

Chapter 2

Marketing, Producing, Monitoring and the Assignment of Principals

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

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

¹Huang (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”.

²Robinson 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.

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

⁴Yang 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.

⁵It 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 costs⁷ 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,

⁶In 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.

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

⁸The 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.

⁹An 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-neutral¹¹; 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.

¹⁰The 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.

¹¹In 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 return¹⁴:

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

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 β ($0 \leq \beta \leq 1$): 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 .

¹²Here 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).

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

¹⁴The 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:

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

We shall assume this is common knowledge; that is, when i chooses $\tilde{b}_i$, both i and j know that j has to choose $a_j^b(\tilde{b}_i)$ and j knows $\tilde{b}_i$ 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:

$$U_i = U_i(Y_i, e_i) = V_i(Y_i) - C_i(a_i, b_i) \quad (2.3)$$

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-

¹⁵The 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 a “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.).

¹⁶Although 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 $(C_j(a_j^b(b_i), \cdot) - C_j(a_j^s, \cdot))$, 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:

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

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

We call $F_j(\cdot)$ "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 $F_j(\cdot)$, 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:

$$\int V_j(Y_j^b) \phi(Y; a_i^s, a_j^b) dY - C_j(a_j^b(b_i), \cdot) \geq \int V_j(Y_j^s) \phi(Y; a_i^s, a_j^s) dY - C_j(a_j^s, \cdot) \quad (2.5)$$

where

$$\begin{aligned} Y_j^b &= w_j^s + F_j(\cdot) + \lambda_j(Y^b - w_i - w_j^s - F_j(\cdot)) \\ Y_j^s &= w_j^s + \lambda_j(Y^s - w_i - w_j^s) \end{aligned}$$

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

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

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

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:

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

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

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

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

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

¹⁷Is 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.

¹⁸When $w_j = w_j^s$, 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

¹⁹We 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.

²⁰It 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):

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

where

$$\begin{aligned}
 EU_P &= \int V_P(w_P + (1 - \beta)(Y - w_M - w_P)) \phi(Y; a_M, a_P) dY - C_P(a_P, b_P) \\
 EU_M &= \int V_M(w_M + \beta(Y - w_M - w_P)) \phi(Y; a_M, a_P) dY - C_M(a_M, b_M)
 \end{aligned}$$

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

²¹I 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 form²²:

$$U_i = Y_i - 0.5a_i^2 - 0.5b_i^2, \quad i = M, P \quad (2.6)$$

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),

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

²³One 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 $Y = f(a_M, a_P)$:

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

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 ($0 \leq \alpha \leq 1$). 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 $a_P \geq 0$ and $a_M \geq 0$; 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 $a_i \geq 0$. For this reason, we define γ as “the degree of teamwork” and assume $0 < \gamma \leq 1$. $\gamma = 0$ implies no teamwork, which is trivial

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

²⁵To 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:

$$\begin{aligned} (i) \quad a_P^b &= \rho b_M \\ (ii) \quad a_M^b &= \mu b_P \end{aligned} \quad (2.8)$$

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; $\rho \rightarrow \infty$ and $\mu \rightarrow \infty$ implies that monitoring is perfect.

With these specifications, the non-cooperative game is reduced to²⁷:

Producing member

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

Marketing member:

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

where

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

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

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

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

The cooperative game is reduced to:

$$\begin{aligned} \text{Max}_{\{\beta\}} \quad & (\alpha (a_M(\beta))^{1-\gamma} + (1-\alpha) (a_P(\beta))^{1-\gamma})^{\frac{1}{1-\gamma}} \\ & - 0.5 ((a_M(\beta))^2 + (b_M(\beta))^2) - 0.5 ((a_P(\beta))^2 + (b_P(\beta))^2) \end{aligned} \quad (2.11)$$

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 $\gamma \rightarrow 1$, $a_P^*(0) = a_P^*(1) = a_M^*(0) = a_M^*(1) = 0$.

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

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

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

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:

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

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

Differentiating Eq. (2.14) gives

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

where

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

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

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

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 $O_P|_{\beta \rightarrow 0} = +\infty$ and $O_P|_{\beta \rightarrow 1} = -\infty$. 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 $\gamma \rightarrow 1$, the returns function reduces to the following Cobb–Douglas form:

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

The first-order conditions are

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

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

The Nash equilibrium solutions are

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

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

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

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

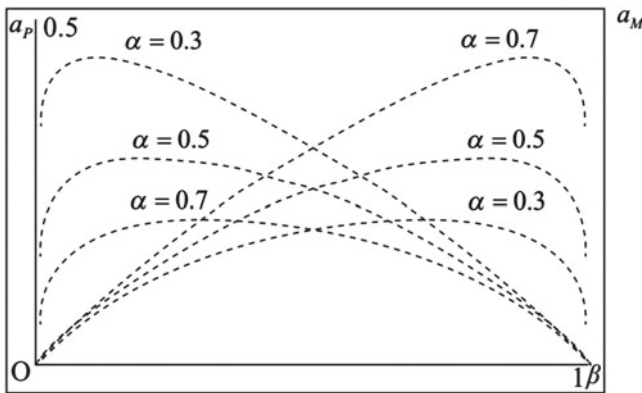


Fig. 2.1 Incentive Functions $a_P(\cdot)$ and $a_M(\cdot)$ where $\alpha = 0.3, 0.5, 0.7$, and $\gamma = 0.8$

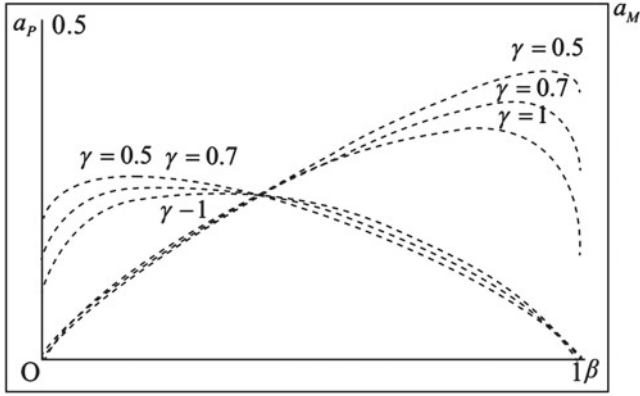


Fig. 2.2 Incentive Function $a_P(\cdot)$ and $a_M(\cdot)$ where $\alpha = 0.6$ and $\gamma = 0.5, 0.7, 1$

of teamwork (γ). Let $\underline{\beta}$ be the residual share at which the producing member's effort is maximized, and $\bar{\beta}$ 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) $\underline{\beta} < \bar{\beta}$ for all $\alpha > 0$ and $\gamma > 0$; (ii) $\underline{\beta}$ is an increasing function of γ and $\bar{\beta}$ is a decreasing function of γ ; (iii) both $\underline{\beta}$ and $\bar{\beta}$ 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:

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

and $\bar{\beta}$ satisfies:

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

Rearranging and differentiating (2.22) gives

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

and

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

Therefore

$$\frac{\partial \underline{\beta}}{\partial \gamma} = - \frac{\frac{\partial P}{\partial \gamma}}{\frac{\partial P}{\partial \beta}} > 0 \quad (2.26)$$

Similarly, since

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

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

Therefore

$$\frac{\partial \bar{\beta}}{\partial \gamma} = - \frac{\frac{\partial M}{\partial \gamma}}{\frac{\partial M}{\partial \beta}} < 0 \quad (2.29)$$

(iii) Since

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

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

So

$$\frac{\partial \underline{\beta}}{\partial \alpha} = - \frac{\frac{\partial P}{\partial \alpha}}{\frac{\partial P}{\partial \beta}} > 0 \quad (2.32)$$

$$\frac{\partial \bar{\beta}}{\partial \alpha} = - \frac{\frac{\partial M}{\partial \alpha}}{\frac{\partial M}{\partial \beta}} > 0 \quad (2.33)$$

(i) By (ii), $\underline{\beta} \leq \underline{\beta}(\gamma = 1)$ and $\bar{\beta} \geq \bar{\beta}(\gamma = 1)$. But, from (2.20) and (2.21), we obtain

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

$$\bar{\beta} = \frac{1+\alpha}{2} \text{ if } \gamma = 1$$

□

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 $\underline{\beta} > \beta > \bar{\beta}$, 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 $[\underline{\beta}, \bar{\beta}]$. 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 $\bar{\beta}$ 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 ($\gamma = 1$), $(1 - \underline{\beta}) = (\frac{1}{2} + \frac{1-\alpha}{2})$ and $\bar{\beta} = (\frac{1}{2} + \frac{\alpha}{2})$, where the common term $(\frac{1}{2})$ can be explained as the team effect while the terms $(\frac{1-\alpha}{2})$ and $(\frac{\alpha}{2})$ 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) $\underline{\beta} < \beta_Y < \bar{\beta}$ and $\underline{\beta} < \beta_\Pi < \bar{\beta}$; (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 $(\bar{\beta}, 1]$. 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 $\beta = \bar{\beta}$, 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 $(\underline{\beta}, \bar{\beta})$, 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 $\beta = \bar{\beta}$ will not only increase Y but also increase Π .

²⁸Note 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,

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

So

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

and

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

Then β_Y satisfies the following first order condition:

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

where

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

By arranging (2.37), we obtain

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

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:

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

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 $\beta_\Pi < \beta_Y$. $\square$

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 $\underline{\beta} < \beta_Y < \bar{\beta}$), 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 is²⁹:

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

where we omit irrelevant terms $(1-\beta)w_M^s$ 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:

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

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

²⁹The 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

$$\begin{aligned} \text{Max}_{\{a_M^s, b_M\}} EU_M &= Y^b - (1 - \beta)Y^s - C_M(a_M^s, b_M) - (C_P(a_P^b(b_M)) - C_P(a_P^s)) \end{aligned} \quad (2.44)$$

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

Two first-order conditions are:

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

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

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

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

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):

$$w_P^s + (1 - \beta) (Y(a_M^\dagger, a_P^\dagger) - w_M^s - w_P^s) - C_P(a_P^\dagger) \quad (2.48)$$

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

³⁰The 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 McMillan (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:

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

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

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

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

³¹In 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.

³²It 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:

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

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

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

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

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 $\beta \in [\beta_\rho, \beta_\mu]$. $\square$

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 $\mu \rightarrow 0$ and $\rho \rightarrow 0$, $\beta_P \rightarrow 0$ and $\bar{\beta} \rightarrow 1$ 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

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

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

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

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

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

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

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

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

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 obtain³³

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

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

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 $\mu\rho \leq 1$, the self-incentives are always greater than the monitored-incentives (if monitoring is imposed) for $\beta \in [\frac{\mu^2}{1+\mu^2}, \frac{1}{1+\rho^2}]$:

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

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 $\mu \rightarrow \infty$ and (or) $\rho \rightarrow \infty$, 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:

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

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

where superscript ‘FB’ denotes the first best.

It is easy to check that: (i) at $\beta = 0$, $a_P^{s*} \rightarrow a_P^{FB}$ and $a_M^* = \mu b_P^* \rightarrow a_M^{FB}$ as $\mu \rightarrow \infty$; (ii) at $\beta = 1$, $a_M^{s*} \rightarrow a_M^{FB}$ and $a_P^* = \rho b_M^* \rightarrow a_P^{FB}$ as $\rho \rightarrow \infty$; and (iii) for $\beta \in [\beta_\rho, \beta_\mu]$, $a_P^* = \rho b_M^* \rightarrow a_P^{FB}$ and $a_M^* = \mu b_P^* \rightarrow a_M^{FB}$ as $\rho \rightarrow \infty$ and $\mu \rightarrow \infty$. The intuition is that as $\mu \rightarrow \infty$ and $\rho \rightarrow \infty$, 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:*

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

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 $\beta = 0$,

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

At $\beta = 1$,

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

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

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

³⁴The 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

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

This is what we need. $\square$

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

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

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

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

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 $\mu^{**} \rightarrow 0$ and $\rho^{**} \rightarrow \infty$ as $\alpha \rightarrow 0$, and $\mu^{**} \rightarrow \infty$ and $\rho^{**} \rightarrow 0$ as $\alpha \rightarrow 1$. 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 $\alpha \rightarrow 0$ (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^{**}). $\square$

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

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

³⁶It 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-principals 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 principals 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-principals 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-principals 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

³⁷When 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 $\bar{Y}$ 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 $= \frac{1}{2} R_P (1 - \beta)^2 V$; the marketing member's risk premium $= \frac{1}{2} R_M \beta^2 V$; and the total risk premium $= \frac{1}{2} (R_M \beta^2 + R_P (1 - \beta)^2) V$. The certain equivalents are respectively

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

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

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

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:

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

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

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 $R_i \leq R_j$.

³⁸The 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

$$CE = \Pi \big|_{R_P=R_M=0} - \frac{1}{2} (R_M \beta^2 + R_P (1 - \beta)^2) V^0 \quad (2.77)$$

We have already shown how $\Pi \big|_{R_P=R_M=0}$ 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 ($\bar{Y}$) 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

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

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

$$\frac{dV}{d\beta} = \frac{\partial V}{\partial a_M} \frac{\partial a_M}{\partial \beta} + \frac{\partial V}{\partial a_P} \frac{\partial a_P}{\partial \beta} \quad (2.79)$$

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 $\beta_{TRC} < \frac{R_P}{R_P + R_M}$.

³⁹This 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 observe his actions, while in our model, the "agent" shares the residual because the "principal" is too risk-averse.

⁴⁰Another 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

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

The optimal work effort and monitoring effort are respectively

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

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

The mean return of the firm is

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

The total certain equivalent is

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

⁴¹Like 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

$$\max_{a_M, b_M} a_M^\alpha (\rho b_M)^{1-\alpha} - 0.5a_M^2 - 0.5(1 + \rho^2)b_M^2 \quad (2.85)$$

The optimal work effort and monitoring effort are respectively

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

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

The mean return of the firm is

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

The total certain equivalent is

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

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

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

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

⁴²We 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 .

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

The optimal work effort and monitoring effort are respectively

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

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

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

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

The total certain equivalent is

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

$\beta = 1$ 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 $R_M \leq 0.233$ (≤ 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,

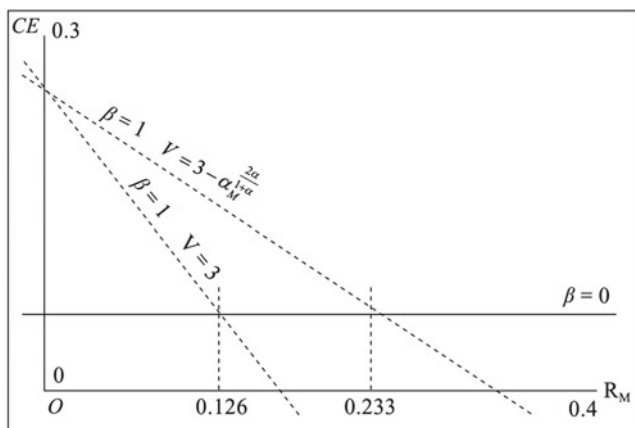


Fig. 2.3 The optimal assignment of principalship when the marketing member has advantages in production and monitoring while the producing member has advantage in risk-tolerance

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.

⁴³It 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.

The Origin of the Capitalist Firm

An Entrepreneurial/Contractual Theory of the Firm

Zhang, W.

2018, XVI, 196 p. 20 illus., Hardcover

ISBN: 978-981-10-0220-5