function as a signal of his marketing ability. Specifically we show that the individual critical ability for being an entrepreneur is increasing with his personal wealth, unless the individual's personal wealth exceeds a certain level. Under the assumption that marketing ability is not observable (or not costlessly observable), it is shown that priority in being an entrepreneur (marketing member) and /or the right of selecting the person to undertake marketing should be given to capitalists because such a contractual arrangement can ensure that only qualified candidates win the competition for being entrepreneurs (or managers). This conclusion implies that imperfect capital markets may be socially optimal. In Chap. 4 (the third step), based on the arguments given in Chaps. 2 and 3, a general equilibrium entrepreneurial model of the firm is set up; the main properties of the equilibrium will be derived; and the partition of the population into entrepreneurs, manager, pure capitalists and workers will be identified. We show that in equilibrium, (a) individuals with high ability, high personal wealth and low risk-aversion become entrepreneurs, (b) individuals with low ability, low personal wealth and high risk-aversion become workers, (c) individuals with high ability but low personal wealth become managers hired by capitalists, and (d) individuals with low ability but high personal wealth become "pure" capitalists to hire managers. Chapter 5 concludes the thesis and directs attention to some promising aspects for our future research.

Chapter 2
Marketing, Producing, Monitoring and the Assignment of Principalship

2.1 Introduction: The Firm as a Cooperative Organization and the Agency Problem

What distinguishes a market economy is that a producer produces goods not directly for his own consumption but for markets through which he sells outputs and buys in inputs. Accordingly, his income and hence his utility do not just depend upon how much he has produced from given inputs but also upon how much he can charge for his outputs and how much he has paid for the inputs. What concerns him is the dot of the price vector and product vector (therefore the net return) rather than the product vector itself. In order to maximize his expected utility, the most important task he has is "deciding what to do and how to do it" (Knight 1921, pp. 268), or in Coase's words, "discovering the relevant prices" (Coase 1937, pp. 390). We define this "primary function" as "marketing", while all other activities involved in executing the decisions are defined as "producing" (mainly physically transforming inputs into outputs).

Of course, even an autarkic farmer (or tenant) has to make decisions of "what to do and how to do it". However, his decision-making has little to do with "discovering the relevant prices" and therefore can be understood as "producing". The reason is that in autarky all the information he needs for decision-making is his own preferences and his own resources, both of which are certain for him; and the only uncertainty he has to face is about something like the weather (e.g., rain or not rain?), which is entirely beyond his control. In Knight's words, this is "risk" rather than "uncertainty". In contrast, in a market economy, the most important information one needs for decision-making is about others' preferences and others' resources, both of which are uncertain. In order to decide what to produce and how to produce, he must acquire some information about how other people will value various products he could choose to produce, i.e., about the relevant prices; he must have some knowledge about the relative efficiency associated with different "production functions"; and he must discover where the market "disequilibrium" (opportunity) is. Because information is too costly to be complete, he has to face some risks. Because

W. Zhang, *The Origin of the Capitalist Firm*, DOI 10.1007/978-981-10-0221-2_2

these risks are associated with his decisions and are in some sense endogenous to his actions, they are uninsurable (Knight 1921).[1] To a very great extent his return fluctuations are dominated by his marketing activities more than by his producing activities. A businessman is more likely to go bankrupt when he produces the "wrong" products at *minimum cost* than when he produces "right" products even *inefficiently*.[2]

Marketing ability can be defined as the ability to decide what to produce and how to produce it (or the ability of discovering the relevant prices). Although everyone may possess some marketing ability, the observation is that individuals differ in their marketing ability. This is so not just because different people face different costs of collecting and processing information, but also because marketing ability greatly depends upon the person's "alertness" (Kirzner), "imagination" (Shackle) and "judgement" (Casson). All these personal characteristics are at least partially innate and ineducable.[3] Furthermore, although individuals also differ in their producing ability, the distribution of producing ability needs not be coincident with the distribution of marketing ability. For simplicity, it is reasonable to assume that individuals are identical in their producing ability. It is the differences in marketing ability that create an opportunity for people to cooperate with each other by setting up a "firm" in which someone who has high marketing ability is responsible for marketing and those who are not good at marketing are responsible for producing, instead of each being an individual businessman. In this sense, the firm is a cooperative organization characterized by division of labour.[4]

However, although it is potentially profitable to set up a firm, the firm as a cooperative organization is confronted by two problems. First, because of uncertainty, the return of the firm is a random variable and business risk is inevitable. The problem is how to distribute risk among the members of the firm by assigning residual claims. Second, because of "team production", each member's contribution to the total return is not costlessly measurable.[5] This creates an incentive problem: one party may take actions (e.g., shirking) which benefit himself but cost others. The problem is how to design an incentive scheme to make each party as responsible for his own actions as

[1] Huang (1973) distinguishes between risks associated with decision making and exogenous or natural risks beyond the party's control. To my understanding, Huang's distinction coincides with Knight's (1921) classical interpretation of the distinction between risk and uncertainty. Roughly speaking, "exogenous or natural risk" in Huang is "risk" in Knight; "risk associated with decision making" in Huang is "uncertainty" in Knight. In the present thesis, for simplicity, we call the first type risk "natural risk" and the second "business risk".

[2] Robinson Crusoe was very disappointed when he found the boat which he had spent four years in building couldn't be moved into the water. However, if he were a businessman, he would have gone bankrupt rather than just be disappointed. Karl Marx referred to the market as a "thrilling jump", failure in which would destroy not just the product but also its producer himself.

[3] For more discussion about entrepreneurial qualities, see Casson (1982), Chap. 2.

[4] Yang and Ng (1995) assume that individuals are ex ante identical in their marketing ability. In their model, the firm exists because of economy of specialization and transaction costs.

[5] It should be noted that although we borrow this term from Alchian and Demsetz (1972), what we emphasize here is that the marginal contribution of producing (marketing) member's effort depends on marketing (producing) member's effort, rather than worker A's contribution depends on worker B's effort.

possible. These two problems cannot be resolved separately because business risk is highly related to members' actions and therefore uninsurable.[6] That is, how big the risk is depends on how the risk is to be distributed. The major purpose of the contractual arrangement is to deal with these two problems simultaneously. Following Alchian and Demsetz (1972), we can identify the issue as assigning principalship: which member(s) should be entitled as the principal to monitor others and hold the residual claim? The three polar options of contractual arrangements are: (i) the principalship is assigned to the marketing member; (ii) the principalship is assigned to the producing member; (iii) they become partners to monitor each other and share the risk.

This chapter is intended to characterize the optimal assignment of principalship between the marketing member and the producing member. Two types of costs are identified as the determinants. One is risk costs and the other incentive costs. Following the literature of uncertainty, risk costs are defined as the difference between the expected income with given uncertainty and its certainty-equivalent income. Given the distribution function of the firm's return and each individual's utility function, the total risk cost (the risk cost for the marketing member plus the risk cost for producing member) is a function of contractual choice. Following Jensen and Meckling (1976), the incentive costs[7] are defined as the difference between the "first-best" expected return (that is, when each member's contribution to the total return of the firm can be perfectly and costlessly measured) and the actual expected return associated with a given contract, including all losses from the incentive problem, such as monitoring expenditures by the principal, bonding expenditures by the agent, and the "residual loss" identified by Jensen and Meckling. Incentive costs exist no matter how principalship is assigned. However, different assignments are associated with different incentive costs. One of our main contributions is identifying these incentive costs as the key factor determining the assignment of principalship itself. In contrast, in most of the existing literature on the agency theory, the incentive costs affect only how the principal designs the incentive scheme for the agent. In particular, given that the risk-cost approach has been well exploited by economists, our attention is almost exclusively focused on how the assignment of principalship is related to the incentive costs.

For analysis, we make the following assumptions. First, both the marketing member and the producing member are taken to be single units. In reality, because of economies of scale in marketing, one marketing member may be responsible to a number of producing members. Because all these producing members are functionally identical, taking them as a single unit allows us to focus on the role of functional asymmetry between marketing and producing in assigning principalship.[8] Secondly,

---

[6]In Alchian and Demsetz (1972), only the incentive problem has been identified. In their model, although the monitor claims the residual, the returns to monitoring are certain for given monitoring effort. Here, following Knight (1921), we relate the incentive problem to the risk problem.

[7]We use "incentive costs" instead of "agency costs". In the thesis, the agency costs are defined as the sum of risk costs and incentive costs.

[8]The prisoners' dilemma problem among the producing members (workers) in monitoring may enable a single marketing member to have advantages in competing for principalship. There

in this chapter, we assume that each individual's marketing ability is known to all others as well as himself, and therefore the marketing function and producing function are correctly allocated to members of the firm. Thirdly, we ignore the capital problem which will be the focal point of the next chapter. Under the above assumptions, the problem can be formulated as follows. The firm consists of two types of "workers", one is the marketing member and the other the producing member; the return of the firm is jointly determined by the actions taken by both members as well as the state of nature; the distribution of the return is associated with incentive costs as well as risk costs; and the principalship is assigned so as to minimize the sum of the risk costs and the incentive costs. In the next chapter, we turn to the role of capital in assigning principalship by dropping the second and the third assumptions. Our analytic strategy is first to demonstrate why principalship should be assigned to the marketing member, and then to show why capitalists should be entitled to select the marketing member.[9]

The major argument of this chapter is that assigning principalship to the marketing member is preferred because marketing activities dominate uncertainty, and because the marketing member's behaviour is more difficult to monitor. This provides a rationale for the asymmetric relationship between the entrepreneur and workers. The chapter is arranged as follows. In Sect. 2.2, the basic model is set up. In Sect. 2.3, under the risk-neutral assumption, we will be concerned with how the degree of teamwork, relative importance, and monitoring technology determine the optimal assignment of principalship through their effects on incentive problem. The first part of Sect. 2.3 deals with the optimal assignment when monitoring is technically impossible; and the second part deals with the optimal assignment when monitoring is possible. Section 2.4 discusses two commonly observed forms of firm (the classic capitalist firm and partnerships) and a theoretically-created form of firm (the Alchian–Demsetz firm). In Sect. 2.5 we introduce risk-attitudes to see how risk costs are associated with the assignment of principalship and what kind of effect it may impose. Section 2.6 concludes the chapter.

2.2 The Model

The firm consists of two types of members, the marketing member M and the producing member P. Both are assumed to be the expected utility-maximizers. The task

(Footnote 8 continued)
are two reasons why we do not emphasize this argument. First, our proposition must hold even if the firm consists of only one producing member and one marketing member. Second, this argument implies that principalship held by the marketing member is "deepening" as the number of producing members increases. We have no evidence to support this prediction.

[9] An alternative is first to show why capitalists should hold the principalship, given that the marketing party is entitled to monitor the producing party, and then to demonstrate why the marketing party rather than the producing party should be the monitor. We prefer our approach because it is more logical as well as historical.

of each member is well-defined. The return to the firm is jointly determined by both members' actions (as well as a "state of nature"). Let A_i be the set of actions available to i and denote a generic element of A_i by a_i, where $i = M, P$. In particular, we identify a_i with a continuous, one-dimensional effort variable, called i's "work effort", which might be thought of as an *aggregated* measure of all i's actions for his task. Let Y be the total return to the firm. Then Y is a stochastic function of a_M and a_P. Following Mirrlees (1974, 1976) and Holmstrom (1979), we assume that there is a distribution function over Y, conditional on a_M and a_P, denoted by $\Phi(Y; a_M, a_P)$.[10] Uninsurable uncertainty defined by Frank Knight implies that $\frac{\partial \Phi}{\partial a_M} \neq 0$ and $\frac{\partial \Phi}{\partial a_P} \neq 0$. We make the following assumption for $\Phi(Y; a_M, a_P)$:

Assumption 1 (i) $\frac{\partial \Phi}{\partial a_i} \leq 0$, with strict inequality for at least some Y; (ii) $\frac{\partial^2 \Phi}{\partial a_i^2} \geq 0$; (iii) $\frac{\partial^2 \Phi}{\partial a_P \partial a_M} \neq 0$.

(i) implies that $\Phi(Y; a_M, a_P)$ satisfies the first-order stochastic dominance condition over a_M and a_P; (ii) implies that $\Phi(Y; a_M, a_P)$ satisfies the convexity of the distribution function condition (CDFC) over a_M and a_P, or stochastic diminishing return to scale; (iii) is the assumption of team-work in the presence of uncertainty.

One of the main implications of Assumption 1 (iii) is that it is impossible for each member to be *fully* and *only* responsible for the uncertain outcome of his own actions, even if both members are risk-neutral[11]; and therefore the relationship between M and P cannot be solved by a complete contract in the sense that there must be some transfer of responsibility between the two members. Here by transfer of responsibility, we mean that i's interests are affected by j's actions.

Our major concern is: What is the optimal arrangement of transfer of responsibility? We identify this problem with the assignment of principalship under the following definition:

Definition 1 Member i is called the principal of member j if he has to take full or partial responsibility for the uncertain outcome of j's actions; correspondingly j is called the agent, where $i, j = M, P$; $i \neq j$.

This definition seems quite coincident with the conventional conception of the principal. But it allows for the case in which responsibility is mutually shared between the two members (partnership).

Note that what the principal is responsible for is the *uncertain outcome* of the agent's actions, rather than the agent's actions per se. Therefore 'risk-bearing' might be a more proper terminology. In the case that the agent's actions are perfectly observable, an action-contingent payment contract would make the agent fully responsible for his own actions, while the principal is still a principal because he has to bear risk for the agent's actions, unless there is no uncertainty.

---

[10] The main advantage of this so-called parameterized distribution formulation is to allow us to capture Frank Knight's uncertainty which says that business risks are dominated by actions.

[11] In a simple agency model, it is always possible for the agent to take full responsibility for his own actions by letting him be the residual claimant, because the distribution function of the outcome is only conditional on the agent's actions. The problem is that such a full responsibility system cannot be efficient if the agent is strictly risk-averse (regardless of the principal's risk attitude).

The essence of principalship is risk-bearing. In return for this risk-bearing, the principal is also entitled to "authority to monitor", by which he can require (enforce) the agent to work more than otherwise within a limit to be defined later.[12] Therefore the assignment of principalship is a two-dimensional contract between the marketing member and the producing member: it defines a distribution of the return (risk-bearing) and an allocation of authority of monitoring. However, we will see that, under some standard assumptions about the utility functions and a *reasonable* assumption about limitation to authority to monitor, the incentive to monitor is *uniquely* determined by the distribution of the return. For this reason, we shall make no further restriction on the allocation of authority of monitoring, apart from acceptance by the monitored member. In fact, a remarkable property of the model is that the allocation of authority to monitor is endogenous; that is, anyone is authorized to monitor the other, subject to the latter accepting his monitoring.[13]

With the above arguments in mind, we characterize the assignment of principalship by the following linear distribution system of the return[14]:

$$\begin{aligned} Y_M &= w_M + \beta(Y - w_M - w_P) \\ Y_P &= w_P + (1-\beta)(Y - w_M - w_P) \end{aligned} \tag{2.1}$$

where Y_M is the total return distributed to the marketing member and Y_P is the total return distributed to the producing member, and $Y_M + Y_P = Y$; w_M is the fixed contractual term to M, and w_P is the fixed contractual term to P. Here by 'fixed' we mean their independence of the realized return Y, but they may depend on some other observable variables (see later). To ensure that the fixed terms are riskless, we shall assume that $w_M + w_P \leq \underline{Y}$, where $\underline{Y}$ is the low bound of Y. The most important parameter is $\beta (0 \leq \beta \leq 1)$: β measures the residual share of the marketing member, and $(1-\beta)$ measures the residual share of the producing member. We will see that for given bargaining positions, $\{w_M, w_P\}$ is uniquely determined by β; and therefore we shall quite often identify the assignment of principalship with the one-dimensional variable β. Two special cases are: (i) $\beta = 0$: P is the principal and M is the agent; (ii) $\beta = 1$: M is the principal and P is the agent. When $0 < \beta < 1$, we say that principalship is shared between M and P.

[12] Here we follow Frank Knight who argued that "With human nature as we know it would be impracticable or very unusual for one man to guarantee to another a definite result of the latter's actions without being given power to direct his work. And on the other hand the second party would not place himself under the direction of the first without such a guarantee." (pp. 270).

[13] We will discuss how acquisition of authority of monitoring is constrained by acceptability for the other later.

[14] The assumption of linearity is made only for simplicity.

The assignment of principalship matters because it affects incentives to work. Intuition suggests that someone chooses to work harder for one of two reasons: either because he *wishes to* or because he *has to.* To accommodate this intuition, we make a distinction between the *self-interested work effort (self -interested incentive)* and the *monitored-work effort (monitored-incentive),* denoted by a_i^s and a_i^b respectively.

Definition 2 Work effort is said to be self-interested if it is chosen even without being monitored; work effort is said to be monitored if it would not be chosen without being monitored. For any given assignment of principalship, if the self-interested effort is greater than the monitored effort (i.e., what one wishes to do is more than what one has to do: $a_i^s \geq a_i^b$), we say the monitoring is *non-binding*; otherwise monitoring is *binding*.

Exercising the authority of monitoring requires time and energy. Let $b_i \in B \in [0, \infty)$ denote member i's effort expended on monitoring member j, called i's "monitoring effort", which makes no direct contribution to the firm's return but may affect member j's work effort through a monitoring technology. A monitoring technology is defined as a map from i's monitoring effort into j's monitored work effort:

$$\begin{aligned} &(i)\ \ a_P^b = a_P^b(b_M) \\ &(ii)\ a_M^b = a_M^b(b_P) \end{aligned} \tag{2.2}$$

We shall assume this is common knowledge; that is, when i chooses $\widetilde{b}_i$, both i and j know that j has to choose $a_j^b(\widetilde{b}_i)$ and j knows $\widetilde{b}_i$ has been chosen. Therefore the contract can be contingent on $a_j^b(b_i)$. We make the following assumption:

Assumption 2 (i) $\frac{\partial a_j^b}{\partial b_i} \geq 0$, and $\frac{\partial^2 a_j^b}{\partial b_i^2} \leq 0$; (ii) $a_j^b(0) = a_i^b(0) \equiv 0$.

Assumption 2 (i) says that j's monitored effort is an increasing but concave function of the i's monitoring effort; in other words, the "marginal productivity of monitoring effort" is positive but diminishing. (ii) implies that neither i nor j can force the other to work unless he chooses a positive monitoring effort. This assumption seems quite sensible. In addition, we define:

Definition 3 Monitoring is said to be *technically* impossible if $a_j^b(b_i) \equiv 0$ for all $b_i > 0$.

The key point is that monitoring has a positive effect on work efforts. (2.2) can be understood as a reduced form of some more complex monitoring mechanisms. One possibility is that the principal expends time and energy to *directly* force the agent to work more than otherwise. Alternatively, the principal can observe how much work effort the agent chooses and then reward the agent on the basis of the observed working effort; because the work effort positively depends on the time and energy spent on observing, monitoring can *indirectly* induce the agent to work more. The third possibility is that the principal only detects whether or not the agent is shirking and punishes him if shirking; because the probability of being caught shirking is

increasing with the principal's monitoring effort, then optimal shirking is decreasing with the monitoring effort.[15,16]

It is worth noting the difference in modeling monitoring between the present model and a standard agency model. In a standard agency model, such as in Holmstrom (1979), monitoring is treated as an exogenous system to provide some *costless* signals about the agent's actions (but not direct observation of the actions), and its value depends on the informativeness of its signals in inferring the actions. In contrast, in the present model, monitoring is an endogenous choice variable to provide direct information on the agent's actions at the cost of monitoring effort. In this sense, the present model follows the procedure of Alchian–Demsetz theory rather than agency theory. It is this difference that allows us to model explicitly the allocation of the authority to monitor and hence the assignment of principalship itself, rather than the optimal incentive scheme under a predetermined assignment of principalship. Furthermore, we shall see that the standard agency model is a special case of the present model when monitoring is technically impossible.

We now characterize the utility functions of both members. For simplicity, endowments are assumed equal to zero. Then the von Neumann–Morgenstern utility functions are described as follows:

$$U_i = U_i(Y_i, e_i) = V_i(Y_i) - C_i(a_i, b_i) \tag{2.3}$$

Through the chapter, we shall call $V_i(Y_i)$ "the benefit (utility) of income", and $C_i(a_i, b_i)$ "the cost (disutility) of effort". We adopt the following standard assumptions:

Assumption 3 (i) $\frac{\partial V_i}{\partial Y_i} > 0$, and $\frac{\partial^2 V_i}{\partial Y_i^2} \leq 0$; (ii) $\frac{\partial C_i}{\partial a_i} > 0$, and $\frac{\partial^2 C_i}{\partial a_i^2} \geq 0$; (iii) $\frac{\partial C_i}{\partial b_i} > 0$, and $\frac{\partial^2 C_i}{\partial b_i^2} \geq 0$; (iv) $\frac{\partial^2 C_i}{\partial a_i \partial b_i} \geq 0$.

Assumption 3 (i) says that both members are risk-averse or risk-neutral; (ii) and (iii) say that they are not only averse to working, but also averse to monitoring: neither the marketing member nor the producing member enjoy monitoring (or being-

[15]The agent may play a monitoring-anti-monitoring game with the principal. For instance, if the principal checks the agent n times randomly a day, the agent may make use of an "spy-hole" so that he needs to work only when the principal is coming. Then the shirking time will certainly fall as n increases. (In the jargon of games, the agent's strategic space for shirking shrinks as the principal's monitoring effort increases.).

[16]Although positive effects of monitoring on work effort are quite intuitive and widely observed, theoretical views are far from unanimous. In Putterman and Skillman (1988), it is argued that the positive incentive effect of monitoring depends critically on the compensation scheme employed, the risk preferences of the agents, and the informational content of increased monitoring. In particular, they show that: when monitoring is understood to produce a noisy signal of working effort with a first- or second-order stochastic dominance condition, the positive effect cannot be guaranteed in general under either the share payment scheme or the wage payment scheme, and moreover under some reasonable assumptions of the risk-preferences, the effect is actually negative; on the other hand, the positive effect is most easily guaranteed (under both compensation schemes) if monitoring produces an accurate signal of working effort with a probability which depends on the level of monitoring intensity.

monitored) per se; (iv) says that work effort and monitoring effort cannot be complementary in the preferences; that is, the marginal cost of work effort (monitoring effort) cannot be decreasing as monitoring effort (work effort) increases.

The utility functions (2.3) have some important implications for the assignment of principalship. First, the contract affects each member's expected utility only through the distribution of the return and the choices of work effort and monitoring effort; because both working effort and monitoring effort incur costs, i will choose $a_i^s > 0$ and $b_i > 0$ only if Y_i is not independent of Y; in other words, i cannot have (self-) incentives either to work or to monitor unless he shares some residual return. Second, because monitoring effort makes no direct contribution to output but incurs costs, member i will choose $b_i > 0$ if and only if $a_j^b(b_i) > a_j^s$.

The third, and perhaps the most significant implication is that under a given contract, if i chooses $b_i > 0$ such that $a_j^b(b_i) > a_j^s$, he imposes on j an extra cost equal to $\left(C_j(a_j^b(b_i), .) - C_j(a_j^s, .)\right)$, called "the external cost" of monitoring. This external cost sets a limit to the authority of monitoring. The question is: under what kind of contract is i's monitoring acceptable to j?

We deal with this problem by making the fixed term w_j contingent on a_j^b as follows:

$$w_j = \begin{cases} w_j^s & \text{if } b_i = 0 \\ w_j^s + F_j(a_j^b) & \text{if } \mathrm{b}_i > 0 \end{cases} \tag{2.4}$$

where w_j^s is a constant, $F_j(.) \geq 0$.

We call $F_j(.)$ "the rule governing the acceptability of monitoring", which defines how the agent's fixed return (w_j) should vary with the monitored-work effort (a_j^b). To characterize $F_j(.)$, we make the following assumption:

Assumption 4 (*acceptability of monitoring*) For any given a_i^s, the authority of monitoring with b_i by member i, which forces member j to choose $a_j^b(b_i) > a_j^s$, is acceptable for member j if and only if the following condition holds:

$$\int V_j(Y_j^b)\phi(Y; a_i^s, a_j^b)dY - C_j(a_j^b(b_i), .) \geq \int V_j(Y_j^s)\phi(Y; a_i^s, a_j^s)dY - C_j(a_j^s, .) \tag{2.5}$$

where

$$Y_j^b = w_j^s + F_j(.) + \lambda_j(Y^b - w_i - w_j^s - F_j(.))$$
$$Y_j^s = w_j^s + \lambda_j(Y^s - w_i - w_j^s)$$

where Y^b and Y^s are the total return to the firm respectively when j is and is not monitored by i; λ_j is the j's residual share ($\lambda_j = \beta$ for $j = M$, and $\lambda_j = 1 - \beta$ for $j = P$).

Assumption 4 says that member j will accept monitoring by member i if and only if his expected utility under i's monitoring is no less than that under no monitoring, for any given self-incentive work effort by i. Because there is no need to compensate

j more than necessary for i to acquire the authority to monitor, we shall assume the equality holds.[17]

To give some intuition of $F_j(.)$, let us consider the case of $\lambda_j = 0$ and the case of risk-neutrality.

If $\lambda_j = 0$, condition (2.5) reduces to:

$$V_j(w_j^s + F_j(.)) - V_j(w_j^s) \geq C_j(a_j^b(b_i), .) - C_j(a_j^s, .)$$

That is, the agent's 'wage' when he is monitored to work more should be increased to such a level that the additional utility from the increased income offsets the additional disutility from being monitored.

In the case of risk-neutrality, condition (2.5) reduces to:

$$(1 - \lambda_j)F_j(.) + \lambda_j(Y^b - Y^s) \geq C_j(a_j^b(b_i), .) - C_j(a_j^s, .)$$

where Y^b and Y^s denote the expected values.

This implies that the total external cost $\left(C_j(a_j^b(b_i), .) - C_j(a_j^s, .)\right)$ is compensated by two parts: the fixed term $(1 - \lambda_j)F_j(.)$ and the residual term $\lambda_j(Y^b - Y^s)$. In the case of $\lambda_j = 0$,

$$F_j(.) \geq C_j(a_j^b(b_i), .) - C_j(a_j^s, .)$$

This is the standard compensation rule when the agent's payment can be contingent on his actions. The rule governing the acceptability of monitoring defined by (2.5) can be understood as a generalization of this standard rule to the case that the assignment of principalship is endogenous.

Under Assumption 4, member i has to take into account two costs of monitoring effort when he chooses whether or not to monitor j: the first is his internal cost $(C_i(a_i, b_i) - C_i(a_i, 0))$, and the second is the external cost $\left(C_j(a_j^b(b_i), .) - C_j(a_j^s, .)\right)$. Clearly a contract with this property Pareto dominates a contract in which $w_j \equiv w_j^s$ in the sense that it makes use of available (observed) information on a_j^b.[18]

In summary, an assignment of principalship is characterized by a three-dimensional distribution system $\{w_M^s, w_P^s; \beta\}$ accompanied by a rule governing the acceptability of monitoring which should be read as follows: M claims β share and P claims $(1 - \beta)$ share from the residual; given β, if neither of them chooses to monitor the

[17] Is such an assumption really acceptable in the sense of fairness? Obviously it would be irrational for j to accept monitoring by i which makes him worse-off. The problem is: why cannot j do better? The answer is the authority of monitoring is equally open to j with a symmetric compensation rule: j is free to monitor i as long as his monitoring would not make i worse-off. Under such a symmetric treatment, the condition seems fair.

[18] When $w_j = w_j^s$, i would choose too much monitoring effort since the externality of monitoring cost is not fully internalized.

other, M claims the fixed term $w_M = w_M^s$ and P claims the fixed term $w_P = w_P^s$; if M chooses to monitor P with $b_M > 0$, P's fixed term will be $w_P = w_P^s + F_P(a_P^b)$; if P chooses to monitor M with $b_P > 0$, M's the fixed term will be $w_M = w_M^s + F_M(a_M^b)$.

The distribution system $\{w_M^s, w_P^s, \beta\}$ defines a status quo for each agent, which in turn forms a constraint for any would-be monitor. However, such a status quo is not constant with a_i^s unless $\lambda_j = 0$. The reason is that when i increases his self-incentive a_i^s, j's residual term $(\lambda_j Y^s)$ will automatically increase through the effect of a_i^s on Y^s, which implies that j's status quo is improving with an increase in i' s self-incentive! We call this effect "the residual share effect", which is a potential rationale for a full principalship contract (i.e., $\lambda_j = 0$) strictly Pareto dominating a partnership contract.[19,20]

A non-constant status quo is not the standard assumption in agency theory. This leads to one of the main differences in dealing with the so-called "participation constraint" between the existing literature of agency models and the present model. The existing literature, in general, considers only *partial* equilibrium in that one party's expected utility at equilibrium (it may be the agent's as in most models, or the principal's as in the Jensen-Meckling model) is equal to the participation (or reservation) level determined by "markets" and therefore all surplus from contracting is distributed to the second party (in general the principal). Under such an assumption the optimization problem is to design a contract which maximizes the principal's expected utility subject to the agent's participation constraint and the incentive compatibility constraint. At equilibrium the agent's participation constraint is binding. This is a natural assumption for the case where the authority of designing a contract is held by the principal and the only choice left to the agent is "take-it-or-leave-it". However this assumption is not satisfactory for the present model in which the allocation of principalship is to be determined. Although we consider a firm consisting of one marketing member and one producing member, this firm is a representative firm and both M and P are representatives of their respective professions. The contract arrangement derived from the model should be a characteristic of all firms. In other words, what we are concerned with is the *general* rather than *partial* equilibrium. To induce a person to join the firm, his expected utility from joining the firm must be not less than that from doing business individually. But this participation constraint cannot be binding in general, unless the total number of his type is in surplus. Suppose there are n identical producing members and n identical marketing members, and any pair of the marketing member and the producing member can form a coalition to set up a firm. Then there is no reason to assume that one member's participation constraint should be binding. It is more reasonable to assume that the distribution of the surplus from the firm between the two members is determined by a Nash-bargaining solution. Formally we assume

[19] We will discuss this point later. Briefly, the residual share effect implies that the principal can never fully internalize the benefit from his effort unless he is the only residual claimant.

[20] It is interesting to note that under Assumption 4, i's monitoring actually improves j's welfare (unless j takes no residual share) compared to the non-monitoring equilibrium (an equilibrium when monitoring is technically impossible). The arguments in Sect. 2.3 will show that Assumption 4 has implicitly incorporated a bargaining procedure into the contract.

Assumption 5 The distribution of surplus from the firm is determined by a Nash-bargaining solution with the threat point being the expected utility levels from doing business individually.

Our interpretation of the marketing function (see Sect. 2.1) implies that the marketing member can do better than the producing member when no firm exists, and this in turn implies that the marketing member will get more total welfare than the producing member from joining the firm.[21] However, insofar as the optimal solution is not affected, we shall normalize both members' reservation utilities to zero.

We now formally define the optimal assignment of principalship as follows:

Definition 4 Let $\Omega = \{w_M^s, w_P^s; \beta\}$ be the set of contracts available (accompanied by the rule governing the acceptability of monitoring) and ω be an element. Then an assignment of principalship $\omega = (w_M^s, w_P^s; \beta) \in \Omega$ is the optimal one, denoted by ω^*, if and only if it solves the following problem (overall game):

$$
\begin{array}{ll}
Max & EU_M EU_P \\
\{w_M^s, w_P^s; \beta\} & \\
s.t.. & \text{Incentive Compatibility Constraints:} \\
& (1)\{a_P, b_P\} \in \text{argmax } EU_P \\
& \qquad s.t.\ (i) \text{ monitoring technology (2.2)} \\
& \qquad\quad (ii) \text{ the rule governing monitoring (2.5)} \\
& (2)\{a_M, b_M\} \in \text{argmax } EU_M \\
& \qquad s.t.\ (i) \text{ monitoring technology (2.2)} \\
& \qquad\quad (ii) \text{ the rule governing monitoring (2.5)}
\end{array}
$$

where

$$EU_P = \int V_P\left(w_P + (1-\beta)(Y - w_M - w_P)\right)\phi(Y; a_M, a_P)dY - C_P(a_P, b_P)$$

$$EU_M = \int V_M\left(w_M + \beta(Y - w_M - w_P)\right)\phi(Y; a_M, a_P)dY - C_M(a_M, b_M)$$

That is, the overall game of assignment of principalship is to select a contract $\{w_M^s, w_P^s; \beta\}$ which maximizes the product of the two members' expected utility levels subject to two incentive compatibility constraints.

The overall game can be decomposed into two sub-games: a non-cooperative game and a cooperative game. In the non-cooperative game, with a given contract $\{w_M^s, w_P^s; \beta\}$, each member chooses his own efforts (a_i, b_i) to maximize his expected utility, subject to monitoring technology (2.2) and the rule governing the acceptability of monitoring (2.5). The solution to the non-cooperative game is a Nash-equilibrium, which defines a relationship between the action set $(A_M \times B_M) \times (A_P \times B_P)$ and

[21] I believe this is one of the major reasons for the observation that the entrepreneur (or manager) have higher expected income than the worker.

the assignment set Ω. The cooperative game is to choose a special assignment ω^* which maximizes the Nash-welfare function $EU_M EU_P$.

What we are interested in is: What determines ω^*? The analysis to follow is aimed at characterizing the solution to this problem.

2.3 Degree of Teamwork, Relative Importance, Monitoring Technology and Optimal Assignment of Principalship

Intuition tells us that there may be a trade-off between the marketing member's incentive and the producing member's incentive associated with assignments of principalship. Loosely speaking, the principal's incentive comes from self-monitoring, while the agent's incentive can only come from being monitored. Different assignments of principalship provide different combinations of the two members' incentives. The Pareto-dominance of one assignment over another depends on the distribution function $\Phi(Y; a_M, a_P)$, monitoring technology, as well as the individuals' utility functions. The purpose of the following analysis is to characterize these dependencies.

For the analysis to be tractable, we shall parameterize all important variables by making some *technical* assumptions.

First, we assume that both the marketing member and the producing member are identical in preferences and risk-neutral, and work effort and monitoring effort enter the utility function symmetrically and additively. In particular, we assume the utility function takes the following form[22]:

$$U_i = Y_i - 0.5a_i^2 - 0.5b_i^2,\ i = M, P \tag{2.6}$$

A popular assumption in economics of information is that the principal is risk-neutral and the agent is risk-averse, although the explanation is not unambiguous.[23] We agree that risk-attitudes play an important role in contractual relationship. But in the context of the firm, risk-neutrality is neither a necessary nor a sufficient condition for someone to be the principal. The incentive problem may be so dominant that an optimal contract may require a member to be the principal even if he is risk-averse, and the other to be the agent even if he is risk-neutral. The assumption of risk-neutrality allows us to focus on the incentive problem, because under this assumption, contracts have no insurance function. We shall discuss the effect of risk-aversion on the optimal assignment of principalship in Sect. 2.5. Another advantage of the risk-neutrality assumption is that when both members are risk-neutral (with additive preferences),

[22]The number can be replaced by parameters. Here the numbers are chosen so as to make the expressions simple.

[23]One argument is that the optimal contract requires that the risk-neutral member becomes the principal and the risk-averse becomes an agent. But the most of literature simply assumes that the principal is risk-neutral.

Nash-equilibrium actions are independent of the fixed payment w_i^s. Because any utility constraint can be satisfied by adjusting w_i^s, this assumption implies that Nash-equilibrium actions are independent of the participation constraints. Therefore our conclusion about the optimal assignment of principalship will not be affected by bargaining power. Because our primary concern is who should be the residual-claimant rather than how much each should receive from the total return, the assignment of principalship can be identified by the one-dimensional variable β.

The symmetric treatment of work effort and monitoring effort in the utility function seems questionable. This is not essential for the results, however, as long as the two members have identical preferences. In addition, $(0.5a_i^2 + 0.5b_i^2)$ can be replaced by $0.5(a_i + b_i)^2$ without much effect on the results. We assume $(0.5a_i^2 + 0.5b_i^2)$ instead of $0.5(a_i + b_i)^2$ because imperfect substitution between working effort and monitoring effort seems more reasonable than perfect substitution.[24]

Second, under risk-neutrality, β affects the two members' respective utility only through the expected return (EY_M and EY_P) and the choice of actions (e_M and e_P).[25] This allows us to be concerned with the firm's expected return function $Y = f(a_M, a_P)$, instead of the distribution function $\Phi(Y; a_M, a_P)$. Assumption 1 reduces to the assumption that the expected return to the firm is an increasing, concave function of a_M and a_P, with $\frac{\partial^2 Y}{\partial a_M \partial a_P} > 0$. In particular, we shall use the following CES (constant elasticity of substitution) form to characterize $Y = f(a_M, a_P)$:

$$Y = f(a_M, a_P) = \left(\alpha a_M^{1-\gamma} + (1-\alpha) a_P^{1-\gamma}\right)^{\frac{1}{1-\gamma}} \tag{2.7}$$

The CES function contains two parameters α and γ, both of which will play an important role in determining the optimal assignment of principalship. Mathematically α and $(1-\alpha)$ are the parameters of the effort elasticities of Y with respect to a_M and a_P, respectively ($0 \le \alpha \le 1$). In this thesis, we interpret α as a measurement of the relative importance of the two members in teamwork. $\alpha = \frac{1}{2}$ implies the two members are equally important; $\alpha > \frac{1}{2}$ implies the marketing member is more important; and $\alpha < \frac{1}{2}$ implies the producing member is more important.

γ is the parameter of elasticity of substitution between the two members' work efforts. It is easy to verify that for any given $0 < \alpha < 1$, the mixed partial derivatives $\frac{\partial^2 Y}{\partial a_M \partial a_P}$ are an increasing function of γ. In particular, when $\gamma = 0$, (2.7) reduces to the linear function $Y = \alpha a_M + (1-\alpha) a_P$ and $\frac{\partial^2 Y}{\partial a_M \partial a_P} \equiv 0$ for all $a_P \ge 0$ and $a_M \ge 0$; when $\gamma = 1$, (2.7) converges to the Cobb–Douglas function: $Y = a_M^{\alpha} a_P^{1-\alpha}$, and $\frac{\partial Y}{\partial a_i} \equiv 0$ at $a_j = 0$ for all $a_i \ge 0$. For this reason, we define γ as "the degree of teamwork" and assume $0 < \gamma \le 1$. $\gamma = 0$ implies no teamwork, which is trivial

[24] In Itoh (1991), imperfect substitution between own effort and helping effort is essential for teamwork to be optimal in designing an incentive scheme.

[25] To avoid complexity of notation, in the following analysis except Sect. 2.5, we use Y instead of EY to denote the expected return.

because in that case the first best can be achieved by dissolving the "firm" into two individual businessmen; $\gamma = 1$ generates *pure* teamwork.[26]

Thirdly, for simplicity, we assume that the monitoring technology takes the following linear forms:

$$\begin{array}{l}(i)\ \ a_P^b = \rho b_M \\ (ii)\ a_M^b = \mu b_P\end{array} \tag{2.8}$$

where ρ and μ measure the effectiveness of monitoring: the easier to monitor P (M), the greater ρ (μ). $\rho = \mu = 0$ implies that monitoring is technically impossible; $\rho \to \infty$ and $\mu \to \infty$ implies that monitoring is perfect.

With these specifications, the non-cooperative game is reduced to[27]:

Producing member

$$\begin{array}{ll} Max \\ \{a_P, b_p\} & \beta w_P + (1-\beta)\left(Y^b - w_M\right) - 0.5a_P^2 - 0.5b_P^2 \\ \text{s.t..} & a_P \geq \rho b_M \\ & a_M \geq \mu b_P \\ & w_M \geq \begin{cases} w_M^s & \text{if } b_P = 0 \\ w_M^s + \frac{1}{1-\beta}\left(0.5(\mu b_P)^2 - 0.5(a_M^s)^2\right) - \frac{\beta}{1-\beta}\left(Y^b - Y^s\right) & \text{if } b_P > 0 \end{cases} \\ & w_P \geq \begin{cases} w_P^s & \text{if } b_M = 0 \\ w_P^s + \frac{1}{\beta}\left(0.5(\rho b_M)^2 - 0.5(a_P^s)^2\right) - \frac{1-\beta}{\beta}\left(Y^b - Y^s\right) & \text{if } b_M > 0 \end{cases} \end{array} \tag{2.9}$$

Marketing member:

$$\begin{array}{ll} Max \\ \{a_M, b_M\} & (1-\beta) w_M + \beta\left(Y^b - w_P\right) - 0.5a_M^2 - 0.5b_M^2 \\ \text{s.t.} & a_P \geq \rho b_M \\ & a_M \geq \mu b_P \\ & w_P = \begin{cases} w_P^s & \text{if } b_M = 0 \\ w_P^s + \frac{1}{\beta}\left(0.5(\rho b_M)^2 - 0.5(a_P^s)^2\right) - \frac{1-\beta}{\beta}\left(Y^b - Y^s\right) & \text{if } b_M > 0 \end{cases} \\ & w_M = \begin{cases} w_M^s & \text{if } b_P = 0 \\ w_M^s + \frac{1}{1-\beta}\left(0.5(\mu b_P)^2 - 0.5(a_M^s)^2\right) - \frac{\beta}{1-\beta}\left(Y^b - Y^s\right) & \text{if } b_P > 0 \end{cases} \end{array} \tag{2.10}$$

where

$$Y^b = \begin{cases} \left(\alpha\left(\mu b_P\right)^{1-\gamma} + (1-\alpha)\left(\rho b_M\right)^{1-\gamma}\right)^{\frac{1}{1-\gamma}} & \text{if } b_P > 0 \text{ and } b_M > 0 \\ \left(\alpha\left(a_M^s\right)^{1-\gamma} + (1-\alpha)\left(\rho b_M\right)^{1-\gamma}\right)^{\frac{1}{1-\gamma}} & \text{if } b_P = 0 \text{ and } b_M > 0 \\ \left(\alpha\left(\mu b_P\right)^{1-\gamma} + (1-\alpha)\left(a_P^s\right)^{1-\gamma}\right)^{\frac{1}{1-\gamma}} & \text{if } b_P > 0 \text{ and } b_M = 0 \\ \left(\alpha\left(a_M^s\right)^{1-\gamma} + (1-\alpha)\left(a_P^s\right)^{1-\gamma}\right)^{\frac{1}{1-\gamma}} & \text{if } b_P = 0 \text{ and } b_M = 0 \end{cases}$$

[26] Strictly speaking, Leontief technology ($\gamma = \infty$) is pure teamwork.

[27] The specific forms of w_M and w_P in (2.9) and (2.10) are derived from (2.5).

$$Y^s = \left(\alpha \left(a_M^s\right)^{1-\gamma} + (1-\alpha)\left(a_P^s\right)^{1-\gamma}\right)^{\frac{1}{1-\gamma}}$$

The cooperative game is reduced to:

$$\begin{aligned} \underset{\{\beta\}}{Max} \; & \left(\alpha \left(a_M(\beta)\right)^{1-\gamma} + (1-\alpha)\left(a_P(\beta)\right)^{1-\gamma}\right)^{\frac{1}{1-\gamma}} \\ & -0.5\left(\left(a_M(\beta)\right)^2 + \left(b_M(\beta)\right)^2\right) - 0.5\left(\left(a_P(\beta)\right)^2 + \left(b_P(\beta)\right)^2\right) \end{aligned} \tag{2.11}$$

That is, the assignment of principalship is chosen to maximize the total expected return net of the total costs of work effort and monitoring effort.

Let β^* solve the above problem. What we need to show is how β^* depends on $(\rho, \mu; \alpha; \gamma)$.

2.3.1 *Optimal Assignment When Monitoring Is Technically Impossible*

As a starting point, let us first consider how the optimal choice of work efforts for each member depends on β and (α, γ) when monitoring is technically impossible. The results are contained in Lemmas 1 and 2 and Theorem 3.

Lemma 1 *Assume that utility functions are given by (2.6), the return function by (2.7) and the monitoring technology by (2.8). Then, if $\rho = \mu \equiv 0$, the optimal choices of efforts have the following properties: (i) $b_P^* = b_M^* \equiv 0$; (ii) $a_P^*(\beta)$ and $a_M^*(\beta)$ are quasi-concave, first increasing and then decreasing in β for $\gamma > 0$; (iii) as $\gamma \to 1$, $a_P^*(0) = a_P^*(1) = a_M^*(0) = a_M^*(1) = 0$.*

Proof (i) The statement in (i) is obvious. If $\mu = 0$, Y is independent of b_P but b_P incurs disutility. Therefore, as a utility-maximizer, the producing member will choose $b_P^* = 0$. Similarly, when $\rho = 0$, the marketing member will choose $b_M^* = 0$.

(ii) From the first-order conditions, we have

$$a_P^{1+\gamma} = (1-\beta)(1-\alpha)\left[\alpha a_M^{1-\gamma} + (1-\alpha)a_P^{1-\gamma}\right]^{\frac{\gamma}{1-\gamma}} \tag{2.12}$$

$$a_M^{1+\gamma} = \beta\alpha\left[\alpha a_M^{1-\gamma} + (1-\alpha)a_P^{1-\gamma}\right]^{\frac{\gamma}{1-\gamma}} \tag{2.13}$$

Equations (2.12) and (2.13) define two reaction functions respectively for the producing member and the marketing member, and the following non-cooperative equilibrium solutions of work efforts:

$$a_P^* = (1-\beta)(1-\alpha)\left[(1-\alpha) + \alpha\left(\frac{\beta\alpha}{(1-\beta)(1-\alpha)}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{\gamma}{1-\gamma}} \tag{2.14}$$

$$a_M^* = \beta\alpha\left[\alpha + (1-\alpha)\left(\frac{(1-\beta)(1-\alpha)}{\beta\alpha}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{\gamma}{1-\gamma}} \tag{2.15}$$

Differentiating Eq. (2.14) gives

$$\frac{\partial a_P^*}{\partial \beta} = (1-\alpha)\triangle^{\frac{\gamma}{1-\gamma}-1}\left[-\triangle + \left(\frac{\gamma}{1+\gamma}\right)\left(\frac{\alpha}{\beta}\right)\left(\frac{\beta\alpha}{(1-\beta)(1-\alpha)}\right)^{\frac{1-\gamma}{1+\gamma}}\right] \tag{2.16}$$

where

$$\triangle = \left[(1-\alpha) + \alpha\left(\frac{\beta\alpha}{(1-\beta)(1-\alpha)}\right)^{\frac{1-\gamma}{1+\gamma}}\right]$$

Then the sign of $\frac{\partial a_P^*}{\partial \beta}$ is equivalent to the sign of

$$O_P \equiv \alpha\left(\frac{\beta\alpha}{(1-\beta)(1-\alpha)}\right)^{\frac{1-\gamma}{1+\gamma}}\left(\frac{\gamma}{(1+\gamma)\beta} - 1\right) - (1-\alpha) \tag{2.17}$$

O_P can be greater than, equal to and less than zero, depending on β for given $\gamma > 0$. This can be verified by noting that $O_P\,|_{\beta\to 0} = +\infty$ and $O_P\,|_{\beta\to 1} = -\infty$. Because $O_P(\beta)$ is a continuous function in (0, 1), there must be some point $\beta \in (0, 1)$ such that $O_P(\beta) = 0$. Because O_P is monotonously decreasing in β, $\beta = (\beta : O_P(\beta) = 0)$ is unique.

Thus we have proved that a_P^* is first increasing and then decreasing with β. Similarly we can also prove the claim for a_M^*.

(iii) When $\gamma \to 1$, the returns function reduces to the following Cobb–Douglas form:

$$Y = a_M^{\alpha} a_P^{1-\alpha}$$

The first-order conditions are

$$a_P = ((1-\beta)(1-\alpha))^{\frac{1}{1+\alpha}}\, a_M^{\frac{\alpha}{1+\alpha}} \tag{2.18}$$

$$a_M = (\beta\alpha)^{\frac{1}{2-\alpha}}\, a_P^{\frac{1-\alpha}{2-\alpha}} \tag{2.19}$$

The Nash equilibrium solutions are

$$a_P^* = (\beta\alpha)^{\frac{\alpha}{2}}\, ((1-\beta)(1-\alpha))^{1-\frac{\alpha}{2}} \tag{2.20}$$

$$a_M^* = (\beta\alpha)^{\frac{1+\alpha}{2}}\, ((1-\beta)(1-\alpha))^{\frac{1-\alpha}{2}} \tag{2.21}$$

It is easy to see that $a_P^*(0) = a_P^*(1) = a_M^*(0) = a_M^*(1) = 0$. □

Remarks Lemma 1 has some important implications for understanding the individual's behavior in the presence of teamwork but in the absence of monitoring. It says that as long as teamwork occurs to some degree, neither the producing member's work effort nor the marketing member's work effort would be maximized by full principalship arrangements ($\beta = 0$, or $\beta = 1$). The intuition is that although one member's effort increases in reaction to any increase in his own residual share for any *given* effort made by the other (this can be seen from the reaction functions), his willingness to increase effort might be undermined by the disincentive of the other member caused by such a move, because in the presence of teamwork the marginal productivity of his effort depends positively on the other's effort. Therefore an increase in his own residual share has two opposite effects: on the one hand, it makes him more interested in the total return of the firm, which induces him to work harder; on the other hand, it makes his effort less valuable if the other's effort falls, which discourages him from working hard. Similarly, when one member's residual share decreases, the above effects work in the opposite directions. The equilibrium result depends on the balance of the two effects in both directions (remember that an increase in one member's residual share implies a decrease in the other's residual share). In particular, as one member becomes the only residual claimant, the second effect is so dominant that his incentive to work will be almost as low as his fixed-wage colleague's, if the degree of teamwork is high. The fundamental reasons are different, however: his incentive is low because the marginal productivity is low, while his fellow's incentive is low because the residual share is low. Figures 2.1 and 2.2 give some intuition about the shapes of incentive functions $a_P(\beta)$ and $a_M(\beta)$ for different γ and different α respectively.

Figures 2.1 and 2.2 also give some intuition about the relationship of the maximum effort share to the members' relative importance (α) in production and to the degree

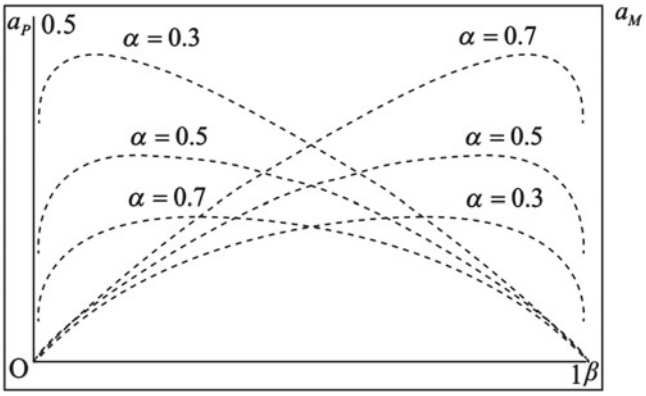

Fig. 2.1 Incentive Functions $a_P(.)$ and $a_M(.)$ where $\alpha = 0.3, 0.5, 0.7$, and $\gamma = 0.8$

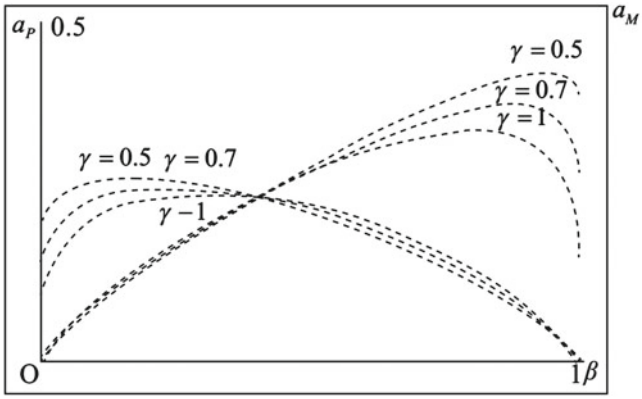

Fig. 2.2 Incentive Function $a_P(.)$ and $a_M(.)$ where $\alpha = 0.6$ and $\gamma = 0.5, 0.7, 1$

of teamwork (γ). Let $\underline{\beta}$ be the residual share at which the producing member's effort is maximized, and $\overline{\beta}$ be the residual share at which the marketing member's effort is maximized. Then we have

Lemma 2 *Assume that utility functions are given by (2.6), the returns function by (2.7), and the monitoring technologies by (2.8), and $\rho = \mu = 0$. Then: (i) $\underline{\beta} < \overline{\beta}$ for all $\alpha > 0$ and $\gamma > 0$; (ii) $\underline{\beta}$ is an increasing function of γ and $\overline{\beta}$ is a decreasing function of γ; (iii) both $\underline{\beta}$ and $\overline{\beta}$ are increasing with α.*

Proof For convenience of presentation, we first prove (ii) and (iii) and then go back to (i).

(ii) From the first order conditions, $\underline{\beta}$ satisfies:

$$O_P = \alpha\left(\frac{\beta\alpha}{(1-\beta)(1-\alpha)}\right)^{\frac{1-\gamma}{1+\gamma}}\left(\frac{\gamma}{(1+\gamma)\beta} - 1\right) - (1-\alpha) = 0 \tag{2.22}$$

and $\overline{\beta}$ satisfies:

$$O_M = (1-\alpha)\left(\frac{(1-\beta)(1-\alpha)}{\beta\alpha}\right)^{\frac{1-\gamma}{1+\gamma}}\left(\frac{\gamma}{(1+\gamma)(1-\beta)} - 1\right) - \alpha = 0 \tag{2.23}$$

Rearranging and differentiating (2.22) gives

$$\frac{\partial O_P}{\partial \beta}|_{\underline{\beta}} = -\left(\frac{\beta(1-\gamma) + 2\gamma^2(1-\beta)}{(1+\gamma)^2(1-\beta)\beta^2}\right)\left(\frac{\beta}{1-\beta}\right)^{\frac{1-\gamma}{1+\gamma}} < 0 \tag{2.24}$$

and

$$\frac{\partial O_P}{\partial \gamma}|_{\underline{\beta}} = \left(\frac{1}{(1+\gamma)^2}\right)\left(\frac{\beta}{1-\beta}\right)^{\frac{1-\gamma}{1+\gamma}}\left[-2\ln\left(\frac{\beta}{1-\beta}\right)\left(\frac{\gamma}{(1+\gamma)\beta}-1\right)+\frac{1}{\beta}\right] > 0 \tag{2.25}$$

Therefore

$$\frac{\partial \underline{\beta}}{\partial \gamma} = -\frac{\frac{\partial P}{\partial \gamma}}{\frac{\partial P}{\partial \beta}} > 0 \tag{2.26}$$

Similarly, since

$$\frac{\partial O_M}{\partial \beta}|_{\overline{\beta}} = \left(\frac{(1-\beta)(1-\gamma)+2\gamma^2\beta}{(1+\gamma)^2(1-\beta)^2\beta}\right)\left(\frac{1-\beta}{\beta}\right)^{\frac{1-\gamma}{1+\gamma}} > 0 \tag{2.27}$$

$$\frac{\partial O_M}{\partial \gamma}|_{\overline{\beta}} = \left(\frac{1}{(1+\gamma)^2}\right)\left(\frac{1-\beta}{\beta}\right)^{\frac{1-\gamma}{1+\gamma}}\left(-2\ln\left(\frac{1-\beta}{\beta}\right)\left(\frac{\gamma}{(1+\gamma)(1-\beta)}-1\right)+\frac{1}{1-\beta}\right) > 0 \tag{2.28}$$

Therefore

$$\frac{\partial \overline{\beta}}{\partial \gamma} = -\frac{\frac{\partial M}{\partial \gamma}}{\frac{\partial M}{\partial \beta}} < 0 \tag{2.29}$$

(iii) Since

$$\frac{\partial O_P}{\partial \alpha}|_{\underline{\beta}} = \left(\frac{2}{(1+\gamma)\alpha^2}\right)\left(\frac{1-\alpha}{\alpha}\right)^{\frac{1-\gamma}{1+\gamma}} > 0 \tag{2.30}$$

$$\frac{\partial O_M}{\partial \alpha}|_{\overline{\beta}} = -\left(\frac{2}{(1+\gamma)(1-\alpha)^2}\right)\left(\frac{\alpha}{1-\alpha}\right)^{\frac{1-\gamma}{1+\gamma}} < 0 \tag{2.31}$$

So

$$\frac{\partial \underline{\beta}}{\partial \alpha} = -\frac{\frac{\partial P}{\partial \alpha}}{\frac{\partial P}{\partial \beta}} > 0 \tag{2.32}$$

$$\frac{\partial \overline{\beta}}{\partial \alpha} = -\frac{\frac{\partial M}{\partial \alpha}}{\frac{\partial M}{\partial \beta}} > 0 \tag{2.33}$$

(i) By (ii), $\underline{\beta} \le \underline{\beta}(\gamma = 1)$ and $\overline{\beta} \ge \overline{\beta}(\gamma = 1)$. But, from (2.20) and (2.21), we obtain

$$\underline{\beta} = \tfrac{\alpha}{2} \text{ if } \gamma = 1$$

$$\overline{\beta} = \tfrac{1+\alpha}{2} \text{ if } \gamma = 1$$

□

Remarks An interpretation of Lemma 2 (i) is that the producing member's maximum effort always comes 'earlier' than the marketing member's along the curve of β. The implications are as follows. First, for $\underline{\beta} > \beta > \overline{\beta}$, both members' efforts increase as the residual becomes more equally shared. Second, it is impossible to find a β at which both members' efforts are maximized. There must exist some incentive trade-off in the interval $[\underline{\beta}, \overline{\beta}]$. Third, for any given α and γ, to induce one member to make a greater effort than his fellow always requires him to be granted a bigger residual share. Lemma 2 (ii) suggests that the distance between $\underline{\beta}$ and $\overline{\beta}$ shrinks as the degree of teamwork increases. In other words, from incentive point of view, the greater degree of teamwork requires more equal residual sharing. The reason is that as the degree of teamwork increases, the interdependence of marginal productivities increases, and therefore the negative effect associated with any switch in the residual share will overtake the positive effect earlier. Lemma 2 (iii) says that one member's maximum effort residual share increases with his relative importance in contribution to the firm's return. Greater importance simply mitigates the negative effect caused by the other member's disincentive. Because the sum of "importance" in production has been assumed to equal to one, greater importance of one member implies less importance of the other, and therefore an increase in α will shift both incentive curves rightwards (see Fig. 2.1). In particular, in the case of pure teamwork ($\gamma = 1$), $(1 - \underline{\beta}) = \left(\frac{1}{2} + \frac{1-\alpha}{2}\right)$ and $\overline{\beta} = \left(\frac{1}{2} + \frac{\alpha}{2}\right)$, where the common term $\left(\frac{1}{2}\right)$ can be explained as the team effect while the terms $\left(\frac{1-\alpha}{2}\right)$ and $\left(\frac{\alpha}{2}\right)$ reflect the effects of relative importance.[28]

We can now characterize the optimal assignment of principalship when monitoring is technically impossible. Denote by β_Y the residual share which maximizes the total expected return of the firm, and by β_Π the residual share which solves the cooperative game (2.11) (i.e., maximizes the total welfare), when $\rho = \mu = 0$. Then we obtain:

Theorem 3 *Assume that utility functions are given by (2.6), the returns function by (2.7), the monitoring technology by (2.8), and* $\rho = \mu = 0$. *Then: (i)* $\underline{\beta} < \beta_Y < \overline{\beta}$ *and* $\underline{\beta} < \beta_\Pi < \overline{\beta}$; *(ii) both* β_Y *and* β_Π *are monotonically increasing with* α *for any given* $\gamma > 0$, *and increasing with* γ *for* $\alpha < \frac{1}{2}$ *and decreasing with* γ *for* $\alpha > \frac{1}{2}$; *(iii)* $\beta_Y < \beta_\Pi$ *for* $\alpha < \frac{1}{2}$ *and* $\beta_Y > \beta_\Pi$ *for* $\alpha > \frac{1}{2}$.

Proof (i) By Lemmas 1 and 2, β_Y and β_Π cannot be in $[0, \underline{\beta})$ and $(\overline{\beta}, 1]$. Beginning with $\beta = \underline{\beta}$, an infinitesimal increase in β incurs only a second-order loss in a_P but has a first-order positive effect on a_M, and therefore Y must increase. Similarly, beginning with $\beta = \overline{\beta}$, an infinitesimal decrease in β incurs only a second-order loss in a_M but has a first-order positive effect on a_P, and therefore Y must increase. To see β_Π also lies in $(\underline{\beta}, \overline{\beta})$, simply note that individual maximization problems imply that at the margin, for any given effort by the other member, the marginal disutility of effort is smaller than the marginal productivity of effort; thus, an infinitesimal change starting from $\beta = \underline{\beta}$ or $\beta = \overline{\beta}$ will not only increase Y but also increase Π.

[28] Note an increase in the degree of teamwork has symmetric effects on both member's incentive functions, while an increase in alpha has asymmetric effects.

(ii) Note that at the equilibrium,

$$a_P^* = \left(\frac{(1-\beta)(1-\alpha)}{\beta\alpha}\right)^{\frac{1}{1+\gamma}} a_M^* \tag{2.34}$$

So

$$Y = \beta\alpha\left[\alpha + (1-\alpha)\left(\frac{(1-\beta)(1-\alpha)}{\beta\alpha}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{1+\gamma}{1-\gamma}} \tag{2.35}$$

and

$$\begin{aligned}\Pi = {} & \beta\alpha\left[\alpha + (1-\alpha)\left(\tfrac{(1-\beta)(1-\alpha)}{\beta\alpha}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{1+\gamma}{1-\gamma}} \\ & -0.5\left\{\beta\alpha\left[\alpha + (1-\alpha)\left(\tfrac{(1-\beta)(1-\alpha)}{\beta\alpha}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{\gamma}{1-\gamma}}\right\}^2\left(1 + \left(\tfrac{(1-\beta)(1-\alpha)}{\beta\alpha}\right)^{\frac{2}{1+\gamma}}\right)\end{aligned} \tag{2.36}$$

Then β_Y satisfies the following first order condition:

$$\frac{\partial Y}{\partial \beta} = \alpha\triangle^{\frac{1+\gamma}{1-\gamma}-1}\left(\triangle - \left(\frac{1-\alpha}{1-\beta}\right)\left(\frac{(1-\beta)(1-\alpha)}{\beta\alpha}\right)^{\frac{1-\gamma}{1+\gamma}}\right) = 0 \tag{2.37}$$

where

$$\triangle = \left[\alpha + (1-\alpha)\left(\frac{(1-\beta)(1-\alpha)}{\beta\alpha}\right)^{\frac{1-\gamma}{1+\gamma}}\right]$$

By arranging (2.37), we obtain

$$\frac{\beta_Y}{1-\beta_Y} = \left(\frac{\alpha}{1-\alpha}\right)^{\frac{1}{\gamma}} \tag{2.38}$$

This is a simple but very interesting result. It is easy to demonstrate that β_Y is a monotonically increasing, first convex and then concave function of α with $\alpha = \frac{1}{2}$ as the turning point. It is also easy to see that β_Y is increasing with γ for $\alpha < \frac{1}{2}$ but decreasing with γ for $\alpha > \frac{1}{2}$.

β_Π satisfies the following first order condition:

$$\begin{aligned}&\left[\alpha + (1-\alpha)\left(\tfrac{(1-\beta)(1-\alpha)}{\beta\alpha}\right)^{\frac{1-\gamma}{1+\gamma}}\left(\tfrac{1-\beta(2+\gamma)}{(1-\beta)(1+\gamma)}\right)\right]\left[\alpha + (1-\alpha)\left(\tfrac{(1-\beta)(1-\alpha)}{\beta\alpha}\right)^{\frac{1-\gamma}{1+\gamma}}\right] \\ &= \beta\alpha\left[\alpha + (1-\alpha)\left(\tfrac{(1-\beta)(1-\alpha)}{\beta\alpha}\right)^{\frac{1-\gamma}{1+\gamma}}\left(\tfrac{1-\beta(1+\gamma)}{(1-\beta)(1+\gamma)}\right)\right]\left[1 + \left(\tfrac{(1-\beta)(1-\alpha)}{\beta\alpha}\right)^{\frac{2}{1+\gamma}}\right]\end{aligned} \tag{2.39}$$

This is the simplest analytic expression available. Formally identifying the relationship between β_Π and α and γ by using this expression is very time-consuming and tedious. However, by applying an argument similar to that in the proof of (i), it can be verified that β_Π is also an increasing function of α, and increasing with γ for $\alpha < \frac{1}{2}$ and decreasing with γ for $\alpha > \frac{1}{2}$. (Also see remarks.)

(iii) For $\alpha = \frac{1}{2}$, the optimal solution has the property of symmetry: $\beta_Y = \beta_\Pi = \frac{1}{2}$ and $a_P^* = a_M^*$. This follows from the assumption of identical preferences. For $\alpha < \frac{1}{2}$, $\beta_Y < \frac{1}{2}$ and $a_P^*(\beta_Y) > a_M^*(\beta_Y)$. Then at $\beta = \beta_Y$, the marginal disutility of a_P is greater than the marginal disutility of a_M; an infinitesimal increase in β from β_Y incurs a second order loss in Y but has a first order gain in reduction of costs in terms of disutility. Therefore $\beta_\Pi > \beta_Y$ for $\alpha < \frac{1}{2}$. Similarly, for $\alpha > \frac{1}{2}$, $\beta_Y > \frac{1}{2}$ and $a_P^*(\beta_Y) < a_M^*(\beta_Y)$; and the marginal disutility of a_P is smaller than the marginal disutility of a_M at β_Y. An infinitesimal decrease in β from β_Y will increase Π and so $\beta_\Pi < \beta_Y$. □

Remarks The results in Theorem 3 are quite intuitive. It cannot be optimal to fully assign the principalship to either of the two members when monitoring is technically impossible. The optimal assignment requires a balance of incentives between the two members. Furthermore the optimal residual share held by each member should be positively related to his relative importance in the returns function. The more important his effort is, the bigger his residual share. However, the relationship is not in general linear. There are two effects working in determining the deviation of the optimal residual share from the relative importance. The first is the output effect. From the point of view of maximization of the total return, the residual share assigned to the more important member should be more than proportional to his relative importance: $\beta_Y < \alpha$ for $\alpha < \frac{1}{2}$ and $\beta_Y > \alpha$ for $\alpha > \frac{1}{2}$, unless $\gamma = 1$ in which case $\beta_Y \equiv \alpha$. The reason is that at $\beta = \alpha$, the marginal productivity of the more important member is greater than the marginal productivity of the less important member; a small shift to favour the more important member will induce him to work more (remember $\underline{\beta} < \beta_Y < \overline{\beta}$), which more than offsets the disincentive of the less important member, as long as the returns function is not pure teamwork. The second is the cost effect. Under the assumption of identical preferences, at $\beta = \alpha$, the more important member incurs higher marginal cost (in terms of disutility) than the less important member in providing effort. Therefore it is desirable to deviate from $\beta = \alpha$ from the point of view of cost reduction. Whether the optimal residual share β_Π is less than, equal to or greater than the relative importance depends on the dominance of one effect over the other, which in turn depends on the degree of teamwork through the interdependence of marginal productivities. A higher degree of teamwork implies higher interdependence and hence a lower productivity advantage of the more important member at $\beta = \alpha$. In particular, when the returns function exhibits pure teamwork ($\gamma = 1$), the two marginal productivities are equalized at $\beta = \alpha$; and therefore the output effect disappears at $\beta = \alpha$ and the cost effect implies that the less important member should be assigned a residual share more than proportional to his relative importance (in other words, the more important member should be assigned a residual share less than proportional to his relative importance).

2.3.2 *Optimal Assignment When Monitoring Is Technically Possible*

We now turn to the case where $\rho > 0$ and $\mu > 0$; that is, monitoring is technically possible. In this case, for any given β, there are four possible outcomes: (i) P monitors M but M does not monitor P; (ii) M monitors P but P does not monitor M; (iii) neither P nor M monitor the other; and (iv) P and M monitor each other. We call the first two "one-sided monitoring regimes" (respectively P's monitoring regime and M's monitoring regime), the third "the monitoring free regime", and the fourth "the mutual-monitoring regime". Lemma 4 characterizes the incentives within the one-sided monitoring regimes; Lemma 5 identifies the conditions for division of the regimes; Lemma 6 and Corollary 7 claim a pure principalship contract Pareto dominates the corresponding monitoring regime and that the mutual-monitoring regime is always Pareto dominated by both the one-sided monitoring regimes; finally Theorems 8 and 12 establish the conditions of the optimal assignment.

Lemma 4 *Assume that utility functions are given by (2.6), the returns function by (2.7), and the monitoring technology by (2.8), where $\mu > 0$ and $\rho > 0$. Then, at equilibrium, (i) if P monitors M but M does not monitor P, $a_P^*(= a_P^s)$ and b_P^* (therefore $a_M^* = a_M^b = \mu b_P^* > a_M^{s*}$) are decreasing with β and maximized at $\beta = 0$; (ii) if M monitors P but P does not monitor M, $a_M^*(= a_M^s)$ and b_M^* (therefore $a_P^* = a_P^b = \rho b_M^* > a_P^{s*}$) are increasing with β and maximized at $\beta = 1$.*

Proof (i) If P monitors M but M does not monitor P, P's problem is[29]:

$$\underset{\{a_P^s,\, b_P\}}{Max}\ EU_P = Y^b - \beta Y^s - C_P(a_P^s, b_P) - \left(C_M(a_M^b(b_P)) - C_M(a_M^s)\right) \quad (2.40)$$

where we omit irrelevant terms (1-β)w_M^s and βw_P^s.

Because of team work, when P chooses his self-incentive a_P^s, he knows that it will indirectly affect a_M^s; but because a_M^s is chosen by M such that:

$$\beta \frac{\partial Y^s}{\partial a_M^s} - \frac{\partial C_M}{\partial a_M^s} = 0, \quad (2.41)$$

by the envelope theorem this effect can be omitted from the optimization problem.

Two first-order conditions are as follows:

$$\frac{\partial Y^b}{\partial a_P^s} - \beta \frac{\partial Y^s}{\partial a_P^s} = \frac{\partial C_P}{\partial a_P^s} \quad (2.42)$$

$$\frac{\partial Y^b}{\partial a_M^b} \frac{\partial a_M^b}{\partial b_P} = \frac{\partial C_P}{\partial b_P} + \frac{\partial C_M}{\partial a_M} \frac{\partial a_M^b}{\partial b_P} \quad (2.43)$$

[29]The objective function is derived from (2.9). βY^s enters the function because of the effect on M's status quo of P's self-incentive effort.

The first term on the LHS of the condition (2.42) is the marginal effect of a_P^s on the output when monitoring is imposed, and the second is the marginal effect of a_P^s on the output when monitoring is not imposed, discounted by β which reflects the residual share effect. That is, when P chooses his self-incentive work effort, he will take into account its effect on the improvement of M's status quo. The optimization requires that the *net* marginal benefit equal the marginal cost. It is obvious that the optimal a_P^{s*} is decreasing with β; in particular, when $\beta = 0$, M's status quo is not affected by P's self-incentive, and (2.42) reduces to

$$\frac{\partial Y^b}{\partial a_P^s} = \frac{\partial C_P}{\partial a_P^s}$$

which implies that a_P^{s*} is maximized at $\beta = 0$.

The condition (2.43) requires that the marginal benefit of monitoring equal the total marginal cost of monitoring (the internal cost plus the external cost). It seems that the rule governing the acceptability of monitoring defined by Assumption 4 fully internalizes both the benefit and cost of monitoring. However, because a_M and a_P are complementary in producing Y, combining (2.42) and (2.43) implies that b_P^* is also decreasing with β and maximized at $\beta = 0$.

(ii) The proof of (ii) is nothing more than repeating a symmetric case. We present it here for the record.

When M monitors P but P does not monitor M, M's problem is

$$\underset{\{a_M^s,\, b_M\}}{Max}\ EU_M = Y^b - (1-\beta)Y^s - C_M(a_M^s, b_M) - \left(C_P(a_P^b(b_M)) - C_P(a_P^s)\right) \tag{2.44}$$

where items (1-β)w_M^s and βw_P^s are omitted for their irrelevances.

Two first-order conditions are:

$$\frac{\partial Y^b}{\partial a_M^s} - (1-\beta)\frac{\partial Y^s}{\partial a_M^s} = \frac{\partial C_M}{\partial a_M^s} \tag{2.45}$$

$$\frac{\partial Y^b}{\partial a_P^b}\frac{\partial a_P^b}{\partial b_M} = \frac{\partial C_M}{\partial b_M} + \frac{\partial C_P}{\partial a_P^b}\frac{\partial a_P^b}{\partial b_M} \tag{2.46}$$

where $(1-\beta)\frac{\partial Y^s}{\partial a_M^s}$ is the residual share effect.

Thus both a_M^{s*} and b_M^* are increasing with β and maximized at $\beta = 1$.

□

Remarks What Lemma 4 says is that the monitor's incentives to monitor as well as to work depend positively on his residual share; and that a fixed-wage member cannot have an incentive to monitor, while a full residual claimant has the greatest incentive to monitor. This kind of proposition is nothing new! Indeed it has been known since Alchian and Demsetz (1972) classic paper. What distinguishes the present model from the conventional view is that we find the proposition crucially depends on the

assumption of the status quo (Assumption 4) of no-monitoring which generates the residual share effect. To see this, let us consider an alternative definition of the status quo. For convenience of discussion, we will take the case of M-monitoring-P.

Suppose that when monitoring is technically impossible, the equilibrium is $a_M^s = a_M^\dagger$ and $a_P^s = a_P^\dagger$; and that when monitoring is technically possible, M finds it in his interest to monitor P, $a_M^s = a_M^{s*}$ and $b_M = b_M^*$, and $a_P^b = a_P^b(b_M^*)$. Given that $a_M^s = a_M^{s*}$, P will "choose" $a_P^s = a_P^{s*}(a_M^{s*}) > a_P^\dagger$ because of the teamwork effect; that is, P's self-incentive is greater when monitoring is technically possible than when monitoring is technically impossible (but $a_P^{s*}(a_M^{s*}) < a_P^b(b_M^*)$, otherwise monitoring is non-binding).

According to Assumption 4, P's status quo is equal to

$$w_P^s + (1-\beta)\left(Y(a_M^{s*}, a_P^{s*}) - w_M^s - w_P^s\right) - C_P(a_P^{s*}) \tag{2.47}$$

It is this assumption that guarantees the residual share effect which underlies the claim in Lemma 4.

Alternatively one might argue that P's status quo should be defined as follows, on the assumption that P could not do better if M *really* chose $b_M = 0$ (not monitoring):

$$w_P^s + (1-\beta)\left(Y(a_M^\dagger, a_P^\dagger) - w_M^s - w_P^s\right) - C_P(a_P^\dagger) \tag{2.48}$$

This kind of definition would rule out the residual share effect since M would fully internalize both benefit and cost of monitoring (i.e., $Y(a_M^\dagger, a_P^\dagger)$ is independent of a_M^{s*}).

The problem is: Why adopt (2.47) instead of (2.48)? The reason is that a contract with (2.48) as the status quo cannot be self-enforceable (unless the allocation of authority of monitoring is predetermined). Under such a contract, the member who plays monitoring captures the whole surplus while the other gains nothing, and therefore the incentive to be a monitor is always greater than the incentive to be monitored—nothing can be worse than being monitored! M's threat of not monitoring is not credible. Nor is P's! That is why we claim that Assumption 4 accurately reflects the bargaining problem between M and P. Under Assumption 4, monitoring is self-selected and the contract is self-enforcing.

It is interesting to note that even under Assumption 4, the monitoring incentive is independent of the residual share *if* the firm's return does not depend on the monitor's work effort. The intuition is that when the monitor does not work, the status quo of the monitored member defined by Assumption 4 is invariant to the monitor's monitoring effort for any given (w_j^s, λ_j); thus the monitor can fully internalize the net surplus of the monitoring effort.[30] This implies that a *professional* monitor's incentive to

[30]The argument cannot be extended to the case when the monitor takes no residual share at all. The reason is that in that case there is no channel through which the 'monitor' can capture any surplus generated by monitoring, unless the contract specifies the monitoring effort. But this case need not trouble us because the working member has already obtained a full incentive to work.

monitor is not sensitive to his residual share if the payment contract can be contingent on the monitored effort of the member monitored. This kind of argument contradicts the assertions of Alchian and Demsetz(1972), but is coincident (in spirit) with Harris and Holmstrom (1982) and McAfee and McMillian (1991).[31,32]

Lemma 4 specifies how the monitor's incentive to monitor changes with his residual share. The remaining question is: Who will be the monitor? We now seek to answer this question.

Lemma 5 *Assume that utility functions are given by (2.6), the returns function by (2.7), and the monitoring technology by (2.8), where* $\mu > 0$, $\rho > 0$. *Then, at equilibrium, there are a* $\beta_\mu < \frac{\mu^2}{1+\mu^2}$ *and a* $\beta_\rho > \frac{1}{1+\rho^2}$, *such that: (i) if* $\mu\rho \leq 1$, *the whole residual share interval [0,1] is divided into* $[0,\beta_\mu]$, (β_μ, β_ρ) *and* $[\beta_\rho, 1]$, *where* $[0,\beta_\mu]$ *is P′s monitoring regime,* (β_μ, β_ρ) *the monitoring-free regime, and* $[\beta_\rho, 1]M'$*s monitoring regime; (ii) if* $\mu\rho \gg 1$, *where "*$\gg$*" means "sufficiently larger than", the whole residual share interval [0,1] is divided into* $[0,\beta_\rho)$, $[\beta_\rho, \beta_\mu]$ *and* $(\beta_\mu, 1]$, *where* $[0,\beta_\rho)$ *is P′s monitoring regime,* $[\beta_\rho, \beta_\mu]$ *the mutual-monitoring regime, and* $[\beta_\mu, 1]M'$*s monitoring regime.*

Proof First note that direct observation of both members' objective functions suggests that at $\beta = 0$, P has incentive to monitor M but M cannot have incentive to monitor P; the reverse is true at $\beta = 1$.

When P monitors M but M does not monitor P, by substituting $a_M^b = \mu b_P$ into (2.42) and (2.43), we have:

$$\begin{aligned} a_P^{s*} : \left(a_P^s\right)^{1+\gamma} = (1-\alpha)\left[\alpha\mu^{1-\gamma}b_M^{1-\gamma} + (1-\alpha)(a_P^s)^{1-\gamma}\right]^{\frac{\gamma}{1-\gamma}} \\ -\beta(1-\alpha)\left[\alpha(a_M^s)^{1-\gamma} + (1-\alpha)(a_P^s)^{1-\gamma}\right]^{\frac{\gamma}{1-\gamma}} \end{aligned} \tag{2.49}$$

$$b_P^* : b_P^{1+\gamma} = \tfrac{\alpha\mu^{1-\gamma}}{1+\mu^2}\left[\alpha\mu^{1-\gamma}b_M^{1-\gamma} + (1-\alpha)(a_P^s)^{1-\gamma}\right]^{\frac{\gamma}{1-\gamma}} \tag{2.50}$$

where a_M^s is defined by M's reaction function (2.13):

$$a_M^{s*} : \left(a_M^s\right)^{1+\gamma} = \beta\alpha\left[\alpha(a_M^s)^{1-\gamma} + (1-\alpha)(a_P^s)^{1-\gamma}\right]^{\frac{\gamma}{1-\gamma}} \tag{2.51}$$

---

[31] In both Holmstrom and McAfee and McMillan, the principal is not a member of team. Under this assumption, Holmstrom argues that the principal's primary role is to break the budget-balancing constraint so as to create group incentive to work, while McAfee and McMillan argue that the role of monitoring is to discipline the monitor himself instead of the team. Holmstrom has clearly realized that "it is important that the principal not provide any (unobservable) productive inputs or else a free-rider problem remains." (pp. 328) We say our argument is coincident with theirs only in spirit because they do not explicitly model the choice of monitoring effort.

[32] It is widely known that the incentive problem depends on the degree to which the net surplus of a decision can be internalized by the decision-maker. A full internalization does not necessarily require a full residual claim; but if it does, the residual share is important.

Comparing (2.50) with (2.13), we see that when $\beta \geq \frac{\mu^2}{1+\mu^2}$, M's self-incentive will not be less than the monitored-incentive, and therefore $b_P^* = 0$. This implies that there must be a $\beta_\mu < \frac{\mu^2}{1+\mu^2}$ such that $b_P^* > 0$ iff $\beta \leq \beta_\mu$.

Similarly, when M monitors P but P does not monitor M, the following first order conditions hold:

$$\begin{aligned} a_M^{s*} : \left(a_M^s\right)^{1+\gamma} = \alpha \left[\alpha (a_M^s)^{1-\gamma} + (1-\alpha)\rho^{1-\gamma} b_M^{1-\gamma}\right]^{\frac{\gamma}{1-\gamma}} \\ -(1-\beta)\alpha \left[\alpha (a_M^s)^{1-\gamma} + (1-\alpha)(a_P^s)^{1-\gamma}\right]^{\frac{\gamma}{1-\gamma}} \end{aligned} \tag{2.52}$$

$$b_M^* : b_M^{1+\gamma} = \tfrac{(1-\alpha)\rho^{1-\gamma}}{1+\rho^2} \left[\alpha (a_M^s)^{1-\gamma} + (1-\alpha)\rho^{1-\gamma} b_M^{1-\gamma}\right]^{\frac{\gamma}{1-\gamma}} \tag{2.53}$$

where a_P^s is defined by P's reaction function (2.12):

$$a_P^{s*} : \left(a_P^s\right)^{1+\gamma} = (1-\beta)(1-\alpha) \left[\alpha (a_M^s)^{1-\gamma} + (1-\alpha)(a_P^s)^{1-\gamma}\right]^{\frac{\gamma}{1-\gamma}} \tag{2.54}$$

Comparing (2.53) with (2.12) suggests that when $\beta \leq \frac{1}{1+\rho^2}$, $b_M^* = 0$. Therefore there must be a $\beta_\rho > \frac{1}{1+\rho^2}$ such that $b_M^* > 0$ iff $\beta \geq \beta_\rho$.

Since $\mu\rho \leq 1$ is equivalent to $\left(\frac{\mu^2}{1+\mu^2} + \frac{\rho^2}{1+\rho^2}\right) \leq 1$ which in turn implies $\beta_\mu < \beta_\rho$, there will be no monitoring for $\beta \in (\beta_\mu, \beta_\rho)$.

To prove part (ii), note that at $\beta = \frac{\mu^2}{1+\mu^2}$, $a_M^{s*} > 0$ and a monitoring effort b_P, which satisfies $\mu b_P = a_M^{s*}$, will incur an discontinuous increase in monitoring cost. This implies that although at $\beta = \frac{\mu^2}{1+\mu^2}$, P can force M to work as much as M wishes, he will not do so because of monitoring cost, therefore β_μ must be sufficiently smaller than $\frac{\mu^2}{1+\mu^2}$. A similar argument applies to β_ρ: β_ρ must be sufficiently larger than $\frac{1}{1+\rho^2}$. However, if $\left(\frac{\mu^2}{1+\mu^2} + \frac{\rho^2}{1+\rho^2}\right)$ is sufficiently larger than one, we will have that $\beta_\mu > \beta_\rho$ such that $b_P^* > 0$ and $b_M^* > 0$ for $\beta \in [\beta_\rho, \beta_\mu]$. □

Remarks We call β_μ and β_ρ "switching points of monitoring regimes". The lemma can be interpreted as follows. When $\beta = 0$, P has a self-incentive to work as well as to monitor M who would shirk otherwise. As β increases from zero, P's self-incentive falls while M's self-incentive rises. But what M wishes to do is still less than he has to do until $\beta = \beta_\mu$. If $\mu\rho \leq 1$, once $\beta > \beta_\mu$, P loses interest in monitoring M while M is not ready to monitor P. They leave P's monitoring regime and enter the monitoring-free regime: each works as much as he likes; but P works less and less while M works more and more as β increases. Once $\beta \geq \beta_\rho$, M finds it in his interest to force P to work more than otherwise: they enter M's monitoring regime until $\beta = 1$. On the other hand, if $\mu\rho \gg 0$, as β increases from zero, P's self-incentive to work falls faster than the incentive to monitor M while M's self-incentive to work rises slower than the incentive to monitor P. Once $\beta \geq \beta_\rho$, M finds it in his interest to monitor P while P has not lost his interest in monitoring M. They enter the mutual-monitoring regime: each has to do more than he wishes to. When $\beta > \beta_\mu$,

P is no longer interested in monitoring M while $M's$ monitoring incentive becomes stronger and stronger. They enter M's monitoring regime, until $\beta = 1$.

The most important result is that the division of regimes depends upon the monitoring technology parameters μ and ρ. The scope of monitoring regimes are increasing with the effectiveness of monitoring, while the monitoring-free regime shrinks as monitoring technology improves. For instance, if both $\mu < 1$ and $\rho < 1$, monitoring cannot be implementable when the residual is equally shared between the two members (i.e., $\beta = \frac{1}{2}$). This implies that a symmetric contractual arrangement is more likely to generate a free-rider problem when monitoring is not very effective. In particular, as $\mu \to 0$ and $\rho \to 0$, $\beta_P \to 0$ and $\overline{\beta} \to 1$ and we go back to the case when monitoring is technically impossible. On the other hand, if both μ and ρ are sufficiently greater than unity, the mutual-monitoring regime will occur in which although neither M and nor P has (sufficient) incentive to work, they do have incentives to monitor each other such that at the equilibrium each of them has to work more than he wishes.

As a confirmation of the claim, we make the following comparisons.

When $\beta = 0$, solving (2.49) and (2.50), we obtain

$$a_P^{s*} = (1-\alpha)\left[(1-\alpha) + \alpha\left(\frac{\alpha\mu^2}{(1-\alpha)\mu^2}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{\gamma}{1+\gamma}} \tag{2.55}$$

$$b_P^{*} = (1-\alpha)\left(\frac{\alpha\mu^{1-\gamma}}{(1-\alpha)(1+\mu^2)}\right)^{\frac{1}{1+\gamma}}\left[(1-\alpha) + \alpha\left(\frac{\alpha\mu^2}{(1-\alpha)\mu^2}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{\gamma}{1+\gamma}} \tag{2.56}$$

When $\beta = 1$, solving (2.52) and (2.53), we obtain

$$a_M^{s*} = \alpha\left[\alpha + (1-\alpha)\left(\frac{(1-\alpha)\rho^2}{\alpha(1+\rho^2)}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{\gamma}{1-\gamma}} \tag{2.57}$$

$$b_M^{*} = \alpha\left(\frac{(1-\alpha)\rho^2}{\alpha(1+\rho^2)}\right)^{\frac{1}{1+\gamma}}\left[\alpha + (1-\alpha)\left(\frac{(1-\alpha)\rho^2}{\alpha(1+\rho^2)}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{\gamma}{1-\gamma}} \tag{2.58}$$

For $\beta \in (0, 1)$, if monitoring is not imposed, solving (2.51) and (2.54), we obtain

$$a_P^{s*} = (1-\beta)(1-\alpha)\left[(1-\alpha) + \alpha\left(\frac{\beta\alpha}{(1-\beta)(1-\alpha)}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{\gamma}{1-\gamma}} \tag{2.59}$$

$$a_M^{s*} = \beta\alpha\left[\alpha + (1-\alpha)\left(\frac{(1-\beta)(1-\alpha}{\beta\alpha}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{1}{1-\gamma}} \tag{2.60}$$

If the two members choose to monitor each other, solving Eqs. (2.50) and (2.53) (where a_P^s is replaced by μb_M and a_M^s by ρb_P), we obtain[33]

$$b_P^* = \alpha\left(\frac{\mu^{1-\gamma}}{1+\mu^2}\right)^{\frac{1}{1+\gamma}}\left(\alpha\left(\frac{\mu^2}{1+\mu^2}\right)^{\frac{1-\gamma}{1+\gamma}} + (1-\alpha)\left(\frac{(1-\alpha)\rho^2}{\alpha(1+\rho^2)}\right)^{\frac{1-\gamma}{1+\gamma}}\right)^{\frac{\gamma}{1-\gamma}} \tag{2.61}$$

$$b_M^* = (1-\alpha)\left(\frac{\rho^{1-\gamma}}{1+\rho^2}\right)^{\frac{1}{1+\gamma}}\left(\alpha\left(\frac{\alpha\mu^2}{(1-\alpha)(1+\mu^2)}\right)^{\frac{1-\gamma}{1+\gamma}} + (1-\alpha)\left(\frac{\rho^2}{1+\rho^2}\right)^{\frac{1-\gamma}{1+\gamma}}\right)^{\frac{\gamma}{1-\gamma}} \tag{2.62}$$

Note that within the mutual-monitoring regime, both b_P^* and b_M^* are *conditionally* independent of β. This should not be a surprise since mutual-monitoring makes both members' (monitored) work efforts contractible and under the compensation rules defined by Assumption 4, two individual maximizations will result in a collective maximization solution. In fact a mutual-monitoring regime is always available in the sense that monitoring can make all working efforts contractible.

Comparing (2.61)–(2.62) with (2.55)–(2.56), and (2.57)–(2.58) respectively, we find that for any $\mu < \infty$ and $\rho < \infty$, the work incentives within the mutual-monitoring regime (if there is such a regime) are strictly less than at both $\beta = 0$ and $\beta = 1$:

$$a_P^* = a_P^b(b_M^*)\Big|_{\beta\in[\beta_\rho,\beta_\mu]} < a_P^* = a_P^{s*}\Big|_{\beta=0}$$

$$a_M^* = a_M^b(b_P^*)\Big|_{\beta\in[\beta_\rho,\beta_\mu]} < a_M^* = a_M^{b*}\Big|_{\beta=0}$$

$$a_P^* = a_P^b(b_M^*)\Big|_{\beta\in[\beta_\rho,\beta_\mu]} < a_P^* = a_P^{b*}\Big|_{\beta=1}$$

$$a_M^* = a_M^b(b_P^*)\Big|_{\beta\in[\beta_\rho,\beta_\mu]} < a_M^* = a_M^{s*}\Big|_{\beta=1}$$

This confirms that the mutual-monitoring regime cannot include $\beta = 0$ and $\beta = 1$.

Comparing (2.61)–(2.62) with (2.59)–(2.60), we find that if $\mu\rho \le 1$, the self-incentives are always greater than the monitored-incentives (if monitoring is imposed) for $\beta \in [\frac{\mu^2}{1+\mu^2}, \frac{1}{1+\rho^2}]$:

$$a_P^{s*} \ge a_P^b = \rho b_M^* \text{ and } a_M^{s*} \ge a_M^b = \mu b_P^*$$

[33] A mathematical problem with the mutual-monitoring is that the two compensation rules cannot be simultaneously binding. One way to deal with this problem is to assume the two members play a non-cooperative monitoring game; that is, each member chooses his own monitoring effort subject to the compensation rule for the other. Alternatively we can assume that the two members play a cooperative monitoring game; that is, they maximize the joint output net of the total effort costs. The reader can check to confirm the two methods give the same solution.

where the strict inequalities hold if $\mu\rho = 1$. This confirms that the mutual-monitoring cannot occur if $\mu\rho \leq 1$. On the other hand, if $\mu\rho > 1$, the monitored-incentives (if imposed) will be greater than the self-incentives for $\beta \in [\frac{1}{1+\rho^2}, \frac{\mu^2}{1+\mu^2}]$:

$$a_P^{s*} < a_P^b = \rho b_M^* \text{ and } a_M^{s*} < a_M^b = \mu b_P^*$$

This confirms that if $\left(\frac{\mu^2}{1+\mu^2} + \frac{\rho^2}{1+\rho^2}\right)$ is sufficiently larger than one, the mutual-monitoring regime will occur.

A remarkable result is that when $\mu \to \infty$ and (or) $\rho \to \infty$, that is, when working effort is perfectly observable (monitoring is perfect), the equilibrium solution converges to the first best solution, regardless of β. To see this, note that at the first best equilibrium, $\frac{\partial Y}{\partial a_P} = \frac{\partial C_P}{\partial a_P}$ and $\frac{\partial Y}{\partial a_M} = \frac{\partial C_M}{\partial a_M}$, which give the solutions:

$$a_P^{FB} = (1-a)\left[(1-\alpha) + \alpha\left(\frac{\alpha}{1-\alpha}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{\gamma}{1-\gamma}} \tag{2.63}$$

$$a_M^{FB} = \alpha\left[\alpha + (1-\alpha)\left(\frac{1-\alpha}{\alpha}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{\gamma}{1-\gamma}} \tag{2.64}$$

where superscript '*FB*' denotes the first best.

It is easy to check that: (i) at $\beta = 0$, $a_P^{s*} \to a_P^{FB}$ and $a_M^* = \mu b_P^* \to a_M^{FB}$ as $\mu \to \infty$; (ii) at $\beta = 1 a_M^{s*} \to a_M^{FB}$ and $a_P^* = \rho b_M^* \to a_P^{FB}$ as $\rho \to \infty$; and (iii) for $\beta \in [\beta_\rho, \beta_\mu]$, $a_P^* = \rho b_M^* \to a_P^{FB}$ and $a_M^* = \mu b_P^* \to a_M^{FB}$ as $\rho \to \infty$ and $\mu \to \infty$. The intuition is that as $\mu \to \infty$ and $\rho \to \infty$, the whole interval is going to be covered by the monitoring regimes while an epsilon monitoring effort can induce sufficient work effort.

Lemma 6 *Assume that utility functions are given by (2.6), the returns function by (2.7), and the monitoring technology by (2.8), where* $\rho > 0$, $\mu > 0$. *Then: (i)* $\beta = 0$ *Pareto dominates all* $\beta \leq \beta_\mu$; *and (ii)* $\beta = 1$ *Pareto dominates all* $\beta \geq \beta_\rho$.

Proof The claims are a corollary of Lemmas 4 and 5. □

Corollary 7 *The mutual-monitoring regime* $[\beta_\rho, \beta_\mu]$ *is Pareto dominated by the one-sided monitoring regimes* $[0, \beta_\rho)$ *and* $(\beta_\mu, 1]$.

Remarks The reason underlying Corollary 7 is that the mutual-monitoring regime incurs two monitoring costs, while one-sided monitoring regimes incur only one monitoring cost. A marginal switch from the mutual-monitoring regime to an one-sided monitoring regime will have little effect on either members' working efforts but will reduce the monitoring cost drastically since one member's monitored-incentive is replaced by self-incentive. This strong result suggests that a residual share contract

might be not preferred in the case of risk neutrality even if it induces both members to monitor each other.[34]

With Lemma 6, identifying the optimal assignment of principalship reduces to comparing the following three contracts: $\beta = 0$, $\beta = 1$ and $\beta = \beta_\Pi$ if $\beta_\mu < \beta_\rho$ (here β_Π is used to denote the residual share which maximizes the total welfare within the monitoring free regime). We show this in two steps. First we characterize the conditions under which $\beta = 1$ Pareto dominates (or is Pareto dominated by) $\beta = 0$. We then analyze the conditions under which $\beta = 1$ ($\beta = 0$) Pareto dominates (or is Pareto dominated by) $\beta = \beta_\Pi$.

Theorem 8 *Assume that utility functions are given by (2.6), the returns function by (2.7), and the monitoring technology by (2.8), where $\rho > 0$ and $\mu > 0$. Then $\beta = 1$ Pareto dominates $\beta = 0$ if and only if the following inequality holds:*

$$\left(1 - \left(\frac{\rho^2}{1+\rho^2}\right)^{\frac{1-\gamma}{1+\gamma}}\right) \leq \left(\frac{\alpha}{1-\alpha}\right)^{\frac{2}{1+\gamma}} \left(1 - \left(\frac{\mu^2}{1+\mu^2}\right)^{\frac{1-\gamma}{1+\gamma}}\right) \tag{2.65}$$

Proof We claim that $\Pi(1) \geq \Pi(0)$ iff the inequality (2.65) holds. Because at $\beta = 0$ and $\beta = 1$, the net surplus of both work and monitoring efforts is fully internalized, M's monitoring produces a greater output if and only if M has a higher *overall* productivity than P. This implies that $\Pi(1) \geq \Pi(0)$ is equivalent to $Y(1) \geq Y(0)$. Therefore we can demonstrate the claim by comparing $Y(1)$ with $Y(0)$.

Substituting (2.55)–(2.58) into Y, we obtain at $\beta = 0$,

$$Y(0) = (1-\alpha)\left[(1-\alpha) + \alpha\left(\frac{\alpha\mu^2}{(1-\alpha)(1+\mu^2)}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{1+\gamma}{1-\gamma}} \tag{2.66}$$

At $\beta = 1$,

$$Y(1) = \alpha\left[\alpha + (1-\alpha)\left(\frac{(1-\alpha)\rho^2}{\alpha(1+\rho^2)}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{1+\gamma}{1-\gamma}} \tag{2.67}$$

Suppose $Y(1) \geq Y(0)$; that is,

$$\alpha\left[\alpha + (1-\alpha)\left(\frac{(1-\alpha)\rho^2}{\alpha(1+\rho^2)}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{1+\gamma}{1-\gamma}} \geq (1-\alpha)\left[(1-\alpha) + \alpha\left(\frac{\alpha\mu^2}{(1-\alpha)(1+\mu^2)}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{1+\gamma}{1-\gamma}} \tag{2.68}$$

[34]The argument may not hold when both members are risk-averse. See Sect. 2.5.

Denoting $\xi = \frac{\alpha}{1-\alpha}$ and rearranging (2.68) gives

$$\xi\left[1+\frac{1}{\xi}\left(\frac{\rho^2}{\xi(1+\rho^2)}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{1+\gamma}{1-\gamma}} \geq \left[\frac{1}{\xi}+\left(\frac{\xi\mu^2}{1+\mu^2}\right)^{\frac{1-\gamma}{1+\gamma}}\right]^{\frac{1+\gamma}{1-\gamma}}$$

or

$$\left(1-\left(\frac{\rho^2}{1+\rho^2}\right)^{\frac{1-\gamma}{1+\gamma}}\right) \leq \xi^{\frac{2}{1+\gamma}}\left(1-\left(\frac{\mu^2}{1+\mu^2}\right)^{\frac{1-\gamma}{1+\gamma}}\right) \tag{2.69}$$

This is what we need. □

Remarks The key point of Theorem 8 is that the dominance of $\beta = 1$ over $\beta = 0$ depends on the interactions between relative importance in production, monitoring technology and the degree of team work. Under the assumption of identical preferences, parameter α and parameters μ and ρ fully discriminate between the marketing member and the producing member. $\alpha \geq (\leq)\frac{1}{2}$ implies that $M(P)$ has an advantage in production; $\rho \geq (\leq)\mu$ implies that $M(P)$ has an advantage in monitoring. Then the following three possibilities may be considered. First, M (or P) has advantages in both production and monitoring; second, M has an advantage in production but P has an advantage in monitoring; and third, P has an advantage in production and M has an advantage in monitoring. In the first case, we say M (or P) has absolute advantages; in the latter two cases, we say M (P) has a relative advantage in productivity or monitoring, but not in both. Then the theorem can be decomposed into the following three corollaries.

Corollary 9 *Assigning principalship to a member with absolute advantages is always preferred to assigning to a member with absolute disadvantages; that is, $\alpha \geq \frac{1}{2}$ and $\rho \geq \mu$ implies that $\beta = 1$ Pareto dominates $\beta = 0$.*

Corollary 9 is quite obvious since the condition (2.65) always holds when $\alpha \geq \frac{1}{2}$ and $\rho \geq \mu$. Two special cases are as follows. If $\alpha = \frac{1}{2}$, (2.65) holds if and only if $\rho \geq \mu$. This implies that when the two members are equally important in production, assigning principalship to M is preferred to P if and only if M enjoys a advantage in monitoring technology. Second, if $\rho = \mu$, (2.65) reduces to $\alpha \geq \frac{1}{2}$. That is, when monitoring is equally effective for both members, M's principalship is preferred if and only if he is more important in production.

However, M's absolute advantages in production and in monitoring are a sufficient but not a necessary condition for M's monitoring to Pareto dominate P's monitoring. A relative advantage may also ensure that M is the principal. Observation of (2.65) suggests

Corollary 10 *For any α and μ, there is a $\rho^*(\alpha, \mu)$ such that $\beta = 1$ Pareto dominates $\beta = 0$ iff $\rho \geq \rho^*$, where $\rho^* \geq (\leq)\mu$ if $\alpha \leq (\geq)\frac{1}{2}$, and $\frac{\partial\rho^*}{\partial\alpha} < 0$ and $\frac{\partial\rho^*}{\partial\mu} > 0$; and for any ρ and μ, there is a $\alpha^*(\rho - \mu)$ such that $\beta = 1$ Pareto dominates $\beta = 0$ iff $\alpha \geq \alpha^*$, where $\alpha^* \geq (\leq)\frac{1}{2}$ if $\rho \leq (\geq)\mu$, and $\frac{\partial\alpha^*}{\partial(\rho-\mu)} < 0$.*

That is, the optimal assignment may require that a less important member be the principal if he enjoys a big enough relative advantage in monitoring, or a less effective monitor to be the principal if his role is dominant enough in production.

Condition (2.65) also suggests that both ρ^* and α^* depend on the degree of teamwork (γ). In particular, we have

Corollary 11 *If two members have different advantages in monitoring and in production, an increase in γ will favour the member with an advantage in monitoring but disfavour the member with an advantage in production.*

In other words, an increase in the degree of teamwork will strengthen the role of monitoring advantage but weaken that of production advantage in determining the optimal assignment of principalship; it is more likely to be optimal to let the less important member monitor the more important member when the degree of teamwork is high than when it is low, given that the first member has advantage in monitoring technology. The intuition is that with a higher degree of teamwork, output is more sensitive to the work effort of the member who is less important in production but has an advantage in monitoring, and less sensitive to the work effort of the member who is the reverse; this implies that the first member's monitoring is more favourable because that incurs less diminution of both members' work incentives. To demonstrate the argument, we compare two special cases: $\gamma = 0$ and $\gamma = 1$.

When $\gamma = 0$ (no teamwork), condition (2.65) reduces to

$$\frac{1+\mu^2}{1+\rho^2} \leq \left(\frac{\alpha}{1-\alpha}\right)^2 \tag{2.70}$$

When $\gamma = 1$ (pure teamwork), condition (2.65) reduces to

$$\left(\frac{\rho^2}{1+\rho^2}\right) \geq \left(\frac{\mu^2}{1+\mu^2}\right)^{\frac{\alpha}{1-\alpha}} \tag{2.71}$$

Suppose that it is technically impossible for P to monitor M but it is possible for M to monitor P, i.e., $\mu = 0$ and $\rho > 0$, and $\alpha < \frac{1}{2}$; that is, M has advantages in monitoring while P has advantages in production. Then, for all $\alpha > 0$, and all $\rho > 0$, $\beta = 1$ is always preferred to $\beta = 0$ when $\gamma = 1$, while this may be not true when $\gamma = 0$. For instance, at $\gamma = 0$, if $\rho = 0.2$, $\Pi(1) \geq \Pi(0)$ iff $\alpha \geq 0.495$; if $\alpha = 0.45$, $\Pi(1) \geq \Pi(0)$ iff $\rho \geq 0.703$ (in the case of $\gamma = 0.5$, $\rho \geq 0.554$); if $\alpha \leq 0.414$, it is impossible to find a $\rho < 1$ such that $\Pi(1) \geq \Pi(0)$. The reason is that $\mu = 0$ implies that when the principalship is assigned to P, monitoring cannot be imposed and M will have no incentive to work at all ($a_M^* = b_P^* = 0$); but M's zero incentive has different consequences when $\gamma = 1$ from when $\gamma = 0$: At $\gamma = 1$, $a_M^* = 0$ makes P's effort useless and therefore P also loses the incentive to work ($a_P^* = 0$); and on the other hand, at $\gamma = 0$, P's marginal productivity will not be (entirely) ruined by $a_M^* = 0$ and therefore P still has incentive to work ($a_P^* > 0$). As a result, although

$\rho > 0$ (given $\mu = 0$) will guarantee that $\beta = 1$ Pareto dominates $\beta = 0$ when $\gamma = 1$, the guarantee may not hold when $\gamma = 0$.[35]

We now turn to analyze the conditions for $\beta = 0$ and $\beta = 1$ to Pareto dominate β_Π.

Theorem 12 *Assume that utility functions are given by (2.6), the returns function by (2.7), and the monitoring technology by (2.8), where $\rho > 0$ and $\mu > 0$. Denote by μ^{**} and ρ^{**} the minimum requirements of monitoring effectiveness such that $\mu \geq \mu^{**}$ implies $\Pi(0) \geq \Pi(\beta_\Pi)$ and $\rho \geq \rho^{**}$ implies $\Pi(1) \geq \Pi(\beta_\Pi)$. Then: (i) $\alpha < \frac{1}{2}$ implies $\mu^{**} < \rho^{**}$ and $\alpha > \frac{1}{2}$ implies $\mu^{**} > \rho^{**}$; (ii) μ^{**} is increasing with α and ρ^{**} is decreasing with α; (iii) μ^{**} is increasing with γ for $\alpha < \frac{1}{2}$ and decreasing with γ for $\alpha > \frac{1}{2}$, and ρ^{**} is decreasing with γ for $\alpha < \frac{1}{2}$ and increasing with γ for $\alpha > \frac{1}{2}$.*

Proof Part (i) can be derived from Corollary 10 since $\Pi(0)\big|_{\mu^{**}} = \Pi(1)\big|_{\rho^{**}}$. To prove Part (ii), note that β_Π is a monotonic increasing function of α with $\beta_\Pi(0) = 0$ and $\beta_\Pi(1) = 1$ (by Theorem 3), and $\mu^{**} \to 0$ and $\rho^{**} \to \infty$ as $\alpha \to 0$, and $\mu^{**} \to \infty$ and $\rho^{**} \to 0$ as $\alpha \to 1$. Part (iii) can be derived from Corollary 11. A less accurate but more intuitive proof is as follows.

$\beta_\Pi \in [0,\beta_\mu]$ and (or) $\beta_\Pi \in [\beta_\rho, 1]$ are sufficient for $\Pi(0) \geq \Pi(\beta_\Pi)$ and $\Pi(1) \geq \Pi(\beta_\Pi)$ respectively. Obviously the possibility of $\beta_\Pi \in [0,\beta_\mu]$ and $[\beta_\rho, 1]$ is increasing with μ and ρ. However, the dominance of $\beta = 0$ $(= 1)$ over β_Π does not require β_Π *belongs* to, but only *sufficiently close to* the regimes $[0,\beta_\mu]$ ($[\beta_\rho, 1]$). Since $\beta_\Pi < (>) \frac{1}{2}$ if $\alpha < (>) \frac{1}{2}$, only a lower μ^{**} (ρ^{**}) is required for β_Π to be as close to β_μ as to β_ρ. Since β_Π contracts to 0 (1) as $\alpha \to 0(1)$, μ^{**} should contract (expand) and ρ^{**} should expand (contract). Since β_Π increases with γ for $\alpha < \frac{1}{2}$ and decreases with γ for $\alpha > \frac{1}{2}$, so does μ^{**} (reverse for ρ^{**}). □

Remarks If monitoring is sufficiently effective so that there is a mutual-monitoring regime, β_Π is always Pareto dominated by either $\beta = 0$ or $\beta = 1$ or both.[36] We now consider a restriction on monitoring technology: $\mu \leq 1$ and $\rho \leq 1$. Such a restriction is interesting both because it means a less-unit transformation from monitoring effort to work effort and because it guarantees a monitoring-free regime (i.e., $\beta_\mu < \beta_\rho$). Under this restriction, we have the following corollary:

Corollary 13 *Assume $\mu \leq 1$ and $\rho \leq 1$. Then: (i) there is a $\mu^{**} \leq 1$ if and only if $\alpha \ll \frac{1}{2}$; and (ii) there is a $\rho^{**} \leq 1$ if and only if $\alpha \gg \frac{1}{2}$, where $\ll$ ($\gg$) denotes "sufficiently smaller (larger) than".*

Remarks A simple calculation shows that at $\alpha = \frac{1}{2}, \Pi\left(\beta_\Pi = \frac{1}{2}\right) = \frac{3}{16}, \Pi(\beta = 0) = \frac{1}{4}\left(\frac{1}{2} + \frac{1}{2}\left(\frac{\mu^2}{1+\mu^2}\right)^{\frac{1-\gamma}{1+\gamma}}\right)^{\frac{1+\gamma}{1-\gamma}}$ and $\Pi(\beta = 1) = \frac{1}{4}\left(\frac{1}{2} + \frac{1}{2}\left(\frac{\rho^2}{1+\rho^2}\right)^{\frac{1-\gamma}{1+\gamma}}\right)^{\frac{1+\gamma}{1-\gamma}}$. It is easy to check that $\mu^{**} = \rho^{**} = 1$ if $\gamma = 0$ and $\mu^{**} = \rho^{**} > 1$ if $\gamma > 0$.

[35] We use $\gamma = 0$ only for comparison. In fact, in the case of $\gamma = 0$, no firm exists.

[36] It is still possible that β_Π does not belong to the mutual-monitoring regime.

The message from Theorem 12 and Corollary 13 is that for a pure-principalship contract to Pareto dominate an optimal residual sharing contract, the monitoring technology must be sufficiently effective; in other words, impossibility of monitoring in a technical sense is not the only rationale for a residual sharing contract to be optimal, since a technically possible but economically less effective monitoring technology can also call for a residual sharing contract. Such an argument seems a common sense but has not, to our knowledge, been formally modeled previously. Furthermore, a residual sharing contract is more likely to be preferred when the production advantage and the monitoring advantage are held separately by the two members than when they are held simultaneously by one member.

2.4 Discussion: Classical Capitalist, Partnership and Alchian–Demsetz Firms

In the previous section, we analyzed in the abstract how the optimal assignment of principalship depends on the interaction between the relative importance of each member, the effectiveness of monitoring technology, and the degree of teamwork. The conclusions are that: (i) A residual sharing contract is more likely to be preferred when monitoring is technically impossible or less effective, and this proposition is strengthened by a high degree of teamwork. (ii) A pure-principalship contract is more likely to be preferred when a single member possesses advantages in both production and monitoring technology. (iii) The minimum requirement of effectiveness of monitoring for a pure-principalship contract to Pareto dominate a residual sharing contract is decreasing with one member's relative importance, but may be increasing or decreasing with the degree of teamwork depending on whether he has an advantage or disadvantage in production. We now apply these results to the three commonly observed forms of firm: the classic capitalist firm, partnership firm, and a theoretically-created form of firm, the Alchian–Demsetz firm.

2.4.1 The Capitalist Firm

In this thesis our aim is to characterize the contractual arrangements of the capitalist firm. For this purpose, we start with the classical form in which the marketing member is the principal who claims the residual return and possesses monitoring authority while the producing member is an agent who receives a fixed payment and is monitored by the marketing member. In the conventional terminology, the former is called "the entrepreneur" and the latter "workers". The present model legitimizes the positional transformations from "marketing member" to "entrepreneur" and from "producing member" to "workers", under the following two assumptions:

Assumption A: Marketing-Dominated Uncertainty (MDU): The distribution function $\Phi(Y; a_M, a_P)$ is more sensitive to the marketing actions a_M than to the producing actions a_P. In the parameterized model developed in the previous section, this implies that $\alpha > \frac{1}{2}$; that is, the marketing member has an advantage in production.

Assumption B: Asymmetry of Monitoring Technology (AMT): The marketing actions are more difficult and therefore more costly to monitor than the producing actions. In the parameterized model, this implies that $\mu < \rho$; that is, the marketing member also holds an advantage in monitoring.

These two assumptions seem quite realistic. First, as we pointed out in Sect. 2.1, marketing activities originate from uncertainty and the main function of the marketing member is to deal with uncertainty. In contrast, although producing activities are also conducted under uncertainty, the degree of uncertainty involved is much lower. Taking into account that in our model one marketing member is matched by one producing member while in reality one marketing member is matched by more than one producing member, it is reasonable to assume that $\alpha > \frac{1}{2}$. Second, because the marketing member decides what to do and how to do it, and his activities are inventive and creative, while the producing member implements the decisions made by the marketing member by transforming input into output physically and his activities are almost routine, it is reasonable to assume that it is more effective for the marketing member to monitor the producing member than otherwise. After all, a glance at the producing member will reveal whether he is working, while a stare at the marketing member may tell us little about what he is thinking.

These two assumptions together with teamwork ensure that assigning principalship to the marketing member is preferred to assignment to the producing member. In order for M's principalship to Pareto-dominate the 'optimal' residual sharing contract (β_Π), there is the further requirement that monitoring of M over P is sufficiently effective: $\rho \geq \rho^{**}$. We assume that this condition is indeed satisfied in reality. By Theorem 12, we know that for $\alpha \gg \frac{1}{2}$, $\rho^{**} \leq 1$ and $\mu^{**} > 1$. Based on Assumptions A and B, we conjecture that $(\beta = 1) \succ (\beta = \beta_\Pi) \succ (\beta = 0)$. That is, an "optimal" residual sharing contract is Pareto dominated by M's principalship but Pareto dominates P's principalship.

Note that although we have been concerned so far only with risk-neutral preferences, uncertainty also plays a crucial role in explaining the contractual arrangements of the capitalist firm. Without uncertainty, there would be no need for a marketing member and all activities would be reduced to producing; without uncertainty, there would be no reason to assume that the marketing member is more difficult to monitor than the producing member.

The Worker-in-the-Dark and the Worker-in-the-Light: Our major argument for the capitalist firm can be sharpened by the following example. Suppose that there is a working team of two persons A and B. They work only at night when there is moonlight. The production technology requires that Person A works in the light while Person B works in the shadow. The output cannot be attributed to each individual's marginal effort. Person A cannot see whether Person B works hard or is lazy while Person B can see how hard Person A works. In this case, Person A has more incentive to let Person B be residual claimant even than Person B has. The suggestion made by

A is that "I (Person A) volunteer to accept your (Person B) authority of monitoring my behaviour if you are willing to pay me such-and-such fixed sum from our joint output." A contract is born: Person B becomes the principal and Person A becomes the agent. Obviously if Person B volunteers to be the agent, Person A will reject such an offer, because he knows it is impossible to monitor B. In the context of the firm, the entrepreneur is the worker-in-the-dark, whereas the workers are in the light. Marketing activities can be done only in the shadow while producing activities take place in the light. Few people can know what the entrepreneur is doing while it is quite easy to see how hard the workers work.

2.4.2 Partnership Firm

Although our major concern is about the capitalist firm, the results are quite general in terms of their power to explain non-capitalist firms. One such kind of firm is a partnership which is characterized by a symmetric residual sharing contract between its members.

The model predicts that partnership is more likely to be preferred in a firm with the following characteristics: (i) members are equally important in production: $\alpha = \frac{1}{2}$; (ii) members are equally difficult to monitor: $\mu = \rho < \mu^{**} = \rho^{**}$; (iii) the degree of teamwork is high. In terms of our definitions of marketing and producing, partnership might be optimal in a firm consisting of two marketing members (two workers-in-the dark).

In reality, partnership is common in industries such as law, accountancy, consultancy or academic research. This observation is quite consistent with the predictions from the model. In all these industries, "marketing" is the main work for all members and "producing" is trivial; the outcome depends on efficient uses of all members' intelligence much more than on the hours they spend in the office. This makes monitoring difficult. As a result, a partnership provides higher *overall* incentives than a principal-agent contract.[37]

2.4.3 Alchian–Demsetz Firm

An Alchian–Demsetz firm is characterized by $\alpha = \frac{1}{2}$ and $\mu = \rho \geq \mu^{**} = \rho^{**}$. That is: (i) the firm's members are identical in production; (ii) monitoring is symmetrically effective. The original example given by Alchian and Demsetz (1972) is where two persons jointly lift heavy cargo into trucks. An Alchian–Demsetz firm consists of

[37] When should a professor treat her research assistant as a co-author or only acknowledge him in a footnote? Observation is that when the research requires assistance from brains, the research assistant appears as a co-author; on the other hand, when the research requires assistance mainly from hands (collecting and calculating data), the research assistant will be "gratefully acknowledged".

two producing members or two workers-in-the-light. For such a firm, it is important to assign either of them to monitor the other, but it does not matter who monitors whom.

Alchian and Demsetz attempted to provide the rationale for the capitalist firm by analyzing the incentive problem associated with teamwork. They correctly show that monitoring may be more efficient than the residual share in providing incentives in presence of team production. However, because of their failure to distinguish between different firm members, they cannot answer who should be the monitor. Furthermore, in their model, the monitor alienates himself from the team by specializing in monitoring; that is, the monitor is no longer a team member and has no direct contribution to the return of the firm. This "outside" principal assumption has become a standard starting point in most agency models. But once this stage is reached, legitimacy of monitoring itself is in doubt. For instance, Harris and Holmstrom (1982) argues that the principal's primary role is not essentially one of monitoring, but instead is to break the budget-balancing constraint so that group incentives can work well; McAfee and McMillan (1991) argue that monitoring is not needed to prevent shirking by the team members and a principal can do as well when he observes only total output as when he observes the individual contributions. My argument is that the assumption of an outside principal should not be a starting point in modelling the capitalist firm. Instead we seek to identify a principal within the firm.

2.5 Risk-Attitudes and the Assignment of Principalship

To focus on the incentive problem, it has been assumed so far that both the marketing member and the producing member are risk-neutral. We have shown that the marketing member should be the principal because he has advantages in both production and monitoring. In this section, we relax the risk-neutral assumption and discuss how risk-attitudes may affect the optimal assignment of principalship through interaction with the relative importance and the effectiveness of monitoring technology. The discussion is informal. Our purpose is to show that although risk-aversion may have some impact on the optimal assignment of principalship at margin, risk-neutrality is not a necessary condition for the marketing member to be the principal. In particular, even if the marketing member is more risk-averse, his advantages in production and monitoring may be so dominant that assigning principalship to him is still preferred. This is particularly true when the variance of the return is dependent on actions. In general, given that we have no sound reason to assume which member is more risk-averse, our previous propositions can hold.

The simplest way to tackle the problem is to use the *value maximization principle*, which assumes that individuals care only for the *certain equivalent* which is equal to the expected return minus the risk premium associated with the random return (as well as the cost of effort if the effort is valuable), where the *risk-premium* is defined as one-half times the coefficient of absolute risk aversion times the variance of the return. According to the *value maximization principle*, an arrangement is optimal

(efficient) if and only if it maximizes the total certain equivalent of all the parties involved. To legitimize this approach, we assume that (i) the variance is not too large relative to the individual's risk aversion, (ii) the coefficient of absolute risk aversion is independent of the expected return, and (iii) the cost of the effort is equivalent to its monetary value.[38] These are strong and often unrealistic assumptions, but they greatly simplify the analysis.

Denote by $\overline{Y}$ the expected return, V the variance of Y, R_i the coefficients of absolute risk aversion, CE_i the certain equivalent, $i = M, P$. Then, the producing member's risk premium $= \frac{1}{2}R_P(1-\beta)^2V$; the marketing member's risk premium $= \frac{1}{2}R_M\beta^2V$; and the total risk premium $= \frac{1}{2}\left(R_M\beta^2 + R_P(1-\beta)^2\right)V$. The certain equivalents are respectively

$$CE_P = \beta w_P + (1-\beta)\left(\overline{Y} - w_M\right) - \frac{1}{2}R_P(1-\beta)^2V - C(a_P, b_P) \tag{2.72}$$

$$CE_M = (1-\beta)w_M + \beta(\overline{Y} - w_P) - \frac{1}{2}R_M\beta^2V - C(a_M, b_M) \tag{2.73}$$

$$CE = \overline{Y} - \frac{1}{2}\left(R_M\beta^2 + R_P(1-\beta)^2\right)V - C(a_P, b_P) - C(a_M, b_M) \tag{2.74}$$

A contract is optimal if it maximizes (2.74), given that each member maximizes his own certain equivalent subject to the monitoring technology and the compensation rules governing the acceptability of monitoring defined as follows:

$$w_P \geq \begin{cases} w_P^s & \text{if } b_M = 0 \\ w_P^s + \frac{1}{\beta}\left(C_P^b - C_P^s\right) - \frac{1-\beta}{\beta}\left(\overline{Y}^b - \overline{Y}^s\right) + \frac{1}{2}R_P\frac{(1-\beta)^2}{\beta}\left(V^b - V^s\right) & \text{if } b_M > 0 \end{cases} \tag{2.75}$$

$$w_M \geq \begin{cases} w_M^s & \text{if } b_P = 0 \\ w_M^s + \frac{1}{1-\beta}\left(C_M^b - C_M^s\right) - \frac{\beta}{1-\beta}\left(\overline{Y}^b - \overline{Y}^s\right) + \frac{1}{2}R_M\frac{\beta^2}{1-\beta}\left(V^b - V^s\right) & \text{if } b_P > 0 \end{cases} \tag{2.76}$$

where, as before, superscript b and s denote the values respectively when monitoring is imposed and when it is not imposed.

As in the previous sections, we shall assume that the two members are identical in the cost function of efforts. But they are allowed to differ in their risk-aversion degrees. To parallel the previous analysis, we will say member i has advantages in risk-tolerance (less risk averse) if $R_i \leq R_j$.

[38] The exponential form of utility function is one under which the value maximization principle (or the certain equivalent approach) is perfectly satisfactory. See Holmstrom and Milgrom (1991). For more discussion of this approach, see Milgrom and Roberts (1992), Chap. 7.

In the case of risk-neutrality, the relative importance of each member in contributing to the return of the firm was identified with the parameter of the effort elasticity of the expected return. In the case of risk-aversion, we also need to consider the effects of effort on the variance, if any. We shall assume that the ranking of relative importance in terms of the effects on the expected return is the same as the ranking of relative importance in terms of the effects on the variance, unless the variance is constant; that is, the elasticity of the variance of returns to i's effort cannot be smaller than to j's if the elasticity of the expected return to i's effort is greater than to j's. This seems quite reasonable. Then the implication of risk-aversion for the optimal assignment of principalship does not only depend on the relative tolerance of risk, but also on the relationship between the variance and effort. The analysis can be conducted with different cases

Case 1: The variance is independent of action:

It is easy to show that when the variance is independent of action, that is, $V^b = V^s \equiv V^0$, the compensation rule is independent of the variance and the risk attitudes, and the optimal choices of efforts are also independent of the variance and the risk-attitudes. As a result, the risk premium can be separately calculated and its effect on the optimal contract is additive. The optimal contract is simply to maximize

$$CE = \Pi \left|_{R_P=R_M=0} - \frac{1}{2}\left(R_M\beta^2 + R_P(1-\beta)^2\right)V^0 \right. \tag{2.77}$$

We have already shown how $\Pi \left|_{R_P=R_M=0} \right.$ changes with β for a given $(\alpha, \gamma; \mu, \rho)$. It is easy to show that the risk-cost is a convex function of β with the minimum value at $\beta = \frac{R_P}{R_P+R_M}$. Assume that $\alpha \geq \frac{1}{2}$ and $\rho \geq \mu$. Then the implication of risk-attitudes for the optimal assignment of principalship can be summarized as follows:

(1) If the marketing member is risk-neutral ($R_M = 0$) and the producing member is risk-averse ($R_P > 0$), the risk cost is minimized (zero) at $\beta = 1$. Then assigning principalship to the marketing member is always preferred since the marketing member has advantages in all of production, monitoring and risk-tolerance;

(2) If the producing member is risk-neutral ($R_P = 0$) and the marketing member is risk-averse ($R_M > 0$), the risk cost is minimized to zero at $\beta = 0$. This producing member's advantage in risk-tolerance offsets some effect of the marketing member's advantages in productivity and monitoring. However, assigning principalship to the marketing member may be still preferred to assigning to the producing member, if the former's advantages in production and monitoring are sufficiently strong such that the total certain equivalent is greater at $\beta = 1$ than at $\beta = 0$;

(3) If the marketing member and the producing member are identical in degrees of risk-aversion with $R_P = R_M > 0$, the risk cost is minimized at $\beta = \frac{1}{2}$. This offsets some of the marketing member's advantages in production and monitoring, and of the producing member's disadvantages. But assigning principalship to the producing member cannot be optimal;

(4) In both cases (2) and (3), and in general for $R_M > 0$, a mutual-monitoring partnership might be preferred to either member's principalship, if the marketing

member is sufficiently risk-averse. Note that although we have shown that in the risk-neutral case, mutual-monitoring cannot be optimal, we now argue that risk-aversion may justify mutual-monitoring.[39]

Case 2: The variance is dependent on actions:

In this case, an increase in work effort not only increases the expected return ($\overline{Y}$) but also decreases the variance (V)[40]; the residual share affects the total risk cost not only through its direct effect on the risk premium for given variance but also through its indirect effect on variance *per se*. Because of this, each member's optimal work effort (and monitoring effort) for any given residual share depends also on their degrees of risk-aversion, and therefore the incentive effect and the risk cost effect of the residual share are not as additive as in case 1.

It is not easy to analyze the overall relationship between the total certain equivalent and the residual share as defined by (2.74) when the variance is dependent on actions. Fortunately, a focus on the risk cost alone can generate insights.

Differentiating the total risk cost (TRC) with respect to the residual share β, we obtain

$$\frac{\partial TRC}{\partial \beta} = \frac{1}{2}\left(2\beta R_M - 2(1-\beta)R_P\right)V + \frac{1}{2}(\beta^2 R_M + (1-\beta)^2 R_P)\frac{dV}{d\beta} \tag{2.78}$$

The first term of (2.78) is the direct effect of a marginal changes in β on the total risk cost, and the second term is the indirect effect which would disappear when $\frac{dV}{d\beta} = 0$. It is clear that $\beta = \frac{R_P}{R_P+R_M}$ cannot be a point which minimizes the total risk cost as long as $\frac{dV}{d\beta} \neq 0$ (unless $R_P = 0$ or $R_M = 0$). $\frac{dV}{d\beta} \neq 0$ is equivalent to $\frac{\partial V}{\partial a_i} \neq 0$, $i = M, P$, since

$$\frac{dV}{d\beta} = \frac{\partial V}{\partial a_M}\frac{\partial a_M}{\partial \beta} + \frac{\partial V}{\partial a_P}\frac{\partial a_P}{\partial \beta} \tag{2.79}$$

Denote by β_{TRC} the residual share which minimizes TRC. Then at $\beta = \frac{R_P}{R_P+R_M}$, if $\frac{dV}{d\beta} > 0$, $\beta_{TRC} < \frac{R_P}{R_P+R_M}$; if $\frac{dV}{d\beta} < 0$, $\beta_{TRC} > \frac{R_P}{R_P+R_M}$. Particularly, assume that $R_P = R_M > 0$ and $\beta = \frac{R_P}{R_P+R_M} = \frac{1}{2}$ belongs to the monitoring-free regime. Our assumptions about the relative importance of each member in production implies that the (negative) first term of (2.79) dominates the (positive) second term such that $\frac{dV}{d\beta} < 0$, and therefore $\beta_{TRC} > \frac{1}{2}$. More generally, if $\beta = \frac{R_P}{R_P+R_M}$ belongs to M-monitoring regime, $\frac{dV}{d\beta} < 0$ and therefore $\beta_{TRC} > \frac{R_P}{R_P+R_M}$; if $\beta = \frac{R_P}{R_P+R_M}$ belongs to P-monitoring regime, $\frac{dV}{d\beta} > 0$ and therefore $\beta_{TRC} < \frac{R_P}{R_P+R_M}$.

[39]This reflects a difference of mode of thinking between our model and the standard principal-agent model. In the standard principal-agent model, the "agent" shares the residual because the principal cannot observes his actions, while in our model, the "agent" shares the residual because the "principal" is too risk-averse.

[40]Another possibility is that an increase in work effort increase both the mean and the variance subject to the condition that the first-order stochastic dominance still holds.

Combination of the incentive problem with the risk cost problem implies that, compared to case 1 where the variance is independent of action, in case 2 the marketing member's advantages in production and monitoring are more likely to play a dominant role in determining the optimal residual share. Assigning principalship to the marketing member might be strictly preferred even if the marketing member is risk averse and the producing member is risk neutral, if the mean and the variance are sufficiently sensitive to the marketing member's actions and the marketing member's monitoring is sufficiently effective.[41] The argument can be shown by the following example.

An Example: Optimal assignment of principalship when the marketing member has advantages in production and monitoring, while the producing member has advantage in risk-tolerance.

Assume that the (mean) returns function is given by the Cobb–Douglas form of (2.7), monitoring technology by (2.8), and the cost of effort function by $0.5a_i^2 + 0.5b_i^2$. Assume that the producing member is risk-neutral ($R_P = 0$) and the marketing member is risk-averse ($R_M > 0$).

(1) The variance is independent of actions: $V \equiv V^0$.

(1.a) Principalship is assigned to the producing member: $\beta = 0$.

The producing member's problem is

$$\underset{a_P, b_P}{\text{Max}} (\mu b_P)^{\alpha} a_P^{1-\alpha} - 0.5a_P^2 - 0.5(1+\mu^2)b_P^2 \tag{2.80}$$

The optimal work effort and monitoring effort are respectively

$$a_P^* = \alpha^{\frac{\alpha}{2}} (1-\alpha)^{\frac{2-\alpha}{2}} \left(\frac{\mu^2}{1+\mu^2}\right)^{\frac{\alpha}{2}} \tag{2.81}$$

$$b_P^* = \alpha^{\frac{1+\alpha}{2}} (1-\alpha)^{\frac{1+\alpha}{2}} \frac{\mu^{\alpha}}{(1+\mu^2)^{\frac{1+\alpha}{2}}} \tag{2.82}$$

The mean return of the firm is

$$\overline{Y} = \alpha^{\alpha}(1-\alpha)^{1-\alpha} \left(\frac{\mu^2}{1+\mu^2}\right)^{\alpha} \tag{2.83}$$

The total certain equivalent is

$$\begin{aligned} CE &= \overline{Y} - \tfrac{1}{2}(\beta^2 R_M V + (1-\beta)^2 R_P V - C_P(a_P, b_P) - C_M(a_M, b_M) \\ &= \overline{Y} - C_P(a_P, b_P) - C_M(a_M, b_M) \\ &= 0.5\alpha^{\alpha}(1-\alpha)^{1-\alpha} \left(\tfrac{\mu^2}{1+\mu^2}\right)^{\alpha} \end{aligned} \tag{2.84}$$

[41] Like in case 1, mutual-monitoring may be efficient if the marketing member is sufficiently risk-averse.

(1.b) Principalship is assigned to the marketing member: $\beta = 1$.
The marketing member's problem is

$$\underset{a_M, b_M}{\text{Max}}\ a_M^{\alpha}(\rho b_M)^{1-\alpha} - 0.5a_M^2 - 0.5(1+\rho^2)b_M^2 \tag{2.85}$$

The optimal work effort and monitoring effort are respectively

$$a_M^* = \alpha^{\frac{1+\alpha}{2}}(1-\alpha)^{\frac{1-\alpha}{2}}\left(\frac{\rho^2}{1+\rho^2}\right)^{\frac{1-\alpha}{2}} \tag{2.86}$$

$$b_M^* = \alpha^{\frac{\alpha}{2}}(1-\alpha)^{\frac{2-\alpha}{2}}\frac{\rho^{1-\alpha}}{(1+\rho^2)^{\frac{2-\alpha}{2}}} \tag{2.87}$$

The mean return of the firm is

$$\alpha^{\alpha}(1-\alpha)^{1-\alpha}\left(\frac{\rho^2}{1+\rho^2}\right)^{1-\alpha} \tag{2.88}$$

The total certain equivalent is

$$\begin{aligned} CE &= \overline{Y} - \tfrac{1}{2}(\beta^2 R_M V + (1-\beta)^2 R_P V - C_P(a_P, b_P) - C_M(a_M, b_M) \\ &= \overline{Y} - \tfrac{1}{2}R_M V - C_P(a_P, b_P) - C_M(a_M, b_M) \\ &= 0.5\alpha^{\alpha}(1-\alpha)^{1-\alpha}\left(\tfrac{\rho^2}{1+\rho^2}\right)^{1-\alpha} - \tfrac{1}{2}R_M V \end{aligned} \tag{2.89}$$

Comparing (2.89) with (2.84), we see that $\beta = 1$ is preferred to $\beta = 0$ iff the following condition holds:

$$\alpha^{\alpha}(1-\alpha)^{1-\alpha}\left(\left(\frac{\rho^2}{1+\rho^2}\right)^{1-\alpha} - \left(\frac{\mu^2}{1+\mu^2}\right)^{\alpha}\right) \geq R_M V \tag{2.90}$$

Let $\alpha = 0.8$, $\mu = 0.4$, $\rho = 0.8$, and $V = 3$. Then $\beta = 1$ is preferred to $\beta = 0$ as long as $R_M \leq 0.126$.

(2) The variance is independent of producing actions but dependent on marketing actions: $V = V^0 - a_M^{\frac{2\alpha}{1+\alpha}}$.[42]

(2.a) Principalship is assigned to the producing member: $\beta = 0$.

This is exactly the same as (1.a) since the total risk cost is equal to zero and the producing member does need to take into account the effect on the variance when choosing monitoring effort.

(2.b) Principalship is assigned to the marketing member: $\beta = 1$.

The marketing member's problem is

[42] We choose this particular form for simplifying the calculation. V^0 is chosen such that V will not become negative for the relevant range of effort. Note V is convex in a_M.

$$\underset{a_M, b_M}{\text{Max}}\ a_M^{\alpha}(\rho b_M)^{1-\alpha} - \frac{1}{2}R_M(V^0 - a_M^{\frac{2\alpha}{1+\alpha}}) - 0.5a_M^2 - 0.5(1+\rho^2)b_M^2 \tag{2.91}$$

The optimal work effort and monitoring effort are respectively

$$a_M^* = \left(\alpha(1-\alpha)^{\frac{1-\alpha}{1+\alpha}}\left(\frac{\rho^2}{1+\rho^2}\right)^{\frac{1-\alpha}{1+\alpha}} + \frac{\alpha}{1+\alpha}R_M\right)^{\frac{1+\alpha}{2}} \tag{2.92}$$

$$b_M^* = (1-\alpha)^{\frac{1}{1+\alpha}}\left(\frac{\rho^{1-\alpha}}{1+\rho^2}\right)^{\frac{1}{1+\alpha}}\left(\alpha(1-\alpha)^{\frac{1-\alpha}{1+\alpha}}\left(\frac{\rho^2}{1+\rho^2}\right)^{\frac{1-\alpha}{1+\alpha}} + \frac{\alpha}{1+\alpha}R_M\right)^{\frac{\alpha}{2}} \tag{2.93}$$

We see that the optimal efforts are increasing with the degree of risk aversion. The mean return of the firm is

$$\overline{Y} = (1-\alpha)^{\frac{1-\alpha}{1+\alpha}}\left(\frac{\rho^2}{1+\rho^2}\right)^{\frac{1-\alpha}{1+\alpha}}\left(\alpha(1-\alpha)^{\frac{1-\alpha}{1+\alpha}}\left(\frac{\rho^2}{1+\rho^2}\right)^{\frac{1-\alpha}{1+\alpha}} + \frac{\alpha}{1+\alpha}R_M\right)^{\alpha} \tag{2.94}$$

The total certain equivalent is

$$\begin{aligned} CE &= \overline{Y} - \tfrac{1}{2}(\beta^2 R_M V + (1-\beta)^2 R_P V - C_P(a_P, b_P) - C_M(a_M, b_M) \\ &= \overline{Y} - \tfrac{1}{2}R_M(V^0 - a_M^{\frac{2\alpha}{1+\alpha}}) - C_P(a_P, b_P) - C_M(a_M, b_M) \\ &= 0.5\left(\alpha(1-\alpha)^{\frac{1-\alpha}{1+\alpha}}\left(\tfrac{\rho^2}{1+\rho^2}\right)^{\frac{1-\alpha}{1+\alpha}} + \tfrac{\alpha}{1+\alpha}R_M\right)^{\alpha} \\ &\quad \times \left((1-\alpha)^{\frac{1-\alpha}{1+\alpha}}\left(\tfrac{\rho^2}{1+\rho^2}\right)^{\frac{1-\alpha}{1+\alpha}} + \tfrac{\alpha}{1+\alpha}R_M + R_M\right) - \tfrac{1}{2}R_M V^0 \end{aligned} \tag{2.95}$$

$\beta = 1$ is preferred to $\beta = 0$ iff (2.95) is greater than (2.84). For example, if $\alpha = 0.8$, $\mu = 0.4$, $\rho = 0.8$, and $V^0 = 3$. Then $\beta = 1$ is preferred to $\beta = 0$ as long as $R_M \leq 0.233 (\leq 0.126$ for the case where $V \equiv V^0 = 3)$. Figure 2.3 shows comparison between (1.a), (1.b) and (2.b).

2.6 Conclusion

In this chapter we have discussed the optimal assignment of principalship within a firm consisting of a marketing member and a producing member. We have argued that differences in marketing ability among individuals are the origin of the firm and have identified the relative importance and the effectiveness of monitoring as the two key elements determining the optimal assignment of principalship between the marketing member and the producing member. It has been shown that assigning principalship to the marketing member is optimal because marketing activities dominate uncertainty,

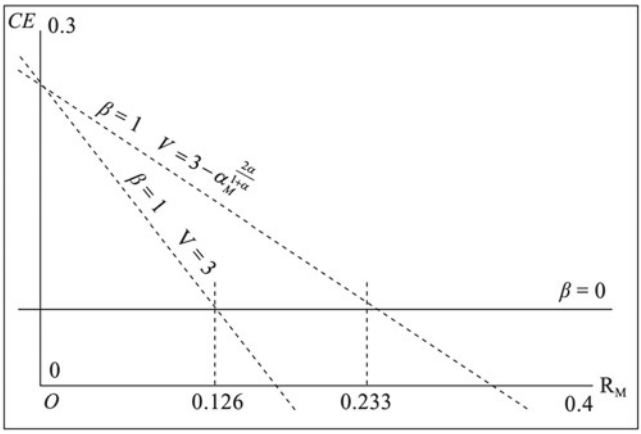

Fig. 2.3 The optimal assignment of principalship when the marketing member has advantages in production and monitoring while the producing member has advantage in risk-tolerance

and because the marketing member's behaviour is more difficult to monitor. This provides a rationale for the asymmetric relationship within the firm between the entrepreneur and workers; that is, the former holds authority over the latter and the latter agree to obey that authority within some limits. We have also analyzed the problem of risk-aversion. However, our basic argument is that although risk-aversion may have some impact on the optimal assignment of principalship at margin, advantages in risk-bearing are not a necessary condition for the marketing member to be the principal. In particular, even if the marketing member is more risk-averse than the producing member, his advantages in production and monitoring may be so dominant that assignment of principalship to him is still preferred. Given that we have no sound reason to assume that either member is more risk-averse, it is more relevant to argue that the marketing member obtains the entrepreneurial status by bearing risk because he is the major "risk-maker" and because his actions are the most difficult to monitor, and not necessarily because he is less risk-averse.[43] Our next task is to analyze the intrinsic relationship between capitalists and entrepreneurs. This is the task of Chap. 3.

[43]It should be pointed out that although our formal propositions are derived under some special technical assumptions about production function, preferences and monitoring technology, the basic arguments should hold under more general assumptions, because what is important is the extent to which external costs of a action are internalized, rather than concrete parameters.

Chapter 3
Marketing Ability, Personal Wealth, and Capital-Hiring-Labour

3.1 Introduction

The purpose of Chap. 2 was to demonstrate the optimal assignment of principalship between the marketing member and the producing member. It has been shown that assigning principalship to the marketing member is optimal because marketing activities dominate uncertainty, and because the marketing member's behaviour is more difficult to monitor. This provides a rationale for the asymmetric relationship within the firm between the entrepreneur (or management) and workers; that is, the former holds authority over the latter and the latter agree to obey that authority within some limits. However, our story cannot stop here. In reality and historically, entrepreneurship has been inherently related to capitalists: an individual of high marketing ability can become an entrepreneur if and only if he is also a capitalist. In particular, although the person who is rich in marketing ability but poor in personal wealth may undertake marketing as a manager of the firm, he can do so, in general, only as the agent of capitalists who are principals. In other words, what we find is not that one kind of labour hires another kind of labour, but that capital hires labour. For our theory to be complete, we need to show: Why is the choice of the one who is to become an entrepreneur constrained by his personal wealth? Why is the authority to select the management assigned to capitalists? Is such a "capital-hiring-labour" system *socially* optimal? In this chapter attempts will be made to answer such questions.

Our basic argument is that priority in being the entrepreneur or in having authority to select the management is given to capitalists because the entrepreneurial choice of the rich is more informative than that of the poor in signalling marketing ability. The underlying assumption in this chapter is that observing a person's marketing ability is much more difficult and much more costly than observing his personal wealth. To focus on the functional asymmetry between the marketing member and the producing member, in Chap. 2 we assumed that each individual's marketing ability was known to all others as well as himself, and therefore the marketing function was exercised by those most qualified for it. In reality, marketing ability is at most only partially observable. Although some information such as education, background,

W. Zhang, *The Origin of the Capitalist Firm*, DOI 10.1007/978-981-10-0221-2_3

work experience, may be available, a person's marketing ability cannot be accurately judged until he has been marketing for some years. What he says about his own ability may not be very useful unless convincing evidence is produced. In contrast, personal wealth is easy to observe and to reveal. It is almost impossible for the poor to pose as the rich; similarly, it is difficult and very costly (if not impossible) for the rich to evade their responsibility (e.g., for paying debts) by hiding personal wealth. Given that marketing ability is less observable than personal wealth, we show that free entry in the entrepreneurial market will make capitalists the winners of competition for entrepreneurship, and that capital-hiring-labour is *socially* desirable because only such a mechanism can guarantee that marketing work will be allocated to the qualified candidates. In contrast, if labour hires capital, the market for entrepreneurs would be full of lemons; that is, there would be too many unqualified people claiming that they can do marketing.

The assumption of non-negative consumption is crucial for the results. The intuition is as follows. Because of the non-negative consumption constraint, the opportunity cost of being an entrepreneur is higher for the rich than for the poor, and therefore, for a given marketing ability, a poor person has more incentive to be an entrepreneur than a rich one. However, other people are more reluctant to follow a poor would-be entrepreneur since the market reads his low personal wealth as a signal of a low (expected) marketing ability, given that marketing ability is private information. As a result, the rich would-be entrepreneurs are selected by the market while the poor would-be entrepreneurs are rejected by the market.

It should be pointed out that, in this thesis, *personal wealth* takes a *value* form, not necessarily *physical* forms.[1] Because of this, we reject the explanation given by Alchian and Demsetz (1972) that monitoring goes with the ownership of capital because the cost of monitoring the use of capital by the non-monitor owner would be too high. We shall make a distinction between the owner of *physical* capital and the owner of financial capital. If I borrowed one hundred pounds from you and bought a machine, the owner of the machine is me, not you. So far as running the machine is concerned, it is not necessary for you to monitor me. If the borrowing contract allows you some say about my activities with the machine (e.g., what I produce with the machine), there must be reasons other than because you worry that I might abuse the machine physically. Furthermore, intuition tells us that for any borrowing transaction, the lender's incentive to interfere with the borrower's affairs and the borrower's willingness to accept such interference depend on the borrower's personal wealth. When a rich person asks to borrow from me, I am more likely to meet his request without any hesitation; on the other hand, if such a request comes from a poor person, I am more likely to ask him what he will do with this money before I decide whether to meet his request. In a more general sense, given that capital *itself* has no incentive problem, we need to show: Why should capitalists have priority in being entrepreneurs or in selecting management, rather than be merely rentiers?

This chapter is organized as follows. In Sect. 3.2, the basic model will be set up. Section 3.3 will be concerned with the relationship between personal wealth and

[1] The argument will be strengthened rather than weakened if it takes physical forms.

the critical ability for someone to choose being an entrepreneur. In Sect. 3.4, we discuss how the market infers a would-be entrepreneur's ability from his personal wealth so that the rich become the winners of the competition for entrepreneurship. Wealth-dependent interest rates and wages as a mechanism of separating high ability from low ability are discussed in Sect. 3.5, where we also show why this mechanism may not work when bankruptcy incurs verification costs. Section 3.6 shows why the capital-hiring-labour is socially desirable. Section 3.7 concludes this chapter.

3.2 The Model

The economy consists of many individuals differing in their marketing ability $\theta \in [0, 1]$ and personal wealth $W_0 \geq 0$. We assume that W_0 is known to all individuals of the economy but θ is known only to an individual himself.[2] Each individual is assumed to be a risk-neutral expected utility maximizer with a utility function $U = W_1$, where W_1 is his final wealth. There are two types of occupations available for all individuals to choose: an entrepreneur or a worker. An entrepreneur runs the firm and earns the residual return, while a worker earns the contractual market wage in return for his service in the firm. Being a capitalist is not an occupation which can be chosen by anyone since it depends on personal wealth endowment.[3] We make a distinction between *active* capitalists and *passive* capitalists. A capitalist is called *active* if he chooses to be an entrepreneur, and *passive* if he chooses to be a worker. The capital owned by the active capitalist earns a residual return, while the capital owned by the passive capitalist earns a contractual market interest rate. We shall assume that an individual with W_0 is free to guarantee himself a *riskless* return equal to W_0 by holding the money without depreciation.[4]

Although our conclusion is that the entrepreneur will be selected from capitalists, we begin with two assumptions which impose no capital constraint on being an entrepreneur.

Assumption 1 *Free Choice of Occupation (FCO)*: There is no institutional restriction to stop an individual from being an entrepreneur. In other words, an individual is always free to set up a firm.

Assumption 2 *Perfect Capital Market (PCM)*: The capital market is perfect in the sense that an individual can borrow as much as he wants for his business investment at given market interest rate (i.e., there is no credit-rationing) if he chooses to be an entrepreneur, or he can lend as much as he has at the market interest rate if he chooses to be a worker (and therefore a passive capitalist).

---

[2] We shall assume that θ is drawn from a common distribution which is known to all individuals in the economy.

[3] "Capitalist" is loosely used in the text since we assume that personal wealth is continously distributed between zero and a large amount. The reader can easily understand its different meaning in the different context.

[4] This assumption can be replaced by a riskless interest rate.

Assumption 1 implicitly incorporates the underlying assumption that marketing ability is private information and cannot be directly observed by the outsiders; otherwise, some kind of professional certificate might be required so that only those whose marketing ability is greater than some particular level would be allowed to be entrepreneurs.[5] Assumption 2 is made for analytic reasons.[6] A perfect capital market is widely assumed in neoclassical economics. But the assumption turns out to be inconsistent with the capitalist firm because it (together with FCO) is equivalent to a labour-hiring-capital system. Our analytic strategy is first to show how an individual makes his choice of being an entrepreneur or a worker under *PCM*, and then to demonstrate why the market will itself reject *PCM*. The analysis will show that capital-hiring-labour is imposed by the market rather than by some exogenous forces.

The following assumption is crucial for our results:

Assumption 3 *Unlimited Liability with Non-Negative Consumption (ULNGC)*: An entrepreneur has a liability for repaying all his debts to lenders and the contractual wages to workers of the firm until his personal wealth becomes zero (in an one-period model, we must assume that he cannot repay debts by further borrowing).

Enforceability of liability is dependent on observability of personal wealth. It seems natural to impose a non-negative consumption constraint upon unlimited liability. In fact most modern legal systems allow for some insurance in the form of bankruptcy against low income states.[7]

The implications of *ULNGC* assumption are as follows. First, it does not make any sense to distinguish between the residual return from the entrepreneur's marketing function and the residual return from his personal wealth as capital investment, and therefore we shall summarize them into a single term called "profit".[8] Second, although the entrepreneur is called "the residual claimant", he may not need to be fully responsible for all costs of his business in the case of bankruptcy, if his personal wealth is not sufficient to cover all the contractual payments. In other words, there may be a difference between his promised payment and his actual payment. It is this difference that generates both the moral hazard problem and the adverse selection problem in the entrepreneurial choice.[9] Third, related to the second, because the contractual payment cannot be riskless due to the probability of default by the entrepreneur, from the point of view of workers and passive capitalists, it matters with which entrepreneur they match. This is the underlying force of the entrepreneurial selection mechanism in

[5] In reality, for some occupations such as lawyer, teacher, medical doctor and so on, a certificate is needed; but for entrepreneur, it is not. We conjecture that the reason for this difference is that entrepreneurial ability is much more difficult to observe than other ability.

[6] Note that although we use the term "perfect capital market", we exclude consumption borrowing.

[7] Non-negative consumption can be replaced with minimum subsistence without affecting the argument. In addition, the analysis can be carried over to limited liability by replacing the total wealth with equity share (one may like to call ULNGC itself "limited liability".).

[8] This might be the source of the long-running debate over what the profit is.

[9] Limited liability is the underlying assumption of most agency-type models on capital markets; e.g., see Stiglitz and Weiss (1981) credit-rationing model, Eswaran and Kotwal (1989) capital-hiring-labour model, among others.

the market. Given the market wage and the market interest rate, a passive capitalist worker's expected return depends negatively on the probability of default by the entrepreneur he matches with. Intuition suggests that other things being equal, the more wealthy is his matched entrepreneur, the more secure is a passive capitalist worker's contractual payment, and therefore he should choose to match with the rich rather than the poor. But our results are much stronger than this. Because other things are not equal, wealth itself may not suffice for a low probability of default. In particular, given that marketing dominates the uncertainty of the firm's return, we may assume that the entrepreneur's marketing ability is crucial for business success. If people prefer to follow the rich to join the firm, there must be something linking personal wealth with marketing ability, from outsiders' point of view.

An individual faces the choice first of whether he should be an entrepreneur or a passive capitalist worker, and second, if the latter, to which entrepreneur he should lend his capital (if he has any) as well as to whom he should sell his labour.[10] A complete analysis of the individual choice requires us to model both capital market and labour market. However, most insights of the analysis can be derived from modelling one market alone.[11] Since what we are interested in is the relationship between capitalists and entrepreneurs, we shall limit ourselves to the capital market by assuming that a contractual wage is paid prior to production so that workers face no default.[12] This implies that the entrepreneur must finance the hiring of labour before any physical investment takes place, and his total financial capital requirement is equal to the sum of physical investment and hired labour cost (wage times the number of workers). If his personal wealth is not sufficient for both physical investment and labour cost, he must borrow from some passive capitalists. Passive capitalists cannot avoid the probability of default and therefore it matters which borrower they choose.[13]

Assume that everyone has access to a production technology which requires a fixed amount of *aggregated* capital comprised of both physical capital investment and labour cost, denoted by K.[14] Business can be either a success or a failure. If a success, it will yield a return $y = f(K) > 0$; if a failure, it yields a zero return.[15] Denote by r the market interest rate and by w the market wage. We shall assume that $f(K) \geq (1 + r)K + w$. In other words, we assume that in the case of success, the

[10] A passive capitalist does not need to lend capital and sell labour to the same entrepreneur.

[11] In an earlier version of this chapter, we modelled both the labour market (the choice of workers) and the capital market (the choice of lenders). We found the marginal benefit of modeling more than one is little more than making the description more like reality.

[12] Therefore they do not care about which entrepreneur they should match with. Alternatively, we can assume that the lowest return of the firm (in the worst state) is not less than labour cost.

[13] The assumption of the wage being paid prior to production is equivalent to workers delegating their choice of match to passive capitalists. In reality, workers normally have priority when the entrepreneur cannot pay all contractual payments, even if they are paid at the end of the period. An interesting question is why workers have priority in most cases.

[14] It is convenient to refer to K simply as "capital". If k is physical investment, w is wage per worker and l is the number of workers, $K = k + wl$. We implicitly assume that the entrepreneur always chooses an optimal combination of k and l. In addition, K can be a variable.

[15] Zero return can be replaced by a positive return as long as it is smaller than the total cost.

total return will be sufficiently large to cover both the contractual payment and the entrepreneur's opportunity cost (otherwise there will be nobody choosing to be an entrepreneur). In the following, we normalize $w = 0$.

The importance of marketing ability is that it determines the probability of success p. In particular, for simplicity, we assume that $p = \theta$.[16] This implies that the probability of default by an entrepreneur is uniquely determined by his marketing ability, given that his personal wealth is not sufficient to finance all investment. More notably, because the returns for both the highest ability ($\theta = 1$) and the lowest ability ($\theta = 0$) are certain, but uncertain for all others, if marketing ability is public information, the assumption implies that all individuals with the highest ability would become entrepreneurs, regardless of their personal wealth.

The total expected return of the firm is a linear increasing function of the entrepreneur's marketing ability defined as follows:

$$Ey = \theta f(K) \tag{3.1}$$

The entrepreneur's expected personal return, denoted by W_1^e, depending on his wealth endowment W_0, can be defined as follows[17]:

(i) If $W_0 < K$,

$$W_1^e = \theta \left(f(K) - (1 + r) \left(K - W_0 \right) \right) \tag{3.2}$$

(ii) If $W_0 \geq K$,

$$W_1^e = \begin{cases} \theta f(K) + \delta_K (1 + r)(W_0 - K) & \text{if lending out excess funds} \\ \theta f(K) + (W_0 - K) & \text{if holding excess funds} \end{cases} \tag{3.3}$$

where δ_K denotes the (weighted) expected probability of success of the entrepreneur(s) to whom the excess funds of the entrepreneur concerned are lent.

Note we have implicitly assumed that the entrepreneur makes investment first with his own asset before he can borrow from passive capitalists, and that he will not lend out his excess assets unless $W_0 > K$. This assumption is not necessary for the results, and we make it only for simplifying the analysis.[18]

If an individual with W_0 chooses to be a passive capitalist worker, his expected return, denoted by W_1^l, is

$$W_1^l = \begin{cases} \delta_K (1 + r) W_0 & \text{if lending out his wealth} \\ W_0 & \text{if holding his wealth} \end{cases} \tag{3.4}$$

δ_K can be defined as follows:

$$\delta_K = E\theta^B \tag{3.5}$$

[16] Recall that we have normalized marketing ability to be distributed between zero and one.

[17] In the following analysis, we normalize wage to zero for convenience.

[18] In the literature, the assumption is called "*maximum equity participation*" (MEP) (e.g., Gale and Hellwig 1985).

where E is the expectation operator, and superscript B denotes the entrepreneur to whom the funds are lent ("borrower"). The entrepreneur borrows from the outsiders if and only if $W_0 < K$, which implies that $\delta_K = 1$ if and only if $E\theta^B = 1$. In other words, the lender has to bear risk for default unless he is certain that the borrower has the highest marketing ability ($\theta^B = 1$).

An individual will choose to be an entrepreneur if and only if the following condition holds:

$$W_1^e \geq W_1^l \tag{3.6}$$

where W_1^e and W_1^l are defined by (3.2)–(3.3) and (3.4), respectively.

Given his personal wealth W_0, the individual's choice of being an entrepreneur or a worker depends not only on his own marketing ability θ, but also on his expectations of the potential borrower's marketing ability $E\theta^B$ which determines δ_K. Given δ_K, (3.6) defines a critical value θ^* such that he will choose to be an entrepreneur if and only if $\theta \geq \theta^*$. We call θ^* "the individual critical marketing ability" for being an entrepreneur. How does θ^* depend on W_0? How is $E\theta^B$ related to W_0^B?

3.3 The Critical Marketing Ability and Personal Wealth

In this section, we shall focus on the relationships between an individual's critical marketing ability θ^* and his personal wealth W_0, and between θ^* and δ_K.

Case (i): If $W_0 < K$, θ^* is defined by the following equality[19]:

$$\theta^* \left(f(K) - (1+r)\left(K - W_0\right)\right) \equiv \delta_K (1+r) W_0 \tag{3.7}$$

Rearranging (3.7), we obtain

$$\theta^* = \frac{\delta_K (1+r) W_0}{f(K) - (1+r)\left(K - W_0\right)} \tag{3.8}$$

Differentiating θ^* with respect to W_0 and rearranging gives

$$\frac{\partial \theta^*}{\partial W_0} = \frac{\delta_K (1+r)\left(f(K) - (1+r)\,K\right)}{\left(f(K) - (1+r)\left(K - W_0\right)\right)^2} > 0 \tag{3.9}$$

since $\left(f(K) - (1+r)\,K\right) > 0$.

That is, the individual's critical marketing ability is increasing with his personal wealth.

[19] We assume that δ_K is big enough that the individual prefers to lend out his asset rather than hold it when he chooses to be a worker. If this is not the case, we replace $\delta_K (1+r)$ with 1.

Case (ii): If $W_0 > K$, θ^* is defined by[20]

$$\theta^* f(K) + \delta_K (1+r)(W_0 - K) = \delta_K (1+r) W_0 \tag{3.10}$$

Rearranging gives

$$\theta^* = \frac{\delta_K (1+r) K}{f(K)} \tag{3.11}$$

Therefore

$$\frac{\partial \theta^*}{\partial W_0} = 0 \tag{3.12}$$

In summary, we have

Theorem 14 *Given Assumptions 1–3, (i) an individual will choose to be an entrepreneur if and only if his marketing ability is greater than his individual critical level; and (ii) the individual critical marketing ability is increasing with personal wealth until personal wealth is greater than the capital requirement.*

Figure 3.1 illustrates this result. Roughly speaking, Theorem 14 says that, at any given ability level, a poor person has more incentive to be an entrepreneur than a rich man. The intuition behind this result is that the opportunity cost of being an entrepreneur is higher for the rich than for the poor, given the non-negative consumption constraint. For those with little personal wealth, the opportunity cost of being an entrepreneur is nothing more than the market wage of a worker (normalized to zero), while for those with large personal wealth, being an entrepreneur incurs a large wealth loss if the business is not successful. Because the cost of being an entrepreneur increases with personal wealth, the optimum requires the return to increase too, which implies that the critical marketing ability must be higher as he becomes richer. An implication of the theorem is that the poor person is more likely to overreport his marketing ability than the rich; or to put it differently, the entrepreneurial choice of the rich person is more informative in signalling marketing ability than the choice of the poor. We will see that this is the fundamental reason why capitalist would-be entrepreneurs succeed in the competition for entrepreneurship.

We now turn to the relationship between θ^* and δ_K. It is easy to show that:

(i) If $W_0 < K$,

$$\frac{\partial \theta^*}{\partial \delta_K} = \frac{(1+r) W_0}{f(K) - (1+r)(K - W_0)} > 0 \tag{3.13}$$

(ii) If $W_0 \geq K$,

$$\frac{\partial \theta^*}{\partial \delta_K} = \frac{(1+r) K}{f(K)} > 0 \tag{3.14}$$

[20]Here we assume that the individual faces the same expected probability of success of the potential borrowers regardless of whether he is lending out excess funds (when he himself is also an entrepreneur) or lending out all funds (when he chooses to be a worker).

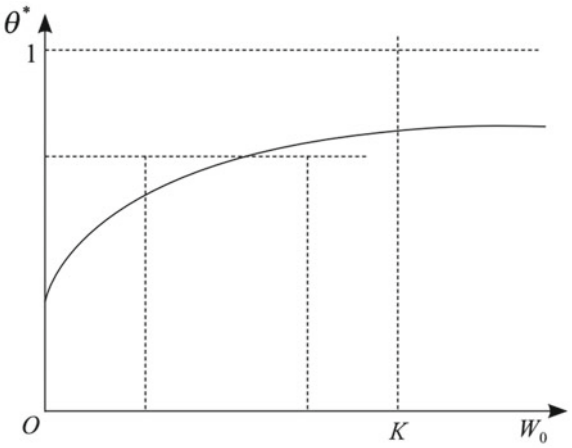

Fig. 3.1 The critical marketing ability and personal wealth

Theorem 15 *Given Assumptions 1–3, the individual critical marketing ability for being an entrepreneur is increasing with the expected probability of success of the potential borrower.*

Theorem 15 says that an individual is more likely to choose to be an (self-employed) entrepreneur when otherwise he has to lend to an entrepreneur(s) with low probability of success than when he can lend to those with high expected probability of success. The argument is quite intuitive. The probability of success of the matched entrepreneur determines the riskiness of the contractual return for being a passive capitalist worker (or more generally the expected return of the contractual return). Higher expected probability of success implies a higher expected contractual return, which in turn implies that it is less necessary for someone to be a self-employed entrepreneur.

3.4 The Expected Marketing Ability of the Would-Be Entrepreneurs and Personal Wealth

Under Assumptions 1–3, the population is divided into two sets: the set of would-be entrepreneurs (active capitalists) and the set of would-be workers (passive capitalists). In an economy where individuals have free choice of which entrepreneur to match with, a would-be entrepreneur can become an actual entrepreneur if and only if he can successfully raise the required capital. With Theorems 14 and 15, we now show why the rich would-be entrepreneurs are more likely to be successful than their poor fellows (to put it differently, why passive capitalists are reluctant to lend their capital to the poor would-be entrepreneurs), given that marketing ability is private information. The basic argument is that although an individual's actual marketing ability might be independent of his personal wealth, from the point of view of

outsiders the expected marketing ability of a would-be entrepreneur is not independent of his personal wealth.

Denote by $\phi(\theta)$ and $\Phi(\theta)$ the density function and the distribution function of marketing ability among population, with support [0, 1], which are assumed independent of the distribution of personal wealth W_0.[21] Then, from the point of view of outsiders, the expected marketing ability of a would-be entrepreneur, conditional on his personal wealth W_0^B, can be defined as follows[22]:

$$E\theta^B(W_0^B) = E\left(\theta^B \left| W_0^B\right.\right) = \frac{\int_{\theta^*}^1 \theta\phi(\theta)d\theta}{1-\Phi(\theta^*)} \tag{3.15}$$

where θ^* is defined by (3.8), or (3.11), depending on $W_0^B < (\geq) K$.

Differentiating (3.15) with respect to W_0^B and rearranging, we have

$$\frac{\partial E\theta^B\left(W_0^B\right)}{\partial W_0^B} = \frac{\phi\left(\theta^*\right)\frac{\partial\theta^*}{\partial W_0^B}\int_{\theta^*}^1\left(1-\Phi(\theta)\right)d\theta}{\left(1-\Phi(\theta^*)\right)^2} \tag{3.16}$$

Then, by (3.9), and (3.12), we have

$$\frac{\partial\theta^B(W_0^B)}{\partial W_0^B}\begin{cases} > 0 & \text{if } W_0^B < K \\ = 0 & \text{if } W_0^B \geq K \end{cases} \tag{3.17}$$

Therefore, we have

Theorem 16 *The expected marketing ability of a would-be entrepreneur is an increasing (or non-decreasing) function of his personal wealth.*

Theorem 16 says that although outsiders have no accurate information about the marketing ability of a particular would-be entrepreneur, they can be sure that, on average, a would-be entrepreneur with large personal wealth has higher marketing ability than a would-be entrepreneur with small personal wealth. It is rational to infer marketing ability according to personal wealth. Immediately, we have

Corollary 17 *The expected probability of default by the borrower is a strictly decreasing function of his personal wealth.*

---

[21] One may like to argue that the distribution of ability and the distribution of personal wealth is positively correlated either because of dynamic effects (today's wealthy people are yesterday's successful businessmen) or because the wealthier people have better opportunities for good education. If this is the case, wealth itself signals ability.

[22] Since θ^* is dependent on δ_K, an outsider must base his judgment of an would-be entrepreneur's θ^* on the δ_K in the latter's expectation (that is, to know person A's θ^*, an outsider has to know A's expectation of his potential borrower's probability of success if he chooses to be a worker). But given that the only available information is personal wealth, rational expectations imply that the outsider will hold the same expectation of all would-be entrepreneurs' δ_Ks. In the following, we shall make this assumption.

Note that here the link between personal wealth and the probability of default is not direct, but rather indirect: personal wealth affects the individual's choice of being an entrepreneur which in turn determines the probability of default.

A would-be entrepreneur's personal wealth not only affects his perceived marketing ability and therefore his attractiveness to a potential lender, but also affects others' entrepreneurial choices at the margin. By Theorem 15, we know that an individual's critical marketing ability is increasing with the expected probability of success of his potential borrower. Combining this with Corollary 17, we have

Theorem 18 *(i) An individual's critical marketing ability for becoming an entrepreneur is increasing with the potential borrower's personal wealth; and (ii) the slope of this relation depends positively on his own personal wealth.*

Part (i) says that given his personal wealth and marketing ability, an individual is more likely to choose to be a passive capitalist worker when he can lend his wealth to a wealthier person than when he can only lend it to a less wealthy person; part (ii) says that the rich are more sensitive to potential borrowers' personal wealth than the poor in making choices between being an entrepreneur or a worker. The intuition is that larger personal wealth of a potential borrower or employer signals a higher expected marketing ability and a lower expected probability of default, and therefore a higher expected contractual payment. The argument can be easily shown with a diagram.

In Fig. 3.2, we draw the implied relationship between an individual's critical marketing ability θ^* and his potential borrower's (or employer's) personal wealth W_0^B

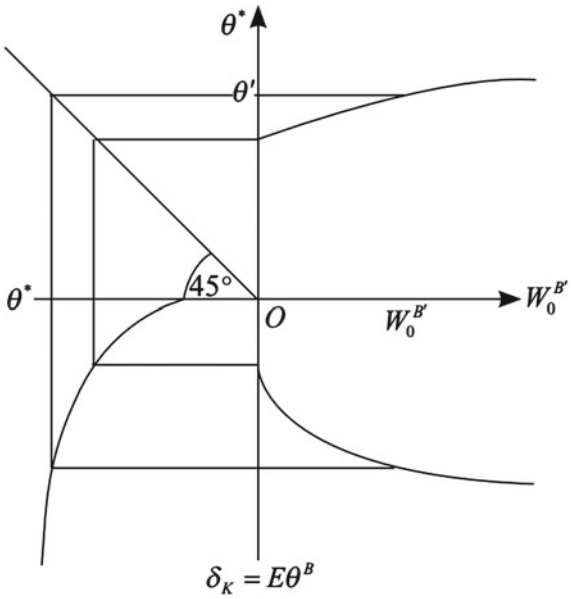

Fig. 3.2 The expected probability of success and personal wealth

in the first quadrant. In the fourth quadrant, the relationship is drawn between the expected probability of success of the potential borrower or employer and his personal wealth W_0^B, following Corollary 17. The third quadrant describes the relationship between θ^* and δ_K, following Theorem 15. The 45-degree line in the second quadrant allows us to derive the curve in the first quadrant. In the case shown, an individual with marketing ability θ' (given W_0), for example, will choose to be an entrepreneur if the potential borrower's or employer's personal wealth is smaller than $W_0^{B'}$, while he will choose to be a worker when the latter's personal wealth is greater than $W_0^{B'}$.

The strong implication of the preceding discussion is that although an individual with lower personal wealth has greater incentives to choose being an entrepreneur, other people are more reluctant to accept him, since they read his low personal wealth as a signal of a low (expected) marketing ability. From the point of view of would-be lenders, a rich would-be entrepreneur is always more attractive than a poor one; and it is always in their self-interest to lend to the former rather than the latter. Because a would-be entrepreneur can become an actual entrepreneur (materialize his dream) only if there are sufficient numbers of lenders (if he needs external funds) who voluntarily lend to him, we predict that only those would-be entrepreneurs who have sufficiently large personal wealth will succeed in the competition for entrepreneurship.

Theorem 19 *Given that marketing ability is private information and personal wealth is public information, market competition for entrepreneurship implies that a would-be entrepreneur can become an actual entrepreneur only if his personal wealth is greater than some specified level.*

To be concrete, assume that marketing ability is uniformly distributed in the population. It is easy to show that:

$$E\left(\theta^B \left| W_0^B \right.\right) = \frac{1}{2} + \frac{1}{2}\theta^* = \begin{cases} \frac{1}{2} + \frac{1}{2}\frac{\delta_K(1+r)W_0^B}{f(K)-(1+r)(K-W_0^B)} & \text{if } W_0^B < K \\ \frac{1}{2} + \frac{1}{2}\frac{\delta_K(1+r)K}{f(K)} & \text{if } W_0^B \geq K \end{cases} \tag{3.18}$$

That is, in the case of uniform distribution of ability, the expected marketing ability of a would-be entrepreneur is a weighted average of the highest ability ($\theta = 1$) and the critical ability (θ^*), with equal weights.

What does the market mechanism for entrepreneurial selection look like? If we rank all would-be entrepreneurs in terms of their personal wealth from the highest to the lowest, it is like a "pecking order". The first group of would-be entrepreneurs to be successfully selected by the market are those whose personal wealth is sufficiently large to cover both physical investment and the *riskless* contractual payment for workers, that is, $W_0 \geq K$.[23] This group of entrepreneurs are perceived by the market to be those with the highest expected marketing ability among all would-be

[23]These would-be entrepreneurs do not depend on external funds, and are "selected" by workers.

entrepreneurs, equal to[24]

$$E\theta = \frac{1}{2} + \frac{1}{2}\frac{\delta_K(1+r)K}{f(K)} \tag{3.19}$$

Since capital itself is productive, an economy in which entrepreneurship is restricted to this group cannot be in equilibrium. The second group selected for entrepreneurship consists of would-be entrepreneurs whose personal wealth is sufficiently high to cover labour costs, but is not sufficient for covering the full costs (physical investment plus labour costs). The third group consists of those would-be entrepreneurs who need to borrow both for physical investment and labour payment. The last two groups are the most interesting cases since the existence of these groups is a precondition for capital markets to occur.[25]

A general result is that the set of entrepreneurs is defined by a lower-bound of personal wealth. To plot the equilibrium lower-bound, we need a general equilibrium model. Nevertheless, the following partial equilibrium analysis can provide some insights.

First note that since the decision to be an entrepreneur is made after comparing with the expected return from being a passive capitalist worker, the following inequality must hold:

$$E\theta^B = \frac{1}{2} + \frac{1}{2}\frac{\delta_K(1+r)W_0^B}{\left(f(K) - (1+r)(K - W_0^B)\right)} \leq \frac{1}{2} + \frac{1}{2}\frac{(1+r)W_0^B}{\left(f(K) - (1+r)(K - W_0^B)\right)} \tag{3.20}$$

That is, the expected marketing ability of this group cannot be greater than in the case where the contractual return for lending is riskless ($\delta_K = 1$).

For a potential lender, the possibility of holding wealth instead of lending out implies that the following condition must hold for lending to take place:

$$\delta_K(1+r) = E\theta^B(1+r) \geq 1, \text{ or } E\theta^B \geq \frac{1}{1+r} \tag{3.21}$$

Thus, a necessary (but not sufficient) condition for a potential lender to meet the would-be entrepreneur's borrowing requirement is

$$\frac{1}{2} + \frac{1}{2}\frac{(1+r)W_0^B}{\left(f(K) - (1+r)(K - W_0^B)\right)} \geq \frac{1}{1+r} \tag{3.22}$$

By rearranging (3.22), we have

[24] In the following analysis, for concreteness, we assume that marketing ability is uniformly distributed among population.

[25] In the previous analysis, we have implicitly assumed the existence of these groups; otherwise, we should replace $\delta_K(1+r)$ with 1. Since we have assumed that workers are paid before production, we shall not make a distinction mathematically between the second and the third groups.

$$W_0^B \geq \frac{(1-r)}{2r(1+r)}\left(f(K)-(1+r)K\right) \tag{3.23}$$

This is the lower-bound of personal wealth of this group imposed by the potential lenders. Potential lenders will reject the borrowing request if the would-be entrepreneur's personal wealth is smaller than this bound.

To give a concrete example, let us assume that $K = 50, r = 0.1$, and $f(50) = 60$. Then, the bound imposed by potential lenders is:

$$W_0^B \geq 20.5$$

That is, a lender will never lend to a would-be entrepreneur whose personal wealth is less than 20.5. If capital of $K = 50$ is necessary for the firm to be profitable, we shall expect that there will be no person in the entrepreneurial team whose personal wealth is smaller than 20.5.

3.5 Interest Rates (and Wages) as Mechanisms for Capital-Hiring-Labour

So far we have assumed that the interest rate (and wage) is fixed at a uniform level. The preceding analysis shows that the uniform rate cannot be in equilibrium, since that implies that different lenders earn different expected returns (different borrowers are perceived with different expected probability of default). In this section, we relax this assumption to discuss how changes in the interest rate (and wage) affect the critical marketing ability for someone to choose to be an entrepreneur, and in particular how the interest rate (and wage) may be used to some extent as mechanisms to restrict low-wealth people from being entrepreneurs.[26] Discussions are focused on the case of $W_0^B < K$.

First consider effects of changes in the interest rate on the critical marketing ability. Differentiating θ^* with respect to r, we have

$$\frac{\partial \theta^*}{\partial r} = \frac{\delta_K W_0 f(K)}{\left(f(K)-(1+r)(K-W_0)\right)^2} > 0 \tag{3.24}$$

We have

Theorem 20 *The critical marketing ability is increasing with the interest rate for all individuals.*

The reason behind this argument is simple: increases in the interest rate will increase both direct costs and opportunity costs of being an entrepreneur, and therefore raise the marginal level of marketing ability at which being an entrepreneur is more profitable than being a worker.

[26]The following arguments about changes in the interest rate also apply to changes in wages.

Although a change in the interest rate affects the *average* marketing ability of the pool of would-be entrepreneurs, it does not change the fact that among all would-be entrepreneurs, those with low personal wealth have lower average marketing ability than those with high personal wealth, since Theorem 14 applies to all levels of interest rates (up to a upper-bound.[27]) Therefore we claim that a *uniform* interest rate (and wage) cannot be an effective mechanism for separating low ability would-be entrepreneurs from high ability would-be entrepreneurs.

Second, we show how *wealth-dependent* interest rates (and wages) may work as a mechanism to stop penniless lemons from choosing being entrepreneurs. By wealth-dependent, we mean less wealthy people have to pay higher interest rates (and higher wages) than the wealthier people if they choose to be entrepreneurs.[28]

By Theorems 14 and 19, for a given critical (or expected) marketing ability, the following condition holds[29]:

$$\left.\frac{\partial r}{\partial W_0}\right|_{\theta^*} = -\frac{\frac{\partial \theta^*}{\partial W_0}}{\frac{\partial \theta^*}{\partial r}} < 0 \tag{3.25}$$

Theorem 21 *A necessary condition for maintaining the same critical marketing ability among all people is that the interest rates to be charged depend negatively on the borrower's personal wealth.*

The essence of wealth-dependent interest rates is that under such a system, the less wealthy incur a higher borrowing cost for being entrepreneurs so that penniless lemons will "voluntarily" withdraw from being would-be entrepreneurs. This kind of discrimination is one of the most important characteristics of capital markets. In the literature, it has been called an "imperfection" of capital markets. But this imperfection should be understood as a mechanism for "capital-hiring-labour", since it operates against high ability-low wealth would-be entrepreneur.

The mechanism is introduced by high ability-but-low-wealth would-be entrepreneurs as well as potential lenders and workers. Under the system of uniform interest rates and wages, the would-be entrepreneurs whose personal wealth is below some critical level will be rejected by potential lenders (and workers), regardless of their individual marketing ability. Since the rejected high ability people have a greater (expected) loss, it is worthwhile for them to pay higher interest rates (and wages) in order to be entrepreneurs instead of workers. By so doing, they can partially separate themselves from rejected low ability people since the latter can not afford to mimic

[27] In the present model, this upper-bound is an interest rate $\overline{r}$ (or wage $\overline{w}$) at which only those with the highest marketing ability ($\theta = 1$) can be indifferent between being entrepreneurs and being workers and all others strictly prefer being workers, that is, $f(K, L) - (1 + \overline{r})(K - W_0) - \overline{w}L - \overline{w} \equiv 0$. This requirement is too strong to hold in reality.

[28] Since, given capital investment, one's demand for borrowing is decreasing with initial personal wealth, this means the interest rate charged to a borrower is an increasing function of the borrowing amount.

[29] Technically we shall assume that the particular individual's expected return on lending and expected wage from being a passive capitalist worker are given.

them. On the other hand, for potential lenders (and workers), what matters is the *expected* returns ($\delta_K(1+r)$ and $\delta_L w$). Although matching with the less wealthy would-be entrepreneurs incurs a higher probability of default, the expected return may not be lower if the interest rate and wage to be paid are sufficiently higher in the case of success. Therefore it may pay to trade off with a high probability of default. As a result, the average marketing ability of the group of low wealth would-be entrepreneurs is also increased.

One problem is that if the wealth-dependent interest rates and wages can be effective in stopping low-ability people from being entrepreneurs, why in reality are some would-be entrepreneurs rejected by potential lenders or workers even when they wish to pay higher interest rates and higher wages? There are several possible reasons for this. One reason, provided by Stiglitz and Weiss (1981), is that an increase in the interest rate may affect the quality of projects itself through both adverse selection and moral hazard effects on the borrower's side so that the lender's expected return may decrease rather than increase as a result of interest rate increases.[30] Another reason, provided by Eswaran and Kotwal (1989), is that an increase in the interest rate may have a negative effect on the entrepreneur's (borrower's) work incentives and therefore increase the probability of default.[31] In our simple model, to focus on the relationship between marketing ability and personal wealth, we have ignored these two effects. Although we believe that the informativeness of wealth in signalling marketing ability is more fundamental in explaining capital-hiring-labour, the two arguments above can be complementary to our model.[32] Nevertheless, by extending our model to a more general case in which the number of states is more than two, we can offer an alternative explanation for why wealth-dependent interest rates and wages may eventually force all the poor to drop out from entrepreneurship because of bankruptcy costs for the lenders.

Consider a continuum of states of nature $s \in [0, 1]$. Assume that the return of the firm is strictly increasing with s for any given capital input and labour input: $\frac{\partial f(K,s)}{\partial s} >$

[30]The Stiglitz–Weiss model is based on asymmetric information between borrowers and lenders about the risk quality of investment projects. They argue that because of the adverse selection problem and the moral hazard problem, the investment projects become riskier as the interest rate increases; and hence an increase in the interest rate may decrease rather than increase the total expected return to lenders (under limited liability). This provides the incentive for the lenders to ration credit rather than raise the interest rate when there is an excess demand for loanable funds.

[31]The argument is as follows. The probability of default is a decreasing function of the entrepreneur's work effort. Since an increase in the interest rate not only decreases the range of realization of states over which the entrepreneur is the residual claimant, but also decreases his marginal return in that range, the entrepreneur will work less hard following the interest rate increase. As a result, the probability of default increases. For the lender, this effect may more than offset the direct increases of return when good states occur.

[32]In fact, these two arguments can be incorporated into our model simply by assuming that the distribution funtion $\Psi(y)$ of the return is a function of marketing ability θ, work effort a as well as a parameter of riskness α: $\Psi(y;\theta,a,\alpha)$. If we assume that $\Psi(.)$ satisfies the first-order stochastic condition over θ and a, and α is the mean-preserving parameter (that is, a higher α represents higher riskness), we can show that: (i) the critical marketing ability is an increasing function of W_0, r; (ii) the optimal effort is increasing with W_0, but decreasing with r (given marginal disutility of effort increases); (iii) the choice of α increases in r.

0 for all K, s. Let $G(s, \theta)$ be the distribution of states of nature parameterized by marketing ability θ. Assume that $G(s, \theta)$ satisfies the first-order stochastic dominance condition in θ, i.e., $\frac{\partial G}{\partial \theta} < 0$ for $s \in [0, 1)$, which means that the high marketing ability makes states in the upper tail of probability distribution more likely.[33] Denote by s^* the critical state of bankruptcy such that

$$f(K, s^*) \equiv (1 + r)(K - W_0)$$

$$f(K, s) \leq (1 + r)(K - W_0) \text{ for all } s \leq s^*$$

Then, s^* is an increasing function of the interest rate paid by the entrepreneur. So is the probability of bankruptcy $G(s^*, \theta)$. Since under wealth-dependent interest rates, within a group of entrepreneurs of a given ability, those with low wealth pay higher interest rates than those with high wealth, the probability of bankruptcy by the former will be further increased.[34] Assume that when the firm goes bankrupt, it costs x for the lender to verify. Thus the expected bankruptcy cost for the lender, $G(s^*, \theta)x$, is increased by the interest rate increase. As a result, if x is big, the potential lender may prefer to simply reject lending to the less wealthy borrower rather than charge a higher interest rate. Alternatively, even if the lender can be compensated by further increasing the interest rate, the high ability-less wealthy people may find it no longer profitable to be entrepreneurs. That is, the wealth-dependent interest rate itself may force all less wealthy people to drop out from entrepreneurship.[35]

In summary, we can predict that capital markets are characterized by both wealth-dependent interest rates and credit-rationing. This prediction is coincident with casual observation.

3.6 The Market Solution and Social Optimum

The preceding analyses suggest that market competition assigns priority in being entrepreneurs to the rich people; the would-be entrepreneurs with low personal wealth are either rejected or have to pay higher interest rates and higher wages than their rich fellows. Is this market solution socially desirable?

For this problem, we first need to define the social cost of capital. Since the market interest rate r is not riskless, $(1 + r)$ is not the social cost. However, denoting by $\delta = E\delta_K$ the average probability of success (or equivalently the average marketing

[33] In the two-state case, this simply means that the probability of success is increasing with marketing ability.

[34] The effect on the probability of bankruptcy of higher debt has already been taken into account by the lenders.

[35] Theoretically credit-rationing can be interpreted as that the borrower has to pay an extremely high interest rate so that even if the best state occurs, the return cannot cover the cost.

ability) of entrepreneurs, we can interpret $\delta(1+r)$ as the social cost of unit investment (equivalent to a riskless return).

Denote by $\theta^{\dagger}$ the social critical marketing ability for someone to be an entrepreneur (i.e., an individual should be selected for being an entrepreneur if and only if his marketing ability is equal to or greater to $\theta^{\dagger}$). Then $\theta^{\dagger}$ can be defined as follow:

$$\theta^{\dagger} f(K) = \delta(1+r)K \tag{3.26}$$

That is, at $\theta^{\dagger}$, the expected return of the firm is equal to the social cost. Note that condition (3.26) can be justified from the point of view of social justice in the sense that it makes no discrimination between the rich and the poor. Those with $\theta < \theta^{\dagger}$ should not be selected for entrepreneurship not because they are poor in wealth but because they are poor in ability.

Rearranging (3.26) gives

$$\theta^{\dagger} = \frac{\delta(1+r)K}{f(K)} \tag{3.27}$$

This is nothing but the individual critical marketing ability of those with $W_0 \geq K$. This implies that capital-hiring-labour is *socially* desirable in the sense that the resulting assignment of entrepreneurship in the population is closer to the social optimum than otherwise.

3.7 Concluding Remarks

The major conclusion of this chapter is that priority in being entrepreneurs is given to capitalists because the entrepreneurial choices of the wealthy are more informative than that of the poor in signalling marketing ability. The model can also explain why capital markets cannot be *perfect* in the sense that wealth-dependent-interest rates and credit-rationing, instead of uniform rates and free borrowing, are present, and why would-be workers are keen to embrace rich rather than poor would-be entrepreneurs. Since the capitalists' priority in being entrepreneurs comes from the information asymmetry about marketing ability, an implication of the model is that high-ability people whose ability has been revealed through their previous successes are less constrained by their personal wealth endowments when they want to expand their businesses. This implication is consistent with casual observation.

More importantly, although we have focused on the classical capitalist firm, the model can explain the occurrence of joint-stock companies in an economy. Assume that the distribution of ability and the distribution of wealth in the population are not symmetric, that is, rich people are not necessarily high ability and high ability people are not necessarily rich.[36] Then there inevitably exist two potential earning gaps, one between different providers of capital, and the other between different

[36] We have assumed that they are independent.

abilities. The capital owned by the more able will earn its factor price plus a "pure" rent from signalling, while the capital owned by the less able will earn only its factor price because its owner has no ability to signal; furthermore, the ability of the rich will yield a residual rent, while the ability of the poor will yield only a "market wage", because the poor have no capital with which to signal. In particular, "entrepreneurs" may use their monopoly market power to exploit other capital and other ability by pushing down factor prices. These possible earnings gaps make it profitable for both ability and capital to look for possible cooperation with each other. In particular, possession of some personal information about others' ability might be profitable for capitalists. Although a rich person with low ability cannot make a profit by marketing himself, he may increase his return by using his capital to signal other people's ability, if he knows some high ability people (e.g., his relatives), or if search for high ability is not too costly; on the other hand, a high ability person can also increase his return if he can convince some rich person that he is really good at marketing. Furthermore, the incentive for each side to search is an increasing function of their respective resources (ability or wealth), because the more personal wealth (ability) someone has, the more rent he can earn, if search is successful. As a result, they become a *joint* entrepreneur: the high ability person is called the manager by doing marketing, and the wealthy are called "shareholders" by claiming the residual and taking responsibility for selection of the qualified manager.

A point needs mentioning about the non-pecuniary penalty for bankruptcy. In the literature, some models of capital structure are based on the idea that the manager or entrepreneur has to bear the non-pecuniary penalty for bankruptcy which will bond him not to take gamble.[37] We believe that insofar as marketing ability is concerned, this penalty mechanism does not work. The reason is that for a penniless person, bankruptcy costs nothing more than going back to be a worker; but if he is lucky, he becomes a rich man. So it pays to take such a gamble.[38]

[37] e.g., see Ross (1977), Grossman and Hart (1982), Diamond (1984) and Gale and Hellwig (1985).

[38] Of course, a death sentence in case of bankruptcy could be helpful in preventing low ability people from being entrepreneurs!

Chapter 4
A General Equilibrium Entrepreneurial Model of the Firm

4.1 Introduction

The free choice of occupations is said to be one of the major virtues of a market economy. However, the fact is that the choice of becoming an entrepreneur is considerably constrained by personal wealth, both explicitly and implicitly. Evidence shows that capital is essential for entrepreneurial choice. Many would-be entrepreneurs fail to start in business because they are short of start-up capital.[1] In this sense, the choice is not *free* and is relevant only for capitalists, as many critics of capitalism have pointed out. Mainstream economists have simply accepted this fact as an institutional constraint, and provided little rationale for it. Particularly, economists frequently take the capital constraint to be synonymous with an imperfect capital market, without investigating further. A principal theme of this thesis is to provide such a rationale. In the last chapter, it was argued that the capital constraint on becoming the entrepreneur is socially desirable because otherwise there would be too many inferior people entering the entrepreneurial market.

The purpose of this chapter is to incorporate this capital constraint into a general equilibrium entrepreneurial model of the firm. In the model, each individual is identified by a three-dimensional vector with marketing ability, personal wealth and risk-attitude as its elements. These three variables are the determinants of the occupational choice. The occupational choice equilibrium is characterized by partition into a set of entrepreneurs, a set of workers, a set of managers and a set of capitalists. We will show that in equilibrium, (a) individuals with high ability, high personal wealth and low risk-aversion become entrepreneurs, (b) individuals with low ability, low personal wealth and high risk-aversion become workers, (c) individuals with high ability but low personal wealth become managers selected by capitalists, and

[1] Although this is even commonsense, empirical studies are scarce and have only appeared just recently. For an econometric study for the United States' case, see Evans and Jovanovic (1989), and for Britian, see Blanchflower and Oswald (1990).

W. Zhang, *The Origin of the Capitalist Firm*, DOI 10.1007/978-981-10-0221-2_4

(d) individuals with low ability but high personal wealth become (pure) capitalists to select managers. The relationship between a worker and an entrepreneur (or the joint entrepreneur) is identified by a market wage rate, the relationship between an entrepreneur and other capitalists is identified by a market interest rate plus a borrowing constraint rule; and the relationship between an employee-manager and an employer-capitalist is identified by a residual-sharing rule plus a managerial selection rule. We shall show that the equilibrium relationship between different parties is determined by the joint distribution function of marketing ability, personal wealth and risk-attitudes.

The intuition of the model is as follows. Given that the entrepreneur performs two functions —marketing and risk-bearing— a high ability and less risk-averse person is more likely to be an entrepreneur than a low ability and more risk-averse person; and given that the entrepreneurial choice is constrained by personal wealth, *ceteris paribus* a wealthier person finds it easier to be an entrepreneur if he wishes. If marketing ability, personal wealth and risk-attitudes in the population are perfectly correlated with each other, that is, the high ability person is also rich and less risk-averse, the equilibrium would be very simple: there would be a cut-off point which partitions the total population into two groups, one capitalist-entrepreneur group and one penniless-worker group, with the second group hired by the first group. The problem is that correlations between marketing ability, personal wealth and risk-attitude are far from perfect. A high ability person is not necessarily wealthy or less risk-averse; similarly a rich person is not necessarily high ability or more risk-loving; and so on. Given such imperfect correlations, if the choice existed only between being an independent entrepreneur or a worker, many high-ability people would be excluded from marketing, and many low-ability capitalists would lose the opportunities to make the best use of their capital. In response to this problem, the managerial occupation is created through cooperation between high-ability-less-wealthy people and low-ability-more-wealthy people. This cooperation allows those high-ability people who fail to be entrepreneurs because of personal wealth constraints to have a chance to do marketing by becoming professional managers. The cooperation benefits not only high-ability-less-wealthy people and low-ability-more-wealthy people, but also people with low ability and low personal wealth, because otherwise their market wage would be much lower. The joint distribution of marketing ability, personal wealth and risk-attitude affects the equilibrium relations between different occupations through its effects on the supply-demand relations.

The underlying assumption is still that marketing ability is private information while personal wealth is public information. However, in this chapter, the assumption is made concrete as a personal wealth constraint rule imposed by markets on entrepreneurial choices. This rule discriminates against less wealthy people in being entrepreneurs. Those high-ability but less-wealthy people can become managers if and only if there exists *ex ante* information asymmetry of ability between *outsiders*, or it is possible to acquire knowledge of others' ability through searching activities and search costs are not prohibitively high.

This thesis is not the first to study entrepreneurial choice in a general equilibrium framework. In fact, some pioneer work has been done in Kihlstrom and Laffont (1979), Lucas (1978), Kanbur (1979) and Evans and Jovanovic (1989), among others. However, compared with our model, all these models have the following problems. First, the authors do not provide a sound reason why the entrepreneur is a residual claimant and why the choice of being the entrepreneur is constrained by personal wealth in the first place. In all these models, these two arguments are simply assumed rather than deduced. Second, each of these models is focused on only one aspect of the problem. In Kihlstrom-Laffont's model, individuals are assumed identical in their ability and personal wealth and different only in their risk-attitudes; the equilibrium has the property that less risk averse individuals become entrepreneurs, while the more risk averse become workers. In the Lucas model, there is no uncertainty and the capital constraint is simply ignored[2]; individuals are risk-neutral, homogeneous in their productivity as workers but differ in their managerial ability as entrepreneurs; the equilibrium is characterized by a cut-off point of ability such that the individuals whose ability is greater than the cut-off become entrepreneurs and others become workers.[3] Lucas' model may be understood as an equivalent of Kihlstrom–Laffont's model if one substitutes ability for risk-attitudes. Kanbur is more concerned with how national risk-attitudes affect the personal income distribution than with how the individual risk-attitudes affect the entrepreneurial choice. He shows that in an economy with homogeneous risk attitudes, in equilibrium all individuals are indifferent between being a worker or an entrepreneur, and the more risk averse a society is, the smaller the proportion of entrepreneurs, but the relationship between risk-attitude and inequality of income distribution is not monotonic. When he turns to an economy with heterogeneous risk-attitudes, his findings are not much different from Kihlstrom–Laffont. In his model, the capital constraint is ignored; entrepreneurial ability is assumed to differ across individuals, but because nobody knows his own ability, it makes no difference from a model in which individuals are identical in ability but face an uncertain environment. Evans and Jovanovic were the first to model explicitly the capital constraint on entrepreneurial choice. Using US data, they find that capital is essential for starting a business and liquidity constraints tend to exclude those with insufficient personal assets from being entrepreneurs. In particular, they find that a person cannot use more than 1.5 times his or her own initial assets for starting a new venture.[4] However, because their primary concern is to test whether the liquidity constraint hypothesis proposed by Knight is consistent with empirical data, rather than to explore theoretically why liquidity constraints are there in the first place or how capital distribution may affect the entrepreneurial choice, they do not characterize the general equilibrium. Third, each of these models (except Evans-Jovanovic) has provided some comparative statics results about

[2]This is natural since the capital constraint is irrelevant when there is no uncertainty.

[3]In the paper, Lucas uses terms "manager" and "managerial" insteead of "entrepreneur" and "entrepreneurial". He may be aware that with no uncertainty entrepreneurship is irrelevant.

[4]Some of their other findings are also interesting, including (i) entrepreneurs may be relatively poor wage workers; and (ii) entrepreneurial ability and personal assets are negatively correlated.

the equilibrium relationship between worker and entrepreneurs: for example, in the Kihlstrom–Laffont model, economy-wide increases in risk-aversion reduce the equilibrium wage; in the Lucas model, an increase in capital per capita raises wages relative to marginal entrepreneurial rents; and in the Kanbur model, an increase in risk aversion raises the mean income of the entrepreneur and reduces the wage. However, little has been said about the equilibrium relations between managers and capitalists.

The lack of a comprehensive general equilibrium entrepreneurial model of the firm may be attributed to the complexity of the problem itself. The entrepreneurial choice is influenced by many factors, including some social-cultural variables. For economic analysis, no doubt, marketing ability, personal wealth and risk-attitude are seen as the three major factors. Modelling any one of them ignoring others has advantages mathematically. However, in so doing, important insights are lost, since the interrelation between different factors can not be characterized in such an one-dimensional model. This consideration is the main motivation for the present model in which all three factors are taken into account. We hope its imperfections in formulation can be offset by the interest of the results.

The chapter is organized as follows. The basic model will be set up in Sect. 4.2. In Sect. 4.3, we shall characterize the entrepreneurial choice and the set of entrepreneurs. Section 4.4 will be concerned with the existence of equilibrium. Comparative statics will be discussed in Sect. 4.5. Section 4.6 will show how our model can encompass other models. Having established the basic results, we shall extend the model to take into account cooperation between high-ability individuals and capitalists in Sect. 4.7. Section 4.8 concludes the chapter with a heuristic story.

4.2 The Model

The economy consists of a continuum of individuals each of whom is identified by a three-dimensional vector $\nu = (\theta, W_0, \rho)$, where $\theta \in [0, \infty)$ is marketing ability, $W_0 \in [0, \infty)$ is initial personal wealth endowment (to be used as capital input), and $\rho \in [0, \infty)$ is the index of risk-aversion. Individuals A and B are said to be identical if and only if $\nu_A \equiv \nu_B$; otherwise they are different. Individual ν's von Neumann–Morgenstern utility function is represented by a parameterized function $U = U(W_1, \rho)$, where $W_1 \in [0, \infty)$ is his final income. We make the following assumptions.

Assumption 1 $U(W_1, \rho)$ is twice continuously differentiable in W_1 with $U_W > 0$ and $U_{WW} \leq 0$.[5]

Assumption 2 The Arrow–Pratt measure of absolute risk-aversion is non-decreasing in ρ in the sense that

[5]Here subscriptions denote the first- and the second derivatives respectively. Note that we assume that the utility function is independent of marketing ability.

$$\rho_1 > \rho_2 \Rightarrow -\frac{U_{WW}(W_1, \rho_1)}{U_W(W_1, \rho_1)} \geq -\frac{U_{WW}(W_1, \rho_2)}{U_W(W_1, \rho_2)} \text{ for all } W_1 \in [0, \infty) \quad (4.1)$$

or

$$-\frac{\partial^2 \ln U_W}{\partial \rho \partial W_1} \geq 0 \quad (4.2)$$

That is, high ρ represents more risk-averse.[6] In particular, we define $\rho = 0 \Leftrightarrow U_{WW} = 0$. Note that we assume that individual ν is either risk-neutral ($\rho = 0$) or risk-averse ($\rho > 0$), but not risk-loving.

At the first stage, assume that there are only two occupations available for an individual to choose: being an entrepreneur with residual return or being a worker with contractual return (to be defined later). If an individual becomes an entrepreneur, his total entrepreneurial output (Y) is a stochastic function of the employed labour (L), capital input (K) and his marketing ability (θ), defined as follows:

$$Y = f(L, K, \theta, s) \quad (4.3)$$

where $s \in [\underline{s}, \overline{s}]$ is a state of nature drawn from the distribution function $\Phi(s)$ (the density function denoted by $\phi(s)$). Note that L does not include the entrepreneur's own labour input which is treated as a set-up cost of the firm.

Assumption 3 $f(L, K, \theta, s)$ is twice continuously differentiable in L, K, θ, and s with the following properties: (i) $f_L > 0$, $f_{LL} < 0$ for all $L \geq 0$; $f_K > 0$, $f_{KK} < 0$ for all $K \geq 0$; $f_\theta > 0$, $f_{\theta L} > 0$, $f_{\theta K} > 0$; and $f_s > 0$, $f_{sK} > 0$, $f_{sL} > 0$; and (ii) $f(0, K, \theta, s) \equiv f(L, 0, \theta, s) \equiv f(L, K, 0, s) = f(L, K, \theta, \underline{s}) \equiv 0$.

Part (i) is standard and self-explanatory. Part (ii) says that both labour and capital are necessary for output to be positive, the lowest ability person cannot be a productive entrepreneur, and the worst that can happen is zero output, regardless of labour, capital and marketing ability.[7]

Assume that marketing ability is private information while personal wealth is public information. In addition, we also assume that an individual's final wealth cannot be below zero level. Then, based on arguments derived in Chap. 3, we have

Assumption 4 The entrepreneur's choices of labour input and capital input (L, K) are constrained by his personal wealth endowment.

Note that, unlike Assumptions 1–3 which are technical assumptions, Assumption 4 is an institutional assumption. The assumption is based on both observation and our theoretical justification of why capital hires labour (provided in Chap. 3). The major difference between the present model and standard models

[6] I am adopting this formulation from Kihlstrom and Laffont (1979).

[7] This is made only for simplicity. Whether the total return in the worst state is zero or positive affects only the boundary point, not the direction of the relationship.

of general equilibrium in treatment of the personal wealth constraint is that in the present model, the personal wealth constraint is one of starting assumptions and characteristics of equilibrium, while in standard models, it is a characteristic of disequilibrium.

In reality, the constraint may take two forms: wealth-dependent interest rates and wages, and limited maximum borrowing; that is, both prices and quantities depend on the entrepreneur's personal wealth. However, for analysis to be tractable, we consider only the quantity constraint by assuming that all entrepreneurs face the same interest rate and the same wage rate, and that only the size of the firm is limited by personal wealth. This simplicity is not implausible if the quantity constraint rule is such that all entrepreneurs incur the same probability of bankruptcy.[8] In particular, we assume that the personal wealth constraint rule (PWCR) takes the following linear form:

$$wL + rK \leq \lambda W_0 \tag{4.4}$$

where w is the market wage of labour, r is the market rental price of capital (equal to one plus the interest rate), and $\lambda > 1$ is the parameter of PWCR which summarizes all information of *non-price* constraints on the entrepreneurial choice. We will see that λ plays an important role in determining equilibrium, together with (w, r).

Some comments on (4.4) are in order. First, λ is assumed to be equal for all would-be entrepreneurs. That is, an individual's budget constraint is uniquely determined by his personal wealth, regardless of his other personal characteristics. This follows the underlying assumption that individual marketing ability (drawn from a common distribution) is private information, not observable to outsiders. PWCR can be only based on W_0 since W_0 is the only publicly observable personal characteristic. In reality, if there is other information related to marketing ability, PWCR might be also based on this "other information". In particular, a high ability person may be more able to convince others of his ability, which implies that PWCR may directly depend on one's ability. Nevertheless, as long as ability cannot be perfectly observable, the above simplification may be justified.

Second, in (4.4), labour and capital are treated symmetrically. An alternative way is to put different weights on labour and capital, or even consider capital only. The advantage of symmetric treatment is that no matter whether the constraint is binding or not, the combination of labour and capital is always optimal for the given production technology $f(L, K, \theta, s)$ and the distribution function $\Phi(s)$.[9] For example, if the output obeys a Cobb–Douglas function $Y = \theta s L^{\alpha} K^{1-\alpha}$, where $0 < \alpha < 1$, the

[8] One may claim that this assumption is at least a crude approximation to reality. An observation is that a significant proportion of loan applications are simply rejected, even if borrowers are willing to pay a higher rate.

[9] Since utility depends only on final income irrespective of how the income is produced, the optimal combination of capital and labour is independent of utility function. However, in general, the optimal combination depends on output level, and therefore the optimal $\left(\frac{K}{L}\right)$ when PWCR is binding may be different from the optimal $\left(\frac{K}{L}\right)$ when PWCR is not binding.

optimal $\left(\frac{K}{L}\right)$ will satisfy $\frac{K}{L} = \frac{(1-\alpha)w}{\alpha r}$ for all output levels. Substituting this into (4.4), we obtain PWCR as follows:

$$K \leq \lambda \frac{(1-\alpha)}{r} W_0$$

Third, we shall assume that $K \leq W_0$ is always feasible, that is, PWCR cannot be binding as long as the entrepreneur's capital investment does not exceed his wealth endowment. This implies that λ is strictly greater than one. In the above example, that is, $\lambda \geq \frac{r}{1-\alpha} > 1$.[10] In addition, we shall assume that wages are paid at the end of period, and therefore the borrowing requirement is $(K - W_0)$.[11]

Having elaborated PWCR, we now come to characterize the decision problem facing an entrepreneur.

Denote by $s^*(\geq \underline{s})$ "the critical bankruptcy state". $s^* \equiv \underline{s}$ if $(wL + r(K - W_0)) \leq 0$ (i.e., $rW_0 \geq wL + rK$, which can be the case only if $W_0 > K$.) Otherwise, s^* satisfies the following conditions:

$$f(L, K, \theta, s^*) - (wL + r(K - W_0)) = 0 \tag{4.5}$$

$$f(L, K, \theta, s) - (wL + r(K - W_0)) \leq 0 \text{ for all } s \leq s^* \tag{4.6}$$

Equation (4.5) defines

$$s^* = s^*(L, K, w, r, \theta, W_0) \tag{4.7}$$

$$\Phi(s^*) = \Phi(s^*(L, K, w, r, \theta, W_0)) \tag{4.8}$$

Then, under the assumption of non-negative consumption and with normalization of $U(0, \rho) = 0$, the entrepreneur's problem is

$$\begin{aligned} &\text{Max}_{(L,K)} \int_{s^*}^{\bar{s}} U\left(f(L, K, \theta, s) - (wL + r(K - W_0), \rho\right) \phi(s) ds \\ &\text{S.T.} \qquad wL + rK \leq \lambda W_0 \end{aligned} \tag{4.9}$$

where $\pi = f(L, K, \theta, s) - (wL + r(K - W_0)$ is the entrepreneurial profit at state s.

Our assumptions about the utility function and the production function guarantee that solutions to (4.9) exist. Denote by L^* and K^* the optimal labour and capital inputs respectively. Then,

$$L^* = L(\theta, W_0, \rho; w, r, \lambda) \tag{4.10}$$

[10] Since we take wealth as capital, the capital market cannot be at equilibrium if λ is not strictly greater than one.

[11] If wages are paid at the beginning of period, the borrowing requirement is $(K + wL - W_0)$.

$$K^* = K(\theta, W_0, \rho; w, r, \lambda) \tag{4.11}$$

Equations (4.10) and (4.11) say that the entrepreneur's demands for labour and capital depend on three individual characteristics (θ, W_0, ρ) and three market parameters (w, r, λ). Their particular relations will be formally identified in the next two sections.

Substituting (4.10) and (4.11) back into the utility function, we obtain the entrepreneur's *indirect* expected utility function as follows:

$$V = EU(\pi, \rho) = V(\theta, W_0, \rho; w, r, \lambda) \tag{4.12}$$

If individual ν chooses to be a worker (passive capitalist), he simply sells his labour and capital endowment in markets to earn contractual returns (wage and rental price) at the end of period. Possibility of bankruptcy incurred by potential buyers (entrepreneurs) implies that his contractual return cannot be *riskless*. So he has to make a decision about to which entrepreneur his labour and capital should be sold in order to maximize his expected utility. Simultaneous modelling of both the entrepreneurial choice and a (passive capitalist) worker's choice within a general equilibrium model is tempting, but less tractable. For analysis to be tractable, we eliminate the problem of a worker's choice by assuming that passive capitalist workers "delegate" their choices to a diversified wage insurance company and a diversified financial intermediary (bank) so that contractual return for each individual are riskless.[12] This assumption also implies that all passive capitalist workers earn the same wage rate and the same interest rate. With the above arguments in mind, we write a passive capitalist worker's certain utility as follows[13]:

$$U = U\left(w + rW_0, \rho\right) \tag{4.13}$$

In a competitive market, individual ν makes his decision of being an entrepreneur or a worker taking (w, r, λ) given. He will choose to be an entrepreneur if and only if

$$V(\theta, W_0, \rho; w, r, \lambda) \geq U\left(w + rW_0, \rho\right) \tag{4.14}$$

He will be a worker if and only if the inequality in (4.14) is reversed.

[12]This assumption is not implausible at least for capital markets. In reality, most passive capitalists deposit their financial assets into banks who in turn lend funds to entrepreneurs. It is banks who impose PWCR on borrowers. In addition, since workers have a priority claim in case of bankruptcy in most cases, risk related to wages is small, compared to risk related to loans.

[13]The participation constraint of the insurance companies and the banks implies that the wage and the interest rate paid by entrepreneurs should be higher than received by workers and passive capitalists. Technically this is equivalent to a taxation levied on riskless wages and interest rates. Because this does not affect the main arguments, we ignore it.

Because (w, r, λ) are identical for all individuals in equilibrium, the decision to be an entrepreneur or a worker is entirely determined by individuals' personal characteristics (θ, W_0, ρ).

Let E be the set of entrepreneurs and Z the set of workers defined by

$$(i)\ E = \{(\theta, W_0, \rho) :\ V(\theta, W_0, \rho; w, r, \lambda) \geq U\ (w + rW_0, \rho)\} \tag{4.15}$$

$$(ii)\ Z = \{(\theta, W_0, \rho) :\ V(\theta, W_0, \rho; w, r, \lambda) < U\ (w + rW_0, \rho)\} \tag{4.16}$$

Denote the joint distribution function of (θ, W_0, ρ) in population by $\Psi = \Psi(\theta, W_0, \rho)$: $[0,\infty) \times [0, \infty) \times [0, \infty) \longrightarrow [0, 1]$. The marginal distributions are denoted by $\Psi^{\theta} = \Psi(\theta)$, $\Psi^{W} = \Psi(W_0)$, and $\Psi^{\rho} = \Psi(\rho)$ respectively. Then equilibrium is a set of (w, r, λ) and a partition (E, Z) of the population, denoted by $\Re = \{(w^*, r^*, \lambda^*); (E, Z)\}$, which satisfies $E \cap Z = \emptyset$ and $E \cup Z = \Omega = [0,\infty) \times [0, \infty) \times [0, \infty)$, such that

$$1 - \iiint_E d\Psi(\theta, W_0, \rho) = \iiint_Z d\Psi(\theta, W_0, \rho) = \iiint_E L(\theta, W_0, \rho; w^*, r^*, \lambda^*) d\Psi(\theta, W_0, \rho) \tag{4.17}$$

$$\iiint_E K(\theta, W_0, \rho; w^*, r^*, \lambda^*) = \int_0^{\infty} W_0 d\Psi(W_0) \tag{4.18}$$

where (a) $\iiint_E d\Psi(\theta, W_0, \rho)$ is the proportion of entrepreneurs in the population, (b) $\iiint_Z d\Psi(\theta, W_0, \rho)$ is the proportion of workers, (c) $\iiint_E L(\theta, W_0, \rho; w, r, \lambda)$ $d\Psi(\theta, W_0, \rho)$ is *per capita* labour demand by entrepreneurs, (d) $\iiint_E K(\theta, W_0, \rho;$ $w, r, \lambda)$ is *per capita* capital demand by entrepreneurs, and (e) $\int_0^{\infty} W_0 d\Psi(W_0)$ is *per capita* initial personal wealth (capital endowment). Condition (4.17) says that labour supply by workers equals labour demand by entrepreneurs; and condition (4.18) says that capital supply by the population equals capital demand by entrepreneurs.

In a given society at a given time, the joint distribution of (θ, W_0, ρ) as well as individual's characteristics are given. Thus the equilibrium can be reached only through adjustments in the market wage (w) and the market capital price (r) and the personal wealth constraint rule parameter (λ). A different $\Psi(\theta, W_0, \rho)$ will generate a different equilibrium $\Re = \{(w, r, \lambda); (E, Z)\}$. In other words, general equilibrium of entrepreneurial choice is determined by the joint distribution of marketing ability, personal wealth and risk-attitudes.

4.3 The Characterization of the Entrepreneurial Choices

We leave the proof of the existence of equilibrium to next section. In this section, we shall characterize the entrepreneurial choice in equilibrium. What we are concerned with is how, given a parameter set (w, r, λ), an individual's choice of being

an entrepreneur or a passive capitalist worker is related to his personal characteristics (θ, W_0, ρ)? In other words, what makes an entrepreneur?

The first step is to investigate the demand functions for labour and capital (4.10) and (4.11) with respect to (θ, W_0, ρ); that is, *if* individual ν becomes an entrepreneur, how do his optimal demands L^* and K^* depend on his personal characteristics (θ, W_0, ρ)?

First consider how an entrepreneur's optimal demands for labour and capital depend on his personal wealth. To facilitate analysis, define

$$G(L, K) = \int_{\underline{s}}^{\overline{s}} U\left(f(L, K, \theta, s) - (wL + r(K - W_0), \rho\right) \phi(s) ds \tag{4.19}$$

$$H(L, K) = \int_{s^*}^{\overline{s}} U\left(f(L, K, \theta, s) - (wL + r(K - W_0), \rho\right) \phi(s) ds \tag{4.20}$$

where s^* is defined by $f(L, K, \theta, s) - (wL + r(K - W_0)) \equiv 0$.[14]

Then the objective function is

$$EU(L, K) = \begin{cases} G(L, K) \text{ for } wL + rK \leq rW_0 \\ \\ H(L, K) \text{ for } wL + rK \geq rW_0 \end{cases} \tag{4.21}$$

Let (L^G, K^G) maximize (4.19), and (L^H, K^H) maximize (4.20). Denote by (L^{u*}, K^{u*}) the unconstrained demands (that is, (L^{u*}, K^{u*}) which maximize (4.21) when PWCR is not imposed). We have

Lemma 22 *Assume that Assumptions 1–3 hold and individuals have constant absolute risk-aversion. Then, if $H(L, K)$ has a unique local maximum point, (i) there is an W_0' such that $(L^{u*}, K^{u*}) = (L^G, K^G)$ for all $W_0 \geq W_0'$, and $(L^{u*}, K^{u*}) = (L^H, K^H)$ for all $W_0 < W_0'$, where $(L^H, K^H) > (L^G, K^G)$; (ii) (L^H, K^H) is decreasing with W_0 for all $W_0 < W_0'$.*

Proof See appendix A.

Figure 4.1 is a diagrammatic demonstration of Lemma 22. The lemma says that the entrepreneur's *unconstrained* demands for labour and capital are decreasing with his personal wealth until W_0' after which $(L^{u*}, K^{u*}) \equiv (L^G, K^G)$ with a constant Arrow–Pratt measure of absolute risk aversion.[15] More importantly, the unconstrained demands when $W_0 < W_0^{'}$ are unambiguously greater than when $W_0 \geq W_0^{'}$. The intuition is that the expected marginal cost and the expected marginal benefit of labour and capital depend *asymmetrically* on his personal wealth unless $W_0 \geq W_0'$. For an entrepreneur with $W_0 < W_0'$, an increase in labour and capital input at the margin have two different effects: the effect on the marginal product and the effect

[14] Technically, $s^* \leq 0$ is allowed.

[15] Note that the constant absolute risk-aversion is sufficient but not necessary for the result.

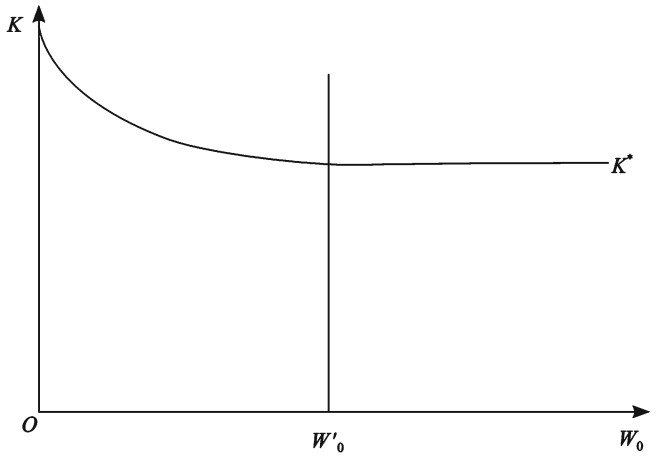

Fig. 4.1 The unconstrained optimal investment

on probability of bankruptcy. The first effect is negative and the second effect is positive. This can be seen from the second derivatives of the objective function with respect to capital (labour)[16]:

$$\frac{\partial^2 EU}{\partial K^2} = -\frac{\partial s^*}{\partial K}(f_K(s^*) - r)\phi(s^*) + \int_{s^*}^{\bar{s}} f_{KK}\phi(s)ds \tag{4.22}$$

The second term of (4.22) is standard and strictly negative. The first term can be zero only if there is no bankruptcy, which can be the case only if W_0 is sufficiently large. For a relatively small W_0, implicitly differentiating (4.5)

$$\frac{\partial s^*}{\partial K} = -\frac{(f_K(s^*) - r)}{f_{s^*}} \tag{4.23}$$

Therefore the first term of (4.22) is positive

$$-\frac{\partial s^*}{\partial K}\left(f_K(s^*) - r\right)\phi(s^*) = \frac{\phi(s^*)}{f_{s^*}}\left(f_K(s^*) - r\right)^2 > 0 \tag{4.24}$$

This positive effect partially offsets the negative effect on the marginal product such that the entrepreneur's optimum implies a larger demand than when only the negative effect occurs. Furthermore, since the positive effect is decreasing with personal wealth, the gap between (L^{u*}, K^{u*}) and (L^G, K^G) decrease with W_0 and vanishes at $W_0 = W_0'$.

[16]Here we take a risk-neutral case.

This provides further legitimacy for the liquidity constraint. It shows that if the access to capital markets were not restricted by personal wealth endowments, the total capital would be exhausted by less wealthy people.[17] Since $G(L, K)$ generates no externality and is actually the objective function of an social planner, the argument also shows that the constraint is socially desirable. This leads us to a story different from the traditional one which says that the individual investment is smaller than the social optimum when the borrowers cannot borrow as much as they like at given market interest rates. They cannot borrow as much as they like because they would like to borrow too much![18]

We now turn to consider the constrained demands, that is, when PWCR is imposed.

Theorem 23 *Assume that Assumptions 1–3 hold, individuals have constant absolute risk-aversion and $H(L, K)$ has a unique local maximum point. Then, for given θ and ρ, there is $W_0^b (< W_0')$, such that: (i) PWCR is binding if and only if $W_0 \leq W_0^b$; (ii) (L^*, K^*) are increasing with W_0 for all $W_0 \leq W_0^b$; (iii) (L^*, K^*) are first decreasing and then constant with W_0 for $W_0 \geq W_0^b$.*

The proof is quite straightforward. Since the unconstrained demands are decreasing with W_0 while the feasible set defined by PWCR is expanding with W_0 from zero, there must be a switch point W_0^b. $W_0^b < W_0'$ comes from the assumption that $K \leq W_0$ is always feasible. That is, the switch point of the PWCR not binding always comes before the switch point of no-bankruptcy. With Lemma 22, this fact implies (ii) and (iii). This is illustrated in Fig. 4.2.[19]

The proceeding analyses have important implications for the relationships between the size of the firm, the entrepreneur's personal wealth and the binding of the PWCR. If the PWCR is not imposed, low-wealth entrepreneurs would run bigger firms than high wealth entrepreneurs. However, imposition of the PWCR will greatly complicate the picture. With the PWCR, the bigger firms are run by some "moderately" rich entrepreneurs.

Theorem 24 *Under Assumptions 1–3, L^* and K^* can be decreasing or increasing with θ, depending on the balance of three effects: the effect on probability of bankruptcy, the effect on the marginal product and the risk-aversion effect.*

Proof The sign of $\frac{\partial K^*}{\partial \theta}$ when the PWCR is not imposed depends on the sign of the derivative of the first-order condition with respect to θ:

[17] The argument also applies to labour markets. Without personal wealth constraint, the less wealthy people would hire more workers.

[18] A good example is the "investment hunger" phenomenon commonly observed in socialist countries where individual benefits from investment, and the cost of failure of investment are asymmetric.

[19] In the present model. the relationships between personal wealth and the demands for labour and capital are not monotonic. However, in reality, since some personal wealth is not productive but can be used as collateral, the switch point of non-bindingness may come after the switch point of non-bankruptcy point. If this is a case, the demands are monotonically increasing until W_0'.

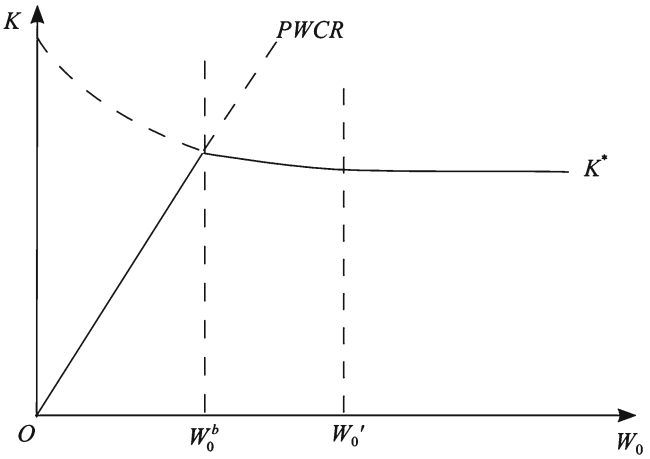

Fig. 4.2 The constrained optimal investment

$$\frac{\partial^2 EU}{\partial\theta\partial K} = -\frac{\partial s^*}{\partial\theta}U_\pi(s^*)(f_K(s^*)-r)\phi(s^*) + \int_{s^*}^{\overline{s}}(U_{\pi\pi}f_\theta(f_K-r)+U_\pi f_{\theta K})\,\phi(s)ds \tag{4.25}$$

The first term of (4.25) is the effect on probability of bankruptcy, which is either negative or zero, depending on W_0. The first term of the integral is the risk-aversion effect, which cannot be positive since $U_{\pi\pi} \leq 0$. The second term of the integral is strictly positive by assumptions. Therefore the sign of (4.25) is indeterminate. □

Theorem 24 implies that a higher ability entrepreneur does not necessarily run a bigger firm even without consideration of the PWCR. The intuition is that although an increase in ability raises the marginal product of labour and capital which has a positive effect on L^* and K^*, it also raises the probability of states over which the entrepreneur has liability to pay the contractual wage and interest to workers and lenders and therefore raises the effective cost of labour and capital. In addition, since the increase in ability raises the expected residual of given labour and capital inputs, a risk-averse entrepreneur may prefer to invest less rather than more. However, among those who are risk-neutral ($U_{\pi\pi} = 0$) and are sufficiently wealthy such that $s^* = \underline{s}$, it is unambiguous that high ability entrepreneurs run bigger firms than low ability entrepreneurs since in this case the two negative effects disappear. This also implies that ability and the size of the firm are more likely to be positively correlated for the more wealthy and less risk-averse entrepreneurs than for the less wealthy and more risk-averse entrepreneurs. Furthermore, since s^* is lower-bounded by $\underline{s}$, we shall expect that the net effect of ability on the size of the firm is more likely to be positive for very high ability entrepreneurs, unless the PWCR becomes binding in which case the constrained optimal labour and capital inputs are constant with ability.

Theorem 25 *Under Assumptions 1–4, L^* and K^* are non-increasing with ρ.*

Proof The proof is to show that the unconstrained first-order derivatives are non-increasing with ρ.

$$\begin{aligned}
\frac{\partial^2 EU}{\partial\rho\partial K} &= \int_{s^*}^{\overline{s}} U_{\pi\rho}(\pi,\rho)(f_K - r)\phi(s)ds \\
&= \int_{s^*}^{\overline{s}} \left(\frac{U_{\pi\rho}(\pi,\rho)}{U_\pi}\right) U_\pi(f_K - r)\phi(s)ds \\
&= \left[\left(\frac{U_{\pi\rho}(\pi,\rho)}{U_\pi}\right)\left(\int_{s^*}^{s} U_\pi(f_K - r)\phi(s)ds\right)\right]_{s^*}^{\overline{s}} \\
&\quad - \int_{s^*}^{\overline{s}} \frac{\partial}{\partial\rho}\left(\frac{\partial \ln U_\pi}{\partial\pi} f_s\right)\left(\int_{s^*}^{s} U_\pi(f_K - r)\phi(s)ds\right)\phi(s)ds \\
&= \int_{s^*}^{\overline{s}} \frac{\partial}{\partial\rho}\left(-\frac{\partial \ln U_\pi}{\partial\pi} f_s\right)\left(\int_{s^*}^{s} U_\pi(f_K - r)\phi(s)ds\right)\phi(s)ds \le 0
\end{aligned}$$

since $\frac{\partial}{\partial\rho}\left(-\frac{\partial \ln U_\pi}{\partial\pi} f_s\right) \ge 0$ and $\int_{s^*}^{s} U_\pi(f_K - r)\phi(s)ds < 0$ for all $s < \overline{s}$. Note that to derive the second equality, we made use of integration by parts.[20]

If, for instance, (L', K') is the optimal solution for a less risk-averse entrepreneur, a reduction in (L, K) from (L', K') will reduce the risk premium more than expected profit decreases for a more risk-averse entrepreneur. Therefore more risk-averse entrepreneurs hire less labour and capital inputs. □

Theorem 25 implies that for given W_0 and θ, less risk averse entrepreneurs will run bigger firms than the more risk averse. This is also a result of Kihlstrom and Laffont (1979). With a combination of Theorems 24 and 25, it might be safe to say that the firms run by high ability and less risk averse entrepreneurs would not be smaller than the firms run by the low ability and the more risk-averse for given W_0. The effect of the PWCR on the relationship between L^* and K^* and θ and ρ is straightforward: L^* and K^* are insensitive to θ and ρ when the PWCR is binding. Note that bindingness of the PWCR depends on not only W_0, but also θ and ρ. From Lemma 22, for given θ and ρ, the low wealth entrepreneur is more likely to be constrained not only because his budget line is low but also because his unconstrained demand is too high. Theorem 25 implies that, for given W_0 and θ, the PWCR is more likely to be binding for high ability and less risk-averse entrepreneurs than for the low ability and more risk-averse.

After characterizing the entrepreneur's demand functions, we now turn to analyze the choice of an individual of being an entrepreneur or a worker. To facilitate the analysis, we define the "entrepreneurial utility rent" Π as follows:

$$\Pi = V(\theta, W_0, \rho; w, r, \lambda) - U(rW_0 + w, \rho) \tag{4.26}$$

[20]This is actually an application of Theorem 4 in Diamond and Stiglitz (1974). With our notation, their theorem says that if increases in ρ represent increases in risk aversion, then the control variable L^* (K^*) decreases (increases) with ρ if there exists a s' such that $\pi_L = (f_L - w) \le (\ge)0$ for $s \le s'$ and $\pi_L = (f_L - w) \ge (\le)0$ for $s \ge s'$. In the present model, the first order conditions guarantee that such a s' does exist.

where $V(.)$ is defined by (4.12). Then an individual will choose to be an entrepreneur if and only if his entrepreneurial utility rent is greater than or equal to zero; otherwise he will be a worker. Characterizing the entrepreneurial choice is equivalent to finding a function $\Pi(\theta, W_0, \rho) \equiv 0$.

Define $\nu^0 = \Pi^{-1}(0) = \{(\theta, W_0, \rho) : \Pi(\theta, W_0, \rho) \equiv 0\}$. We call ν^0 "the hypersurface of marginal entrepreneurs". Letting $(\theta^0, W_0^0, \rho^0)$ be a point in ν^0, we have

Theorem 26 *(i)* $\Pi(\theta, W_0^0, \rho^0) \geq 0$ *if and only if* $\theta \geq \theta^0$*; (ii)* $\pi(\theta^0, W_0^0, \rho) \geq 0$ *if and only if* $\rho \leq \rho^0$*; and (iii) for given* θ^0 *and* ρ^0*: (a) if there exist two different* W_0*, denoted by* W_0^{0-} *and* W_0^{0+} *respectively, where* $W_0^{0-} < W_0^{0+}$*, such that* $\Pi(\theta^0, W_0^{0-}, \rho^0) = \Pi(\theta^0, W_0^{0+}, \rho^0) = 0$*,* $\Pi(\theta^0, W_0, \rho^0) \geq 0$ *for all* $W_0 \in [W_0^{0-}, W_0^{0+}]$*, and* $\Pi(\theta^0, W_0, \rho^0) \leq 0$ *for all* $W_0^{0-} \geq W_0 \geq W_0^{0+}$*; and (b) if there exists only one* W_0^0 *satisfying* $\Pi(\theta^0, W_0^0, \rho^0) = 0$*,* $\Pi(\theta^0, W_0, \rho^0) \geq 0$ *for all* $W_0 \geq W_0^0$.

Proof (i) The proof is equivalent to showing that $\Pi(.)$ is a non-decreasing function of θ. Differentiating (4.26) with respect to θ, we obtain

$$\frac{\partial \Pi}{\partial \theta} = \frac{\partial V}{\partial \theta} + \frac{\partial V}{\partial L^*}\frac{\partial L^*}{\partial \theta} + \frac{\partial V}{\partial K^*}\frac{\partial K^*}{\partial \theta} \tag{4.27}$$

Since $\frac{\partial V}{\partial \theta} = \frac{\partial EU}{\partial \pi}\frac{\partial f}{\partial \theta} > 0$, $\frac{\partial V}{\partial L^*} = \frac{\partial V}{\partial K^*} = 0$ if PWCR is not binding, and $\frac{\partial L^*}{\partial \theta} = \frac{\partial K^*}{\partial \theta} = 0$ once the PWCR becomes binding, we have

$$\frac{\partial \Pi}{\partial \theta} = \frac{\partial V}{\partial \theta} > 0 \tag{4.28}$$

(ii) See Kihlstrom and Laffont (1979).

(iii) The claim can be best shown with a diagram in (W_0, θ) space.

From Theorem 14 of Chap. 3 and Lemma 22 of this chapter, we know that the critical marketing ability θ^0 without PWCR is an increasing function of personal wealth W_0 until W_0', as represented by the curve AA' in Fig. 4.3. PWCR implies that $\Pi(.)$ cannot be positive unless W_0 is greater than a minimum, denoted by W_0^{0-}, such that $(L, K) = \{(L, K) : wL + rK = \lambda W_0^{0-}\}$ are sufficiently profitable. By part (i), W_0^{0-} is a decreasing function of θ, as represented by the curve BB' (see later discussion for deriving BB'). Then clearly, a given θ^0 may correspond to two different W_0^0s. If this is a case, $\Pi(\theta^0, W_0, \rho^0) \geq 0$ for all $W_0 \in [W_0^{0-}, W_0^{0+}]$, and $\Pi(\theta^0, W_0, \rho^0) \leq 0$ for all $W_0^{0-} \geq W_0 \geq W_0^{0+}$. On the other hand, if θ^0 is corresponding to only one $W_0^0 (= W_0^{0-})$, $\Pi(\theta^0, W_0, \rho^0) \geq 0$ for all $W_0 \geq W_0^0$. □

Roughly speaking, Theorem 25 says that in equilibrium high-ability/more-wealthy/less-risk-averse individuals become entrepreneurs while low-ability/less-wealthy/more-risk-averse individuals become workers. However, part (iii) implies that this statement may not be universally true. It is possible that, within the group of individuals whose marketing ability belongs to some "moderate" range, only those who are "moderately" wealthy become entrepreneurs, while neither the relatively poor nor the relatively rich become entrepreneurs.

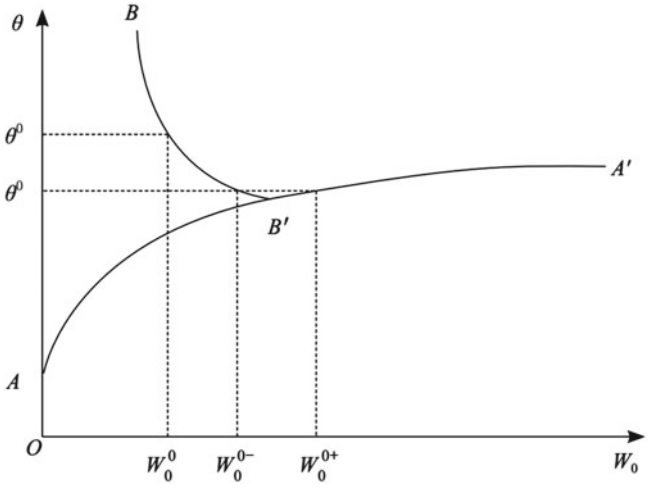

Fig. 4.3 The marginal entrepreneurs

More importantly, Theorem 25 implies that the set of (marginal) entrepreneurs consists of different types in terms of personal characteristics. In particular, there are some substitutions between the three personal characteristics in selecting a marginal entrepreneur. For example, if we take a marginal entrepreneur characterized by "mean" ability, "moderate" risk-aversion, and "average" personal wealth as a representative marginal entrepreneur, we will find that many marginal entrepreneurs deviate from the representative: one with less-than-moderate risk aversion but lower-than-mean ability, another with higher-than-mean ability but more-than-moderate-risk aversion, and the third one with higher-than-mean ability but less-than-average wealth. The purpose of the following analysis is to present a more rigorous demonstration of the set of entrepreneurs and particularly the hypersurface of marginal entrepreneurs. Because of complexity, we restrict our demonstration to two-dimensional spaces.

First consider the combination of marketing ability and degree of risk-aversion. In (θ, ρ), for a given W_0 (assumed big enough that there exist some θ and ρ with $\Pi \geq 0$), $\Pi \geq 0$ defines a set of entrepreneurs, denoted by $E(\theta, \rho)$. First, there must exist a $\theta_{\min}$ for all $\rho \geq 0$ and a $\rho_{\max}$ for all $\theta \geq 0$ such that

$$\{(\theta, \rho) : \theta < \theta_{\min} \cup \rho > \rho_{\max}\} \notin E(\theta, \rho) \tag{4.29}$$

Equation (4.29) is quite intuitive. It says that an individual will never choose to be an entrepreneur once either his marketing ability is lower than a minimum level or his risk-aversion index is greater than a maximum level. $\theta = 0$ or $\rho = \infty$ is such an example. Second, $\Pi = 0$ defines a marginal entrepreneur curve which has the slope

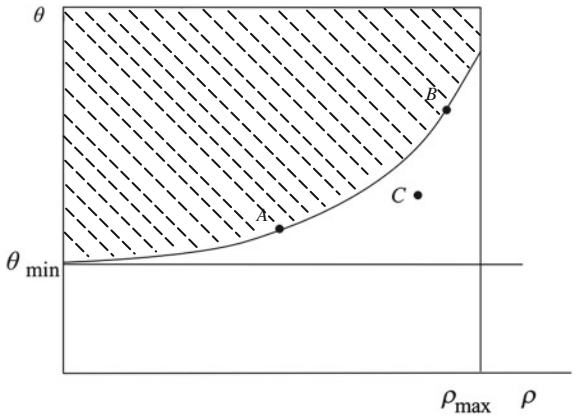

Fig. 4.4 The set of entrepreneurs in (θ, ρ) space

$$\frac{d\theta^0}{d\rho^0} = -\frac{\frac{\partial \Pi}{\partial \rho^0}}{\frac{\partial \Pi}{\partial \theta^0}} \geq 0 \tag{4.30}$$

since $\frac{\partial \Pi}{\partial \rho^0} \leq 0$ and $\frac{\partial \Pi}{\partial \theta^0} > 0$. Equation (4.30) says that there is some substitution between marketing ability and risk-attitudes in making a marginal entrepreneur: along the marginal curve, a more risk-averse entrepreneur has higher marketing ability (in other words, the critical marketing ability for an individual to choose to be an entrepreneur increases as his risk-aversion degree increases.) The intuition is that to be indifferent between being an entrepreneur and being a worker, more risk-averse people require higher expected returns, which can only be met by a higher marketing ability. In Fig. 4.4, both A and B are marginal entrepreneurs, but B's ability is much higher than $A's$ because B is much more risk averse than A; C's ability is higher than A's and he is also less risk-averse than B, but he is not an entrepreneur because, compared to A, his ability is not high enough to "compensate" his higher risk-aversion, and compared with B, his risk-aversion is not low enough to "compensate" his lower ability.

Second, consider the set of entrepreneurs in (θ, W_0) space, defined by $E(\theta, W_0) = \{(\theta, W_0) : \Pi \geq 0 \text{ for a given } \rho \leq \rho_{\max}\}$. In Fig. 4.5, $E(\theta, W_0)$ is a half-open space lower-bounded by AA' and BB'. The AA' curve is the *un*constrained marginal entrepreneur curve based on Theorem 14 of Chap. 3, and the curve BB' is the constrained marginal entrepreneur curve which is derived as follows. First, for a given marketing ability θ (assumed sufficiently large), there is a minimum personal wealth requirement W_0^{0-} depending on θ such that $\Pi \geq 0$ if and only if $W_0 \geq W_0^{0-}$ (probably just locally). Second, since at W_0^{0-}, PWCR is binding

$$\frac{d\theta^0}{dW_0^{0-}} = -\frac{\frac{\partial \Pi}{\partial W_0}}{\frac{\partial \Pi}{\partial \theta}} < 0 \tag{4.31}$$

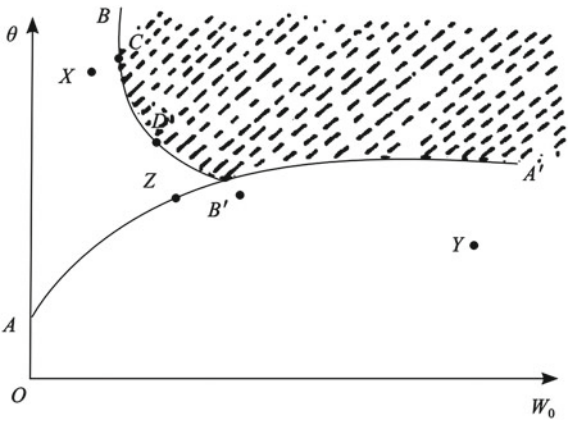

Fig. 4.5 The set of entrepreneurs in (θ, W_0) space

that is, BB' is negatively sloped. The intuition is that compared with a low ability person, a high ability person needs less investment to "break even" in being an entrepreneur and therefore requires lower personal wealth for a given λ. In the figure, both C and D are marginal entrepreneurs, but C's ability is higher than D's while D is richer than C.

Figure 4.5 illustrates that both marketing ability and personal wealth are essential for an individual to be an entrepreneur. Someone may be excluded from entrepreneurship, either because of low ability or because of low personal wealth or both. For example, in the figure, X is good at marketing, but because he is not wealthy enough, he cannot be an entrepreneur; on the other hand, Y is rich, but because his marketing ability is not high enough, he will not choose to be an entrepreneur; and Z is short of both. The figure also shows that for given personal wealth, a higher ability would-be entrepreneur is more likely to survive the PWCR than a low ability one.

It is worth emphasizing that although X, Y and Z all become workers, the mechanism behind their occupational choices are different. Y is a *voluntary* worker out of self-selection, while X and Z are *compulsory* workers because of PWCR. Although both X and Z are excluded from entrepreneurship, changes in personal wealth may generate different effects on the choice of each of them. An increase in W_0 may switch X from *compulsory* worker into *constrained* entrepreneur, while it just transfers Z from *compulsory* worker into *voluntary* worker. The reason is that X is a superior would-be entrepreneur, who *failed* to become an entrepreneur only because he is poor and whose incentive will remain even when he becomes rich, while Z is an inferior one, who *wants* to be an entrepreneur only because he is poor and whose incentive to be an entrepreneur would disappear once he becomes rich.

The effect of risk-attitudes on the entrepreneurial choice can also be incorporated into the (θ, W_0) space. This is shown in Fig. 4.6. An increase in ρ will shift AA' curve upwards and BB' curve rightwards because the more risk-averse person requires higher ability to compensate.

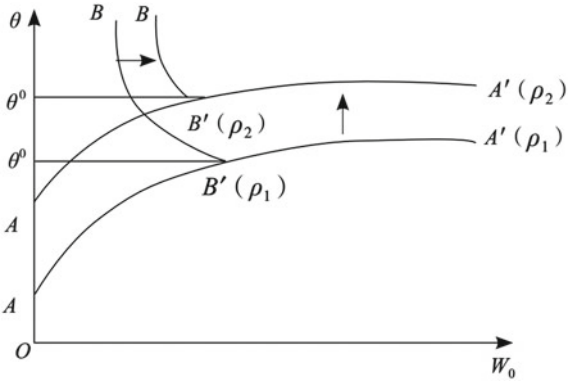

Fig. 4.6 The effect of risk-aversion on the set of entrepreneurs

4.4 The Existence of Equilibrium

In this section we shall prove that a general equilibrium exists. The first step is to investigate the entrepreneur's demand functions $L^*(w, r, \lambda)$ and $K^*(w, r, \lambda)$, and how the hypersurface of marginal entrepreneurs changes in response to changes in (w, r, λ). The results are established in the following lemma.

Lemma 27 *(i) L^* is decreasing with w, K^* is decreasing with r, and both L^* and K^* are increasing or non-decreasing with λ. (ii) Both θ^0 and W_0^{0-} are increasing with (w, r) and non-increasing with λ; ρ^0 is decreasing with (w, r) and non-decreasing with λ.*

Proof (i) The proof of (i) for w and r is a standard exercise in microeconomics.[21] For λ, when the PWCR is not binding, a small change in λ will not affect L^* and K^*; if the PWCR is binding, an increase in λ (i.e., easing the constraint) will increase both L^* and K^*, and a decrease in λ (i.e., tightening the constraint) will decrease both L^* and K^*.

(ii) Given Theorem 26 of the last section, it suffices to show that the entrepreneurial rent Π is a decreasing function of w and r, and an increasing (or non-decreasing) function of λ. It is easy to see that $U(rW_0 + w, \rho)$ is increasing with both w and r and independent of λ.

$$\begin{aligned}\frac{\partial V}{\partial r} &= \int_{s^*}^{\bar{s}} U_\pi(.)\left(\left(\frac{\partial \pi}{\partial K}\frac{\partial K^*}{\partial r} + \frac{\partial \pi}{\partial L}\frac{\partial L^*}{\partial r}\right) - (K^* - W_0)\right)\phi(s)ds \\ &= \int_{s^*}^{\bar{s}} U_\pi(.)\left(\frac{\partial \pi}{\partial K}\frac{\partial K^*}{\partial r} + \frac{\partial \pi}{\partial L}\frac{\partial L^*}{\partial r}\right)\phi(s)ds - \int_{s^*}^{\bar{s}} U_\pi(.)(K^* - W_0)\phi(s)ds\end{aligned} \tag{4.32}$$

[21] When PWCR is binding, the problem is similar to a consumer choosing to maximize his expected utility subject to his budget constraint. An increase in w (r) has two effects: "income" effect and substitution effect. Since both labour and capital are normal goods, the arguments hold. However, we cannot say much about the cross effect of changes in w (r).

If the PWCR is not binding,

$$\begin{aligned}&\int_{s^*}^{\bar{s}} U_\pi(.)\left(\frac{\partial \pi}{\partial K}\frac{\partial K^*}{\partial r}+\frac{\partial \pi}{\partial L}\frac{\partial L^*}{\partial r}\right)\phi(s)ds\\&=\frac{\partial K^*}{\partial r}\int_{s^*}^{\bar{s}} U_\pi(.)(f_K-r)\phi(s)ds+\frac{\partial L^*}{\partial r}\int_{s^*}^{\bar{s}} U_\pi(.)(f_L-w)\phi(s)ds=0\end{aligned} \tag{4.33}$$

since, from the first-order conditions,

$$\int_{s^*}^{\bar{s}} U_\pi(.)(f_K-r)\phi(s)ds=\int_{s^*}^{\bar{s}} U_\pi(.)(f_L-w)\phi(s)ds=0 \tag{4.34}$$

Therefore

$$\frac{\partial V}{\partial r}=-\int_{s^*}^{\bar{s}} U_\pi(.)(K^*-W_0)\phi(s)ds\le 0 \text{ if } K^*\ge W_0 \tag{4.35}$$

In the case of $K^*<W_0$, $\frac{\partial V}{\partial r}=\int_{\underline{s}}^{\bar{s}} U_\pi(.)(W_0-K^*)>0$. But

$$\begin{aligned}\frac{\partial \Pi}{\partial r}=\frac{\partial V}{\partial r}-\frac{\partial U}{\partial r}&=\int_{\underline{s}}^{\bar{s}} U_\pi(.)(W_0-K^*)\phi(s)ds-\int_{\underline{s}}^{\bar{s}} U_\pi(.)W_0\phi(s)ds\\&=-K^*\int_{\underline{s}}^{\bar{s}} U_\pi(.)\phi(s)ds<0\end{aligned} \tag{4.36}$$

If the PWCR is binding, (4.34) does not hold (strictly greater than zero). However, since the capital-labour ratio is always optimal, the following condition holds:

$$f_L=\left(\frac{w}{r}\right)f_K \tag{4.37}$$

Substituting (4.37) into (4.32), we have

$$\frac{\partial V}{\partial r}=\left(\frac{\partial K^*}{\partial r}+\left(\frac{w}{r}\right)\frac{\partial L^*}{\partial r}\right)\int_{s^*}^{\bar{s}} U_\pi(.)(f_K-r)\phi(s)ds-\int_{s^*}^{\bar{s}} U_\pi(.)(K^*-W_0)\phi(s)ds \tag{4.38}$$

Since when PWCR is binding, $wL^*+rK^*\equiv\lambda W_0$, the following condition holds

$$w\frac{\partial L^*}{\partial r}+K^*+r\frac{\partial K^*}{\partial r}=0 \tag{4.39}$$

Substituting (4.39) into (4.38),

$$\frac{\partial V}{\partial r}=-\frac{K^*}{r}\int_{s^*}^{\bar{s}} U_\pi(.)(f_K-r)\phi(s)ds-\int_{s^*}^{\bar{s}} U_\pi(.)(K^*-W_0)\phi(s)ds<0 \tag{4.40}$$

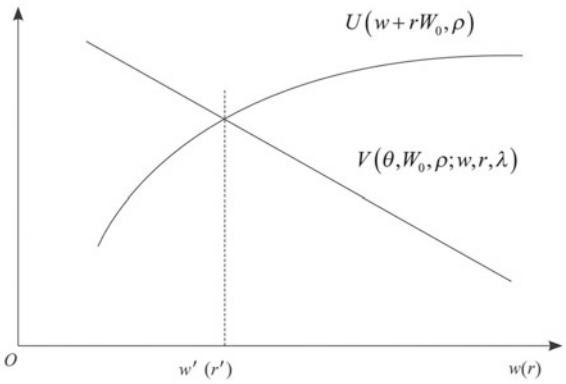

Fig. 4.7 The utilities of being a worker and of being an entrepreneur with $w(r)$

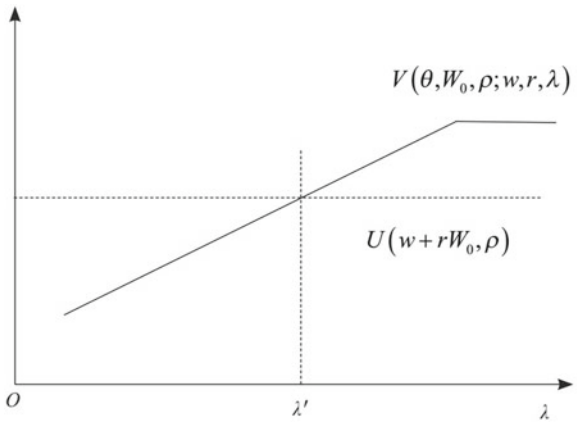

Fig. 4.8 The utilities of being a worker and of being an entrepreneur with λ

Similarly, we can show that $\frac{\partial V}{\partial w} < 0$ and therefore $\frac{\partial \Pi}{\partial w} < 0$.

For λ, when the PWCR is binding, an increase in λ will raise both L^* and K^* and therefore V because L^* and K^* are less than the unconstrained optimal levels. When PWCR is not binding, V is independent of λ at margin. Therefore Π is increasing or non-decreasing with λ.

The above arguments are shown in Figs. 4.7 and 4.8. In Fig. 4.7, utility from being a worker is increasing with $w\,(r)$, and utility from being an entrepreneur is decreasing with $w(r)$; an individual becomes an entrepreneur only if $w \leq w'(r \leq r')$. In Fig. 4.8, utility from being a worker is constant with λ and utility from being an entrepreneur is first increasing and then constant with λ; an individual will become an entrepreneur only if $\lambda \geq \lambda'$. □

Theorem 28 *For a given joint distribution function* $\Psi(\theta, W_0, \rho)$*, there exists at least an equilibrium* (w, r, λ)*, denoted by* (w^*, r^*, λ^*)*, such that*

$$1 - \iiint_E d\Psi(\theta, W_0, \rho) = \iiint_Z d\Psi(\theta, W_0, \rho) = \iiint_E L(\theta, W_0, \rho; w^*, r^*, \lambda^*) d\Psi(\theta, W_0, \rho) \tag{4.41}$$

$$\iiint_E K(\theta, W_0, \rho; w^*, r^*, \lambda^*) d\Psi(\theta, W_0, \rho) = \int W_0 d\Psi^W(W_0) \tag{4.42}$$

Proof Lemma 27 (i) says that $L(\theta, W_0, \rho; w, r, \lambda)$ and $K(\theta, W_0, \rho; w, r, \lambda)$ are a decreasing function of w and r respectively, and increasing or non-decreasing functions of λ; Lemma 27 (ii) says that the set of entrepreneurs E is shrinking as w and r increase and expanding as λ increases, and the set of workers Z is expanding as w and r increase and shrinking as λ increases. Therefore $\iiint_E L(\theta, W_0, \rho; w, r, \lambda) d\Psi(\theta, W_0, \rho)$ (the per capita labour demand) is a decreasing function of w and an increasing function of λ, and $\iiint_Z d\Psi(\theta, W_0, \rho)$ (the per capita labour supply) is an increasing function of w and a decreasing function of λ; $\iiint_E K(\theta, W_0, \rho; w, r, \lambda) d\Psi(\theta, W_0, \rho)$ (the per capita capital demand) is a decreasing function of r and an increasing function of λ (note: $\int W_0 d\Psi^W(W_0)$, the per capita capital supply, is constant). In addition, the substitution between labour and capital for any given λ implies that the capital-labour ratio will be adjusted in response to relative changes in w and r. Hence there must exist a vector (w^*, r^*, λ^*) which partitions the whole population into an entrepreneurial set $E = \{(\theta, W_0, \rho) : \Pi(\theta, W_0, \rho; w^*, r^*, \lambda^*) \geq 0\}$ and a worker set $Z = \{(\theta, W_0, \rho) : \Pi(\theta, W_0, \rho; w^*, r^*, \lambda^*) < 0\}$ such that (4.41) and (4.42) hold. □

The intuition for the theorem is that changes in the market wage, the capital price or the PWCR parameter λ affect not only how much labour and capital each entrepreneur hires, but also how many people will become entrepreneurs. Disequilibrium can occur either because the entrepreneur set E is too big or too small, or because the capital-labour ratio demanded is inconsistent with the supply ratio, or both. In the first case, if E is too big, there is excess demand in both the labour market and the capital market, w and r will go up and λ will go down, which will drive some marginal entrepreneurs into Z (and also reduce intra-marginal entrepreneurs' demand for labour and capital); if E is too small, there are excess supplies in both labour and capital markets, w and r will go down and λ will go up, which drives some marginal workers into E (and also increases intra-marginal entrepreneurs' demands for labour and capital). In the second case, λ may remain unchanged (remember we assume that the PWCR has symmetric effects on labour and capital), while w will go up and r will go down if labour market is in excess demand and capital market is in excess supply, or w will go down and r will go up if otherwise. Adjustment in the third case is much more complicated. One possibility is that although there are too many entrepreneurs in E, the labour market is in excess supply. This happens only when w is too high while r is too low. Then adjustment may proceed in the following way: λ goes down to drive some marginal entrepreneurs into Z; w goes down and r goes up such that all remaining entrepreneurs increase demand for labour and reduce demand for capital. The adjustment will continue until equilibrium obtains.

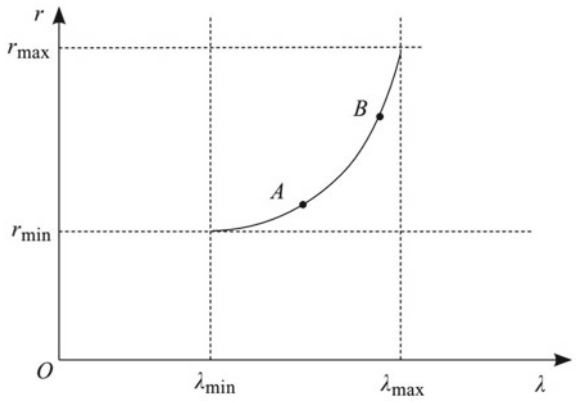

Fig. 4.9 Different equilibria

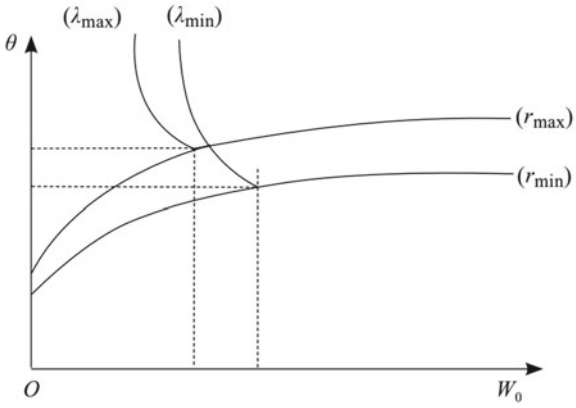

Fig. 4.10 Different sets of entrepreneurs

Is the equilibrium unique? More than likely, it is not. The reason is that because there are only two factors but three "prices", there may exist different combinations of (w, r, λ) consistent with the equilibrium. In particular, within limits, substitution between r and λ may exist. This is shown in Fig. 4.9, where r^* is assumed to be positively related to λ^*. $\lambda^*_{\min} > 1$ is necessary for any capital market to exist; λ^* is upper-bounded by $\lambda^*_{\max}$ since otherwise the personal wealth constraint would not exist. In the interval of $[\lambda^*_{\min}, \lambda^*_{\max}]$, any point on $r^*(\lambda^*)$ curve may be possible. $r^*(\lambda^*)$ is upward sloping because an increase in λ^* would generate an excess demand for capital which can be offset only by an increase in r^*. In the figure, economies A and B are assumed identical in the sense that $\Psi_A(\theta, W_0, \rho) \equiv \Psi_B(\theta, W_0, \rho)$, but A's equilibrium is different from B's. Since a marginal entrepreneur can be selected by different characteristics and different characteristics are differently sensitive to r and λ, different equilibria (w^*, r^*, λ^*) may generate different partitions of the population into entrepreneurs and workers, —although the differences between them cannot be very big. In Fig. 4.10, two extreme cases of equilibria are drawn. The entrepreneurial

set associated with $(r^*_{\min}, \lambda^*_{\min})$ is different from that associated with $(r^*_{\max}, \lambda^*_{\max})$. In particular, $(r^*_{\min}, \lambda^*_{\min})$ favors personal wealth while $(r^*_{\max}, \lambda^*_{\max})$ favors marketing ability. The intuition is that a high-ability-but-less-wealthy individual is more likely to be excluded from entrepreneurship, in the economy with a PWCR.[22]

4.5 Comparative Statics

The above analysis shows that for a given $\Psi(\theta, W_0, \rho)$, there exists at least one equilibrium $\Re = \{(w^*, r^*, \lambda^*); (E^*, Z^*)\}$ such that entrepreneurs' demands for labour and for capital are equal to workers' labour supply and social capital stock. Since the equilibrium $\Re$ represents not only a partition of the entrepreneur set and worker set, but also the relationship between the entrepreneurs and workers, between labour and capital, and between ability and wealth, an interesting question is: how does a change in the joint distribution function $\Psi(\theta, W_0, \rho)$ affect $\{(w^*, r^*, \lambda^*); (E^*, Z^*)\}$? We now turn to this question. Because of complexity of the problem, the following analysis is very preliminary and informal; a more satisfactory analysis has to be postponed to future work.

Theorem 29 *Assume that θ, W_0 and ρ are independently distributed. Then, (i) an improvement in ability distribution Ψ^θ in the first-order stochastic dominance sense will increase w^* and r^*, reduce λ^* and move the E frontier against ability; (ii) an improvement in personal wealth distribution Ψ^W in the first-order stochastic dominance sense will reduce r^*, increase w^* and λ^*, and move the E frontier in favour of ability; (iii) an economy-wide increase in ρ will reduce w^* and r^*, and increase λ^*.*

Proof (i) Under independence,

$$\psi(\theta, W_0, \rho) = \psi(\theta)\psi(W_0)\psi(\rho) \tag{4.43}$$

and the equilibrium conditions become

$$\iiint_E L(\theta, W_0, \rho; w, r, \lambda)\psi(\theta)\psi(W_0)\psi(\rho)d\theta dW_0 d\rho = \iiint_Z \psi(\theta)\psi(W_0)\psi(\rho)d\theta dW_0 d\rho \tag{4.44}$$

[22] In reality, the difference in equilibria across economies is more or less a reflection of the difference of the degree of information asymmetry of ability. The model predicts that an economy with more asymmetric information of ability is closer to $(r^*_{\min}, \lambda^*_{\min})$, while an economy with less asymmetric information of ability is closer to $(r^*_{\max}, \lambda^*_{\max})$. This means that an improvement of information will improve the position of high-ability-but-less-wealthy individuals.

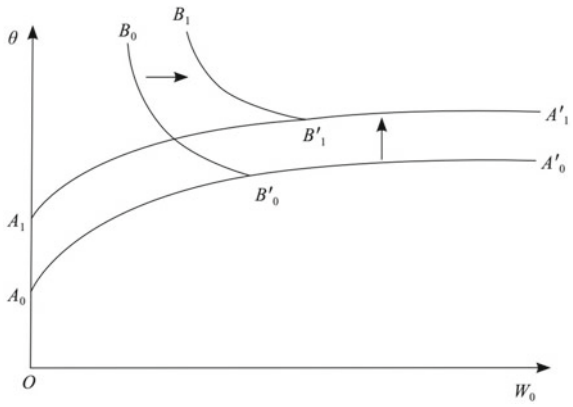

Fig. 4.11 Comparative statics (1)

$$\iiint_E K(\theta, W_0, \rho; w, r, \lambda)\psi(\theta)\psi(W_0)\psi(\rho)d\theta dW_0 d\rho = \int W_0\psi(W_0)dW_0 \tag{4.45}$$

An improvement in $\Psi(\theta)$ implies that the proportion of low-θ population falls and the proportion of high-θ population rises. Because the average θ of the set E is greater than the average θ of the set Z, an improvement in $\Psi(\theta)$ under independence of distributions implies that for a given original equilibrium (w^*, r^*, λ^*), the left hand sides of (4.44) and (4.45) increase, while the right hand side of (4.44) decreases and the right hand side of (4.45) is unchanged. That is, the original equilibrium has been violated. For equilibrium to be restored, w^* and r^* must go up and λ must go down such that either each individual entrepreneur's demands go down, or some marginal entrepreneurs drop out, or both.

To see changes of E, note that a decrease in λ^* shifts BB' curve (representing the PWCR) rightwards, and increases in w^* and r^* shift AA' curve upwards. Hence, following the improvement in the distribution of ability, the set of entrepreneurs "shrinks". In particular, Fig. 4.11 shows that both the marketing ability and personal wealth of a marginal entrepreneur increase. This makes it more difficult for high-ability-less-wealthy people to become entrepreneurs. However, it is ambiguous whether the proportion of entrepreneurs in the population rises, is constant or falls. (See Fig. 4.11)

(ii) Given the original equilibrium (w^*, r^*, λ^*), an improvement in the distribution of personal wealth generates an excess supply in the capital market and an excess demand in the labour market. To see this, first note that because the average W_0 of the set E is higher than the average W_0 of the set Z, as the proportion of high W_0 people goes up, the proportion of entrepreneurs will go up and the proportion of workers will go down.[23] In the capital market, following the improvement in $\Psi(W_0)$, the right hand side of (4.45) goes up (i.e., the total capital supply increases). For given (w^*, r^*, λ^*),

[23]The argument may not hold without the PWCR and independence assumptions.

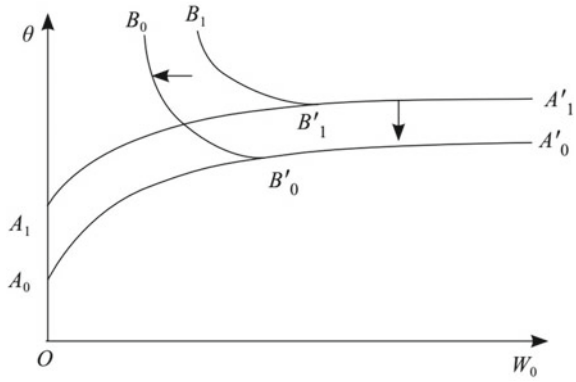

Fig. 4.12 Comparative statics (2)

the left hand side of (4.45) also goes up. However, because some high W_0 people are not in the set E, the increase in LHS must be smaller than the increase in RHS. This implies that the capital market has excess supply. To restore the equilibrium, r^* must go down, and w^* and λ^* must go up.

The effect of the improvement in $\Psi(W_0)$ on E can also be demonstrated diagrammatically. In Fig. 4.12, BB' curve moves backwards because λ^* goes up; AA' curve moves downwards because r^* goes down. Therefore E expands such that θ and W_0 for identifying a marginal entrepreneur fall. That is, high-ability people are less constrained by personal wealth in being entrepreneurs.

(iii) An economy-wide increase in risk-aversion implies that for given (w^*, r^*, λ^*), $\Psi(E)$ decreases and $\Psi(Z)$ increases. This generates excess supply both in the labour and capital markets. For the equilibrium conditions to be satisfied, w^* and r^* go down, and λ^* goes up. □

The underlying mechanism for adjustment is as follows. (i) In equilibrium, an individual can become an entrepreneur only if his ability exceeds a certain level (conditional on his personal wealth and risk-aversion). As the number of high-ability people increase, more people enter the entrepreneurial market and compete for capital and labour. This will push up the market wage and the capital price, which in turn drives some former marginal entrepreneurs out of the entrepreneurial market. Because capital is not increased, as more high-ability would-be entrepreneurs seek to borrow, capitalists will require more collateral to ensure repayment. This implies that relative increases in ability will benefit both workers and capitalists. (ii) On the other hand, if capital increases while the ability of the population remains unchanged, ability becomes more scarce. Competition between capitalists will push the capital price down, and λ and the entrepreneurial rent up. This will induce some marginal workers (previously constrained by their personal wealth) to switch to become entrepreneurs and also increase existing entrepreneurs' demand for labour. Therefore the market wage will go up. (iii) Because on average entrepreneurs are less risk-averse than workers, an economy-wide increase in risk-aversion will reduce the proportion

of entrepreneurs and increase the proportion of workers. This tends to lower the equilibrium wage and the equilibrium capital price and raise the equilibrium λ^*.[24]

The above arguments are based on the assumption of independence of distributions. This assumption may not hold in reality. When distributions are not independent, the effect of a change in the joint distribution function is much more complicated and not easy to analyze.

However, one thing that might be certain is the effect on λ^* of correlation between marketing ability and personal wealth. Denote by $\gamma_{\theta W_0} \in [0, 1]$ the correlation coefficient.[25] Then our basic rationale for personal wealth constraint implies that

$$\frac{\partial \lambda^*}{\partial \gamma_{\theta W_0}} \leq 0 \tag{4.46}$$

That is, as the correlation between marketing ability and personal wealth in the population increases, the PWCR will be tighter and therefore it is more difficult for high-ability but less wealthy people to become entrepreneurs. There are two explanations for this. First, an increase in the correlation makes wealth more *informative* about ability so that the maximum borrowing one can get is more dependent on personal wealth. For instance, if $\gamma_{\theta W_0} = 0$, the probability of a poor person being a high ability person is just the same as the probability of a rich person being a high ability person; on the other hand, if $\gamma_{\theta W_0} = 1$, the first probability is 0, while the second probability is one. Clearly, a less wealthy person is more likely to get external funds in the first case than in the second case. A second explanation for (4.46) is that as more rich people become high ability, more capital will be invested by wealth holders and therefore less will be left to the other users. It is interesting to note that an increase in the individual's marketing ability will lessen the constraint for him to be an entrepreneur while an increase in economy-wide ability will worsen the constraint; on the other hand, both increases in the individual's personal wealth and in national wealth will relax the constraint.[26]

4.6 Discussions

If we assume that all individuals are identical in marketing ability θ and personal wealth W_0 and different only in risk-attitude ρ, and in addition, assume that production does not depend on capital and $wL \leq W_0$ holds, (4.17) reduces to

[24] However the effect is not so strong as in Kihlstrom and Laffont (1979) because of the feedback effect. As w and r decrease and λ increases, some marginal workers because of the personal wealth constraint or ability constraint will switch to become entrepreneurs.

[25] We exclude the possibility of negative correlation.

[26] An implication of the above argument is that if marketing ability can be improved to some extent, development of public education (funded by the state) will lessen the wealth constraint for an individual to become an entrepreneur, provided that it has symmetric effects on both economy-wide ability and economy-wide wealth.

$$1 - \Psi^{\rho}(E) = \Psi^{\rho}(Z) = \int_{E} L(\rho, w) d\Psi(\rho)$$

This is the equilibrium condition of the Kihlstrom–Laffont model (1979) in which the partition $\{E, Z\}$ of population is identified by a cut-off point of ρ such that $E = \{\rho : \rho \leq \rho^0\}$ and $Z = \{\rho : \rho \geq \rho^0\}$, where ρ^0 is the cut-off point.

Now assume that all individuals are risk-neutral but different in marketing ability, and assume that $\lambda = \infty$. Then (4.17) and (4.18) reduce to

$$1 - \Psi^{\theta}(E) = \Psi^{\theta}(Z) = \int_{E} L(\theta, wr) d\Psi(\theta)$$

$$\int_{E} K(\theta, w, r) d\Psi(\theta) = \int W_0 d\Psi(W_0)$$

This is the equilibrium condition of the Lucas model (1978) where the partition $\{E, Z\}$ of population is identified by a cut-off point of θ such that $E = \{\theta : \theta \geq \theta^0\}$, and $Z = \{\theta : \theta \leq \theta^0\}$, where θ^0 is the cut-off point.

If individuals differ in their marketing ability but no one knows his own θ and each takes the distribution of θ to be the relevant risk, and if they are identical in ρ, (4.17) reduces to

$$1 - \Psi^{\theta}(E) = \Psi^{\theta}(Z) = \Psi^{\theta}(E) \int L(\theta) d\Psi(\theta)$$

This is Kanbur's (1979) equilibrium condition. In the Kanbur model, equilibrium does exist, but entrepreneurs are picked at random because nobody minds whether he is an entrepreneur or a worker, given there is no private information about ability. (This is no longer relevant when Kanbur turns to "heterogeneous risk-attitudes", where, as in the Kihlstrom–Laffont model, there is a cut-off point of ρ.)

Finally, if we are only concerned with testing whether the liquidity constraint is binding for would-be entrepreneurs, we can simply assume that individuals are risk-neutral and λ is a fixed parameter. Then from PWCR, it is easy to see that a high ability person is more likely to be a constrained entrepreneur because his unconstrained optimal capital investment is higher. This is the result of Evans–Jovanovic (1989).

4.7 Cooperation Between Ability and Wealth: "Professional Managers"

In the above model, it has been assumed that there are only two occupations for an individual to choose: either being an entrepreneur or being a worker. Although "capitalists" have been mentioned from time to time, they are not a separate set. Because there are only two occupations to choose, a capitalist has been either an

entrepreneur or a worker. Although we have made a distinction between *active* capitalists and *passive* capitalists, a capitalist has never been *active* unless he becomes an entrepreneur. Of course, the personal wealth constraint for being an entrepreneur has been assumed to be implicitly imposed by capitalists (as well as workers). However, even in this case, capitalists are quite passive because their role is simply superseded by a parameter λ which is determined in markets. A worker-capitalist lends his capital endowment to any entrepreneur-capitalist whose PWCR is satisfied; an entrepreneur-capitalist can borrow capital from any worker-capitalist within his PWCR. In addition, for convenience of analysis, we have assumed that there is a continuum of capitalists distributed between zero to huge personal wealth. Our basic arguments are first that an entrepreneur must be a wealthy capitalist but a wealthy capitalist will not necessarily be an entrepreneur; second, that an entrepreneur must be endowed with high marketing ability but a high ability person will not necessarily be an entrepreneur.

The underlying assumption for the above arguments is that personal wealth is public information while marketing ability is not. We have assumed that all outsiders are equally ignorant of a person's marketing ability. We have also excluded any possibility for an outsider to acquire knowledge of a particular person's ability through search activities. This seems too strict and unrealistic. In reality, among outsiders, some may be more informed of a particular person's ability than others; and some incomplete knowledge of a particular person's ability may be acquirable through costly communication. In this section, we take into account this fact to discuss cooperation between high-ability-less-wealthy people and low-ability-more-wealthy people, and the occurrence of professional managers.

Let us start with the personal wealth constraint rule and its effect on the partition of occupations. The main function of the PWCR is to exclude inferior candidates from entrepreneurship. However, the PWCR is double-edged. This can be shown diagrammatically. In Fig. 4.13, AA' curve divides the population into two sets: the

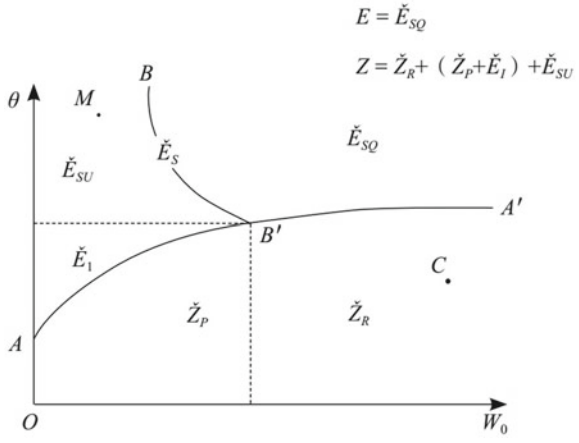

Fig. 4.13 The effect of personal wealth on the classification of population

would-be entrepreneur set $\check{E}$ and the would-be worker set $\check{Z}$. Let $\theta^{\sharp}$ be the "social" critical marketing ability which divides $\check{E}$ into two subsets: set $\check{E}_I$ (inferior) and set $\check{E}_S$ (superior). Let E be the actual entrepreneur set and Z be the complementary set of E. Then, without the PWCR, $E = \check{E}$ and $Z = \check{Z}$. The purpose of the PWCR is to ensure that $\check{E}_I \notin E$. In the figure, the BB' curve successfully excludes $\check{E}_I$ from E. However, the BB' curve also cuts $\check{E}_S$ into two subsets: $\check{E}_{SQ}$ (qualified superior would-be entrepreneurs) and $\check{E}_{SU}$ (*un*qualified superior would-be entrepreneurs). Thus, $E = \check{E}_{SQ}$ and $Z = \check{Z} + \check{E}_I + \check{E}_{SU}$. Furthermore, $\check{Z}$ can be divided into two subsets in terms of personal wealth: $\check{Z}_P$ (poor) and $\check{Z}_R$ (rich). Then $Z = \check{Z}_R + (\check{Z}_P + \check{E}_I) + \check{E}_{SU}$. That is, the non-entrepreneurial set Z contains of three different types: $\check{Z}_R$—rich but low ability; $\check{E}_{SU}$—high ability but poor; $(\check{Z}_P + \check{E}_I)$ —low ability and poor. From now on, people in $\check{Z}_R$ are called " *pure* capitalists" (distinguished from "entrepreneur-capitalists).

There is no way for people in $(\check{Z}_P + \check{E}_I)$ to obtain any entrepreneurial rent. However opportunities may exist for pure capitalists in $\check{Z}_R$ and people in $\check{E}_{SU}$. Take $M \in \check{E}_{SU}$ and $C \in \check{Z}_R$ as examples. M cannot be an entrepreneur because he is poor; C does not want to be an entrepreneur because he is low ability. But if M and C bond themselves together, the personal wealth constraint and the marketing ability constraint for entrepreneurship will no longer bind. Of course, this kind of cooperation can occur only when C has (or is able to obtain) some knowledge about M's ability.

As the result of cooperation, M and C become a "joint entrepreneur". Authority in making business decisions of "what to do and how to do it" is assigned to M for his high marketing ability while authority in selecting "M" is held by C due to his high personal wealth. C's main function is to identify M and to ensure that M belongs to $\check{E}_{SU}$ not $\check{E}_I$. When C decides to "sponsor" M, he in fact signals to outsiders (potential workers and potential lenders) that M is a high ability person at least in his judgment. However, from the point of view of outsiders, C's message can be credible only when he bears an above-average-risk for his selection decision, which provides him a higher incentive to find a high quality manager.[27] This implies that C should be a residual claimant, which in turn implies that C should have some authority to monitor M's activities. M becomes an agent-manager, and C becomes a principal-shareholder.

Cooperation between M and C generates not only the entrepreneurial rent but also agency costs. Agency costs are a summation of all costs associated with cooperation between M and C, which would not occur in the case of a single entrepreneur, including costs of acquiring information of θ_M (but not θ_C), bargaining for division of entrepreneurial rent, and monitoring M's performance and so on. In reality, agency costs may depend on many variables. For example, agency costs would be lower if M is C's brother than if he was just introduced to C by C's brother. Cooperation is profitable only when the entrepreneurial rent exceeds agency costs. However agency costs *per se* are partially endogenous to contractual arrangements concerning

[27] Here the average risk means the risk borne by a lender because of bankruptcy.

distribution of residual claim and control rights between M and C.[28] In particular, since M's activities are not easy to monitor, it is necessary to provides him some residual-claim incentives. This implies that there is a trade-off between providing an incentive for C to select a high quality manager and providing incentives for M to work hard.

Formally, let us denote individuals M and C by $(\theta_M, W_{0M}, \rho_M)$ and $(\theta_C, W_{0C}, \rho_C)$. For simplicity, assume both M and C are risk-neutral (i.e., $\rho_M = \rho_C = 0$). In addition, assume that C has to spend all his time in finding and monitoring M when they cooperate. Then their joint entrepreneurial problem is

$$\begin{aligned} &\text{MAX}_{\{L,K\}}\ \overline{\pi}(L, K) = \int_{s^*}^{\overline{s}} (f(L, K, \theta_M, s) - wL - r(K - (W_{0M} + W_{0C}))\,\phi(s)ds \\ &\text{S.T.} \qquad\quad wL + rK \leq \lambda(W_{0M} + W_{0C}) \end{aligned} \tag{4.47}$$

Denote by L^* and K^* the solutions to (4.47). Then

$$L^* = L(\theta_M, W_{0M} + W_{0C}; w, r, \lambda) \tag{4.48}$$

$$K^* = K(\theta_M, W_{0M} + W_{0C}; w, r, \lambda) \tag{4.49}$$

Note the role of W_{0C} is to relax the personal wealth constraint faced by M such that θ_M can be used for marketing.

Assume that the distribution of the entrepreneurial return between M and C takes the following linear forms:

$$\begin{aligned} W_{1M} &= \alpha_M + \beta(\pi - \alpha_M - \alpha_C) \\ W_{1C} &= \alpha_C + (1 - \beta)(\pi - \alpha_M - \alpha_C) \end{aligned} \tag{4.50}$$

where W_{1M} and W_{1C} are the final returns to M and C respectively; $\alpha_M, \alpha_C \geq 0$ is the fixed term and $0 \leq \beta \leq 1$ is the residual share for M (so $(1 - \beta)$ is the residual share for C).

Denote by c the agency costs. Then c is a function of $(\alpha_M, \alpha_C, \beta)$. Since π is a function of θ_M which depends on C's incentive, π must be a function of $(\alpha_M, \alpha_C, \beta)$ too. An important problem for M and C is to choose $(\alpha_M, \alpha_C, \beta)$ to maximize $(\overline{\pi} - c)$.

If M and C do not cooperate, they both become workers. Their respective certain returns are

$$\begin{aligned} W_{1M} &= w + rW_{0M} \\ W_{1C} &= w + rW_{0C} \end{aligned} \tag{4.51}$$

[28] One might like to assume that the agency costs are a decreasing function of $\varpi = \frac{W_{0M}}{W_{0M}+W_{0C}}$, where ϖ is "equity share" held by M.

Assume that agency costs are directly borne by C. Then M and C will choose to be a joint entrepreneur if and only if the following conditions are satisfied

$$\begin{array}{ll} \bar{\pi} - c & \geq (w + rW_{0M}) + (w + rW_{0C}) = 2w + r(W_{0M} + W_{0C}) \\ \alpha_M + \beta(\pi - \alpha_M - \alpha_C) & \geq w + rW_{0M} \\ \alpha_C + (1 - \beta)(\pi - \alpha_M - \alpha_C) - c & \geq w + rW_{0C} \end{array} \tag{4.52}$$

Otherwise they will choose to be workers. The first condition is the collective condition that the total entrepreneurial return net of agency costs must not be less than the total certain return when they are workers; the last two conditions are individual conditions that each party's entrepreneurial return must not be less than his return as a worker. Since $W_{0C} \geq W_{0M}$, C's *status quo* is higher than M's. This gives C greater bargaining power over the distribution of the total entrepreneurial return.

Denote M's entrepreneurial rent for cooperation by

$$\Pi_M = \alpha_M + \beta(\pi - \alpha_M - \alpha_C) - (w + rW_{0M}) \tag{4.53}$$

and C's entrepreneurial rent from cooperation by

$$\Pi_C = \alpha_C + (1 - \beta)(\pi - \alpha_M - \alpha_C) - c - (w + rW_{0C}) \tag{4.54}$$

Then it is easy to prove that both Π_M and Π_C are increasing with θ_M (and W_{0C} if the PWCR is binding) for a given contract $(\alpha_M, \alpha_C, \beta)$ with $\beta \neq 0, 1$. In particular, $\frac{\partial \Pi_M}{\partial \theta_M}$ is positively related to W_{0C} when the PWCR is binding, and $\frac{\partial \Pi_C}{\partial \theta_M}$ is always positively related to W_{0C}. This implies that M's incentive to seek cooperation with C is positively dependent on his own marketing ability and C's wealth. So is C's. In other words, high-ability people always like to pursue wealthy capitalists, and wealthy capitalists always like to embrace high ability people.

Whether cooperation will occur in equilibrium depends on agency costs most of which are costs of acquiring knowledge of ability of M-type people and monitoring M's performance. Cooperation would not occur if it is prohibitively expensive to acquire knowledge of M's ability and to monitor M's performance.

Assume that agency costs are sufficiently low such that cooperation occurs in equilibrium. Then the population will be partitioned into four sets: entrepreneur set E, the worker set Z, the manager set M, and pure capitalist set C. This is shown in Fig. 4.14. Note that because of the agency cost problem, not all people in $\breve{E}_{SU}$ can become managers. "Reality" is certainly much more complicated; however, this framework can serve as an abstract description of reality.

The equilibrium relations between these four sets are determined by the properties of the joint density function of (θ, W_0, ρ). In particular, in joint entrepreneurship, to what degree the manager looks like an independent entrepreneur in terms of autonomy depends on the relative ratio between ability and capital. To see this, take a simple example. Suppose the economy consists of 100 persons whose marketing ability and personal wealth are independently two-point-distributed: L (low) and

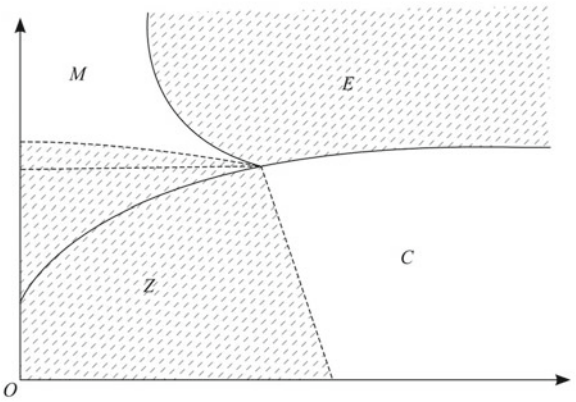

Fig. 4.14 Partition of (E, Z, M, C)

H (high). For simplicity, ignore risk attitudes. Consider the following three cases. Case 1: 10% population with high ability and 10% with high personal wealth. Then, the total high ability persons are 10, the total rich persons are 10, and only one ($10\% \times 10\% = 1\%$) is qualified to be an independent entrepreneur. If the remaining 9 high ability person and 9 rich persons cooperate with each other in pairs, we have 10 firms of which one is run by an independent entrepreneur and the other nine by joint entrepreneurs. Case 2: 10% population with high ability and 20% with high personal wealth. In this case, total high ability persons are 10, total rich persons are 20, and 2 persons ($10\% \times 20\% = 2\%$) are qualified to be independent entrepreneurs; cooperation between the high ability and the rich will create 8 joint entrepreneur firms. Case 3: 20% population with high ability and 10% with high personal wealth. In this case, total high ability are 20, total rich are 10, and 2 persons are qualified to be independent entrepreneurs; cooperation between the high ability and the rich will create more than 8 but less than 18 (depending on investment amount per firm) joint entrepreneur firms. Compared with Case 1 where each capitalist has a managerial partner, in Case 2, 18 capitalists compete for 8 potential managers, whereas in case 3, 18 potential managers compete for 8 capitalists. Obviously, in equilibrium, the managerial position is strongest in Case 2 and weakest in Case 3.

4.8 Concluding Remarks: An Example

Imagine a competitive entrepreneurial market where an individual can either *sell* his marketing ability (intending to be an entrepreneur by hiring workers and capital) or *buy* others' marketing ability (choosing to be a worker and providing capital to the market). All potential participants in the market are identical in their preferences —particularly risk-neutral, and in their producing ability but differ in their marketing ability and personal wealth. They all know that being an entrepreneur implies bearing risk but guarantees a zero income; on the other hand, being

a worker implies earning a *fixed* return which may not be paid in full if the firm's total return fails to reach the total fixed payments promised. Suppose there is a state-owned bank which accepts deposits from capitalists and provides loans to individual entrepreneurs. A capitalist who chooses not to be an entrepreneur can either deposit his personal wealth in the bank for a certain return or lend it directly to individual entrepreneurs for a fixed return with positive probability of default. Imagine that the market area has two large platforms, A and B. Platform A accommodates *would-be* entrepreneurs, and Platform B accommodates *would-be* workers (or passive capitalists). Suppose that there is a certificating officer who stamps each individual's personal wealth value on his face. Now the market is open. We see many people coming in, each with a personal wealth stamp on the face. Some are going to Platform A, and the others to Platform B. Let us look at what kind of people are going to which platform. The first finding might be very confusing because both Platform A and Platform B have all types of people: from penniless to millionaires. However, on reflection, we can discriminate between A and B: it is clear that a person comes to Platform B either because he is so poor at marketing that even in the best state his entrepreneurial return would be very low, or because he is so rich in personal wealth that it is not worthwhile for him to take risks. However, the picture of Platform A is not so transparent. The only certain thing is that the rich people who come to A must be good at marketing —otherwise they would be irrational. A penniless person who comes to A may be really talented or just trying to make a fortune. Nobody knows. All would-be entrepreneurs compete for workers and capital by promising fixed payments. For concreteness, suppose that Mr. W (worker) is on Platform B waiting to be hired, and Mr. R (rich) and Mr. P (poor) are on platform A competing for hiring Mr. W. Mr. P promises to pay Mr. W double what Mr. R promises. With whom should Mr. W make a contract? Mr. W has very strong reason to suspect Mr.P might be a plunger and his promise is not reliable. In contrast, Mr.R is worth trusting. A simple calculation tells him, say, $1 \times 90\% > 2 \times 30\%$. So he accepts Mr. R's offer with little hesitation. Because all Platform-B persons are as rational as Mr. W, all rich would-be entrepreneurs do very well in hiring resources, while few poor would-be entrepreneurs succeed. After the failure of their attempts, some poor would-be entrepreneurs switch to Platform B to be hired workers. However, those whose marketing ability is high may not give up. Suppose our poor Mr.P is such a person. Once he realizes that workers follow capital, he begins to search for a capitalist to help. He may come to Platform B to lobby passive capitalists and say: "Mr. C (capitalist), trust me. I am really good at marketing. You can ask Mr.X and Mrs.Y about me. They know me very well. They are friends of yours. They won't be lying. If you allow me to use your capital, I can make you a lot of money, much more than your current interest paid by the bank. My plan is ...." Mr.C has been convinced. He thinks it is worthwhile taking such a gamble. He withdraws his deposit from the bank and says to Mr.P: "All right, Mr.P. Let us set up a firm. You do the marketing and I will not interfere too much. However, listen! If I find what you just said is not right, I will sack you. Do you understand?" Mr.P knows that under such an agreement he will not be a full entrepreneur like Mr.R but just a employed manager. But he accepts this offer because otherwise he has no chance to do marketing to earn a return higher than

the market wage. When other would-be workers and passive capitalists find Mr.C is willing to finance Mr.P's marketing activities, what comes to their minds is that Mr. P must be very talented, for otherwise, how could Mr.C take such risk? Now Mr. P has no problem in hiring resources, because Mr. C's capital tells others he is reliable. By signaling Mr.P's marketing ability, the capital does not just earn Mr.C a pure profit above the interest rate but also gives him a principalship. Alternatively, the story may proceed in the following way. At the beginning, Mr.C rejects any offer from P-type persons because he does not trust them. So he lends his capital to Mr.R. But he finds Mr.R's own capital earns much more than his capital. He realizes that this is so mainly because some high marketing ability people are not the entrepreneurs because of capital constraints. The temptation to make big money with his capital leads him to Platform A to chat with P-type would-be entrepreneurs. After search, he finds Mr.P might be good at marketing. He says to Mr. P: "Nobody will trust you unless you cooperate with me. You do marketing and I provide capital." ...We may find many pairs of Mr.P and Mr.C emerging in the entrepreneurial market.

Appendix: Proof of Lemma 22

First note that although the integrand $U(.)$ is concave, $H(L, K)$ may not be concave since the lower limit s^* also depends on L and K, and therefore $EU(L, K)$ may not be concave either. However, since s^* is upper-bounded, the optimum must exist.[29] Because of analytical symmetry between L and K, we shall present the proof only for K (all arguments apply to L).

(i) By definition, at $wL + rK = rW_0$, the following equations hold

$$\begin{array}{c} G(L, K) \equiv H(L, K) \\ \frac{\partial G(K)}{\partial K} \equiv \frac{\partial H(K)}{\partial K} \end{array} \tag{A.55}$$

But for all $wL + rK > rW_0$, and therefore $s^* > \underline{s}$,

$$\begin{aligned} \triangle(L, K) = H(L, K) - G(L, K) = -\int_{\underline{s}}^{s^*} U(f(L, K, \theta, s) \\ - (wL + r(K - W_0)), \rho)\phi(s)ds > 0 \end{aligned} \tag{A.56}$$

Combination of (A.55) and (A.56) implies that the curve of $H(K)$ is tangent to the curve of $G(K)$ at $wL + rK = rW_0$ from above, as shown in Fig. 4.15.

For $W_0 \leq \left(\frac{w}{r}\right) L^G + K^G$, since $G(K)$ is strictly concave and at $K = K^G$, the following first-order condition holds

[29] Intuitively, as L and K increase such that $s^* \to 1$, $EU(L, K) \to 0$, and therefore the maximum must be reached for some $s^* < 1$.

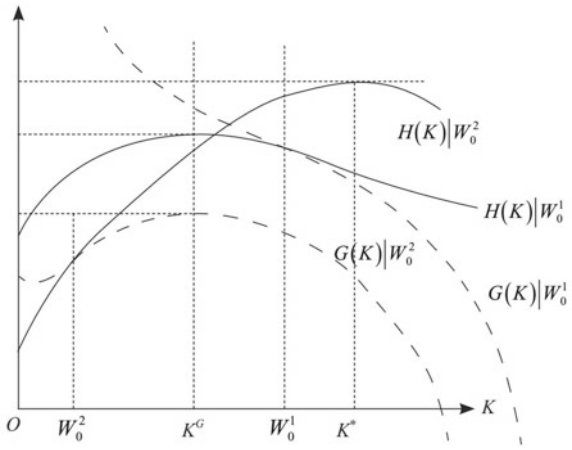

Fig. 4.15 The optimal demand for capital

$$\frac{\partial G}{\partial K} = \int_{\underline{s}}^{\overline{s}} U_{\pi}(.)\left(f_K - r\right)\phi(s)ds = 0 \tag{A.57}$$

we have

$$K^{u*} = K^H \in \arg\max H(K) \tag{A.58}$$

Furthermore, since

$$\frac{\partial \triangle}{\partial K} = -\int_{\underline{s}}^{s^*} U_{\pi}(.)\left(f_K - r\right)\phi(s)ds > 0 \tag{A.59}$$

that is, the gap between $H(K)$ and $G(K)$ is monotonically increasing,[30]

$$K^{u*} = K^H > K^G \in \arg\max G(K) \tag{A.60}$$

For $W_0 \geq \left(\frac{w}{r}\right)L^G + K^G$, at $\left(\frac{w}{r}\right)L + K = W_0$, $\frac{\partial H}{\partial K} = \frac{\partial G}{\partial K} \leq 0$. If $H(K)$ is monotonical for all $\left(\frac{w}{r}\right)L + K \geq W_0$, that is,

$$\frac{\partial H}{\partial K} = \int_{s^*}^{\overline{s}} U_{\pi}(.)\left(f_K - r\right)\phi(s)ds \leq 0 \tag{A.61}$$

then,

$$K^{u*} = K^G \in \arg\max G(K) \tag{A.62}$$

[30]Diagrammatically, the curve of $H(K)$ cannot turned down before $G(K)$.

Although Assumption 1–3 cannot guarantee monotony of $H(K)$ for all $W_0 \geq \left(\frac{w}{r}\right) L^G + K^G$, $U_{\pi\pi} < 0$ and f_{KK} imply that there must exist W_0' such that for all $\left(\frac{w}{r}\right) L + K \geq W_0' \geq \left(\frac{w}{r}\right) L^G + K^G$, (A.61) holds. Part (ii) implies that if (A.61) holds for W_0', it must hold for all $W_0 \geq W_0'$ such that[31]

$$(L^{u*}, K^{u*}) = (L^G, K^G) \in \arg\max G(K) \text{ for all } W_0 \geq W_0' \tag{A.63}$$

(ii)

$$\frac{\partial^2 H}{\partial W_0 \partial K} = -\frac{\partial s^*}{\partial W_0} U_\pi(s^*)(f_K(s^*) - r)\phi(s^*) + \int_{s^*}^{1} U_{\pi\pi}(.)\, r\, (f_K - r)\, \phi(s) ds \tag{A.64}$$

Implicitly differentiating (4.5)

$$\frac{\partial s^*}{\partial W_0} = -\frac{r}{f_{s^*}} < 0 \tag{A.65}$$

Substituting (A.65) into (A.64)

$$\frac{\partial^2 H}{\partial W_0 \partial K} = \frac{r}{f_{s^*}} U_\pi(s^*)(f_K(s^*) - r)\phi(s^*) + \int_{s^*}^{\bar{s}} U_{\pi\pi}(.)\, r\, (f_K - r)\, \phi(s) ds < 0 \tag{A.66}$$

since, under constant absolute risk aversion,

$$\int_{s^*}^{\bar{s}} U_{\pi\pi}(.)\, r\, (f_K - r)\, \phi(s) ds = r \frac{U_{\pi\pi}}{U_\pi} \int_{s^*}^{\bar{s}} U_\pi(.)\, (f_K - r)\, \phi(s) ds = 0$$

Given that the local maximum is unique, inequality (A.66) guarantees the result. □

[31] We can use the argument of part (ii) since its proof does not need the present argument.

Chapter 5
Conclusions

A complete theory of the firm must at least deal with the following three interrelated problems: (i) Why does the firm exist in the first place? (ii) How is principalship (residual claim and authority) assigned among the different members of the firm? (iii) What are the optimal contracts that the principal should use to control agents? While much of the literature on the theory of the firm has so far focused on the first and the third problems, the present thesis is intended to make a contribution to understanding the second problem. The main arguments can be summarized as follows.

The firm is a cooperative organization of different participants (factor-owners). From the point of view of functions, all participants can be grouped into three types of members: the marketing member, producing members and capitalists. The marketing member is to make decisions on "what to do and how do it" (Knight 1921), or "discovering the relative prices" (Coase 1937); the producing members are to execute these decisions by transforming inputs into outputs physically; and the capitalists are to provide capital. Because of the separability property of capital, the capitalists need not stand by their capital and may therefore become "outside members". In contrast, both the marketing member and the producing members are always "inside members". A necessary condition for a capitalists to be an insider is that he also acts either as the marketing member or as a producing member. In other words, an inside capitalist must play dual functions. For obvious reason, we often refer to the marketing member as the decision-maker and the right to do marketing as the decision right.

The importance of marketing comes from uncertainty facing the firm (Knight 1921). In fact, without uncertainty, there would be no need for the firm. Uncertainty makes marketing or decision-making play the dominant role in determining the returns to the firm. The firm is more likely to go bankrupt when it produces a "wrong" product at low cost than when it produces a "right" one at high cost. Although everyone may possess some marketing ability, the observation is that individuals differ in their marketing ability. This is so not just because different people face

W. Zhang, *The Origin of the Capitalist Firm*, DOI 10.1007/978-981-10-0221-2_5

different costs of collecting and processing information, but mainly because marketing ability greatly depends upon the person's "alertness" (Kirzner), "imagination" (Shackle), and "judgment" (Casson). All these personal characteristics are at least partially *innate* and *ineducable*. It is the difference of marketing ability among individuals that creates an opportunity for people to cooperate with each other by setting up a "firm" in which someone who has high marketing ability is responsible for marketing and those who are not good at marketing are responsible for producing. However, the problem is that marketing ability is not an easily observable variable. Given this constraint, for the firm to survive and to be profitable, there must be a mechanism which can ensure that only the sufficiently (if not the most) qualified people will be the marketing members.

The dominance of the marketing member does not mean that the producing members and the capitalists are irrelevant or not important. The return of the firm is a joint stochastic outcome of actions and services supplied by all members. Because of uncertainty and teamwork (Alchian and Demsetz 1972), it is impossible to reward all members with fixed contractual payments corresponding to their respective contributions to the total return. This creates an incentive problem: some party may take an action (e.g., shirking) which benefits himself but costs others. To deal with this problem, there must a mechanism which makes each member as responsible for his actions as possible.

The above two problems interact with each other, because the return of the firm is jointly determined by both ability and actions. The observed organizational structure of the capitalist firm can be understood as an optimal response to these two problems. Briefly, the two problems are solved by assignment of principalship (the residual claim accompanied by authority of monitoring). As the term suggests, the residual claim is an entitlement to claim the residual (the total return minus the contractual payments). Because the contractual payments are independent of the total return (in normal cases), the residual claimant has to bear a risk (responsibility) for the uncertain outcome of all members' actions. In return for this, he is entitled to the authority of monitoring others (Knight 1921).

Our analysis of the optimal assignment of principalship within the firm consists of three steps. In the first step (Chap. 2), we argue that from the incentive point of view, the residual claim should be assigned to the marketing member. This is not only because the marketing member plays the dominant role in determining the residual, but also because his behaviour is more difficult to monitor than others (asymmetry of monitoring).[1] The dominance role implies that the loss of the marketing member's incentive is more costly than that of any other members' incentives, and therefore it pays to sacrifice the latter for the former. The asymmetry of monitoring implies that assigning the residual to the marketing member will incur much lower "aggregated" incentive losses.[2] The two factors together ensure that the welfare loss when the

[1] As we have argued, asymmetry of monitoring is quite intuitive. A glance at the producing members will reveal whether they are working, while a stare at the marketing member may reveal little about what he is thinking.

[2] Let us just repeat the following example to sharpen this argument: Suppose that there is a working team of two person A and B. They work only during the night when the moonlight shines.

marketing member is the residual claimant is lower than when the producing members are the residual claimants. Thus the marketing member becomes the entrepreneur and the producing members become wage-workers.[3] This is the entrepreneurial firm.

Our second step is to demonstrate why priority in being entrepreneurs is given to capitalists, or why entrepreneurship is normally provided by capitalists (Chap. 3). We have shown that, given that the marketing ability is not easily observable, the free choice of occupations implies that there would be too many unqualified people claiming they can do marketing. The reasoning is as follow. Because of the non-negative consumption constraint, the lower-bound net residual, and therefore the net expected return of being the entrepreneur instead of being a worker, are higher when one's personal wealth is lower. This implies that the person with lower personal wealth is more likely to "over-report" his marketing ability than the person with high personal wealth. In other words, insofar as marketing ability is concerned, the rich are more likely to be honest, credible, when they choose to be entrepreneurs. Priority in being the entrepreneur is given to capitalists because the choice of the rich is more informative than the choice of the poor in the sense of signaling marketing ability, such that other people rationally follow capitalist would-be entrepreneurs. This legitimizes the institutional characteristics of the classical capitalist firm: an entrepreneur is also a capitalist and the residual becomes returns to capital. Thus we have an institution called "capital-hiring-labour".

Although our formal models of Chaps. 2 and 3 have been focused on the classical capitalist firm, the theory developed can also explain the occurrence of joint-stock companies in an economy. The argument is as follows (see Sects. 3.7 and 4.7). The function of capital-hiring labour is to exclude inferior candidates from entrepreneurship. However, the capital constraint is double-edged. Because the distribution of ability and the distribution of personal wealth in the population are not symmetric in reality, the capital constraint also excludes those with high ability but low assets from being entrepreneurs. As a result, on the one hand, the capital owned by the high ability people earns its factor price plus a pure profit (rent) from signaling, while the capital owned by the low ability people can earn only its factor price because it has no ability to signal; on the other hand, the high ability of the rich earns entrepreneurial rent, while the high ability of the poor can only earn the market wage. This implies that there is a profitable opportunity for cooperation between the high-ability-low-capital people and the low-ability-high capital people. Although a rich person with low ability cannot make profit by directly marketing, he may increase his return by using his capital to signal someone else's ability, if he knows some high ability people (e.g., his relatives), or if searching for high ability is not too costly. Similarly,

(Footnote 2 continued)
The production technology requires that person A works in the light while person B works in the shadow. The output cannot be attributed to each individual's marginal effort. Then, obviously, it is preferred to let person B claim the residual than person A, because person A cannot see what person A does while person B can easily see whether person A works hard or shirks. In the context of the firm, the marketing member is a worker-in-the-dark, whereas the producing is a worker-in-the-light.

[3]Here following Knight (1921), we understand that the entrepreneur has dual functions: making decisions and bearing risks.

although the high ability-low wealth person cannot make the entrepreneurial rent independently, he may be able to increase his return by using his ability for some capitalists if he can convince the latter that he is a high ability person if the costs of convincing them is not too high. Furthermore, the incentive for each party to search for the other party is an increasing function of their respective resources (ability or wealth), because the more personal wealth (marketing ability) someone has, the more rent he can earn, if search is successful. As a result, they become a *joint* entrepreneur: the high ability person is called the manager by doing marketing, and the rich are called "shareholders" by claiming the residual and taking the responsibility for selection of the qualified manager. This is the origin of the joint-stock company.

In the third step we set up a general equilibrium entrepreneurial model of the firm to show the properties of the equilibrium partition of the population into four different occupations (entrepreneurs, managers, pure capitalists and workers) and to link the equilibrium relationship between different members of the firm with the joint distribution function of marketing ability, personal wealth and risk attitudes in the population (Chap. 4). We have show that in equilibrium, (a) individuals with high ability, high personal wealth and low risk-aversion become entrepreneurs, (b) individuals with low ability, low personal wealth and high risk-aversion become workers, (c) individuals with high ability but low personal wealth become managers hired by capitalists, and (d) individuals with low ability but high personal wealth become "pure" capitalists to hire managers. We have also shown that: (a) an improvement in marketing ability of the population benefits wealthy capitalists and workers but disadvantages high ability-low wealth people (professional managers), (b) an improvement in personal wealth distribution will favour high ability-low wealth people as well as workers but disfavour wealthy capitalists, and (c) an economy-wide increase in risk-aversion will reduce both the market wage and the interest rate and therefore harm both workers and wealthy capitalists.

In its complete form, the contractual arrangements of the firm are that the capitalist is the principal who "delegates" the decision rights to the manager (called the agent) who in turn "employs" workers (sub-agents); the relationships between them become "the principal designs a incentive scheme (or monitoring mechanism) to induce (or enforce) the agents to work in the best interest of the principal, subject to the condition that the agent will not move to a different principal or simply quit (participation constraint) and will voluntarily do what the principal wants him to do because he cannot do better (incentive compatibility constraint)".[4] This disguises many insights about the capitalist firm and its original morphology. Economists have seldom asked why the capitalist should be the principal, which has led to many confusions in understanding the evolution of the institutional structure of the firm.

Our theory has a very important implication. Since advantages of capital over labour result from asymmetry of information on marketing ability, we may predict that these advantages will be diminishing as other signals about ability become available. Education is one such signal, which may reveal some information on

[4]This is the basic framework of the principal-agent literature. See Hart and Holmstrom (1987) for a survey.

marketing ability and therefore help some MBA-holders to become managers.[5] In the extreme case, if information on ability becomes perfect, capital would become a pure production factor and would lose all its advantages over labour, and capitalists would be deprived of principalship.[6] However, if we believe that marketing ability is some kind of innate ability which is not entirely educable, capital will still enjoy advantages over labour in signaling information about a person's ability. If someone is unhappy with the capitalists' social status, he should ask the government to do something else (such as improving education) to make individuals' ability more socially observable, rather than ask the government to wipe out capitalists by nationalization.[7]

The theory presented in this thesis is quite abstract. But much of our argument is based on intuitions about the observed historical evolution of capitalist economies. We started with a simple explanation of why the firm develops from division of labour between a marketing member and a producing member, explained why the marketing member gets the status of the entrepreneur and the producing member becomes a wage-earner, explored why the entrepreneur is also a capitalist, and ended up with a rationale for the occurrence of joint-stock companies. This story seems quite consistent with the history of capitalist firms. While, as Stiglitz pointed out, in modeling economic relationships, if you improve in one area you may sometimes make things worse overall, we hope our theoretical models will improve economists' *overall* understanding of the institution of capitalist firms by offering insights on the origin of the capitalist firm and on its evolution.

It should be pointed out that although our major purpose is to explain the asymmetric contractual arrangements of the capitalist firm, the theory throws light on other observed firms. For instance, the argument developed in Chap. 2 can explain why a partnership firm is more likely to be preferred in industries where members of the firm are equally important in production and are equally difficult to monitor (see Sect. 2.4); and the argument developed in Chap. 3 implies that "labour-hiring-capital" is more likely to be preferred in industries where dominant ability is easy to identify by some easily observable signals such as education certificates. We believe that law, accountancy, consultancy and academic research are such industries where partnership/labour-managed firms are common.

[5] Education is a signal not necessarily because it improves one's ability, but because the cost of education is lower for high ability people than for low. See Spence (1973).

[6] But in this case, nobody has any advantages over others in marketing, and thereafter the firm itself becomes redundant in Coase's sense.

[7] My research work on this topic was partly motivated by my personal experience of socialist China where because of lack of capitalists, most management posts of the state-owned enterprises were occupied by lemons. Fortunately, more and more capitalists are emerging in China as the reform program proceeds, which is certainly helpful in improving the average quality of managers.

So much has been done. However, there is still much left to be done. In particular, we have not yet formally modeled contracts between shareholders and managers in joint-stock companies. As we have pointed out, the contracts must deal with the trade-off between providing the manager with incentives to work and providing the shareholders with incentives to select high ability managers. We believe that many of observed characteristics of these contracts can be attributed to this problem. Given that most of the existing literature is focused on the incentive problem on the manager's side, it is very important to emphasize the incentive problem on the shareholders' side. Future research will be conducted along with this line. We offer the following informal arguments to conclude the thesis.

The cooperation between ability and capital (or separation of the decision right from the residual right) of a joint-stock company is accompanied by several agency-type problems. First, because of imperfection in observation as well as the cost of revelation of ability, a capitalist inevitably makes some mistakes in picking a manager. Someone who was initially thought to have high ability may prove a lemon as the cooperation proceeds! If this is the case, a chance should be given to the capitalist to correct his mistake (of course, correction of the mistake can only minimize rather than eliminate the cost of the mistake, otherwise nobody cares about mistakes). The mistake can also occur the other way: a high-ability manager may be blamed for being a lemon by the capitalist's misjudgment. Because sacking a manager sends on average bad news about his ability, the high-ability manager will be unfairly harmed. There should be a mechanism to prevent the manager from such mis-treatment. Second, because of the importance of marketing activities and the difficulty of monitoring them, there is a serious incentive problem on the manager's side. This suggests that the managerial payment should be more closely linked to the performance of the firm, rather than fixed by contract. In other words, the manager should share some residual! Thirdly, when the capitalist is an outside member of the firm, capital itself is more vulnerable to abuse.[8] Because abuse of capital can benefit the manager in various ways, it is necessary for the capitalist to have some voice in respect of the use of capital. Fourthly, when capital demand is large, the shareholders will be many. This creates an incentive problem of monitoring on the capitalist side, because the cost of monitoring is concentrated while the benefit of monitoring is spread. There should be some mechanisms to mitigate this problem.

How serious the above problems are depends on the degree of overlapping of the decision right with the residual claim, which can be defined as the percentage of the manager's own stake in the equity capital, and which is in turn affected by availability and effectiveness of other mechanisms. The manager holding a higher stake can always mitigate the agency problem.[9] If there are no other effective mechanisms

[8]Capital abuse by management can take various forms, one of which is "overinvestment" for career concerns (See Holmstrom and Ricart 1986).

[9]The agency theory literature of the capital structure initiated by Jensen and Meckling (1976) decomposes the agency problems into the conflicts between the manager who holds less than 100% share and outside shareholders and the conflicts between shareholders and debt-holders. It is argued that the capital structure is determined by minimization of the agency costs. For a comprehensive survey of this literature, see Harris and Raviv (1991).

available, the entrepreneurial firm (full overlapping of management with the residual claim) would be the only form of the firm. The pervasiveness of firms with a low degree of overlap of marketing and capital suggests that there are indeed such mechanisms. Control by a board of directors ("voting-with-hands") and the stock market ("voting-with-feet") are identified as two major mechanisms in dealing with the agency problems. They are complementary but also substitutes. The decision to replace the incumbent by "voting-with-hands" is generally based on the score (share price) from "voting-with-feet". An efficient stock market surely makes the direct control less important. The analogy is that frequent patrol by the police makes the prison less crowded! The stock market is not only a mechanism to constrain managerial behaviour but also a mechanism to constrain capitalist behaviour. For instance, transferability of shares ensures that the capitalist can easily correct his mistakes in judging the manager's ability, while inability to withdraw real capital can protect the high ability manager from unfair assessment by a share-holder; the market valuation of stocks not only values the performance of the manager, but also values the performance of the shareholders. The replacement of management is often preceded by replacement of the shareholders; the shareholders are harmed before the manager. It is the shareholders' responsibility to select a talented and industrious manager. If they do not pay for their careless mistakes, who pays? The evidence of a strong correlation between the managerial payment and the firm's performance suggests that the actual residual stake held by the manager is more than proportional to his nominal stake.[10]

[10]For a survey and synthesis, see Rosen (1992).

Appendix A
A Principal-Agent Theory of the Public Economy and Its Applications to China

A.1 Introduction

This paper analyzes the principal-agent relationship and its associated monitoring-incentive problems under the public ownership economy. Section A.2 characterizes the public ownership economy as a dual hierarchical principal-agent chain between original principals (owners) and ultimate agents (managers). Section A.3 analyzes the effect of the degree of publicness and the size of economy on monitoring incentives of the original principals and work incentives of the ultimate agents in a canonical public economy. Here "the canonical public economy" is one in which the original principals are the residual claimants of the firm, "the degree of publicness" is defined by the number of the original principals, and "the size of the public economy" is defined by the number of public-owned enterprises. Our basic result is that both the monitoring incentive of the original principals and the work incentive of the ultimate agents decrease with the degree of publicness and the size of economy. This result implies that one cannot make simple analogy between the public enterprises owned by small population (such as Singapore's SOEs or Chinese TVEs) and those owned by large population (such as Chinese SOEs). The degree of publicness matters! In the extreme, if a "state" consisted of a single person, there would be no difference between the "public" and the "private". A naive argument which echoes very often in China is that "Singapore's state-owned enterprises (such as Singapore Airline) are efficient, and so Chinese state-owned enterprises can also be efficient." The result also implies that it makes a difference how many enterprises the public owns: the more, the worse. One cannot make analogy between, say, French state enterprises and Chinese state enterprises. The underlying reason behind this result is that increases in the degree of publicness and the size of economy increase layers

The paper was published in *Economics of Planning*, Vol. 31: 231–251, 1998. The author is very grateful for Donald Hay and James Mirrlees for their helpful comments on the early version of this paper.

W. Zhang, *The Origin of the Capitalist Firm*, DOI 10.1007/978-981-10-0221-2

of the hierarchy and therefore the distance between the original principals and the ultimate agents, which makes monitoring less effective. After analyzing the canonical public economy in Sects. A.3 and A.4 turns to a corrupt public economy where the agents (bureaucrats), instead of the principals, claim *de facto* the residual. Our analysis shows that a corrupt public economy can Pareto-dominate a canonical public economy in which the original principals are the residual claimants. In other words, in the public economy, corruption can be a Pareto improvement over noncorruption. The reason is that when the agents claim the residual, the total cost of monitoring is reduced, and work incentive is enhanced. This finding can explain why all socialist economies have evolved into corrupt economies: bureaucrats afford to buy out. In Sect. A.5, we apply our theory to the Chinese economy and analyze the following four questions: (1) Why are the TVEs more efficient than the SOEs? (2) How has the fiscal contract system boomed the Chinese economy? (3) How has the shifting of the residual claim and decision rights from the central government to the management improved the performance of the SOEs? (4) How has corruption promoted high growth of the Chinese economy?

This paper is related to three kinds of literature: the hierarchical theory, the principal-agent theory and the theory of economic transition. Qian (1994) generalizes the hierarchical models of Williamson (1967), Calvo–Wellisz (1978, 1979) and Keren–Levhari (1979), with a focus on the incentive problem within the hierarchy. The current paper also focuses on the incentive problem. However, since what we are concerned with is the hierarchical structure of the whole economy of a socialist country, rather than that of a single firm of a capitalist economy, the optimal design of the hierarchy is not much relevant. Therefore we take the hierarchical structure given, and study how the number of the owners and of the firms of a public economy affect monitoring and work incentives through the principal-agent chain. The literature of the principal-agent theory is concerned with how the principal designs a incentive scheme for the agent (see Hart and Holmstrom 1987 for a survey). This paper is more concerned with how the free-rider problem among principals is worse as the number of the principals increases, and how the agents can efficiently bribe the principals in a public economy. The literature of economic transition is very divided about how important privatization is for transforming a planned economy into a market economy. In particular, many economists are puzzled by the fact that the Chinese economy has sustained a long time high growth without mass privatization, while Russia has been suffering from a sharp fall in output after mass privatization. Some economists even cite the Chinese case to argue that private property rights are not a necessary condition for market efficiency (e.g., see Stiglitz 1994). This paper can throw some lights on understanding the Chinese economy.

A.2 Characterization of the Principal-Agent Relationship of the Public Economy

We define ownership of the firm by residual claimancy. [1] Public ownership of the firm can be defined as such an institutional arrangement under which: (i) the residual claim is assigned to a community ("public") consisting of more than one individuals; (ii) each individual has an equal share in the residual claim; and (iii) no individual has right to transfer his residual claim to someone else in exchange for a personal payment. This definition distinguishes the public firm from the joint stock company which is also owned by the "public" consisting of many individuals but each of them claims the residual proportional to his share and is free to transfer his residual share to somebody else for a personal payment without approval of any other shareholders (Alchian, 1965). For this reason, I shall call an individual owner of the public firm the "co-owner" distinguished from the share-holder of the joint stock company.

Like a private firm, from the point of view of functioning, a public firm also consists of the decision-making member, the producing members (workers) and capital-owner members. The producing members are always insiders and capital-owner members can be either insiders or outsiders. However, unlike a private firm, the decision-making member of the public firm may not be an insider.[2] In fact, an important problem for the public firm is where the decision-making member should be.

The incentive problem originating from teamwork and uncertainty is also the main concerns of the organizational design of a public firm. However, by definition, public ownership itself is a major constraint to the organizational design. Given that the residual claim is equally shared, the only mechanism with freedom lies at directly monitoring.

Theoretically, public ownership of the firm does not necessarily correspond to public ownership of physical capital. For instance, capital used by a public firm may be hired from private owners (as in today's China where more than 80% of investment of SOEs is debt-financed) or from a different community who is paid a fixed interest. Similarly, capital owned by the public can be leased to a private firm. However, in reality, public ownership of the firm is typically integrated with public ownership of capital. A possible reason is that the public firm was invented by imitating the capitalist firm. A full integration means that residual claimants of the public firm are the same as the owners of its capital. An important implication is that if the community is large (owners are many), there will be many public firms owned by many individuals of the same community. All these public firms together comprise a public economy.

---

[1] Economists have recognized that residual claim and control rights are two major components of ownership. Here we omit the control right by assuming that the control right is a derivative of the residual claim. Grossman and Hart (1986) define ownership of the firm by the control right. But they do not investigate the relationship between the residual claim and the control right.

[2] I find it is really difficult to define the boundary of a public firm in terms of membership, although such a problem also exists for the private firm. Should we include any decision-maker as a member of the firm no matter how far he is from the firm as people usually understand?

We now characterize the institutional structure of a public economy. Let us call the residual claimants (co-owners) of the public economy the "principals", and the inside members of the firm paid with the fixed terms the "agents".[3] As in the private firm, the principals of the public firm also hold the authority of monitoring the agents. However, by definition, public ownership implies that "voting-with-hands" is the only way available for the principals to exercise their authority of monitoring. In other words, monitoring is a typical "public choice" in the sense that it can only be made through an aggregating process of individuals' actions based on "one-person-one-vote" or some delegation systems. Because of communication problems and organizational costs, if the community is large, a hierarchical structure of delegation is inevitable. Typically, the community consists of (or is decomposed into) many bottom-communities; each individual belongs to one bottom-community; each bottom-community delegates power of monitoring to their representative; the representatives from several bottom-communities form a committee which in turn delegates power to its representative; the representatives from several committees form a super-committee which in turn delegates power to its representative; and so on, until a "central" committee which is the representative of the whole community taking responsibility for monitoring and controlling all the public firms. Because of the limitation of span of control, if there are many firms, it is impossible for the central committee to directly supervise all these firm, and another hierarchical structure is inevitable: the central committee delegates power to some sub-committees; each sub-committee delegates power to some sub-sub-committees which in turn delegate power to sub-sub-sub-committees, and so on, down to the inside members of the firm.

In summary, the principal-agent relationship of the public economy is typically characterized by two "macro" hierarchies. The first hierarchy is formed via a delegation chain of power from the residual claimants (principals) to the central committee; its direction of principal-agent relation is upward (from bottom to top). The second hierarchy is formed via a delegation chain from the central committee to the inside members of the firm; its direction of principal-agent relation is downward (from top to bottom). Note that apart from the two types of extreme players (the residual claimants and the inside members), each player plays two roles: he is the agent of the principal and the principal of the agent. For convenience, we call the residual claimants the *original* principals, the inside members of the firm *ultimate* agents, and the central committee the central agent. "Up-stream agent" and "down-stream agent" will also be used to refer to the agent before and after the concerned party respectively.

[3] Normally, in a public economy, an individual plays two roles: as a residual claimant, he is a principal, and as a wage-worker, he is an agent. But conceptually these two roles should be distinguished. In fact, even if there is only one public firm, this conceptual distinction is very important for understanding his behaviour: as the principal he has some self-incentive to work and hopes (may monitor) others to work hard; on the other hand, as an agent, he has an incentive to shirk and hope others not to monitor him.

In reality, the first hierarchy of the public economy is overlapped with the government administration structure which looks not very different from that in the private economy. This overlapping gives the public firm another name called "the state-owned firm". The legitimacy of this overlapping can be found from economizing organization costs. But we shall not explore this problem here.

A.3 The Effect of the Degree of Publicness and the Size of the Economy on Monitoring- and Work-Incentives

We now turn to analyze the effects of the degree of publicness of and the size of the public economy on the incentive of monitoring by the original principals and the constraint on the ultimate agents' behavior. Denote by n the number of the original principals (co-owners) of the public economy (representing the size of the community), m the number of the public firms (representing the size of the economy), $H_1 + 1$ the layers of the first hierarchy ($H_1 \geq 1$), and $H_2 \geq 1$ the layers of the second hierarchy ($H_2 \geq 1$). For simplicity, we assume that the span of representation in the first hierarchy is a constant ($t \leq n$) (i.e., there are $\frac{n}{t}$ bottom communities each consisting of t original principals, and each downstream agent represents t immediate upstream agents, up to the central agent) and the span of control in the second hierarchy is a constant $s(\leq m)$ ($i.e.$, the central agent control s immediate sub-committees each in turn controlling s immediate sub-sub-committees, until that m firms are directly controlled by $\frac{m}{s}$ immediate supervisors)[4]. Then the following relations hold[5]:

$$n = t^{H_1} \text{ and } m = s^{H_2} \tag{A.1}$$

By definition, under public ownership, only the original principal has a self-incentive to monitor the agents and the incentive of all the agents (up to the ultimate agents) to serve the original principals can only come from being-monitored. Suppose that all the original principals and the agents at the same level are identical. Then at equilibrium all the original principals will choose the same monitoring effort to monitor their immediate agents; and all the agents at the same level will have the same response to their immediate principals' (or up-stream agents') monitoring. Denote by I_P the monitoring effort of a representative original principal to monitor his immediate agent, I_h the monitoring effort of a h-level agent to monitor the $h+1$ level agent of the first hierarchy on behalf of the original principals, $h = 1, 2, \ldots, H_1 - 1$, and I_l the monitoring effort of a $l-$level agent to monitor a $l + 1$ level agent, $l = 1, 2, \ldots H_2 - 1$. Normalize the inside members of the firm to a single agent and denote by I_A the work effort of a representative ultimate agent to serve the interest

[4]For a reference of the optimal span of control in the context of the firm, see Qian (1994).

[5]Here we ignore the correlation between n and m.

of the original principals. Note that I_A is the work effort while all other $I_{h(l)}$ are monitoring efforts.

A monitoring technology is defined as an incentive transformation mechanism from the principals (or up-stream agents) to the immediate agents.[6] In particular we assume that monitoring technology takes a linear form at all levels:

$$I_1 = R_0^1 I_P;\ I_h = R_{h-1}^h I_{h-1};\ I_l = D_{l-1}^l I_{l-1};\ I_A = D_{H_2}^A I_{H_2}$$

That is, the agent at the h level of the first hierarchy will choose $I_h = R_{h-1}^h$ if each of his t immediate principals chooses one unit of monitoring effort; each of s agents at the l level of the second hierarchy will choose $I_l = D_{l-1}^l$ if his immediate principal chooses one unit of monitoring effort. We call R_{h-1}^h and D_{l-1}^l "parameters of effectiveness of monitoring". $R_{h-1}^h = 0$ ($D_{l-1}^l = 0$) corresponds to that monitoring is technically impossible and $R_{h-1}^h = \infty$ ($D_{l-1}^l = \infty$) corresponds to that monitoring is perfect.[7] Both R_{h-1}^h and D_{l-1}^l depend on technology, uncertainty of environment, as well as institutional arrangements (e.g., election system, regulations, policy). In this paper we assume they are given.[8]

Note that the above defined monitoring technology has already incorporated the **incentive compatibility constraint** for all agents; that is, given I_h, the agent at the $h+1$ level cannot do better than by choosing $R_h^{h+1} I_h$.

Without loss of generality, in the following analysis, we assume:

$$\begin{aligned} R_0^1 = R_1^2 = \cdots = R_{h-1}^h = \cdots = R_{H_1-1}^{H_1} \equiv R \\ D_0^1 = D_1^2 = \cdots = D_{l-1}^l = \cdots = D_{H_2-1}^A \equiv D \end{aligned}$$

Then, by recursion, the relationship between I_A and I_P can be written as follows:

$$I_A = R^{H_1(n)} D^{H_2(m)} I_P \tag{A.2}$$

Suppose that all m firms are identical in production and for the moment that the decision right of "what to produce and how to produce it" is assigned to the ultimate agent. Let π be the expected output produced by a representative firm to be distributed

[6] Monitoring technology used here is different from that used by other authors. For instance, Shapiro and Stiglitz (1984) and Qian (1994) define monitoring as checking the agent with probability. But our monitoring technology can be reinterpreted as theirs: when the principal makes more effort to monitor the agent, the probability of shirking behaviour being found increases, and for a given level of wage, this induces less shirking.

[7] It is *perfect* in the sense that an ε -monitoring effort can induce the agent to do as much as the principal likes. This can be a case only when there is no uncertainty.

[8] One may define $\frac{R_{h-1}^h}{t}$ as the *average* marginal productivity of the principals' monitoring effort in the first hierarchy and sD_{l-1}^l as the *aggregated* marginal productivity of the principal's monitoring effort in the second hierarchy.

to n original principals. Then π is a function of I_A only. For simplicity, we assume that the production function is linear: $\pi = I_A = R^{H_1} D^{H_2} I_P$.[9]

The output accruing to a representative original principal is equal to[10]

$$\pi_P = \frac{m}{n}\pi = \left(\frac{R}{t}\right)^{H_1} (sD)^{H_2} I_P \tag{A.3}$$

The risk-neutral original principal's marginal benefit of monitoring is $MB = (t^{-1}R)^{H_1}(sD)^{H_2}$, which is decreasing (increasing) with the size of the community if $(t^{-1}R) \leq (\geq)1$, and decreasing (increasing) with the size of the economy if $sD \leq (\geq)1$.

Monitoring does not just bring the output, but also incurs utility costs. When the principal chooses I_P to monitor his immediate agents, he costs himself as well as all his agents through the principal-agent chain. His authority of monitoring is limited by the agents' **participation constraint**, which implies that the principal must pay all the costs. For simplicity, suppose that all individuals have an identical cost function of monitoring effort (or work effort if he is the ultimate agent), which takes a quadratic form of $C(I_i) = 0.5I_i^2$. Then, by recursion and adding up, it can be shown that the total cost of I_P for each original principal is

$$TC(I_P) = 0.5I_P^2 \left(\sum_{i=0}^{H_1} \left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1} \sum_{j=1}^{H_2} \left(sD^2\right)^j \right) \tag{A.4}$$

Note that the denominators appear because the cost is shared by all concerned principals. For instance, the cost for the immediate agent is shared by all t members of the bottom community, and the cost for the central agent is shared by all n members of the community.

The original principal's marginal cost of monitoring is

$$MC(I_P) = I_P \left(\sum_{i=0}^{H_1} \left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1} \sum_{j=1}^{H_2} \left(sD^2\right)^j \right) \tag{A.5}$$

---

[9] As in the standard literature, a realized output depends on both the ultimate agent's work effort and the state of nature. However, since in this paper all individuals are assumed risk-neutral, we only need to consider the expected output which is independent of the state of nature. It should also be emphasized that the production function used here is somewhat different from Qian (1994) who, following Williamson (1967), Beckmann (1977) and Rosen (1982), employs a recursive production technology in which, at any tier t an intermediate product y_{t-1} from the immediate superior is used as an input and combined with effort a_t to produce y_t for the immediate subordinate. In our model, only the ultimate agent's effort is directly productive. Recursive effect comes from monitoring technology rather than production technology.

[10] Note this amount of output is not that eventually enjoyed by the principal because the principal must compensate the agents for their costs caused by his monitoring.

MC is strictly increasing with both the size of the community (n) and the size of the economy (m) for all $R, D > 0$.

The original principal chooses his monitoring effort such that $MB = MC$, which implies that

$$I_P^* = \frac{\left(\frac{R}{t}\right)^{H_1} (sD)^{H_2}}{\sum_{i=0}^{H_1} \left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1} \sum_{j=1}^{H_2} (sD^2)^j} \tag{A.6}$$

where I_P^* denotes the optimal choice of I_P.

If $R \leq t$, the numerator is decreasing with n and the denominator is strictly increasing with n; if $R > t$, the denominator increases faster than the numerator with n. This implies $\frac{\partial I_P^*}{\partial n} < 0$.[11] However, the effect of m can be positive or negative depending on s and D as well as R. A necessary condition for the effect to be positive is that (sD) is sufficiently greater than one and D is sufficiently smaller than one. Otherwise the effective will be negative.

Substituting (A.6) back into (A.2), we obtain:

$$I_A^* = \frac{\left(\frac{R^2}{t}\right)^{H_1} (sD^2)^{H_2}}{\sum_{i=0}^{H_1} \left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1} \sum_{j=1}^{H_2} (sD^2)^j} \tag{A.7}$$

It is easy to see that the effects of both n and m on I_A^* are unambiguously negative.

In summary, we have:

Proposition 1 *(i) Both the original principal's optimal incentive to monitor and the ultimate agent's monitored work effort are strictly decreasing with the expansion of the community size. (ii) The original principal's optimal incentive to monitor can be either increasing or decreasing with the economy size depending the parameters of effectiveness of monitoring and the span of control in the second hierarchy; however, the ultimate agent's monitored work effort is strictly decreasing with the expansion of the economy size.*

The first part of the proposition is quite intuitive. When the size of the community increases, each individual principal can share less and the hierarchical expansion effect implies that the monitoring is more costly. As a result, the optimal incentive to monitor decreases, so does the monitored work effort of the agent. It is worth pointing out that the negative impact of publicness on efforts here comes from two effects: the first is the free-rider effect, and the second is the hierarchical expansion

[11] To differentiate the equation, we need to transform the denominator into an expression of integrals. Direct comparing the speeds of changes between the numerator and the denominator is much more intuitive.

effect. Although the first effect is well-known, the second effect is ignored in the existing literature of the public economy.

To understand the second part, note that although the increase of the size of economy makes monitoring more costly through the hierarchical expansion effect, which induces the principal to choose less monitoring effort, it also makes monitoring more useful in the sense that monitoring now applies to more ultimate agents, which induces the principal to choose more monitoring effort. The total effect of the change in the size of economy is ambiguous depending on which effect is dominant, which is summarized by (sD) and D. A positive effect requires that D is sufficiently small than one given that sD is sufficiently larger than one, which implies that the *aggregated* effectiveness of monitoring in the second hierarchy must be large while the *individual* effectiveness should not be large, which can only be a case when s large, that is, the second effect dominates the first. The reader may be puzzled by the fact that the effect of m on I_P^* is more likely to be negative when D becomes large. The reason is that a larger D implies a larger disutility for the agents for any given monitoring effort by the principal. Because the positive effect would occur only when $D \ll 1$, D^{H_2} shrinks faster as H_2expands, which adds another negative effect on I_A^*. As a result, even if I_P^* is increasing with m, I_A^* is still decreasing with m.

Note that I_P^* is not monotonically increasing with R and D. The reason is that monitoring itself is not productive and its value comes only from affecting the ultimate agents' incentives. The increase in R (or D) does not only increases the effectiveness of monitoring, but also raise the costs of monitoring. When R (or D) is small, the first effect dominates the second, and the principal will increase monitoring effort in response to an increase in R (or D). When R (or D) becomes very large, the second effect will dominate the first, and the principal will decrease his monitoring effort in response to an increase in R (or D). Nevertheless, the analysis of (A.7) shows that the increases in R and D always make the ultimate agents work harder and therefore increase the residual distributed to the original principals and their welfare surplus. In particular, when $R \to \infty$ and $D \to \infty$ (i.e., monitoring goes perfect), $I_A^* \to 1$which is the first best optimal effort.

To get some policy implications from the above analysis, let us consider the effect of splitting up the community on the principal's monitoring incentive and on the agent's incentive to serve the principal. Suppose that the community of n is now split up into k small communities such that each small community consists of $\left(\frac{n}{k}\right)$ original principals and owns $\left(\frac{m}{k}\right)$ firms. Then,

$$I_P^* = \frac{\left(\frac{R}{t}\right)^{H_1\left(\frac{n}{k}\right)} (sD)^{H_2\left(\frac{m}{k}\right)}}{\sum_{i=0}^{H_1\left(\frac{n}{k}\right)} \left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1\left(\frac{n}{k}\right)} \sum_{j=1}^{H_2\left(\frac{m}{k}\right)} \left(sD^2\right)^j} \tag{A.8}$$

$$I_A^* = \frac{\left(\frac{R^2}{t}\right)^{H_1\left(\frac{n}{k}\right)} \left(sD^2\right)^{H_2\left(\frac{m}{k}\right)}}{\sum_{i=0}^{H_1\left(\frac{n}{k}\right)} \left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1\left(\frac{n}{k}\right)} \sum_{j=1}^{H_2\left(\frac{m}{k}\right)} \left(sD^2\right)^j} \tag{A.9}$$

where $H_1\left(\frac{n}{k}\right) = \frac{\ln n - \ln k}{\ln t} < H_1(n) = \frac{\ln n}{\ln t}$, and $H_2\left(\frac{m}{k}\right) = \frac{\ln m - \ln k}{\ln s} < H_2(m) = \frac{\ln m}{\ln s}$.

By applying the preceding arguments in reverse (note that now H_1 and H_2 decrease simultaneously as k increases), we have

Proposition 2 *For any given size of economy, splitting up the community of ownership will strictly increase the principal's monitoring incentive as well as the agent's monitored work incentive.*

The intuition behind this proposition is that although the splitting-up does not increase the real share of each principal's residual, it improves the aggregated effectiveness of monitoring by flattening the hierarchies (i.e., reducing the distance between the principal and the agent). The community can directly benefit from the splitting-up through saving of total monitoring effort, because now there are fewer intermediate professional monitors along the principal-agent chain who has no direct contribution to the residual. As we will see, this gives a strong policy implication.[12]

A.4 The Incentive Problem of the Corrupt Public Economy

From the above analysis, we can predict that the public economy of a huge community cannot be workable. One may argue that this prediction is inconsistent with the observation that there are some huge public economies (like in China and former Soviet Union) which have worked for a long time, although not very efficient. My answer is that the public economy as described in the last section has never existed in reality; what we observed is a corrupt public economy.

A corrupt public economy is one in which although the ownership is legally (or nominally) entitled to public, the residual is actually, to a great extent, claimed by the agents within the hierarchies. Sharing of the residual by the agents may take various forms. Some of the residual may be hidden by the agents such that the observed residual by the principals is smaller than the actual residual. Another form is that the agents spend some of the residual on projects which are claimed to serve the public's interest, but which actually benefit the agents themselves more than the principals.

[12] Proposition 2 actually reproduced Qian's result (1994). In Qian's model, the size of the firm is determined by the stock of capital (capital-worker ratio is fixed). He argues that the problem with socialist economy is that of "bigness": with all capital owned centrally, production is organized by a hierarchy that is necessarily long and inefficient. "It follows that there are potential gains from privatization or making ownership decentralized, simply because the resulting decentralization reduces the amount of capital per 'firm'" (p. 540).

I claim that a corrupt public economy Pareto-dominates an uncorrupted public economy; in other words, the "illegal" enjoyment of the residual by the agents does not harm the original principals' interests. The intuition is that, given monitoring technology, an individual principal cannot do better than by choosing the optimal monitoring effort defined by (A.6) which generates him a maximum welfare surplus (to be defined later). This implies that the agent faces a minimum welfare constraint from the principals. In other words, the principal should be happy as long as the agent provides him with such a level of welfare. This is of course known by the agent himself. Because the upstream agent is authorized by the principal (or up-upstream agent) to monitor the downstream agent, if he can force the latter to do more, there will be some extra residual (net of costs) which can be captured by him. Furthermore, by so doing, the agent does not just increase the residual but also reduce the cost of monitoring because now the principals (or upstream agents) no longer need to monitor him. This cost-saving benefit can also be enjoyed by him. Because nobody is worse-off from such a corruption, it is a Pareto-improvement.

We now formally model this argument. The analytical strategy is that we first compute the optimal monitoring incentive for each level's representative agent *if* he is the residual claimant, and the maximum welfare surplus he could get from monitoring. We then compare one with another sequentially to show that the optimal monitoring effort and the maximum welfare surplus are increasing with the agent's level from the bottom (the original principal) to top (the central committee) in the first hierarchy, which implies that it pays for the downstream agent to "buy" the residual claim from his immediate upstream agents by guaranteeing the latter a "fixed" payoff equal to the maximum surplus the latter could get if he does not "sell" the residual claim. By recursion we find that at equilibrium the actual residual claim eventually goes to the central committee such that all the upstream agents (including the original principals) simply have no intention to monitor the downstream agents (including the central committee) as long as the latter deliver the fixed payoff. We also show that once the residual claim is in hands of the central agent, there will be no space for further Pareto improvement from downward shifting the residual claim *if decision rights are held by the central agent.*

As in the last section, for simplicity, we shall assume that $\pi(I_A) = I_A$ and $C(I_i) = 0.5I_i^2$. By substituting (A.6) into (A.3), we have that in a non-corrupt public economy,

$$\pi_P = \frac{\left(\left(\frac{R}{t}\right)^{H_1}(sD)^{H_2}\right)^2}{\sum_{i=0}^{H_1}\left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1}\sum_{j=0}^{H_2}\left(sD^2\right)^j} \tag{A.10}$$

By deducting the total cost TC from π_P, we have

$$W_P = 0.5\frac{\left(\left(\frac{R}{t}\right)^{H_1}(sD)^{H_2}\right)^2}{\sum_{i=0}^{H_1}\left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1}\sum_{j=0}^{H_2}\left(sD^2\right)^j} \tag{A.11}$$

W_P is the maximum welfare an individual principal can get by exercising his principalship, or the minimum amount each first-level agent must deliver to each original principal of the bottom community.

Now suppose that the agent of the 1-level is the residual claimant. Then his problem is to choose I_1 to maximize

$$\frac{m}{\left(\frac{n}{t}\right)}R^{H_1-1}D^{H_2}I_1 - 0.5I_1^2\left(\sum_{i=0}^{H_1-1}\left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1-1}\sum_{j=1}^{H_2}\left(sD^2\right)^j\right) \tag{A.12}$$

The first-order condition implies

$$I_1^* = \frac{\left(\frac{R}{t}\right)^{H_1-1}(sD)^{H_2}}{\sum_{i=0}^{H_1-1}\left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1-1}\sum_{j=1}^{H_2}\left(sD^2\right)^j} \tag{A.13}$$

$$I_{A(1)}^* = \frac{\left(\frac{R^2}{t}\right)^{H_1-1}\left(sD^2\right)^{H_2}}{\sum_{i=0}^{H_1-1}\left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1-1}\sum_{j=1}^{H_2}\left(sD^2\right)^j} \tag{A.14}$$

where $I_{A(1)}^*$ denotes the ultimate agent's work effort when the 1-level agent claims the residual (accordingly, I_A^* is rewritten as $I_{A(P)}^*$.

Rearranging (A.13) and (A.14) gives

$$I_1^* = \frac{R\left(\frac{R}{t}\right)^{H_1}(sD)^{H_2}}{\sum_{i=0}^{H_1}\left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1}\sum_{j=1}^{H_2}\left(sD^2\right)^j - 1} \tag{A.15}$$

$$I^*_{A(1)} = \frac{\left(\frac{R^2}{t}\right)^{H_1} \left(sD^2\right)^{H_2}}{\sum_{i=0}^{H_1} \left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1} \sum_{j=1}^{H_2} \left(sD^2\right)^j - 1} \tag{A.16}$$

It is easy to check that $I^*_{A(1)} > I^*_{A(P)}$ and $I^*_1 > RI^*_P$. This implies that the 1-level agent's incentive to monitor the 2-level agent is greater than that imposed optimally by the original principal, and therefore there is no need for the original principal's monitoring.

The per capita residual from monitoring by the 1-level agent is

$$\pi_1 = t \frac{\left(\left(\frac{R}{t}\right)^{H_1} (sD)^{H_2}\right)^2}{\sum_{i=0}^{H_1} \left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1} \sum_{j=0}^{H_2} \left(sD^2\right)^j - 1} \tag{A.17}$$

The per capita welfare surplus is

$$W_1 = 0.5t \frac{\left(\left(\frac{R}{t}\right)^{H_1} (sD)^{H_2}\right)^2}{\sum_{i=0}^{H_1} \left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1} \sum_{j=0}^{H_2} \left(sD^2\right)^j - 1} \tag{A.18}$$

It is easy to see that $\pi_1 > t\pi_P$ and $W_1 > tW_P$. This implies that if the 1-level agent becomes the residual claimant, the residual and the welfare surplus are more than offset that when the original principal is the residual claimant. By delivering each original principal the residual equal to W_P, a representative 1-level agent can capture the remaining residual equal to

$$\pi_1 - tW_P > 0$$

and after deducting the compensation for the agents as well as for himself, he retains the net welfare surplus equal to

$$W_{1\text{net}} = W_1 - tW_P$$

In general, if the h-level agent is the residual claimant,

$$I_h^* = \frac{R^h \left(\frac{R}{t}\right)^{H_1} (sD)^{H_2}}{\sum_{i=0}^{H_1} \left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1} \sum_{j=1}^{H_2} \left(sD^2\right)^j - \sum_{i=0}^{h-1} \left(\frac{R^2}{t}\right)^i} \tag{A.19}$$

$$I_{A(h)}^* = \frac{\left(\frac{R^2}{t}\right)^{H_1} \left(sD^2\right)^{H_2}}{\sum_{i=0}^{H_1} \left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1} \sum_{j=1}^{H_2} \left(sD^2\right)^j - \sum_{i=0}^{h-1} \left(\frac{R^2}{t}\right)^i} \tag{A.20}$$

That is, $I_h^* > RI_{h-1}^* > \cdots > R^h I_P^*$, and $I_{A(h)}^* > I_{A(h-1)}^* > \cdots > I_{A(p)}^*$.

$$\pi_h = t^h \frac{\left(\left(\frac{R}{t}\right)^{H_1} (sD)^{H_2}\right)^2}{\sum_{i=0}^{H_1} \left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1} \sum_{j=1}^{H_2} \left(sD^2\right)^j - \sum_{i=0}^{h-1} \left(\frac{R^2}{t}\right)^i} \tag{A.21}$$

$$W_h = 0.5t^h \frac{\left(\left(\frac{R}{t}\right)^{H_1} (sD)^{H_2}\right)^2}{\sum_{i=0}^{H_1} \left(\frac{R^2}{t}\right)^i + \left(\frac{R^2}{t}\right)^{H_1} \sum_{j=1}^{H_2} \left(sD^2\right)^j - \sum_{i=0}^{h-1} \left(\frac{R^2}{t}\right)^i} \tag{A.22}$$

That is, $\pi_h > t\pi_{h-1} > \cdots > t^h \pi_P$ and $W_h > tW_{h-1} > \cdots > t^h W_P$. In other words, it pays for the h-level agent to buy the residual claim from all previous agents and t^h original principals.

If the central agent is the residual claimant, we have

$$I_{H_1}^* = \frac{(sD)^{H_2}}{\sum_{j=1}^{H_2} \left(sD^2\right)^j} \tag{A.23}$$

$$I_{A(H_1)}^* = \frac{\left(sD^2\right)^{H_2}}{\sum_{j=1}^{H_2} \left(sD^2\right)^j} \tag{A.24}$$

$$\pi_{H_1} = \frac{(sD)^{2H_2}}{\sum_{j=1}^{H_2} \left(sD^2\right)^j} \tag{A.25}$$

$$\pi_{H_1} = 0.5\frac{(sD)^{2H_2}}{\sum_{j=1}^{H_2} \left(sD^2\right)^j} \tag{A.26}$$

This leads to

Proposition 3 *The original principal being the residual claimant is Pareto-dominated by the 1-level agent being the residual claimant. In general, the upstream agent being the residual claimant is sequentially Pareto-dominated by the downstream agent being residual claimant; and the central agent being the residual claimant Pareto-dominates all its upstream agents being the residual claimants in the first hierarchy.*

The above proposition actually reproduced the well-known result in the principal-agent literature, but in a hierarchical context, that if the agent is risk-neutral, the first-best Pareto-improvement can be achieved by just having the principal "sell the store" to the agent; that is, the agent keeps all profits and becomes the sole residual claimant, but must pay a fixed fee up front to the principal. However, it is still interesting to see that, in a hierarchical public economy, even if the original principal does not deliberately design an incentive scheme for the agent, an implicit incentive system may evolve through the agent's corruption.

Can the argument be carried over to the ultimate agent? The answer is yes *if* the decision rights are held by the ultimate agents, as we have assumed so far. It is easy to show that if the ultimate agent is the residual claimant, the first best result is achievable. However, this cannot hold if the decision rights are held by the central agent such that the ultimate agent is only a producing member (i.e., implementing the decision made by the central agent), as in most pre-reform socialist economies. The reason is that when the central agent makes decisions about what to do and how to do it, the output is a joint outcome of both the central agent's work effort and the ultimate agent's work effort.

To see this, denote by $I_{H_1} = \left(I^a_{H_1}, I^b_{H_1}\right)$ the effort vector of the central agent, and by $I_{H_1} = \left(I^a_{H_1}, I^b_{H_1}\right)$ the effort vector of the ultimate agent, where I^a_i and I^b_i are work effort and monitoring effort respectively. For simplicity, assume that the production function takes a Cobb-Douglas form of $\pi = \left(I^a_{H_1}\right)^{\alpha}\left(I^a_A\right)^{1-\alpha}$, and the cost function takes a quadratic form of $C_i = 0.5\left(I^a_i\right)^2 + 0.5\left(I^b_i\right)^2$. Furthermore, suppose $H_2 = 1$ (that is, $s = m$). If the residual is claimed by the central agent, $I^a_A = DI^b_{H_1}$. If the residual is claimed by the ultimate agent, $I^a_{H_1} = BI^b_A$, where B is the counterpart of D representing the effectiveness of monitoring by the ultimate agent over the central agent (his previous principal but current agent).

In this case, the central agent is similar to the manager and the ultimate agent is similar to a worker of the capitalist firm. We can reasonably assume that (i) $D > B$ (that is, manager-monitoring-worker is more effective than worker-monitoring-manager); and (ii) $1 > \alpha > 1/2$ (that is, the manager is more important than a worker). Then it is easy to show that: When the central agent is the residual claimant, the total welfare surplus is

$$W(H_1) = \frac{1}{2}\alpha(1-\alpha)^{1-\alpha}\left(\frac{D^2}{1+D^2}\right)^{1-\alpha}$$

When the ultimate agent is the residual claimant, the total welfare surplus is

$$W(A) = \frac{1}{2}\alpha(1-\alpha)^{1-\alpha}\left(\frac{B^2}{1+B^2}\right)^{\alpha}$$

It is easy to see that $D > B$ and $\alpha > 1/2$ implies that $W(H_1) > W(A)$.

In general, following Zhang (1994), we have

Proposition 4 *Suppose that the central agent holds the decision rights. Then the central agent being the residual claimant Pareto dominates the ultimate agent being the residual claimant.*

Proposition 4 suggests that the shifting of the residual claim from the central agent to the ultimate agent can be a Pareto-improvement only if it is accompanied by the shifting of the decision rights.

Propositions 3 and 4 together suggest that the observed corruption of the public economy seems a rational response to public ownership since corruption is *ex post* efficient with public ownership given. In such a corrupt public economy, the central agent becomes the acting principal who holds both the residual claim and authority of monitoring over the subordinate agents in the second hierarchy such that his incentive appears as the primary engine of the whole public economy. The total residual of π_{H_1} is first collected by the central agent who in turn delivers tW_{H_1-1} to t immediate upstream agents of the $H_1 - 1$ level and keeps the remaining part of $(\pi_{H_1} - tW_{H_1-1})$ for himself some of which is used to compensate the monitored agents for their monitored effort; each of the $H_1 - 1$ level agents in turn delivers tW_{H_1-2} to t agents of the $H_1 - 2$ level and keeps $\left(W_{H_1-1} - tW_{H_1-2}\right)$ for himself; and so on, until that each original principal receives the residual equal to W_P from his immediate agent.

In a corrupt public economy, the original principals appear very inactive in monitoring their agents. But their legal ownership status is the underlying force for the above described distribution system to be an equilibrium. Would be it possible for the original principals or other agents to share some extra rent resulting from the central agent being the acting principal apart from their respective *status quo*, or for the central agent as well as other intermediate agents to exploit the upstream agents and the original principals by delivering less than W_h? The answer may depend on the relative bargaining powers between the upstream agents and the downstream agents,

which in turn depends on, for example, political systems as well as other institutions (even education level). We shall not analyze this problem. However, the observation suggests that as the public economic system operates year by year, the original principals' sense of ownership degenerates and their bargaining power become rusty. This kind of devolution surely strengthens the central agent's discretionary ability in delivering the residual. As a result, it is possible that what actually delivered by the downstream agent to the upstream agent is smaller than W_h.[13] A possible counterbalance against this devolution is the political competition for being agents (being the central agent in particular). Because being the agent is profitable, there exist some rent-seekers attempting to unseat the incumbents by promising to deliver more to the upstream agents and original principals, even under very monopolistic politics. This hopefully puts some pressure on the incumbents to make self-restrictions on their discretionary behavior. Nevertheless, if politics is too open such that agents have no freedom to claim any residual, the total surplus may be reduced. The situation is much similar to the effect of market competition on technology innovations.

It should be pointed out that in the above analysis, we have implicitly assumed that the amount eventually delivered to the upstream agents (up to the original principals) is fixed *ex ante* before the downstream agent makes his monitoring effort decision, — in other words, W_h is not a fixed percentage of π_{H_1} but a lump-sum independent of π_{H_1}. Otherwise the central agent's incentive would be weakened. However, the conclusion contained in Proposition 3 is applicable even if the residual is implicitly or explicitly "shared", as long as the central agent's *expected* residual share is sufficiently large.

A.5 Applications to the Chinese Economy

In summary, the basic results of this paper are as follows:

(1) The monitoring incentive of the original principals and work incentive of the ultimate agents are decreasing with the degree of publicness and the size of the public economy (with some qualifications).
(2) The splitting-up of a given size public economy will increase both the monitoring incentive of the original principals and work incentive of the ultimate agents.
(3) The corrupt public economy in which the central agent claims *de facto* the residual can be a Pareto-improvement over a canonical public economy.
(4) The shifting of the residual claim from the central agent to the ultimate agents should be done simultaneously with the shifting of the decision rights.

We now apply the above results to the Chinese economy and explain four important observations.

[13] Because the residual delivered to the original principals appears to be a free gift, the central agent might be eventually perceived as a paternal ruler and all the original principals feel very grateful for his delivering the residual.

(1) Comparison of Performance Between SOEs and TVEs.

Considerable studies show that, and almost all economists agree that, in 1980s, the Chinese township and village enterprises (thereafter TVEs) outperformed the Chinese state-owned enterprises (thereafter SOEs) (see, for example, Jefferson and Singh 1994 for a survey). The problem is how to explain this phenomenon. In particular, since TVEs are also "public-owned", and their property rights are also "vaguely-defined", why could one kind of the public enterprises be more efficient than the other? For many observers, there exists a paradox" between the efficiency of TVEs and the standard property rights theory (Weitzman and Xu 1994, and Li 1996, among others). To get rid of this paradox, some western economists argue that TVEs are actually privately owned enterprises and they are disguised as "collective" or "communal" for some ideological and legal reasons. This is surely true for some TVEs, but it is not true for the majority of them. Weitzman and Xu (1994) try to reconcile the performance of TVEs with the standard property rights theory by introducing a culture dimension of cooperation. They argue that the standard property rights theory is culture-free and it is really applicable only to low cooperative culture; the success of Chinese TVEs can be attributed to high cooperative Chinese culture.[14] There are also some economists who use TVEs to defend public ownership. However, according to our theory, the outperformance of TVEs over SOEs presents no real paradox to the standard property rights theory. The owners of TVEs are much fewer than the owners of SOEs, and the size of TVEs economy in each town or village is much smaller than that of SOEs in the national economy. This implies that the principal-agent chain of TVEs is much shorter than that of SOEs, and therefore the original principals (or the central agent —the town government) of TVEs can more effectively monitor the management. Our theory also predicts that as the size of TVEs expands, the principal-agent chain becomes longer, and monitoring by the original principals and the central agent is less and less effective. There is a trend that TVEs converge to the second SOEs. For TVEs to remain efficient, privatization of TVEs is inevitable. Observation suggests that Chinese practitioners are indeed following the standard property rights theory–maybe unconsciously–by privatizing their TVEs in various forms, such as "share-holding cooperatives" or simply "sell-out". For instance, in Zibo municipality of Shandong province, private shareholders owned 30% the share of the TVEs in 1992 and 70% in 1995. By 1996, about one thirds of the TVEs had been privatized in Nanhai of Guangdong. By the first half of 1997, more than 60% of the township enterprises in Shenyang of Liaoning became share-holding companies or share-holding cooperatives; 90% of the township enterprises with assets under 5 million yuan had been privatized in provinces such as Zhejiang and Jiangsu (*South China Morning Post* June 13 and 17, 1997). By 1997, about one half or more TVEs had been privatized in such provinces as Guangdong, Shandong, Zhejiang and even Liaoning, a relatively backward and politically conservative region. Privatization tends to accelerate more quickly in areas where neighboring towns have more private enterprises (Li, Li and Zhang, 2000).

[14]They borrow the folk theorem of repeated games from the game theory but fail to explain why the game in China has led to a high cooperative attitude.

(2) The Fiscal Contract System and the Economic Reform.

An important reform measure in the 1980s was the "fiscal contract system" between adjacent levels of governments, which was first introduced in 1980 and renewed in 1984 and 1988 with some modifications. Under such a system, lower level governments have an obligation to hand over a fixed amount or a fixed proportion of their revenues to higher level governments and to keep all the rest for themselves. It is no longer possible to make arbitrary transfers of profits between different levels of governments and between different regions. This system has been criticized by many leading reform-minded economists as it may have promoted regional protectionism, segmented markets, and increased the local government administrative intervention of enterprises (Wu and Liu 1991). But in my view, this system was the most powerful single policy of reform in the 1980s. Firstly, it splits up the whole Chinese public economy into many small or mini-public economies. This is equivalent to delimiting the property rights between different levels of government such that each local government becomes the "central agent" and the real residual claimant of its own public economy, and each community becomes a "conglomerate". According to Proposition 2, such a splitting-up is naturally a Pareto-improvement: both monitoring and work efforts can be improved. Since local governments are closer to their original principals, they face more pressure to deliver the residual to the latter.[15] This is one of the main reasons for booming of TVEs. Secondly, it forces local governments to compete with each other. Although local governments may still use a planning mechanism to control their own enterprises, it can only do business with other communities through bargaining. The relationships between different provinces, cities, counties, towns, and villages, are more or less marketized. The competition between different governments (communities) makes the central planning system more and more difficult to operate and eventually to evolve into a dual-track system which is now converging into a single-track one. According to Li, Li Zhang (2000), cross-regional competition also triggers the ongoing privatization. Intuition is that when cross-regional competition is sufficiently intense in the product market, each region has to cut production costs significantly in order to maintain a minimum market share for survival. Given that the efforts of managers are hidden, in order to induce managers to reduce costs, local governments may have to grant total or partial residual shares to the managers. It is in the interest of local bureaucrats to forgo more residual shares of profits to the managers since the induced "incentive effect" more likely dominates the "distribution effect" as competition intensifies.

(3) The Management Contract System and SOE Reform.

The Chinese SOE reform first introduced in 1979 can be characterized with a continuously evolutionary process of shifting decision rights and residual claim from the government to the firm level. The reform started with no intention to abolish state ownership. Rather, it was intended to improve efficiency within state ownership. Nevertheless, the reform has been directed by a doctrine which is potentially conflicting with the conventional doctrine of state ownership. I call this new doctrine

[15] Here "residual" should include employment in a practical sense.

"the reform doctrine", according to which, both the decision rights and the residual claim should be shifted to the inside members of the firm (i.e., the manager and workers). The argument for shifting the decision rights to the manager of the firm is based on the assumption that decisions made at the firm level are more efficient than at the central agent level because of the information/communication problem. The argument for shifting the residual claim to inside members of the firm is based on incentive considerations. Although the modern theory of incentives was introduced into China much later, the pre-reform Chinese experience seems sufficient for both Chinese economists and reform-minded leaders to understand how essential the incentive system is for economic performance, although it has come much later for them to understand that the incentive system is primarily dependent on property rights and ownership structure.

In practice, shifting decision rights and residual claim has been conducted through various policies. In the early stage of reform, the basic policy was "*fangquan rangli*" (granting autonomy and sharing profit). From 1986 to the early 1990s, the dominant policy was the management contract system (MCS). The basic content of the MCS is to set profit sharing rules and delimit decision rights through contracts negotiated by the firm and the group of governmental agencies (normally including line department, and financial department; sometimes contracts are signed directly between management and mayors). The contract normally lasts for 3 to 4 years.

Although most Chinese economists argue that the SOE reform is far away from being successful, many empirical studies show that the average increase in the total factor productivity of SOEs is 2–5% in the 1980s (see Jefferson and Singh 1994 for a survey). Our theory can shed some lights upon understanding this improvement of efficiency. Through *fangquan rangli* and the MCS, managers of SOEs obtain considerable residual share of profits as well as decision rights. According to Propositions 3 and 4, when decision rights and residual claim simultaneously shift down from the government to management, efficiency improvement can be accordingly obtained through raising managerial work effort and reducing monitoring costs. As I argued in an early paper (Zhang, 1997), under the management contract system there are two kinds of incentives working for management. One is formal and explicit, and the other is informal and implicit. The formal and explicit incentive comes from the fact that managers (and worker) can legally claim part of the residual according to the signed contract. The informal and implicit incentive comes from manager's illegal expropriation of profits by manipulating accounts. As we argues earlier, this illegal expropriation can be an *ex post* efficient remedy to the *ex ante* inefficient ownership structure. Granting autonomy of business decisions makes the manager become a natural holder of part of control rights. By granting the partial residual to him, the residual claim and control right can be better matched at the firm level. This better matching certainly gives better motivation for the manager to make profits.

(4) Corruption and High Growth.

While, like all other socialist economies, the pre-reform Chinese economy was a corrupt one, corrupt behavior was tightly restricted by the "physical" nature of the economy and the centralized planning. One major consequence of economic reform

is that it has made corruption easier and led to pervasive agency discretion. This is because the industrial bureaucrats and managers have more autonomy and the economy is more commercialized, making it more difficult to have judicial and administrative checks on corruption. Corruption is widely regarded as a malignant tumor of the reform. Many people criticize the dual-track system partly because of its contribution to corruption (Wu and Liu 1991). However, the arguments derived from this paper show that corruption might have helped the Chinese economic growth through various ways, apart from mitigating bureaucrats' resistance to the reform. Firstly, corruption has directly improved the incentive systems for both bureaucrats and managers by giving them more freedom to enjoy rents illegally. To a great extent, how much perks or more generally how much personal benefits one can enjoy depends on how much (economic) profits one can make. As a result, the correlation between the performance and personal payoff is much stronger than official statistics show. Not surprisingly, even bureaucrats have become quite profit-oriented in making their decisions. Secondly, corruption has improved the efficiency of resource allocations. There are two reasons behind this. One is that it is less necessary for bureaucrats to embody their personal enjoyment of the residual into resource allocation as it was in the old system: they can first make a pie and than eat the pie. The other is that when bureaucrats in charge of resource allocation cannot take bribery, they care little about where the resources go. However, when they can take bribery, he cares about who can offer the highest bribery. On average, those who can offer the highest bribery must be the most efficient. In the early 1980s, it was through bribery that TVEs got most of their resources (physical and financial) out of the state sector. Thirdly, corruption has helped hardening the budget constraint on SOEs. The soft budget constraint has been claimed as a major reason for inefficiency. But the budget can be soft only if the state can make arbitrary transfer of profits between profit-makers and loss-makers. After the reform, the government's ability to redistribute profits has been greatly reduced by the profit-makers' ability to manipulate accounting data. Typically SOEs under-report their profits (by over-reporting costs). Although loss-makers may not need go bankrupt, they find more and more difficult to get subsidies from the government. This puts pressure on all enterprises to improve efficiency.

References

Alchian, Armen. 1965. Some economics of property rights. *Il Politico* 30 (4): 816–829.

Beckmann, Martin J. 1977. Management Productions and the Theory of the Firm. *Journal of Economic Theory*, 14: 1–18.

Calvo, G., and S. Wellisz. 1978, Supervision, Loss of Control and the Optimal Size of the Firm. *Journal of Political Economy*, 86: 943–952.

Calvo, G., and S. Wellisz. 1979. Hierarchy, Ability and Income Distribution. *Journal of Political Economy*, 87: 911–1010.

Grossman, Stanford, and Oliver Hart. 1986. The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration. *Journal of Political Economy*, 94: 691–719.

Hart, Oliver, and Begnt Holmstrom. 1987. The Theory of Contracts. In *Advances in Economic Theory*, eds. Bewley, T.

Jefferson, Gary, and Inderjit Singh. 1994. China's Industrial Performance: A Review of Recent Findings. Mimeo, the World Bank Policy Research Department.
Keren, M., and D. Levhari. 1979. The Optimal Span of Control in a Pure Hierarchy. *Management Science* 25: 1161–1172.
Li, David D. 1996. Ambiguous Property Rights in Transition Economies. *Journal of Comparative Economics* 23: 1–19.
Li, Shaomin, Shuhe, Li, and Weiying Zhang. 2000. The Road to Capitalism: Competition and Institutional Change in China. *Journal of Comparative Economics*, 28 (2): 269–292.
Qian, Yingyi. 1994. Incentives and Loss of Control in a Optimal Hierarchy. *Review of Economic Studies* 61: 527–544.
Rosen, Sherwin. 1982. Authority, Control, and the Distribution of Earning. *Bell Journal of Economics* 13: 311–323.
Shapiro, C., and J. Stiglitz. 1984. Equilibrium Unemployment as a Discipline. *American Economic Review* 74: 433–444.
Stiglitz, J. 1994. *Whither Socialism*? Cambridge: MIT Press.
Weitzman, Martin, and Xu Chenggang. 1994. Chinese Township and Village Enterprises as Vaguely Defined Cooperatives. *Journal of Comparative Economics* 18: 121–145.
Williamson, Oliver. 1967. Hierarchical Control and Optimal Firm Size. *Journal of Political Economy* 75: 123–138.
Wu, Jinglian, and Jierui Liu. 1992. *On Competitive Market System*. Beijing: Financial and Economic Press.
Zhang, Weiying. 1994. Entrepreneurial Ability, Personal Wealth, and the Assignment of Principalship: An Entrepreneurial/Contractual Theory of the Firm, D. Phil thesis, Oxford University.
Zhang, Weiying. 1997. Decision Rights, Residual Claim and Performance: A Theory of How the Chinese State Enterprise Reform Works. *China Economic Review* 7 (1): 68–82.

Appendix B
Decision Rights, Residual Claims and Performance: A Theory of How the Chinese State Enterprise Reform Works

B.1 Introduction

State-owned enterprise reform in China so far can be characterized by a continuously evolutionary process of reassignment of decision rights and residual claims from the central agent to the inside members of the firm.[1] The reform started with no intention to abolish public ownership. But it has been directed by a new doctrine which is potentially conflicting with the conventional doctrine of public ownership. I call this new doctrine "the reform doctrine". According to the reform doctrine, both decision rights and residual claims should be shifted to the inside members of the firm (i.e., the manager and workers). The argument for shifting the decision rights to the manager of the firm is based on the assumption that decisions made at the firm level are more efficient than at the central planner level because of information/communication problems. The theoretical legitimacy of this assumption dates back to Hayek (1948), while Chinese economists mainly based their argument on the observed poor performance of the traditional planning system.[2] The argument for shifting the residual claims to the inside members of the firm is based on incentive considerations. Although modern theory of incentives was just recently introduced into China, the pre-reform Chinese experience was sufficient for both Chinese economists and reform-minded leaders to understand how essential the incentive system is for economic performance.

The English version of this paper was published in China Economic Review, Volume 8, No. 1, 1997, pp. 67–82.

[1] Here the central agent refers to the central planner of the old system or loosely the government, or "state" In this paper, the central agent and the government are interchangeably used. Theoretically in a public economy, ordinary citizens are original principals who delegate ownership-authority to the government through a hierarchical structure. For a theoretical description of the structure of a public economy, see Zhang (1993).

[2] For a recent study on colocation of knowledge and decision authority, see Jensen and Meckling (1992).

W. Zhang, *The Origin of the Capitalist Firm*, DOI 10.1007/978-981-10-0221-2

The reform doctrine can be summarized by a popular official slogan that "the goal of reform is to make the firm independent, autonomous, and responsible for profits and losses". If this doctrine were fully implemented, public ownership would no longer exist in any economic sense; the government would be nothing more than a "bond-holder" on behalf of the "owners" of capital assets.[3] No doubt such a reformed system would Pareto dominate the traditional one. However, for reasons we will specify later, the reform doctrine has never been fully implemented. Nevertheless, what is important is that, just as the constitutional status of public ownership has made it possible for ordinary citizens (original principals) to acquire some residual via the central planner who were in the position of the acting principal, the reform doctrine has provided a legitimacy for the inside members of the firm to strive for their status of (self-)principalship. As a result, there are two legitimized principals competing with each other. The conflict has been solved in practice by a compromised principalship sharing arrangement determined by a bargaining process, in which the inside members of the firm have appeared more and more offensive, while the government has appeared more and more defensive.

The bargaining between the central agent and the inside members of the firm is mediated by industrial bureaus. Under the old system, the status of the industrial bureau was clearly defined: the agent of the central planner and the principal of the inside members of the firm. However, once the latter acquired their theoretically legitimized status of the principal, the industrial bureau becomes a double-faced agent: for the firm, it represents the government, and for the government, it represents the firm. The functioning of the industrial bureau is important for both the central agent and the firm to get their best deals. The central agent relies on the industrial bureau to provide information and monitor the firm, while the inside members of the firm rely on the industrial bureau for collectively bargaining with the central agent. As a result, the industrial bureau is in the position to "exploit" both the central planner and the inside members of the firm. This has caused much concern among Chinese economists.

This paper is intended to model the process of shifting decision rights and residual claims from the central agent to the inside members of the firm, and to analyze how the reform has improved performance of the state-owned enterprises. The paper is partly motivated by some empirical studies. Since the early 1980s the reform of state-owned enterprises (specially large- and middle-size enterprises) has been put on the top agenda by the reform policy makers as the core of the whole economic reform program. Although the dominant argument among Chinese economists is that the state enterprise reform has not been successful, Jefferson and Xu (1991) and Hay et al. (1994), among others, based on the structure-conduct-performance paradigm, find that the reformed Chinese state enterprises behave almost like the classical firm both in product markets and factor markets. This finding surprises most Chinese economists for it seems quite inconsistent with their intuition of the

[3] It has been argued by some economists that even ownership of the capital assets should be shifted to the inside members of the firm.

behaviour of the state enterprises.[4] Although one can raise questions about their econometrics methodology and data bias, there seems no reason to reject the finding simply because it is counter-intuitive. I myself am an advocate of private property right and have never believed in a public ownership-based market economy (Zhang 1986). However I do believe that their empirical finding is somewhat true. Because of this, I bear a responsibility for providing a theoretical interpretation for this finding based on property rights and incentive theory.

In contrast with most economists' criticism of bargaining, I find that the bargaining solution of allocations of decision rights and residual claims between the government and the enterprise might be socially preferred to the central-agent-set solution. Perhaps the most remarkable result of the paper is that managerial discretion of the state-owned enterprises has provided great incentives for managers to pursue profit and to improve efficiency. The managerial discretion might not be a good phenomenon in the private economy, as usually argued, but it might be good in a public economy. Given that the state is a legal residual claimant of the enterprises, managerial discretion might be an effective way for managers to become actual residual claimants and therefore is an implicit incentive mechanism for managers to work hard. Since the beginning of reform, autonomy of the firm and marketization of the economy have generated more and more opportunities and freedom for managers to enjoy profit, either legally or illegally. To a great extent how much perks or more generally how much personal benefit a manager can enjoy depends on how much profit he can make. As a result, managerial incentives have been greatly improved. This is the case partly because the managerial discretion has a positive effect on hardening the budget constraint. The soft budget constraint has been argued as a major reason for inefficiency of the state-owned enterprises. But the budget can be soft only if the government can make arbitrary transfer of profit between profit-makers and loss-makers. Since the reform, the government's ability to make such a transfer has been greatly reduced by the profit-makers' ability to manipulate accounting. Although the loss-makers may still not need go bankrupt, they find it more and more difficult to get subsidies from the government since the government itself is close to bankruptcy.[5] This puts pressure on all firms to improve efficiency. The unhappiness of Chinese economists with the behaviour of the reformed state-owned enterprises is mainly from observation of the managerial discretion. Those who still believe in a public ownership-based market economy should be greatly relieved by our finding.

Although shifting of decision rights and residual claims from the government to managers and its associated managerial discretion have greatly improved the managerial incentive mechanism, there is a fundamental problem which has not be solved for the Chinese state-owned enterprises. That is selection of high ability managers. An incumbent manager may have incentive to make profit, but at present there is no mechanism to ensure that only qualified people can be selected for management.

[4] Indeed, I remember that when Hay presented their paper at a seminar in Oxford in October 1991, organized by the Chinese Economic Association in the United Kingdom, he was heavily criticized by his Chinese audience.

[5] More recently the loss-makers do face a bankrupt threat under the new reform program.

The reason is that managers of the state enterprises are selected by bureaucrats rather than capitalists. As Zhang (1994) demonstrated, given that entrepreneurial ability is not easy to observe, the existence of capitalists is crucial for only high ability people to occupy management positions. To further improve efficiency of the Chinese economy, privatization of the state enterprises is not only necessary but also inevitable. Given the status quo of China's institutional structure, which must be respected during the reform process, "capitalistization" of incumbent bureaucrats and managers may be the only feasible way for China to privatize the state-owned enterprises.[6] Fortunately, we have every reason to believe that the reform process is already well on this way.

The paper is organized as follows. In the second section, I present the model. The third section analyzes the bargaining solution of allocation of decision rights and residual claims between the central agent and the inside members of the firm. The fourth section discusses the effect of managerial discretion on hardening the budget constraint of the state-owned enterprises. The fifth section concludes the paper with remarks on further reform process.

B.2 Description of the Model

For simplicity, we normalize the inside members (the manager and workers) of a state enterprise to a single agent who has a well-defined utility function. Define the gross profit function as follows:

$$\pi = \pi(\theta, a, \lambda, k, s) \tag{B.1}$$

where π is the *actually realized* gross profit, θ the managerial ability of the firm, a work effort of the inside members of the firm, λ a parameter of decision rights held by the firm (the total decision rights are normalized to one such that $\lambda = 0$ implies the firm has no autonomy at all, and $\lambda = 1$ implies the firm enjoys the full autonomy), k the capital stock used by the firm but owned by the government (on behalf of the original principals), $s \in [\underline{s}, \bar{s}]$ is the state of nature. Note here decision rights refer to autonomy of deciding what to do and how to do it within given human and capital resources.

We shall assume that π is a monotonically increasing, concave function of all θ, a, λ, k, and increasing with s. That is, $\frac{\partial \pi}{\partial x} > 0$, and $\frac{\partial^2 \pi}{\partial x^2} < 0$ for $x = \theta, a, \lambda, k$; $\frac{\partial \pi}{\partial s} > 0$. In addition, we shall assume that $\frac{\partial \pi}{\partial \lambda} \to 0$ as $\lambda \to 1$, which means that when the manager holds full autonomy of making decision, reducing autonomy by a ε will have little effect on efficiency.[7] Because the importance of the managerial ability depends on autonomy of the firm, we assume that $\frac{\partial^2 \pi}{\partial \theta \partial \lambda} > 0$, and $\frac{\partial \pi}{\partial \theta} = 0$ for $\lambda = 0$.

[6] By capitalistization, I mean transforming incumbent bureaucrats and managers into capitalists.

[7] Note that by directly introducing autonomy into the profit function, we suppress the role of the central agent's effort. That is, the relative importance of the central agent's effort over the inside

Assume that the distribution of profit between the central agent and the inside members of the firm takes a linear form[8]:

$$y_F = f + \beta(\pi - f - g) = (1 - \beta)f + \beta(\pi - g) \tag{B.2}$$

$$y_G = g + (1 - \beta)(\pi - f - g) = \beta g + (1 - \beta)(\pi - f) \tag{B.3}$$

where y_F is the profit retained by the inside members of the firm, and y_G the profit delivered to the central agent; f and g are their respective fixed terms, and β and $(1 - \beta)$ are their respective residual shares after deduction of the fixed terms.

From the standard principal-agent theory,[9] given that the government is less informed of the state of nature, and monitoring is costly, if the inside members of the firm are risk-neutral, the optimal contract is to set $\lambda = 1$, $\beta = 1$, $f = 0$ and $g > 0$; that is, the manager has the full autonomy to make decisions, the inside members are the full residual claimant, and the government receives a fixed term in return for capital services. This is exactly the policy implication of the reform doctrine.

The first constraint which makes such a "first best" contract infeasible (unenforceable) is that, on the one hand, because of the lower-bound constraint of the profit and the wealth constraint of the inside members, the guaranteeable fixed term might be too low to compensate capital services (even negative if the bad state occurs); on the other hand, because of the lower-bound constraint of consumption, it is impossible to set $f \leq 0$.[10]

Because of this, it is inevitable to set $\beta < 1$; that is, the residual has to be shared between the government and the inside members of the firm. This leads to the second constraint. If the inside members of the firm are able to manipulate accounting of profits, what actually shared by the central agent is not $(1 - \beta)$ times the real profit but the reported one. In the extreme case, the reported profit might be nothing more than $(f + g)$or even less, although the actual residual is very large. By doing so, the inside members of the firm enjoy all the residual and leave the central agent

---

members' effort decreases as the firm's autonomy increases. The effect of λ on π can be understood as the net benefit of switch of decision rights from the central agent to the manager of the firm.

[8]The practical contract between the government and the firm is typically piece-wise linear. The linearity assumption should not affect the main arguments. In addition, in this paper, we do not consider taxation problem which equivalently exists in a private economy.

[9]For an excellent survey of principal-agent theory, see Hart and Holmstrom (1987).

[10]It is worthwhile to make a comparison with a capitalist firm. In the capitalist firm, bond-holders are protected by share-holders: the fixed return to the bond-holders is guaranteed by assets of the share-holders when the firm makes loss. This protection relieves the bond-holders from regular monitoring; monitoring by the bond-holders occurs only when the share-holders' assets value becomes zero, i.e., bankruptcy occurs, which is very infrequent. In contrast, for a firm in which the only capital-supplier is the bond-holder, bankruptcy would become frequent and the bond-holder must monitor
(Footnote 10 continued)
the firm regularly. This implies that the bond-holder is actually a "bad-holder": he plays the role of monitoring like a share-holder when the performance is bad, but has no status to claim the residual when the performance is good.

very little.[11] The observation is that ability for the inside members to manipulate accounting is increasing with the degree of autonomy of the firm. The intuition is that when the manager has various decision choices, he can transfer funds from one use to another, can overstate input prices and understate output prices, and even open a private account. It is very difficult for the central agent to judge whether a particular use of funds is reasonable or not.[12] This implies that it is impossible for the central agent to agree to set $\lambda = 1$. The manager cannot hold all the decision rights.

A special case of accounting manipulation is where the manager reports a profit when the firm actually makes loss. This is possible because the manager has the right to make investments. Funds borrowed from banks on the pretext of investment may be simply "eaten" as bonuses. In some cases, the "profit" is just depreciation of the capital stock. Because there is no market value of the assets, eating up the principal is very unnoticeable for outsiders. In fact, some capital asset might be delivered to the central agent as profit.

The third constraint comes from impossibility of separation of exercising decision rights from personal enjoyment of rents by the central agent. In a public economy, although the government bureaucrats are motivated by their personal enjoyment of rents, they have no legal rights to pocket the residual; the personal enjoyment of the rents must be embodied into exercising decision rights (see Zhang 1993 for more discussion). Fully transferring the decision rights to the firm implies depriving bureaucrats of the rent, which is not only impossible but also might not be desirable for the society. Because the bureaucrats are multi-task agents, the loss of their incentives would lead to a collapse of the whole society.

These three constraints are complementary. They help each other in supplying the rationales for the central agent to refuse to transfer fully both decision rights and residual claims to the inside members of the firm. As a result, the only feasible contract is to set $\beta < 1$ and $\lambda < 1$. However, the feasible set defined by these restrictions is still large. There is a room for bargaining.

For simplicity, assume that the fixed terms f and g are set such that $(f + g)$ is just equal to $\underline{\pi}$, the lower-bound of the profit (in the worst case), which is common knowledge. Assume $f > 0$, that is, the central agent guarantees the inside members a "subsistence level income" (If $\underline{\pi} = 0$, $f = -g$, which implies that $g < 0$). Because capital abuse by the inside members of the firm is the most serious when the worst state occurs, we miss an important insight by making this assumption. To remedy this, we explicitly introduce a term $\alpha(\lambda)$ to capture the capital abuse, where $0 < \alpha < 1$, $\frac{\partial \alpha}{\partial \lambda} > 0$.

We now use π to denote the net profit (i.e., the gross profit minus $(f + g)$). Denote by π^0 the reported profit and assume that π and π^0 satisfy the following relation:

$$\pi^0 = \delta(\lambda)\pi, \text{ where } 1 \geq \delta > 0, \frac{\partial \delta}{\partial \lambda} < 0 \text{ and } \delta(0) \equiv 1$$

[11] Accounting manipulation should be understood as a reduced expression of the inside members' consumption of all perks and other forms of managerial discretion.

[12] For a theoretical analysis of the effect of richness of the action space on the form of the contract, see Holmstrom and Milgrom (1987).

That is, ability to manipulate accounting is increasing with the degree of autonomy; and it is impossible to manipulate accounting when the manager has no decision right. Note that we assume that the manager only under-reports profit (over-report of profit is captured by the term of capital abuse.) We assume that this relation is known to the central agent but is unverifiable (otherwise there would be no accounting manipulation).

Because the central agent gets its residual according the reported profit rather than the real profit, the real residual is shared as follows[13]:

$$y_F = \beta\pi^0 + (\pi - \pi^0) = (1 - \delta(1-\beta))\,\pi \tag{B.4}$$

$$y_G = (1-\beta)\pi^0 = \delta(1-\beta)\pi \tag{B.5}$$

We call $(1 - \delta(1-\beta))$ "the real residual share" of the inside members of the firm and $\delta(1-\beta)$ "the real residual share" of the central agent (accordingly, β is called "nominal residual share"). From (4) and (5), we see that for any given nominal residual share, the real residual share held (and profit retained) by the inside members is increasing with the degree of autonomy, while the real residual share held by (and profit delivered to) the central agent is decreasing with the degree of autonomy. For a given λ, the deviation of the real from the nominal depends on accounting manipulation technology, which we do not explore here.

Assume that both the central agent and the inside members of the firm are risk-neutral and their respective utility functions are defined as follows (note: now π denotes the expected profit):

The inside members of the firm:

$$U_F = (1 - \delta(\lambda)(1-\beta))\,\pi(\theta, a, \lambda, k) + \alpha(\lambda)k - C(a) \tag{B.6}$$

where $C(a)$ is the cost function of work effort, $\frac{\partial C}{\partial a} > 0$, $\frac{\partial^2 C}{\partial a^2} > 0$.

The central agent:

$$U_G = \gamma\,(\delta(\lambda)(1-\beta)\pi(\theta, a, \lambda, k) - \alpha(\lambda)k) + (1-\gamma)G(\lambda) \tag{B.7}$$

where $\frac{\partial G}{\partial \lambda} < 0$, $\frac{\partial^2 G}{\partial \lambda^2} < 0$.

The first utility function is self-explanatory. The second can be interpreted as follows. The central agent's utility is a weighted sum of two parts: the first part is monetary residual and the second is non-monetary term which captures the idea that he has to exercise some decision rights directly in order to enjoy rents. We use γ to capture both the ownership constraint from the original principals and the possibility of pocketing the rent. The central agent cares for the residual both because of his personal enjoyment and because of his responsibility for delivering some minimum benefit to the original principals. This implies $\gamma > 0$. However, how important the residual is depends on to what degree he can pocket the residual after delivering the

[13] We drop f and g because they are irrelevant for the optimum.

minimum requirement to the original principals. The easier pocketing is, the more important y_G is, which in turn implies the less necessary to use the decision rights as a tool of enjoying the rent. We assume that $\gamma = 1$ if he can directly pocket the rent, in which case, the decision rights are useless for him. Note we implicitly assume that the central agent has no direct preference for the decision rights *per se*, and that indirectly consuming the rent by exercising the decision rights is less efficient than directly (if possible).[14]

The contract between the central agent and the insider members of the firm is characterized by a set of λ(share of decision rights) and β (residual share). Bargaining between the central agent and the inside members is to set (β, λ). We now turn to discuss this problem.

B.3 The Bargaining Solution of Allocation of Decision Rights and Residual Claims

Before discuss how β and λ are to be set, let us first solve the optimal work effort chosen by the inside members for given β and λ. The first-order condition implies that

$$\frac{\partial C}{\partial a} = (1 - \delta(\lambda)(1 - \beta)) \frac{\partial \pi}{\partial a} \tag{B.8}$$

It is easy to see that the optimal choice of work effort is increasing with both the residual share and the decision rights. It is worth noting that the decision rights affect the optimal work effort through two channels: the first is its direct effect on the profit function and the second is its effect on the real residual share through accounting manipulation. For any given β and λ, an improvement in accounting manipulation ability will surely increase work effort. This implies that from the incentive point of view, managerial discretion is not a bad thing.

Condition (B.8) is the incentive compatibility constraint which the contract (β, λ) must satisfy. We now turn to determination of β and λ.

First, let us consider the case in which the central agent has exclusive authority to set both β and λ. Then the central agent's problem is

$$\begin{aligned} &\max_{\beta,\lambda} \; U_G = \gamma\left(\delta(\lambda)(1-\beta)\pi(\theta, a, \lambda, k) - \alpha(\lambda)k\right) + (1-\gamma)G(\lambda) \\ &\text{s.t. } \frac{\partial C}{\partial a} = (1 - \delta(\lambda)(1-\beta)) \frac{\partial \pi}{\partial a} \end{aligned} \tag{B.9}$$

By rearranging the first order conditions, we obtain (we assume that the interior solutions exist and qualify this assumption later):

[14]More ideal formal of the central agent's utility function is $(\delta(\lambda)(1 - \beta)\pi(\theta, a, \lambda, k) - \alpha(\lambda)k)^{\gamma}(G(\lambda))^{1-\gamma}$.

$$\beta: \ (1-\beta)\frac{\partial \pi}{\partial a}\frac{\partial a}{\partial \beta} = \pi \tag{B.10}$$

$$\lambda: \ \delta(\lambda)(1\text{-}\beta)\left(\frac{\partial \pi}{\partial \lambda} + \frac{\partial \pi}{\partial a}\frac{\partial a}{\partial \lambda}\right) = \frac{\partial \alpha}{\partial \lambda}k - (1-\beta)\frac{\partial \delta}{\partial \lambda}\pi - \frac{1-\gamma}{\gamma}\frac{\partial G}{\partial \lambda} \tag{B.11}$$

where $\frac{\partial a}{\partial \beta}$ and $\frac{\partial a}{\partial \lambda}$ are defined by the first-order condition (B.8).

The LHSs of the equations are the marginal benefits and the RHSs are the marginal costs of increasing β and λ, respectively. The marginal benefit of β comes from the effect on work effort through the incentive compatibility constraint, and the marginal cost of β is the direct reduction of the residual share. The marginal benefit of λ contains two parts: the first is the direct efficiency improvement $\left(\frac{\partial \pi}{\partial \lambda}\right)$ based on information advantages held by the manager, and the second is the indirect effect on work effort through the incentive compatibility $\left(\frac{\partial \pi}{\partial a}\frac{\partial a}{\partial \lambda}\right)$; the marginal cost of λ contains three parts: the first is the effect on capital abuse $\left(\frac{\partial \alpha}{\partial \lambda}k\right)$, the second is the effect on accounting manipulation $\left((1-\beta)\frac{\partial \delta}{\partial \lambda}\pi\right)$, and the third is the effect on the use of decision rights to carry out the rent enjoyment $\left(\frac{1-\gamma}{\gamma}\frac{\partial G}{\partial \lambda}\right)$. The optimal β and λ for the central agent, β_G^* and λ_G^*, are determined by equalization of the marginal benefits and the marginal costs.

The condition for the existence of the interior solution of β seems fairly satisfied. Clearly $\beta = 1$ is not optimal because that implies the central agent gets nothing. $\beta = 0$ can be optimal only if either the direct monitoring by the central agent is sufficiently effective, which has been excluded, or the inside members of the firm has very great ability to manipulate accounting such that even if the central agent is the full nominal residual claimant, the inside members of the firm actually claims a considerable part of the residual.[15] However, if the later is the case, the central agent may simply reject assigning any decision rights to the manager, which implies that the whole situation would go back to the *status quo*. So it seems reasonable to assume that $0 < \beta_G^* < 1$.

The condition for the existence of the interior solution of λ is to be qualified. $\lambda = 0$ can be excluded on basis of information advantage of the manager in making decisions, which has been the major rationale for reform (our technical assumptions are sufficient for excluding $\lambda = 0$.) There are several situations each of which can ensure that $\lambda = 1$ cannot be optimal. The first is our assumption that $\frac{\partial \pi}{\partial \lambda} \to 0$ as $\lambda \to 1$. The second is when the inside members of the firm can fully manipulate accounting when they have the full autonomy, i.e., $\delta(1) = 0$. But $\delta(1) = 0$ seems not realistic and we shall exclude it. The third situation is that $\frac{\partial G}{\partial \lambda} \to -\infty$ as $\lambda \to 1$, which means that holding no decision rights at all would be a disaster for the central agent. This is extreme but can be justified given that the government bureaucrats have no legal right to pocket the rent. When the above three situations do not hold, the existence of the interior solution requires that the marginal costs of λ grow faster

[15] In the corrupt public economy, the original principals are the full nominal residual claimant, while the government officials actually claim a considerable part of the residual.

than the marginal benefits. How fast the marginal costs of λ grow depends on how λ affects the marginal ability of abusing capital, the marginal ability of manipulating accounting, and the marginal effect on the central agent's ability to use the decision rights to enjoying the rent. If all these three marginals are increasing with λ (that is, $\frac{\partial^2 \alpha}{\partial \lambda^2} > 0$, $\frac{\partial^2 \delta}{\partial \lambda^2} < 0$ and $\frac{\partial^2 G}{\partial \lambda^2} < 0$), the condition seems satisfied, because given that $\frac{\partial^2 \pi}{\partial \lambda^2} < 0$ and $\frac{\partial^2 \pi}{\partial a^2} < 0$, the marginal benefits cannot grow very fast even if $\frac{\partial^2 a}{\partial \lambda^2} > 0$ (this might be the case since λ has two positive effects on a); the intuition is that λ drives the marginal cost through three channels while it drives the marginal benefit through only one channels.

Based on the foregoing arguments, we shall assume that $0 < \beta_G^* < 1$ and $0 < \lambda_G^* < 1$. We now do some comparative statics of how β_G^* and λ_G^* are dependent on the managerial ability (θ), capital stock (k), the parameter of possibility of pocketing rents (γ), accounting manipulation ability (δ), and the parameter of capital abuse (α). In other words, from the point of view of the central agent's interest, should the optimal residual share and the optimal autonomy of the firm be increasing, constant or decreasing with the managerial ability, capital stock, the possibility of pocketing rents, accounting manipulation ability, and easiness of capital abuse, respectively?

Since the general form of functions is not tractable, we shall restrict ourselves to the following simple case. Assume that the profit function is $\pi = a\theta^{1/2}\lambda^{1/2}k^{1/2}$,[16] the cost function $C(a) = a^2/2$, $\delta(\lambda) = 1-\tau\lambda$, $\alpha(\lambda) = \alpha\lambda$, and $G(\lambda) = (1-\lambda^2)A$ (Here $0 \leq \tau < 1$ and A is sufficiently large). Then we have

$$\beta_G^* = 1 - \frac{1}{2(1-\tau\lambda_G^*)} \tag{B.12}$$

$$\lambda_G^* = \frac{\gamma(\frac{1}{4}\theta - \alpha)k}{2(1-\gamma)A} \tag{B.13}$$

(Note that for simplifying the expression of β_G^*, we use λ_G^* defined by (B.13) to suppress other parameters in (B.12).)

From (B.12) and (B.13), we have the following results:

(i) The easier the accounting manipulation is, the less nominal residual share the central agent would be willing to give to the inside members of the firm. (In fact, for ensuring that $\beta_G^* > 0$, τ must be sufficiently small.)

(ii) The higher the managerial ability is, the more autonomy but the less nominal residual share would be given. (However, one should be cautious in applying this point to the reality before discussing the managerial selection mechanism.)[17]

(iii) The easier the capital abuse is, the less autonomy but the more nominal residual share would be given. (In fact, for ensuring that $\lambda_G^* > 0$, $\alpha < \theta/4$ must be held.)

[16] This profit function violate our early assumption that $\partial\pi/\partial\lambda \to 0$ as $\lambda \to 1$, but the basic results should not be affected.

[17] We shall briefly discuss the managerial selection mechanism in the concluding section.

(iv) The larger the capital stock is, the more autonomy but the less nominal residual share would be given.
(v) The central agent wishes to give more autonomy but less residual share to the firm as pocketing rents becomes easier (i.e., γ increases).

The above results are quite intuitive. Results (i)–(iii) and (v) are quite consistent with the reality. I believe that result (v) partially explains the observation that in the regions where the government bureaucrats have more opportunities to take bribery as the transactions are more and more monetized, managers enjoy more freedom to make their decisions (the managers simply buy autonomy from the bureaucrats. However, the bureaucrats must hold the essential part of decision rights in order to capture these opportunities, given that directly taking money from the state budget is illegal. The right to appoint management is such an essential right.)

However, result (iv) seems not consistent with the observation that the government has been reluctant to vitalize the large firm. This inconsistency comes from that in our simple case, we ignore the interaction between the capital stock (k) and the ability of capital abuse (α). In reality, the larger firm is more easier to abuse capital and to manipulate accounting profits (because of the big action space.) If we take into account this fact together with result (iii), the effect of an increase in the capital stock on λ is ambiguous and result (iv) may be reversed.

It is interesting to note that the effect of θ, α, γ and k on λ_G^* is just opposite to the effect on β_G^*. This is because that λ_G^* and β_G^* are substitutes in providing a given effort but λ_G^* has two additional cost effects. For example, when capital abuse becomes more serious, the cost of providing incentives through β is cheaper than through λ; as a result, the central agent would like to reduce λ and at the same time increase β.

In the above discussion, we assumed that the central agent has the exclusive authority to set the residual share and the degree of autonomy of the firm. What would happen if the inside members of the firm held the exclusive authority to set β and λ ? Absolutely, they would set $\beta = 1$ and $\lambda = 1$! That is, the optimal β and λ for the inside members, denoted by β_F^* and λ_F^*, are strictly greater than β_G^* and λ_G^*. The conflicts occur. Because in reality, neither of the two sides has the exclusive authority to set β and λ (the central agent's authority has been undermined by the reform doctrine, while the inside members' authority cannot be justified by the conventional doctrine of public ownership which is still constitutionally alive), the conflicts can be solved only through bargaining between the two sides. The resulting solutions of β and λ are between $\left(\beta_G^*, \lambda_G^*\right)$ and $\left(\beta_F^*, \lambda_F^*\right)$.

An important question is: Which is *socially* preferred, $\left(\beta_G^*, \lambda_G^*\right)$ or $\left(\beta_F^*, \lambda_F^*\right)$? Denote by β_S^* and λ_S^* the social optimum. The income distribution aside, if we assume that the social welfare function is equal to the total profit π minus the cost of effort and capital abuse term (capital abuse is *socially* bad because it destroys future productivity), that is

$$\max_{\beta,\lambda} \ \pi\text{-}\alpha(\lambda)k - C(a)$$
$$\text{s.t. } \frac{\partial C}{\partial a} = (1 - \delta(1 - \beta)) \frac{\partial \pi}{\partial a}$$

β_S^* and λ_S^* satisfy the following first-order conditions:

$$\left(\frac{\partial \pi}{\partial a}-\frac{\partial C}{\partial a}\right)\frac{\partial a}{\partial \beta}=0,\ \text{and}\ \frac{\partial \pi}{\partial \lambda}+\left(\frac{\partial \pi}{\partial a}-\frac{\partial C}{\partial a}\right)\frac{\partial a}{\partial \lambda}-\frac{\partial \alpha}{\partial \lambda}k=0$$

It is easy to see that $\beta_G^* < \beta_S^* = \beta_F^* = 1$ and $\lambda_G^* < \lambda_S^* < \lambda_F^* = 1$. That is, the social optimal residual share is equal to that preferred by the inside members of the firm and greater than that preferred by the central agent, while the social optimal autonomy is strictly less than that preferred by the inside members but greater than that preferred by the central agent. The social optimum is not incentive compatible, since given $\beta = 1$, the central agent would have no interest to implement the remaining decision right allocated to him (equal to $1 - \lambda_S^*$). The only possible way to implement the social optimum is to grant the residual claims to the insiders permanently so that they no longer have any incentive to abuse capital.[18] But this would change the whole game: the central agent would become redundant. Given that the social optimum is unimplementable, the resulting solution from bargaining between the two sides is strictly preferred to that exclusively set by the central agent, because the bargaining solution is more close to the social optimum.[19]

We now come to the point about the incentive effect of managerial discretion. From the incentive compatibility constraint (B.8), we see that for given β and λ, the managerial discretion has positive effects on the inside members' incentive. However, from equations (B.12) and (B.13), we see that the managerial discretion has negative effects on β_G^* and λ_G^*. Therefore in the statics model, the total effect is ambiguous. (In the simple case, the real residual share $(1 - \delta(\lambda)(1 - \beta))$ is constantly equal to 1/2 when the central agent has the exclusive right to set the contract.) Nevertheless, the Chinese experience suggest that the positive effects dynamically

[18]The policy proposal based on this kind of argument is to extend the tenure of contract between the government and the firm to 3–5 years. The effect has been positive. In agriculture, the long tenure contract of the land has proved quite successful in preventing land abuse.

[19]In reality, the bargaining over β and λ does not go directly between the central agent and the inside members of the firm, but is mediated by the industrial bureau. There are two questions associated with this mediated bargaining: Does the industrial bureau's intermediation make the pie bigger or simply share the pie which has already existed? Does the involvement of the industrial bureau increase or reduce the residual share and the degree of autonomy? Most Chinese economists give the negative answers. They argue that many decision rights released by the central agent has been hoarded by the industrial bureau rather than passed to the manager as they should be, and the industrial bureau has acquired its rent simply by exploiting the inside members of the firm and even the central agent. This may not be quite correct. From the preceding analysis, we have seen that λ_G^* is negatively affected by the inside members' ability to abuse capital; in other words, the central agent would wish to grant greater autonomy to the firm when the inside members' ability of abusing capital is low than when it is high. Because autonomy is socially productive, restriction of autonomy has a negative effect on the total social surplus. This implies that a Pareto improvement would come if some not very costly information is available so that the inside members are less easy to abuse capital. One role played by the industrial bureau is to collect information of the firm and to monitor the inside members' non-productive activities. By doing so, at least theoretically, the industrial bureau can increase λ_G^*. Furthermore, the information provided by the industrial bureau may make it possible for the central agent to get a higher fixed term g (or less negative) so that a greater residual share might be agreed in bargaining. This is of course a Pareto improvement. In summary, the industrial bureau may make contributions to rent generation through the effect on both β and $\lambda$of information collection and monitoring work.

dominate the negative effects. There are two reasons. First, the government had been long mistaking that the state enterprises typically over-reported their performance, and only recently does it recognize that the situation is just the opposite. The second reason is that the bargaining between the government and the enterprises has been a base-climbing one; that is, any (β, λ) set in the current contract will become a base for further bargaining, and the government can hardly down-adjust (β, λ). As a result, both the nominal and the real terms have been increasing. What is the most important is that when one studies the behaviour of the SOEs, one cannot just look at the explicit incentive provided by formal contracts; one must also look at the implicit incentive provided by the managerial discretion. Today hiding profits is a pervasive phenomenon in China. Although managers cannot easily pocket cash, they have various ways to spend money (some even register companies in other countries). As a result, the correlation between their personal benefit and total real profit is much stronger than official statistics shows. This strong correlation has greatly improved managerial incentive. It also suggests that accounting statement is not a right signal of performance of the SOEs.

B.4 The Effects on the Soft-Budget Constraint of Managerial Discretion

In this section, I relate the preceding discussion to a hot topic in the literature of the socialist economy, that is, the soft-budget constraint, to show how the Chinese economic reform has improved performance of the firm through hardening its budget constraint.

The soft-budget constraint was originally coined by Kornai (1980) to characterize the loose correlation between performance of the firm and pay-off of the insiders in the socialist economy. Theoretically, the conception is not quite correct because the loose correlation is not unique for the socialist firm: even in the capitalist firm, the correlation between payment of workers and performance of the firm is very loose.[20] In any kind of firm, it is the residual claimant (owner) who takes responsibility for performance. If the insiders of the firm are not the residual claimant, there is no rationale for them to take responsibility. On the other hand, the residual claimant can never escape from his responsibility, even under the pre-reform socialist economy. However, this kind of objection might be misleading, because the conception points to a more fundamental flaw of the public economy, that is, the responsibility for performance of the firm is so diversified that there is little pressure on an individual firm to improve its efficiency. An implication is that to harden the budget constraint, the residual claims must be shifted from the central agent to the inside members of the firm.

[20] In a classical capitalist firm, the correlation is zero.

Then the problem to be considered is: Has the Chinese economic reform been successful in the sense of hardening the budget constraint? Clearly, if the contract reached *ex ante* is renegotiation-proof *ex post*, as we have assumed so far, the budget constraint can be said "hard". The problem is that in reality, the contract is at most cases renegotiable *ex post*. Demand for renegotiation can come from either the central agent or the insiders of the firm. Typically when the performance has proved good, the central agent asks for renegotiation to decrease β or increase g; when the performance has proved bad, the insiders ask for renegotiation to increase β or reduce g. Because the performance is a joint outcome of effort and unexpected events some of which are controlled (or affected) by the central agent's actions, it is very hard for one party to reject demand for renegotiation by the other. For example, if a bad performance coincides with a tight macro-policy which has not been fully taken into account at the contracting time, the central agent cannot require the firm to deliver profit according to the contract; and if a good performance coincides with a policy-induced increase of output prices, the firm has to deliver some extra surplus to the central agent. Renegotiation may benefit both sides through its insurance effect or improving *ex post* efficiency of resource allocation, but it generates a big negative effect on incentives. When the inside members of the firm anticipate that they cannot retain the extra profit according to the *ex ante* contract, they will stop working further once a reasonable target is reached; when they anticipate that they do not need suffer from a big loss, they might give up once the target proves more difficult than expected. The prevalence of renegotiation has led most Chinese economists to conclude that the soft-budget constraint problem has been little changed. My argument is much different. I certainly agree that the budget constraint of a state-owned enterprise is much softer than that of a typical capitalist firm. However, it seems to me that it is not only much harder than at the pre-reform stage but also much harder than what the statistics data suggests for the following reasons.

First, the fact that the contract is not renegotiation-proof does not mean that renegotiation is costless. The intuition suggests that the *ex post* renegotiation is quite costly. For the inside members, the costs include both pecuniary expense of bribing the central agent or intermediators and non-pecuniary loss (such as fall of the probability of future promotion). Their intention to reduce g or increase β normally faces resistance from the central agent who would be hurt if the renegotiation takes place. Although the loss-makers can always find some exogenous factors to blame and may make themselves better off than when the contract is executed, they can hardly be as well-off as the profit-makers. Therefore, they will resort to renegotiation only if the cost of working is greater than the cost of renegotiation. The renegotiation demand from the central agent is normally met with strong resistance from the inside members. In particular, given that the insiders have some freedom to manipulate accounting, at most cases, it is almost impossible for the central agent to do better than by carrying out the *ex ante* contract.[21]

[21]This leads to a phenomenon of asymmetric renegotiation: renegotiation is more likely to take place when the firm makes loss than when it makes big profit. Chinese economists summarize this

Second, accounting manipulation (as well as other kinds of agency discretion) of an individual firm does not only improve its own incentive system, but also has the effect on hardening the budget constraint of other firms. The reason is that the central agent's budget constraint cannot be soft! An individual firm's budget constraint can be soft only because the central agent can transfer profits among the different firms. The central agent's ability to transfer profit is constrained by its total revenue which is in turn constrained by agency discretion at the firm level. The more share the profit-making firms can retain, the less available for the central agent to subsidize the loss-making firms. In the extreme case, if there are no profit-makers delivering any profit to the central agent, the central agent can do nothing in helping the loss-makers but let them go bankrupt. When the firms anticipate that they are less likely to get help from the central agents, they have to work for themselves. The argument can be formulated as follows.

Assume that there are m firms, and firm i's net *expected* revenue depends on both its own profit π_i and a subsidy f_i from the central agent as follows:

$$y_i = f_i + (1 - \delta(\lambda)(1 - \beta))\,\pi_i \tag{B.14}$$

Because f_i is normally negatively related to π_i, which implies that f_i is a decreasing function of work effort, we can assume that firm i is endowed with a fixed total effort a_i. The decision facing firm i is to divide a_i into two parts: *ex ante* work effort a_i^W and *ex post* bargaining effort a_i^b, to maximize $f_i(a_i^b) + (1 - \delta(\lambda)(1 - \beta))\,\pi_i(a_i^W)$. In other words, firm needs to decide whether to work *ex ante* for a big π_i or wait for *ex post* bargaining for a big f_i.

The central agent's budget constraint is that[22]

$$\sum_{j=1}^{m} f_j \leq \sum_{j=1}^{m} \delta_j(\lambda_j)(1 - \beta_j)\pi_j \tag{B.15}$$

Because $\sum_{j=1}^{m} \delta_j(\lambda_j)(1 - \beta_j)\pi_j$ is the total funds available for the central agent to transfer between firms, we may define the average (upper-bound) degree of softness of the budget constraint as follows:

(Footnote 21 continued)
phenomenon by "*fu-ying bu fu-kui*" (responsible for profit but not for loss). This may be not fair from the point of view of social justice, but it is certainly preferred to the symmetric irresponsibility (i.e., responsible neither for profit nor for loss) from the point of view of efficiency.

[22]The central agent may use deficit budget to subsidize the loss-makers, but this does not affect the argument because the deficit is bounded. In addition, we assume the central agent does not collect a fixed term.

$$S = \frac{\sum_{j=1}^{m} \delta_j(\lambda_j)(1-\beta_j)\pi_j}{\sum_{j=1}^{m} \pi_j} \tag{B.16}$$

S is decreasing with managerial discretion parameter $\delta_j(\lambda_j)$. Suppose that all firms are identical *ex ante*. Then the *expected* maximal subsidy for a representative firm is constrained as follows:

$$f_i = f \leq \frac{1}{m} \sum_{j=1}^{m} \delta_j(\lambda_j)(1-\beta_j)\pi_j \tag{B.17}$$

Even without explicitly solving the firm's optimal problem, we can see that an individual firm's incentive to wait for *ex post* bargaining falls as all other firms' ability of accounting manipulation increases. In the extreme, if all profit-makers can fully manipulate accounting such that $\delta_j = 0$ for all j with π_j, it would be foolish for firm i to wait for bargaining *ex post* instead of working *ex ante*.[23]

B.5 Concluding Remarks: Capitalistization of Incumbent Bureaucrats and Managers

Shifting of decision rights and residual claims from the government to inside members of the firm and its associated managerial discretion have greatly improved performance of the state-owned enterprises through both direct incentive effects and hardening budget constraints. However, there are still many problems to be solved by further reforms, one of which is the mechanism of selecting managers. Currently the government officials hold rights to select management but bear little responsibility for the consequences of their selection. Therefore they have no right incentive to find and choose high ability people. To ensure that only high ability people would be professional managers, authority of selecting management should be transferred from bureaucrats to capitalists (Zhang 1993). This calls for privatization of the state enterprises. China is well on this way. The observation suggests that privatization of the state enterprises will be a process of capitalistization of (some) incumbent bureaucrats and managers (and even some workers).[24] As the reform proceeds, incumbent bureaucrats find it more and more difficult to capture rents in their current posi-

[23]This kind of effect has been ignored by most economists who, on the one hand, blame the government for the soft budget constraint, and on the other hand, argue that the central agent's budget revenue is too small.

[24]Capitalistization will be accompanied by "debureaucratizatio" because of social pressure. In the following, "capitalistization" should understood as a dual process of capitalistization and debureaucratization. Yang (1988) proposes capitalistization of bureaucrats as a policy suggestions for reform.

tions, because of disappearance of monopolistic profits and managerial discretion. Experience teaches them that they can do much better by directly doing business with their remaining political capital of "connection" (before it fully depreciates). They have to make their minds to *xia hai* (go business). By doing so, they lose little because the rents they used to enjoy can be embedded into profits which may legally accrue to them in various forms. They have no risk to bear because start-up capital comes from the state (initially the firm is "owned" by the state). Before they leave government office, they will grant full autonomy to the firms with which they will work. They will appoint themselves as chairmen of the board, directors, or executives. Once they pocket some profits, they will buy out the firms. They can do this quietly because once the firms are corporatized, they can be easily sold piecemeal instead of as a whole.[25] In addition, the central government may have to sell its shares because of its budget deficit. The state-owned enterprises gradually evolve into private joint-stock companies. In this stage, it is possible for the government to become a bond-holder who can be protected by private shareholders. Once incumbent bureaucrats become capitalists, they will have incentives to select high ability people for management; they themselves will voluntarily step down if unqualified. Capitalistization of incumbent bureaucrats and managers will also automatically solve the problem of principal-agent relationship between the manager and workers, which has been a real headache of the state-owned enterprises.

Acknowledgement This paper is based on Sect. 6 of Zhang (1993). I am very grateful to Donald Hay, Shuhe Li, Jim Mirrlees and Adrian Wood for their helpful comments. Thanks also go to the seminar participants at Oxford University, Peking University and City University of Hong Kong.

References

Hart, Oliver, and Begnt Holmstrom. (1987). The Theory of Contracts. In *Advances in Economic Theory*, ed. T. Bewley. Cambridge: Cambridge University Press.

Hay, Donald, D., Morris, G.S., Liu, and S. Yao. (1994). *Economic Reform and State-Owned Enterprises in China: 1979–1987*. Oxford: Oxford University Press.

Hayek, F. (1948). *Individualism and Economic Order*. University of Chicago Press.

Holmstrom, Begnt, and Joan Ricart i Costa. (1986). Managerial Incentives and Capital Management. *The Quarterly Journal of Economics*, 835–860.

Holmstrom, B., and P. Milgrom. (1987). Aggregation and linearity in the provision intertemporal incentives. *Econometrica*, 53: 303–328.

[25] To my knowledge, Wu and Jin (1985) were the first to propose the state-owned joint-stock reform, according to which, the enterprises are still owned by the government, but ownership is implemented by many competing government institutions who function as shareholders. I was critical of this proposal (see Zhang 1986). How can you transform a zebra into a horse simply by brushing stripes on its back? However, I now have realized that the joint stock system may be a feasible transitional approach to retail the large-size enterprises, although the state-holding company itself cannot be a workable model. A strong objection of privatization is that "nobody can afford to buy". This objection may apply to wholesale, but not to retail.

Jefferson, Gary, and Wenyi Xu. (1991). The Impact of Reform on Socialist Enterprises in Transition: Structure, Conduct and Performance in Chinese Industry. *Journal of Comparative Economics*, 15 (1): 22–44.
Jensen, Michael, and William Meckling. (1992). Specific and General Knowledge, and Organizational Structure. In *Contract Economics*, eds. L. Werin, and H. Wijkander. Oxford: Blackwell.
Kornai, J. (1980). *Economics of Shortage*. North-Holland.
Wu, Jiaxian, and Lizuo Jin. (1985). Joint-Stock Companies: An Approach to Further Reform. *Economic Development and System Reform*, No. 12.
Yang, Xiaokai. (1988). Theoretical Misleading and Systemic Reform. *China: Development and Reform*, No. 12.
Weiying, Zhang. (1986). Ownership and Entrepreneurship. *Research Reports on Economic Reform and Development*, No. 31 (published by The Economic Reform Institute of China).
Weiying, Zhang. (1993). *Decision Rights, Residual claims and Performance: A Theory of How the Chinese Economy Works*. Oxford: Mimeo, Nuffield College.
Weiying, Zhang. (1994). Entrepreneurial Ability, Personal Wealth, and the Assignment of Principalship: An Entrepreneurial/Contractual Theory of the Firm. D. Phil thesis, Nuffield College, Oxford University.

References

Aghion, Philippe, and Patrick Bolton. 1992. An Incomplete Contracts Approach to Financial Contracting. *Review of Economic Studies* 59: 473–494.

Akerlof, G. 1970. The market for 'lemons': Quality and the market mechanism. *Quarterly Journal of Economics* 84: 488–500.

Alchian, Armen, and Harold Demsetz. 1972. Production, Information Costs, and Economic Organization. *American Economic Review* 62 (50): 777–795.

Aoki, Masahiko. 1984. *The Cooperative Game Theory of the Firm*. Oxford: Clarendon Press.

Aoki, Masahiko, Bo Gustafsson, and Oliver Williamson (eds.). 1990. *The Firms as a Nexus of Treaties*. London: Sage Publications ltd.

Azariadis, C. 1975. Implicit Contracts and Unemployment Equilibria. *J.P.E* 83: 1183–1202.

Baily, M. 1974. Wages and Employment with Uncertain Demand. *Review of Economic Studies* 41: 3754.

Barnea, Amir, Robert A. Haugen, and Lemma W. Senbet. 1981. Market imperfections, agency problems and capital structure. *Financial Management* 10 (3): 722.

Baumol, W.J. 1959. *Business Behavior, Value and Growth*. New York: Macmillan.

Berle, A.A., and G.C. Means. 1932. *The modern Corporation and Private Property*. New York: Harcourt, Brace and World Inc. Revised edition 1967.

Blanchflower, David G., and Andrew J. Oswald. 1990. What Makes An Entrepreneur?, Working Paper, Dartmouth College, NBER and Centre for Economic Performance, LSE.

Bolton, Patrick, and David S. Scharfstein. 1990. A Theory of Predation Based on Agency Problems in Financial Contracting. *American Economic Review* 80: 93–106.

Casson, Mark. 1982. *The Entrepreneur: An Economic Theory*. Oxford: Martin Robertson.

Chang, Chun. 1987. Capital Structure as Optimal Contracts, Working Paper, Carlson School of Management, University of Minnesota.

Cheung, Steven N.S. 1969a. Transaction cost, risk aversion and the choice of contractual arrangements. *Journal of Law and Economics* 12: 23–42.

Cheung, Steven N.S. 1969b. *The Theory of Share Tenancy*. Chicago: The University of Chicago Press.

Cheung, Steven N.S. 1983. The contractual nature of the firm. *Journal of Law and Economics* 26 (1): 1–21.

Cheung, Steven N.S. 1992. On the New Institutional Economics. In *Contract Economics*, ed. L. Werin, and H. Wijkander. Oxford: Basil Blackwell Publsihers.

Clark, John Bates. 1899. *The Distribution of Wealth*. New York: Macmillan Co.

Coase, Ronald H. 1937. The Nature of the Firm. *Economica* 4: 368–405.

W. Zhang, *The Origin of the Capitalist Firm*, DOI 10.1007/978-981-10-0221-2

Coase, Ronald H. 1960. The problem of social costs. *Journal of Law and Economics* 3 (1): 1–44.
Coase, Ronald H. 1988. *The Firm, the Market, and the Law*. Chicago: University of Chicago Press.
Coase, Ronald H. 1992. "Comments" (on Cheung's "On the New Institutional Economics"). In *Contract Economics*, ed. L. Werin, and H. Wijkander. Oxford: Basil Blackwell Publsihers.
Deaton, Angus, and John Muellbauer. 1980. *Economics and Consumer Behaviour*. Cambridge: Cambridge University Press.
Diamond, D.W. 1984. Financial intermediation and delegated monitoring. *Review of Economic Studies* 51: 393–414.
Diamond, D.W. 1989. Reputation acquisition in debt markets. *Journal of Political Economy* 97: 828–862.
Diamond, Peter A., and Joseph E. Stiglitz. 1974. Increases In Risk and Risk Aversion. *Journal of Economic Theory* 8: 337–360.
Domar, E. 1966. The Soviet Collective Farm as a Producer Cooperative. *American Economic Review* 48: 734–757.
Dow, Gregory K. 1993a. Democracy versus Appropriability: Can Labour-Managed Firms Flourish in a Capitalist World? In *Democracy and Markets: Problems of Participation and Efficiency*, ed. Samuel Bowles, Herbert Gintis, and Bo Gustafsson. New York: Cambridge University Press.
Dow, Gregory K. 1993b. Why Capital Hires Labour: A Bargaining Perspective. *American Economic Review* 83 (1): 118–134.
Eswaran, Mukesh, and Ashok Kotwal. 1989. Why are capitalists the Bosses? *The Economic Journal* 99: 162–176.
Evans, David S., and Boyan Jovanovic. 1989. An Estimated Model of Entrepreneurial Choice Under Liquidity Constraints. *Journal of Political Economy* 97 (4): 808–827.
Fama, Eugene. 1980. Agency Problem and the Theory of the Firm. *Journal of Political economy* 88: 288–307.
Fama, Eugene, and Michael Jensen. 1983. Separation of Ownership and Control. *Journal of Law and Economics* 26: 301–325.
Fama, Eugene, and Michael Jensen. 1983. Agency Problems and Residual Claims. *Journal of Law and Economics* 26: 327–349.
Farmer, Roger. 1985. Implicit Contracts with Asymmetric Information and Bankruptcy: The Effort of Interest Rates on Layoffs. *Review of Economic Studies* 52: 427–442.
FitzRoy, Felix R., and Dennis Mueller. 1984. Cooperation and Conflict in Contractual Organization. *Quarterly Review of Economics and Business* 24 (4): 24–49.
Fudenberg, Drew, Begnt Holmstrom, and Paul Milgrom. 1990. Shortterm contracts and Longterm Agency Relationship. *Journal of Economic Theory* 51: 1–31.
Gale, D., and M. Hellwig. 1985. Incentive-compatible debt contract: the oneperiod problem. *Review of Economic Studies* 52 (4): 647–663.
Grossman, Sanford J., and Oliver Hart. 1982. Corporate financial structure and managerial incentives. In *The Economic of Information and Uncertainty*, ed. J. McCall. Chicago: University of Chicago Press.
Grossman, Sanford J., and Oliver Hart. 1983. An Analysis of the Principal-Agent Problem. *Econometrica* 51: 7–45.
Grossman, Sanford, and Oliver Hart. 1986. The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration. *Journal of Political Economy* 94 (4): 691–719.
Grossman, Sanford, and Oliver Hart. 1988. One share-one vote and the market for corporate control. *Journal of Financial Economics* 20: 175–202.
Haltiwanger, John, and Michael Waldman. 1986. Insurance and Labour Market Contracting: An Analysis of the Capital Market Assumption. *Journal of Labour Economics* 4 (3): 335–375.
Hansmann, Henry. 1988. Ownership of the Firm. *Journal of Law, Economics and Organization* 4 (2): 267–304.
Harris, Milton, and Begnt Holmstrom. 1982. A Theory of Wage Dynamics. *Review of Economic Studies* 49: 313–333.

Harris, Milton, and Artur Raviv. 1988a. corporate control contests and capital structure. *Journal of Financial Economics* 20: 55–86.
Harris, Milton, and Artur Raviv. 1988b. Corporate Governance: voting rights and majority rules. *Journal of Financial Economics* 20: 203–235.
Harris, Milton, and Artur Raviv. 1989. The Design Of Securities. *Journal of Financial Economics* 24: 255–287.
Harris, Milton, and Artur Raviv. 1991. The theory of capital structure. *The Journal of Finance* 46 (1): 297–355.
Harris, Milton, and Artur Raviv. 1992. Financial Contracting Theory. In *Advances in Economic Theory: Volume 2: Sixth World Congress*, ed. Jean-Jacque Laffont. Cambridge University Press.
Hart, Oliver, and Begnt Holmstrom. 1987. The Theory of Contracts. In *Advances in Economic Theory*, ed. Bewley, t.
Hart, Oliver, and John Moore. 1989. Default and Renegotiation: A Dynamic Model of Debt, Working Paper, MIT, August.
Hart, Oliver, and John Moore. 1990. Property Rights and the Nature of the Firm. *Journal of Political Economy* 98 (6): 1119–1158.
Hart, Oliver, and John Moore. 1990a. A Theory of Corporate Financial Structure Based on the seniority of Claims, Working Paper No.560, MIT.
Hay, Donald. 1990. *The Public JointStock Company: Blessing or Curse?*. Oxford: Jesus College. Unpublished.
Hay, Donald, and Derek Morris. 1991. *Industtrial Economics and Organization: Theory and Evidence*. Oxford: Oxford University Press.
Hayek, Friedrich A. 1948. The Meaning of Competition. In *Individualism and Economic Order*. 1976. London and Henley: Routledge and Kegan Paul.
Heinkel, Robert. 1982. A theory of capital structure relevance under imperfect information. *The Journal of Finance* 37: 1141–1150.
Hirshleifer, Jack, and John G. Riley. 1992. *The Analytics of Uncertainty and Information*. New York: Cambridge University Press.
Holmstrom, Begnt. 1979. Moral Hazard and Observability. *Bell Journal of Economics* 10 (1): 74–91.
Holmstrom, Begnt. 1982. Moral Hazard in Teams. *Bell Journal of Economics* 13: 324–340.
Holmstrom, Begnt, and Joan Ricart. 1986. Managerial Incentives and Capital Management. *The Quarterly Journal of Economics* 101: 835–860.
Holmstrom, Begnt, and Paul Milgrom. 1987. Aggregation and linearity in the provison of intertemporal incentuves. *Econometrica* 55: 303–328.
Holmstrom, Bengt, and J. Tirole. 1989. The theory of the firm. In *Handbook of industrial Organization*, eds. Schmalensee, R. and R. Willig. North Holland.
Holmstrom, Begnt, and Paul Milgrom. 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership and Job Design. *Journal of Law, Economics and Organization* 7: 24–52.
Holmstrom, Bengt, and Paul Milgrom. 1993. The Firm as an Incentive System, mimeo.
Holtz-Eakin, D., D. Joufaian, and H. Rosen. 1994. Sticking It Out: Entrepreneurial Survival and Liquidity Constraints. *Journal of Political Economy* 102 (1): 53–75.
Huang, Y. 1973. Risk, Entrepreneurship, and Tenancy. *J.P.E* 81: 1241–1244.
Israel, Ronen. 1991. Capital structure and the market for corporate control. *The Journal of Finance* 46 (4): 1391–1409.
Itoh, Hideshi. 1991. Incentives to Help in Multi-Agent Situations. *Econometrca* 59: 611–637.
Jensen, Michael C., and William Meckling. 1976. Theory of the firm: managerial behaviour, agency costs, and capital structure. *Journal of Financial Economics* 3: 305–360.
Jensen, Michael C., and William Meckling. 1979. Rights and Production Functions: An Application to Labour-Managed Firms and Codetermination. *Journal of Business* 52: 469–506.
Jensen, Michael, and William Meckling. 1992. Specific and General Knowledge, and Organizational Structure. In *Contract Economics*, ed. L. Werin, and H. Wijkander. Oxford: Basil Blackwell Publsihers.

Jensen, Michael C., and Jerold Warner. 1988. The Distribution of power among corporate managers, shareholders, and directors. *Journal of Financial Economics* 20: 3–24.
Johnson, D. 1950. Resource Allocation under Share Contracts. *J.P.E* 58: 111–123.
Jovanovic, Boyan. 1979a. Job Matching and the Theory of Turnover. *J.P.E* 87 (5): 972–987.
Jovanovic, Boyan. 1979b. Firm specific Capital and Turnover. *J.P.E* 87 (6): 1246–1260.
Kanbur, S.M. 1979. Of Risk Taking and the Personal Distribtuion of Income. *Journal of Political Economy* 87 (4): 769–795.
Kihlstrom, Richard E., and Jean-Jacques Laffont. 1979. A General Equilibrium Entrepreneurial Theory of Firm Formation Based on Risk Aversion. *Journal of Political Economy* 87 (4): 719–748.
Kihlstrom, Richard E., and Jean-Jacques Laffont. 1982. A Competitive Entrepreneurial Model of a Stock Market. In *The Economic of Information and Uncertainty*, ed. J. McCall. Chicago: University of Chicago Press.
Kirzner, Israel. 1973. *Competition and Entrepreneurship*. Chicago, IL: Chicago University Press.
Kirzner, I.M. 1979. *Perception, Opportunity and Profit*. Chicago: Chicago University Press.
Klein, B., R. Crawford, and A. Alchian. 1978. Vertical Integration, Appropriable Rents and the Competitive contracting Process. *Journal of Law and Economics* 21: 297–326.
Knight, Frank. 1964 (1921). *Risk, Uncertainty, and Profit*. New York: A.M. Kelley.
Kreps, David M. 1990. *A Course in Microeconomic Theory*. New York: Harvester Wheatsheaf.
Layard, P.R.G., and A.A. Walters. 1987. *Microeconomic Theory*. Maidenheach: McGraw Hill Book Company. Chapter 13.
Lazear, Edward. 1979. Why Is There Mandatory Retirement. *J.P.E* 87 (6): 1261–1284.
Leland, Hayne, and David Pyle. 1977. Information asymmetry, financial structure, and financial intermediation. *The Journal of Finance* 32: 371–388.
LeRoy, Stephen, and Larry D. Singell. 1987. Knight on Risk and Uncertainty. *The Journal of Political Economy* 95: 394–406.
Lewis, Tracy R., and David E.M. Sappington. 1991. Technological Changes and the Bourndaries of the Firm. *American Economic Review* 81 (4): 887–900.
Lucas, Robert Jr. 1978. On the Size Distribution of Business Firms. *The Bell Journal of Economics* p. 508.
Marris, R. 1964. *The Economic Theory of Managerial Capitalism*. London: Macmillan.
McAfee, R.Preston, and John McMillian. 1991. Optimal Contracts for Teams. *International Economic Review* 32 (3): 561–577.
Malcomson, James M. 1984. Work Incentives, Hierarchy, and Internal Labour Markets. *Journal of Political Economy* 92 (3): 486–507.
Meade, J.E. 1972. The Theory of LabourManaged Firms and of Profit Sharing. *The Economic Journal* 82: 402–428.
Meyer, Margaret. 1992. *The Internal Organization of Firms, presented as the Review of Economic Studies Lecture at the Royal Economic Society Conference, London, March 1992; mimeo*. Oxford: Nuffield College.
Meyer, Margaret, Paul Milgrom, and John Roberts. 1992. Organizational Prospects, Influence Costs, and Ownership Changes. *Journal of Economics and Management Strategy*, vol. 1.
Milgrom, Paul, and John Roberts. 1992. *Economics, Organization and Management*. New Jersey: Prentic-Hall International Inc.
Mirrlees, J.A. 1974. Notes on Welfare Economics, Information and uncertainty. In *Essays on Economic Behaviour under Uncertainty*, ed. M. Balch, D. McFadden, and S. Wu. North Holland: Amsterdam.
Mirrlees, J.A. 1975. *The Theory of Moral Hazard and Unobservable Behaviour, Part I*. Oxford: Nuffield College. Unpublished mimeo.
Mirrlees, J.A. 1976. The Optimal Structure of Incentives and Authority within An organization. *Bell Journal of Economics* 7: 105.
Modigliani, Franco, and H. Miller. 1958. The cost of capital, corporation finance, and the theory of investment. *American Economic Review* 48: 261–297.

Mueller, Dennis. 1976. Information, Mobility and Profit. *Kyklos* 29: 419–448.
Myers, Stewart C., and Nicholas S. Majluf. 1984. Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics* 13: 187–221.
Niskanen, W.A. 1968. Nonmarket decision making: The peculiar economics of bureaucracy. *American Economic Review* 58 (2): 293–305.
Putterman, Louis. 1984. On some recent explanations of why capital hires labour. *Economic Inquiry* 22: 171–187.
Putterman, Louis, and Gil Jr Skillman. 1988. The Incentive Effects of Monitoring Under Alternative Compensation Schemes. *International Journal of Industrial Organization* 6 (1): 109–119.
Rao, C.H. 1971. Uncertainty, Entrepreneurship and Sharecropping in India. *J.P.E* 79: 578–595.
Rees, Ray. 1985. The Theory of Principal and Agent: Part I. *Bulletin of Economic Research*.
Riordan, Michael H. 1990. What Is Vertical Integration. In *The Firm as a Nexus of Treaties*, ed. M. Aoki, Bo Gustafsson, and O. Williamson. London: Sage Publications Ltd.
Rosen, Sherwin. 1985. Implicit Contracts: A Survey. *Journal of Economic Literature* 1144–1175.
Rosen, Sherwin. 1992. Contracts and the Market for Executives. In *Contract Economics*, ed. L. Werin, and L. Wijkander. Oxford: Basil Blackwell.
Ross, Stephen. 1973. The Economic Theory of Agency: The Principal's Problem. *American Economic Review* 63: 134–139.
Ross, Stephen. 1977. The determination of financial structure: The incentive signalling approach. *Bell Journal of Economics* 8: 23–40.
Samuelson, Paul. 1957. Wages and interest: A modern dissection of marxian economic models. *American Economic Review* 47 (6): 884–912.
Schumpeter, J.A. 1934. *The Theory of Economic Development*. Cambridge: Harvard Unversity press.
Schumpeter, J.A. 1942. *Capitalism, Socialism and democracy*. New York: Harper & Brothers.
Schumpeter, J.A. 1954. *History of Economic Analysis*. London: Allen and Unwin.
Shackle, G.L.S. 1979. Imagination, Formalism and Choice. In Rizzo, Mario, J. (ed.), op. cit. p. 19
Spence, M. 1973. Job market signalling. *The Quarterly Journal of Economics* 87: 355–374.
Spence, M., and R. Zeckhauser. 1971. Insurance, Information, and Individual Action. *American Economic Review* 61 (2): 552–579.
Stiglitz, Joseph E. 1974. Incentive and Risk Sharing in Sharecropping. *Rev. Economic Studies* 41: 219–255.
Stiglitz, Joseph E., and Andrew Weiss. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71 (3): 393–410.
Stiglitz, Joseph E. A reexamination of the Modigliani-Miller Theorem. *American Economic Review* LIX (5): 79–493.
Stiglitz, Joseph. 1989. Principal and Agent. In Palgrave.
Stulz, Rene. 1988. Managerial control of voting rights: Financing policies and the market for corporate control. *Journal of Financial Economics* 20: 25–54.
Tirole, Jean. 1986. Procurement and Renegotiation. *Journal of Political Economy* 94: 235–259.
Tirole, Jean. 1988. *The Theory of Industrial Organization*. Cambridge: MIT Press.
Tirole, Jean, and Jean-Jacques Laffont. 1986. Using Cost Observation to Regulate Firms. *Journal of Political economy* 94: 614–641.
Townsend, Robert M. 1979. Optimal Contracts and Competitive markets With Costly State Verification. *Journal of Economic Theory* 21: 265–293.
Vanek, Jaroslav. 1970. *The General Theory of Labour Managed Market Economies*. Ithaca: Cornell University Press.
Vickers, John, and G.K. Yarrow. 1988. *Privatization: An Economic Analysis*. Cambridge: MIT Press.
Waldman, Michael. 1990. Up or Out Contracts: A Signalling Perspective. *Journal of Labour Economics* 8 (2): 230–250.
Ward, B. 1967. *The Socialist Economy*. New York: Random House.

Williams, Joseph. 1989. Monitoring and Optimal Financial Contracts, Working Paper, University of British Columbia.
Williamson, O.E. 1964. *The Economics of Discretionary Behaviour: Managerial Objectives in a theory of the Firm*. Englewood Cliffs: Prentice Hall.
Williamson, O.E. 1975. *Markets and Hierarchies: Analysis and AntiTrust Implications*. New York: The Free Press.
Williamson, O.E. 1979. Transaction Cost Economics: The Governance of Contractual Relations. *Journal of Law and Economics* 22: 233–261.
Williamson, O.E. 1980. Organization of Work: A Comparative Institutional Assessment. *Journal of Economic Behaviour and Organization* 1: 5–38.
Williamson, O.E. 1981. The Modern Corporation: Origins, Evolution, Attributes. *Journal of Economic Literature* 19: 1537–1568.
Williamson, O.E. 1983. Organization form, Residual Claimants, and Corporate Control. *Journal of Law and Economics* 26: 351–66.
Williamson, O.E. 1985. *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. New York: The Free Press.
Williamson, O., M. Wachter, and S. Harris. 1975. Understanding the Employment Relation. *Bell Journal of Economics* 6: 250–278.
Wilson, R. 1969. *The Structure of Incentives for Decentralization under Uncertainty*, ed. M. Guilbaud. La Decision, Paris: CNRS.
Yang, Xiaokai, and J. Borland. 1991. A Microeconomic Mechanism for Economic Growth. *Journal of Political Economy* 99: 460–482.
Yang, Xiaokai, and Yew-Kwang Ng. 1993. *Specialization and Economic Organization: A New Classic Microeconomics*. North-Horland: Elsevier Science Publishers B.N.
Yang, Xiaokai, and Yew-Kwang Ng. 1995. Theory of the Firm and Structure of Residual Rights. *Journal of Economic Behaviour and Organization* 26: 107–128.
Zhang, Weiying. 1992. Entrepreneurial Ability, Personal Wealth and the Assignment of Principalship: An Entrepreneurial/Contractual Theory of the Firm, M. Phil thesis, Oxford University.

Index

W. Zhang, *The Origin of the Capitalist Firm*, DOI 10.1007/978-981-10-0221-2