Collusive Supervision and Organization Design in a Three-tier Agency Model with a Continuum of Types♦

Yutaka Suzuki

Faculty of Economics, Hosei University

Revised, September 10, 2012

ABSTRACT

We apply the Monotone Comparative Statics method and the First Order (Mirrlees) Approach to the continuous-type, three-tier agency model with hidden information and collusion a la Tirole (1986,1992), thereby providing a framework that can address the issues treated in the existing literature (e.g. Kofman and Lawarree 1993) *in a much simpler fashion*. We characterize the nature of equilibrium contract that can be implemented under the possibility of collusion between the supervisor and the agent, and then obtain a general comparison result on the two-tier vs. three-tier organization structures. Next, we introduce the recent behavioral contract theory idea, "shading" (Hart and Moore (2008)) into the model. By combining the two ideas, i.e., *collusion* and *shading*, we can not only enrich the existing collusion model, including a new result on the choice of Collusion-proof vs. Equilibrium Collusion regimes, but also give a micro foundation to ex-post haggling costs, addressed by Transaction Cost Economics (e.g. Coase (1937) and Williamson (1975)). This will contribute to a deep understanding of resource allocation and decision process in inside hierarchical organization.

Key Words: Mechanism Design, Collusion, Supervision, Monotone Comparative Statics, Behavioral Economics, Shading, Corporate Governance

JEL Classification: D82, D86

♦I would like to thank Oliver Hart, Robert Gibbons, Hideshi Itoh, Kathryn Spier, Kenichi Amaya, Charles Angelucci, Anton Kolonin, Hongy Li, audiences at Hosei University, Association for Public Economic Theory (PET) Seoul, First Annual UECE Lisbon Meetings: Game Theory and Applications (2009), Japanese Economic Association at Chiba (2010), Game Theory Conference at Stony Brook (2010), Contract Theory Workshop East (CTWE), and Harvard/MIT Contracts and Organization Lunch for their useful comments. This research was supported by Grant-in-Aid for Scientific Research by Japan Society for the Promotion of Science (C) 20530162 and 23530383.

1 Introduction

Recently, auditing has rapidly been increasing in importance in Japan, as well as in the U.S. and Western countries, to meet the needs of corporate governance. Corporate scandals such as those that rocked Yamaichi Securities, Daiwa Bank, Snow Brand Milk Products, and Kanebo in Japan and Enron and WorldCom in the U.S. are examples of firms that failed to build up the effective corporate governance, and collusive supervision (auditing) and revelation of false information was a common occurrence. Auditors (supervisors) usually have greater access to accurate information on the agents, but are subject to collusive pressure (the collusive offer) from the auditees (agents). The means by which adequate supervision (auditing) is used to enhance the efficiency of corporate governance and by which collusive supervision (auditing) can be deterred are important parts of corporate governance reform.

In a typical framework of the top management organization of Japanese firms, a shareholders' meeting elects a director (or a Board of Directors) and an auditor who audits the execution of the management work and makes a report at the shareholders' meeting. With this auditing system, which has been legally amended several times, it is often said that the auditors have access to a great deal of information inside the firm, including the ability of top managers to perform their jobs, while on the other hand it is doubtful that the auditor can objectively supervise the management while maintaining his independence. Indeed, there is a notion that collusive auditing often exists where an auditor and a manager collude to manipulate information. Thus, corporations should optimally utilize the auditing information in order to increase the shareholders' interests, with an arrangement that the auditor and the manager do not collude. Many Japanese firms, such as Toyota and Canon, do preserve and try to improve this traditional Japanese auditing system. However, some companies with auditors, falling into low performance under collusive auditing, tend to move to those with committees, where the monitoring of the manager is tightened and the independence of supervision is ensured by employing outside directors as a majority of the committee members. Our paper can be viewed as an analysis of a top management organization in a hidden information setting.

Literature exists that deals with the issues associated with corporate governance and auditing in a three-tier agency model with collusion, developed by Tirole (1986, 1992), Laffont and Tirole (1991), and Laffont and Martimort (1997) etc. In particular, Kofman and Lawarree (1993) applied a three-tier agency model—consisting of the two-type (productivity) agent, the internal and external auditors (supervisors), and the principal—to the issue of auditing and collusion.[1] However, this is a rather complicated model whose structure involves a Kuhn-Tucker problem with many IC (Incentive Compatibility) and IR (Individual Rationality) constraints, and is not a simple mathematical model. This mathematical complexity of this model is a disadvantage.

We introduce here the outcomes of "Monotone Comparative Statics" à la Topkis (1978), Milgrom and Roberts (1990), Edlin and Shannon (1998), and Milgrom and Segal (2002) into the analysis of corporate governance in a three-tier agency model with a continuum of types. Our paper provides a framework that can address the issues treated in the existing literature *in a much simpler fashion*, and is indeed beneficial in that we can obtain some clear and robust implications for corporate governance reform.

The basic tradeoff in our model is the benefit from the reduction in information rent by adding the auditor (supervisor) versus the resource cost of adding him into the hierarchy, and this bottom line is basically preserved through the extension and generalization of the model. The optimal collusion-proof contract in the Principal-Supervisor-Agent three-tier regime has the property whereby (1) *Efficiency at the top* (the highest type) and (2) *Downward distortion* for all other types, and the downward distortion is mitigated at the optimum, in comparison with the Principal-Agent two-tier regime.The optimal solution allows simple comparative statics, which shows that downward

[1] Bolton and Dewatripont (2005)'s recent textbook presents a simple version of the collusion models (Tirole (1986), Kofman and Lawarree (1993)).

distortions from the first best output levels diminish when the accuracy of supervision increases and the efficiency of collusion declines. This is a specific contribution to the literature.Whether the principal indeed has an incentive to introduce a supervisor—that is, selects a three-tier hierarchy—depends on the balance between the net benefits from both the improvement of marginal incentives and the reduction in information rent and the resource cost of the auditor (supervisor).We obtain these results by constructing a three-tier model with a mathematically more tractable structure, which exploits the outcome of "Monotone Comparative Statics" à la Topkis (1978) and Edlin and Shannon (1998), and Milgrom and Segal (2002)'s generalized envelope theorem.

Though we first consider a situation where the principal can *commit* to a collusion-proof contract, that is, 'full commitment', we analyze as an extension what happens when the principal cannot fully commit to the mechanism and renegotiation is unavoidable.When the principal commits herself to the supervisor reward scheme, but does not commit to the one for the agent, she will be tempted to modify the initial contract (or the outcome) unilaterally, using the information revealed by the supervisor. This situation is similar to the ratchet problem and the renegotiation problem caused by the lack of the principal's commitment in the dynamics of the incentive contracts, studied early by Laffont-Tirole (1988) and Dewatripont (1988) etc. If the agent anticipates such a modification, since he can benefit from a failure by the supervisor to report his type truthfully, he will offer the supervisor the transfer (side payment) equivalent to his information rent. Thus, the principal must pay the supervisor in opposition to the collusive offer by the agent.Hence, the principal can strictly improve his payoff ex post, but must bear the ex ante incentive cost.

We compare the payoffs between three regimes, that is, the 'No-Commitment' regime (NC), the 'Principal-Supervisor-Agent' Collusion-proof, Commitment regime (S), and the 'No-Supervisor' (standard second best) regime (NS). We then find that under the assumption that the cost of introducing the supervisor (a transaction cost) is zero, the principal prefers the 'No Commitment' regime (NC) most in terms of her expected payoff. Intuitively, since the principal does <u>not</u> commit herself <u>not</u> to adjust the output (quantity) rule as well as the price rule in the "No Commitment" regime (NC), she optimally adjusts both of them and tries to design a "more state-contingent" contract through more efficient use of supervisor's report, which is more efficient than the pooling output (quantity) rule which the principal adopts in the "Collusion-proof, Commitment" regime (S). This may be consistent with a situation in the top management organization in corporate governance, where in companies with committees, the committee (the supervisor in our model) accurately grasps the state (type information) of the agent (operating officer) with a high probability and imposes the first best scheme for the agent. Under the positive cost of introducing the supervisor (a transaction cost), which regime the principal prefers most depends on whether the comparative (relative) advantage of the three-tier structure ('No Commitment' regime (NC)) over the two-tier structure ('No Supervisor regimes (NS)) is greater or not than the cost of introducing the supervisor.

Then, we incorporate behavioral elements ala Fahr and Schmidt (1999) into the model, and examine their effects on the optimal solution in the principal-supervisor-agent hidden information model with collusion. We find that these behavioral elements can change the monetary reward for inducing the true information, and so the virtual surplus for each type is also altered through the change in the information rent (an incentive cost for inducing a truthful information revelation). Thus, the optimal solution with behavioral elements can be different from the one with no behavioral elements. More concretely, we introduce the recent behavioral contract theory idea, "shading" (Hart and Moore (2008)) into our collusion model. By combining the two ideas, i.e., *collusion* and *shading*, we can enrich the existing model and obtain a new result on *Collusion-proof* vs. *Equilibrium Collusion* in that the increase in shading pressure (behavioral element) strengthens the incentive for collusion, thereby makes it difficult to implement the collusion-proof (Supervisor's truth telling) incentive schemes, which leads to the Equilibrium Collusion. That is, the collusion-proof principle does not hold any more in the presence of strong shading pressures (behavioral elements).

Further, by considering shading as a component of ex-post haggling (addressed by Coase (1937) and Williamson (1975), more generally, Transaction Cost Economics (TCE)), we can give a micro foundation (an explicit modeling) to ex-post adaptation costs, where we view rent-seeking associated with collusive behavior and ex-post haggling generated from aggrievement and shading as the two sources of the costs.[2] By using this model, we can analyze the optimal organizational design problem as an optimal response to the trade-off between gross total surplus and ex-post haggling cost. We believe that our model can help a deep understanding of resource allocation and decision process in internal organizations of large firms.

In summary, we apply the Monotone Comparative Statics method and the First Order (Mirrlees) Approach to the continuous-type, three-tier agency model with hidden information and collusion, thereby providing a framework that can address the issues treated in the existing literature *in a much simpler fashion*. We then characterize the nature of equilibrium contract that can be implemented under the possibility of collusion, and obtain a general comparison result on the organization structures. Next, we introduce the recent behavioral contract theory idea, "shading" into our collusion model. By combining the two ideas, i.e., *collusion* and *shading*, we can not only enrich the existing collusion model, including a new result on the choice of Collusion-proof vs. Equilibrium Collusion regimes, but also give a micro foundation to ex-post haggling costs, addressed by Transaction Cost Economics. This will contribute to a deep understanding of resource allocation and decision process in internal hierarchical organizations.

2 Principal-Agent Hidden Information Model with a Continuum of Types

2.1 Setting

We consider two players: a principal (P) and an agent (A). The principal owns the firm and hires the manager (agent) to run it. θ is the manager's ability to run the firm and $C(X,\theta)$ is the effort cost for the manager of type θ to attain the output X. For each θ, $C(X,\theta)$ satisfies $C(X,\theta) > 0$, $\partial C(X,\theta)/\partial X > 0$, $\partial^2 C(X,\theta)/\partial X^2 > 0$, $\forall X \in \mathbb{R}_+$. W is the wage payment the agent receives, and so his utility is $W - C(X,\theta)$. We normalize the agent's reservation utility as 0. The timing of the game is as follows. Prior to contracting, θ is determined randomly by nature and is known only to the manager (agent). The principal proposes a take-it-or-leave-it contract offer to the manager. The contract is written as $W(X)$, where X is the output level by the manager and W is the wage he receives if he generates X. If the manager accepts the offer, a contract is signed and the principal is fully committed. If he rejects the offer, the game ends.

2.2 Preliminary: Single Crossing Property (SCP) and Monotonicity of Agent's Choice

Faced with a wage scheme $W(X)$, the agent of type θ will choose

$$X \in \arg\max_{X \in \mathrm{X}} W(X) - C(X,\theta)$$

Analysis is dramatically simplified when the Agent's types can be ordered so that higher types choose a higher output when faced with any wage. We identify when solutions to the parameterized

[2]Theoretically, our model deals with a situation where bilateral collusive contracts are feasible while the grand contract is not feasible (i.e., an incomplete grand contract situation), which corresponds to a case where the Coase Theorem will not hold since externalities cannot be fully internalized (like in the Coase's 1937 paper). It would be novel to model the situation where the third party who suffers from the negative externality brought by such bilateral, collusive contracts shades ex post the colluding party (especially, the supervisor) by a constant times the aggrievement (the negative externality he suffers from), in the three-tier agency framework.

maximization program $\max_{X\in \mathrm{X}} U(X,\theta) := W(X) - C(X,\theta)$ are strictly increasing in the parameter θ. A key property to ensure monotone comparative statics is the following:

Definition 1 *A function $U : X \times \theta \to \mathbb{R}$ where X, $\theta \subset \mathbb{R}$ has the* ***Single Crossing Property (SCP)*** *if $U_X(X,\theta)$ exists and is strictly increasing in $\theta \in \Theta$.*[3]

$U(X,\theta) = W(X) - C(X,\theta)$ has SCP if $U_X(X,\theta) = W_X(X) - C_X(X,\theta)$ exists and is strictly increasing in $\theta \in \Theta$ for all $X \in \mathrm{X}$. In this case, $U(X,\theta)$ satisfies SCP when the marginal cost of output $C_X(X,\theta)$ is decreasing in type θ, i.e., higher types always have gentler indifference curves. SCP implies that large increases in X are less costly for higher parameters θ.

Theorem 1 (Edlin and Shannon 1998)
Let $\theta'' > \theta'$, $X' \in \arg\max_{X\in \mathrm{X}} U(X,\theta')$, and $X'' \in \arg\max_{X\in \mathrm{X}} U(X,\theta'')$. Then, if U has SCP, and either X' or X'' is in the interior of X, then $X'' > X'$.

We can apply Theorem 1 to the agent's choice when facing a wage scheme $W(\cdot)$, assuming that the agent's cost $C(X,\theta)$ satisfies SCP. To ensure full separation of types, we need to assume that the wage $W(\cdot)$ is differentiable. Then, $U(X,\theta)$ will satisfy SCP, and Theorem 1 implies that interior output choices are strictly increasing in types, i.e., we have *full separation*.

2.3 The Full information Benchmark

As a benchmark, we consider the case in which the Principal observes the Agent's type θ. Given θ, she offers the bundle (X, W) to solve:

$$\max_{(X,W)\in \mathrm{X}\times\mathbb{R}} X - W(X) \text{ s.t. } W(X) - C(X,\theta) \geq \bar{u} \text{ (IR)}$$

(IR) is the Agent's *Individual Rationality* constraint, and binds at an optimal solution. Then, the Principal eventually solves: $\max_{X\in \mathrm{X}} X - C(X,\theta) - \bar{u}$ Discarding the constant $\bar{u}$, it is exactly the total surplus maximization. Let $X^{FB}(\theta)$ denote a solution to this maximization problem, which we call the First Best (FB) solution. Using Theorem 1, we check whether our assumptions ensure that $X^{FB}(\theta)$ is strictly increasing in type θ. If $C(X,\theta)$ satisfies SCP, which implies that total surplus $X - C(X,\theta)$ satisfies SCP, and if $X^{FB}(\theta)$ is in the interior for each θ, we can conclude that $X^{FB}(\theta)$ is strictly increasing in θ.

Now we consider a different contract from the contract $W : X \to \mathbb{R}$ which we have considered so far, where the agent is asked to announce his type $\hat{\theta}$, and receives payment $W\left(\hat{\theta}\right)$ in exchange for an output $X\left(\hat{\theta}\right)$ on the basis of his announcement . This is called a *Direct Revelation Contract*. According to the *Revelation Principle*, any contract $W : X \to \mathbb{R}$ can be replaced with a *Direct Revelation Contract* that has an equilibrium in which all types receive the same bundles as in the original contract $W : X \to \mathbb{R}$.

2.4 Solution with a Continuum of Types

Let the type space be continuous: $\Theta = \left[\underline{\theta}, \bar{\theta}\right]$, with the cumulative distribution function $F(\cdot)$, and with a strictly positive density $f(\theta) = F'(\theta)$. In addition to previous assumptions, we assume that

[3] Edlin and Shannon (1998) introduced this SCP under the name of "increasing marginal returns".

$C(X,\theta)$ is continuously differentiable in θ for all X, and $C_\theta(X,\theta)$ is bounded uniformly across (X,θ). The principal's problem is:

$$\begin{aligned}
&\max_{\langle X(\cdot),W(\cdot)\rangle} \int_{\underline{\theta}}^{\bar{\theta}} [X(\theta) - W(\theta)] f(\theta)\, d\theta \\
&s.t. \quad W(\theta) - C(X(\theta),\theta) \geq W\left(\widehat{\theta}\right) - C\left(X\left(\widehat{\theta}\right),\theta\right) \quad (IC_{\theta\hat{\theta}}) \quad \forall \theta, \widehat{\theta} \quad \in \Theta \\
&\qquad W(\theta) - C(X(\theta),\theta) \geq \bar{u} \qquad\qquad\qquad\qquad (IR_\theta) \quad \forall \theta \in \Theta
\end{aligned}$$

Just as in the two-type case, out of all the participation constraints, only the lowest type's IR binds.

Lemma 1 *At a solution $(X(\cdot), W(\cdot))$, all IR_θ with $\theta > \underline{\theta}$ are not binding, and only $IR_{\underline{\theta}}$ is binding.*

As for the analysis of ICs with a continuum of types, Mirrlees (1971) introduced a widely used way to reduce the number of incentive constraints by replacing them with the corresponding First-Order Conditions.[4] The "trick" is as follows.

(IC) can be written as $\theta \in \arg\max_{\hat{\theta}\in\Theta} U\left(\hat{\theta},\theta\right)$, where $U\left(\hat{\theta},\theta\right) = W(\hat{\theta}) - C\left(X\left(\hat{\theta}\right),\theta\right)$ is the utility that the agent of type θ receives by announcing that his type is $\hat{\theta}$. If $\theta \in \left(\underline{\theta},\bar{\theta}\right)$ and $U\left(\hat{\theta},\theta\right)$ is differentiable in $\hat{\theta}$, then the first order condition $\partial U\left(\hat{\theta},\theta\right)\Big/\partial\hat{\theta}\Big|_{\hat{\theta}=\theta} = 0$ is necessary for the above optimality. We define the Agent's equilibrium utility (the value):

$$U(\theta) \equiv U(\theta,\theta) = W(\theta) - C(X(\theta),\theta)$$

Note that this utility depends on θ in two ways – through the agent's true type and through his announcement. Differentiating with respect to θ, we have $U'(\theta) = U_{\hat{\theta}}(\theta,\theta) + U_\theta(\theta,\theta)$, where the first derivative of U is with respect to the agent's announcement (the first argument) and the second derivative is with respect to the agent's true type (the second argument). Since the first derivative equals zero by $\partial U\left(\hat{\theta},\theta\right)\Big/\partial\hat{\theta}\Big|_{\hat{\theta}=\theta} = 0$, we have $U'(\theta) = U_\theta(\theta,\theta)$. This condition is nothing but the well known *Envelope Theorem*: the full derivative of the value of the agent's maximization problem with respect to the parameter – his type – equals to the partial derivative holding the agent's optimal announcement fixed. More concretely,

$$\frac{dU(\theta)}{d\theta} = \underbrace{W'(\theta) - \frac{\partial C(X(\theta),\theta)}{\partial X(\theta)}\frac{dX(\theta)}{d\theta}}_{\text{Indirect Effect}} \underbrace{-\frac{\partial C(X(\theta),\theta)}{\partial\theta}}_{\text{Direct Effect}}$$

Since $W'(\hat{\theta}) - \frac{\partial C\left(X\left(\hat{\theta}\right),\theta\right)}{\partial X\left(\hat{\theta}\right)} \cdot \frac{dX\left(\hat{\theta}\right)}{d\hat{\theta}} = 0$ at $\hat{\theta} = \theta$ (the agent's optimal announcement is *Truth Telling*), we have $W'(\theta) - \frac{\partial C(X(\theta),\theta)}{\partial X(\theta)}\frac{dX(\theta)}{d\theta} = 0$. That is, the indirect effect equals zero.

Thus, we have the *envelope condition*:

$$U'(\theta) = \frac{dU(\theta,\theta)}{d\theta} = -\frac{\partial C(X(\theta),\theta)}{\partial\theta}.$$

By integrating it, we have the important formula:

$$U(\theta) = U\left(\underline{\theta}\right) - \int_{\underline{\theta}}^{\theta} \frac{\partial C(X(\tau),\tau)}{\partial\tau} d\tau \quad \textbf{(ICFOC)}$$

[4] Fudenberg and Tirole (1991) pp257-268 "Mechanism Design with a Single Agent" reviews the methodology first developed by Mirrlees (1971), i.e., the First Order Approach. While on the other hand, this section introduces the "monotone comparative statics" method into the framework.

(ICFOC) demonstrates that with a continuum of types, *incentive compatibility constraints* pin down up to a constant plus all types' utilities for a given output rule $X(\cdot)$.This remarkable result[5] can be mathematically extended by the generalized Envelope Theorem by Milgrom and Segal (2002).

Intuitively, (ICFOC) incorporates local incentive constraints, ensuring that the Agent does not gain by slightly misrepresenting θ. By itself, it does not ensure that the Agent cannot gain by misrepresenting θ by a large amount. For example, (ICFOC) is consistent with the truthful announcement $\hat{\theta} = \theta$ being a local maximum, but not a global one. It is even consistent with truthful announcement being a local minimum.

Fortunately, these situations can be ruled out. For this purpose, recall that by SCP, Topkis (1978) and Edlin and Shannon (1998) establish that the agent's output choices from any tariff (and therefore in any incentive compatible contract) are nondecreasing in type. Thus, any piecewise differentiable IC contract must satisfy that $X(\cdot)$ is nondecreasing (M)

It turns out that under SCP, ICFOC in conjunction with (M) do ensure that truthtelling is a global maximum, i.e., all ICs are satisfied:

Lemma 2 $(X(\cdot), W(\cdot))$ *is* ***Incentive Compatible*** *if and only if both* ***(ICFOC) and (M) hold,*** *where* $U(\theta) = W(\theta) - C(X(\theta), \theta)$.

Proof: See, Appendix 1

Given (ICFOC), we can express transfers: $\underbrace{W(\theta)}_{\text{Wage Payment}} = \underbrace{C(X(\theta), \theta)}_{\text{Effort Cost}} + \underbrace{U(\theta)}_{\substack{\text{Information Rent} \\ \text{given for type } \theta}}$

3 Collusion and Supervision

3.1 Introduction of a Supervisor and the Collusion-proof Problem

Now, we introduce a supervisor into the model. The principal can have access, at a cost z, to a supervisor who can, for each θ, provide a proof of this fact with probability p, and with $1-p$, is unable to obtain any information. We assume that proofs of θ cannot be falsified. That is, θ is hard information.[6] On the other hand, the agent can potentially benefit from a failure by the supervisor to truthfully report that his type is θ when the supervisor observed the signal θ. A self-interested supervisor colludes with the agent only if he benefits from such behavior. We assume the following collusion technology: if the agent offers the supervisor a transfer (side payment) t, he benefits up to kt, where $k \in [0,1]$. That is, only a fraction, $k \in [0,1]$, of the agent's bribe ends up in the supervisor's hands.The idea is that transfers of this sort may be hard to organize and subject to resource losses. We follow the literature in assuming that side-contracts of this sort are *enforceable* (See, e.g., Tirole 1992).[7]

To avoid collusion, the principal will have to offer the supervisor a reward $W_s(\theta)$ for providing θ, such that the following *coalition incentive compatibility constraint* is satisfied.

$$W_s(\theta) \geq kU(\theta) = k\left[U\left(\underline{\theta}\right) - \int_{\underline{\theta}}^{\theta} \frac{\partial C(X(\tau), \tau)}{\partial \tau} d\tau\right]$$

[5] Our methodology is related to the "Envelope Approach" in auction theory, e.g., the analysis of first price auction by the envelope approach. As for it, e.g., see Milgrom (2004).

[6] Note that the supervisor still can hide the hard informative evidence on θ. We also assume that the agent correctly knows whether the supervisor obtained the hard informative signal on his type information θ or not. This is the same assumption as the early literature, e.g., Tirole (1986), Laffont and Tirole (1991).

[7] This means that we assume that the supervisor and the agent can collude through binding side contracts, thereby achieving a collusive manipulation on the supervisor's signal (hard evidence) in such a way that the agent pays a side payment to the supervisor and the supervisor hides the hard informative evidence on θ.

Indeed, once the information θ is obtained, the principal will reduce the Agent θ's payment $W(\theta)$ to effort cost $C(X(\theta),\theta)$, and not pay the information rent $U(\theta)$ to the agent θ. The agent is thus ready to pay the supervisor an amount of $U(\theta)$, and the value of this side payment to the supervisor is $kU(\theta)$, where $k \in [0,1]$. Therefore, hiring a supervisor and eliciting his information requires the principal to pay $W_s(\theta) = kU(\theta)$, $\forall\theta$ to the supervisor if the (hard) information of θ is provided. Substituting $W_s(\theta) = kU(\theta)$ into the Principal's objective function, the difference between total surplus and the information rent for type θ in the Principal-Supervisor-Agent regime is

$$X(\theta) - C(X(\theta),\theta) - [(1-p)+pk]\,U(\theta)$$

Hence, the program of designing the *optimal collusion-proof contract* can be rewritten as

$$\max_{X(.),U(.)} \int_{\underline{\theta}}^{\bar{\theta}} \left[\underbrace{X(\theta) - C(X(\theta),\theta)}_{\text{Total Surplus}} - [(1-p)+pk] \quad \underbrace{U(\theta)}_{\substack{\text{Information}\\ \text{Rent}}} \right] f(\theta)\,d\theta - z$$

$$\text{s.t. } dX(\theta)/d\theta \geq 0 : X(\theta) \text{ is nondecreasing (M)}$$

$$U(\theta) = U\left(\underline{\theta}\right) - \int_{\underline{\theta}}^{\theta} \frac{\partial C(X(\tau),\tau)}{\partial \tau} d\tau \text{ (ICFOC)}$$

$$U\left(\underline{\theta}\right) = W\left(\underline{\theta}\right) - C\left(X\left(\underline{\theta}\right),\underline{\theta}\right) \geq \bar{u} \text{ (Const.) } (\text{IR}_{\underline{\theta}})$$

Note that the objective function takes the familiar form of the expected difference between total surplus and the Agent's information rent.

3.2 Solving the Relaxed Problem

Thus, the principal's optimization problem can be rewritten as

$$\max_{X(.)} \int_{\underline{\theta}}^{\bar{\theta}} \left[X(\theta) - C(X(\theta),\theta) - [(1-p)+pk] \left(U\left(\underline{\theta}\right) - \int_{\underline{\theta}}^{\theta} \frac{\partial C(X(\tau),\tau)}{\partial \tau} d\tau \right) \right] f(\theta)\,d\theta - z$$

$$\text{s.t.} \qquad dX(\theta)/d\theta \geq 0 \quad (M) \quad \forall\theta$$

where $\int_{\underline{\theta}}^{\bar{\theta}} \left[U\left(\underline{\theta}\right) - \int_{\underline{\theta}}^{\theta} \frac{\partial C(X(\tau),\tau)}{\partial \tau} d\tau \right] f(\theta)\,d\theta$ can be called the *expected information rents*.

Lemma 3

$$\int_{\underline{\theta}}^{\bar{\theta}} \left[U\left(\underline{\theta}\right) - \int_{\underline{\theta}}^{\theta} \frac{\partial C(X(\tau),\tau)}{\partial \tau} d\tau \right] f(\theta)\,d\theta = U\left(\underline{\theta}\right) - \int_{\underline{\theta}}^{\bar{\theta}} \frac{\partial C(X(\theta),\theta)}{\partial \theta} \frac{1-F(\theta)}{f(\theta)} f(\theta)\,d\theta$$

Proof: See, Appendix 2

Substituting these expected information rents into the principal's program, and ignoring the constant $U\left(\underline{\theta}\right)$, the program becomes

$$\max_{X(.)} \int_{\underline{\theta}}^{\bar{\theta}} \left[X(\theta) - C(X(\theta),\theta) + [(1-p)+pk] \frac{\partial C(X(\theta),\theta)}{\partial \theta} \frac{1-F(\theta)}{f(\theta)} \right] f(\theta)\,d\theta - z$$

$$\text{s.t.} \qquad dX(\theta)/d\theta \geq 0 \quad (M) \quad \forall\theta$$

We ignore the Monotonicity Constraint (M) and solve the resulting *relaxed program*. Thus, the principal maximize the expected value of the expression within the square brackets, which is called

the *virtual surplus*,[8] and denoted by $J(X,\theta)$. This expected value is maximized by simultaneously maximizing virtual surplus for (almost) every type θ, i.e.,

$$X^S(\theta) \in \arg\max_{X(\cdot)} X(\theta) - C(X(\theta),\theta) + [(1-p)+pk]\left[\frac{1-F(\theta)}{f(\theta)}\right]\frac{\partial C(X(\theta),\theta)}{\partial\theta}$$

This defines the optimal output rule $X^S(\cdot)$ for the relaxed program. The principal's choice of $X^S(\theta)$ can be understood as a trade-off between maximizing the total surplus for type θ and reducing the information rents of all types above θ, just as in the two-type case. Indeed, **(ICFOC)** says that output choice X for type θ results in additional information rent $-\partial C(X(\theta),\theta)/\partial\theta$ for all types above θ.

In particular, for the highest type $\bar{\theta}$, there are no higher types, i.e., $F\left(\bar{\theta}\right)=1$ and the principal just maximizes total surplus, choosing $X^S\left(\bar{\theta}\right)=X^{FB}\left(\bar{\theta}\right)$. In words, we have *efficiency at the top*. For all other types, the principal will distort output to reduce information rents. To see the direction of distortion, consider the parameterized maximization program

$$\max_{X\in\mathrm{X}} \Psi(X,\xi) = X(\theta) - C(X(\theta),\theta) + \xi\left[\frac{1-F(\theta)}{f(\theta)}\right]\frac{\partial C(X(\theta),\theta)}{\partial\theta}$$

Here $\xi=0$ corresponds to surplus-maximization (first-best), and $\xi=1\,(p=0,k\in[0,1])$ corresponds to the principal's (relaxed) second best program with only one agent.

Note that $\frac{\partial\Psi(X,\xi)}{\partial X\partial\xi}=\left[\frac{1-F(\theta)}{f(\theta)}\right]\frac{\partial^2 C(X(\theta),\theta)}{\partial X\partial\theta}<0$ for $\theta<\bar{\theta}$ since the agent's value $U(X,\theta)=W(X)-C(X,\theta)$ has the single crossing property (SCP), that is, $\partial^2 U(X,\theta)/\partial X\partial\theta=-\partial^2 C(X,\theta)/\partial X\partial\theta>0$. Therefore, $\Psi(X,\xi)$ has SCP in $(X,-\xi)$, and by Theorem 1 (Edlin and Shannon), we have $X^*(\xi=1)\Leftrightarrow X(\theta)<X^{FB}(\theta)\Leftrightarrow X^*(\xi=0)$ for all $\theta<\bar{\theta}$. In words, the principal makes all types other than the highest type underproduce in order to reduce the information rents of types above them. Similarly, by introducing the supervisor, which basically corresponds to $0<\xi<1$, we have

$$X^*(\xi=1)\Leftrightarrow X(\theta)<X^*(\xi\in(0,1))\Leftrightarrow X^S(\theta)\le X^*(\xi=0)\Leftrightarrow X^{FB}(\theta).$$

Hence, in the Principal-Supervisor-Agent regime, the principal can induce more marginal incentives than the second best regime with only one agent through the reduction in total and marginal information rents paid to the supervisor and the agent θ,in other words, reducing the implementation costs for any $X<X^S\left(\bar{\theta}\right)=X^{FB}\left(\bar{\theta}\right)$. This result is a generalization of the two-type case. Thus, we obtain the following proposition.

Proposition 1 *In the Principal-Supervisor-Agent regime with a continuum of types, the optimal collusion-proof contract has the property that*

(1) Efficiency at the top (the highest type $\bar{\theta}$) $X\left(\bar{\theta}\right)=X^{FB}\left(\bar{\theta}\right)$

(2) Downward distortion for all other types $\theta\in\left[\underline{\theta},\bar{\theta}\right)$ is mitigated, that is,

$$X(\theta) \underbrace{\le}_{\substack{\text{Equality}\\ \text{holds at } k=1}} X^S(\theta) \underbrace{\le}_{\substack{\text{Equality holds}\\ \text{either at } p=1,k=0\\ \text{or } \theta=\bar{\theta}}} X^{FB}(\theta).$$

Now, remember that we ignored the monotonicity constraint (M) and solved the *relaxed program*. So, we need to check that the solution $X^S(\theta)$ indeed satisfies the monotonicity constraint (M), that

[8] This concept was first introduced by Myerson (1981).

is, the output rule $X^S(\theta)$ is nondecreasing. We can check it using Theorem1 (Edlin and Shannon (1998))[9]. To simplify expressions, define $h(\theta) \equiv f(\theta)/[1-F(\theta)] > 0$, which is called the *hazard rate* of type θ. Then, the principal's program can be rewritten as

$$\max_{X \in \mathrm{X}} J(X,\theta) = X - C(X,\theta) + \frac{[(1-p)+pk]}{h(\theta)} \frac{\partial C(X,\theta)}{\partial \theta}$$

By Topkis (1978) and Theorem 1, assuming that $C(X,\theta)$ is sufficiently smooth, a sufficient condition for $X^S(\theta)$ to be nondecreasing in θ is for the following derivative to be strictly increasing in θ:

$$\frac{\partial J(X,\theta)}{\partial X} = 1 - \frac{\partial C(X,\theta)}{\partial X} + \frac{[(1-p)+pk]}{h(\theta)} \frac{\partial^2 C(X,\theta)}{\partial X \partial \theta} \qquad (*)$$

Since $-C(X,\theta)$ satisfies SCP, the second term is strictly increasing in θ, and the first term does not depend on θ. The only problematic term, therefore, is the third term. Our result is ensured when the third term is nondecreasing in θ. Since $1/h(\theta)$ is positive and $\partial^2 C(X,\theta)/\partial X \partial \theta$ is negative, this is ensured when $\partial^2 C(X,\theta)/\partial X \partial \theta$ is nondecreasing. That is, we have

Proposition 2 *A sufficiency condition for the optimal collusion-proof solution $X^S(\theta)$ to satisfy the monotonicity constraint (M) is that the following conditions hold.*

1. *$\partial^2 C(X,\theta)/\partial X \partial \theta$ is nondecreasing in θ.*

2. *The hazard rate $h(\theta)$ is nondecreasing.*

Example: The first assumption is satisfied e.g.,in the following cost function forms:

$$C(X,\theta) = (X-\theta)^\alpha \text{ and } C(X,\theta) = (X/\theta)^\alpha,\ \alpha \geq 2$$

The second condition is called the "Monotone Hazard Rate Condition" and satisfied by many familiar probability distributions.[10] Now, we can present the following proposition on the comparative statics.

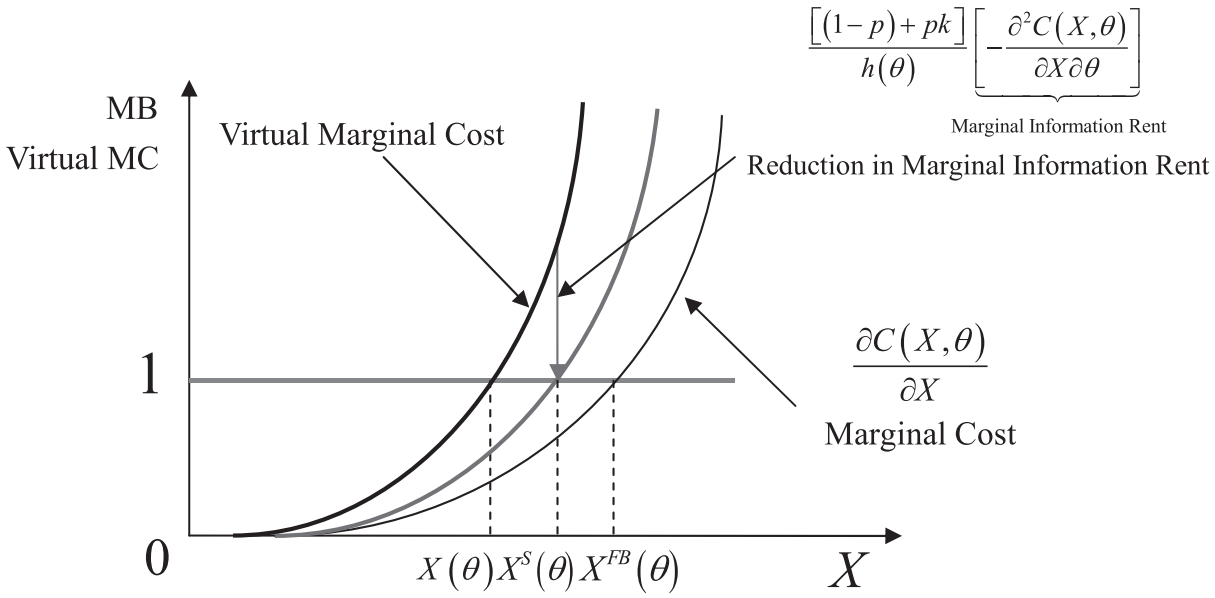
Graphical Explanation

Proposition 1 can be understood by using the Figure 1, which shows that the optimal solution $X^S(\theta)$ is determined such that the marginal benefit 1 equals the marginal *virtual cost* (the marginal cost $\frac{\partial C(X,\theta)}{\partial X}$ plus the marginal information rent $-\frac{[(1-p)+pk]}{h(\theta)} \frac{\partial^2 C(X,\theta)}{\partial X \partial \theta}$). The result of $X(\theta) \leq X^S(\theta) \leq X^{FB}(\theta)$ comes from the reduction in the marginal information rents by the introduction of a supervisor with $k \leq 1$. The point is the reduction in the virtual marginal cost due to $(1-p)+pk \leq 1$, compared with the standard no-supervisor case.

[9] Fudenberg and Tirole (1991) examines when it is legitimate to focus on the *relaxed program*, by using the differentiability approach (i.e., analyzing total differentiation of the first order condition to the relaxed program), not using the monotone comparative statics method. They derive the monotone hazard rate condition, that is, the condition2 of Proposition 2 as the assumption sufficient to satisfy the monotonicity constraint (M).

[10] For example, uniform, normal, logistic, exponential distributions etc. See Fudenberg and Tirole (1991), Bolton and Dewatripont (2005).

Figure 1



The condition 1 of Proposition2 means that the marginal information rent $-\frac{\partial^2 C(X,\theta)}{\partial X \partial \theta}$ is decreasing in θ, that is, shifts downwards as θ increases. Since the marginal cost $\frac{\partial C(X,\theta)}{\partial X}$ is also decreasing in θ, the proposition 2 as a whole refers to a sufficient condition for the virtual marginal cost to decrease in θ, that is, for $X^S(\theta)$ to increase in θ.

Proposition 3 *Suppose that the sufficiency condition in proposition 2 holds. Then, the optimal collusion-proof solution $X^S(\theta)$ is nondecreasing in the parameter p, and nonincreasing in the parameter k.*

Proof: From the equation $(*)$, the derivative $J_X(X,\theta)$ is nondecreasing in the parameter p, because the derivative of $J_X(X,\theta)$ in the parameter p is $-1+k \leq 0$ for $k \in [0,1]$, multiplied by $\frac{1}{h(\theta)}\frac{\partial^2 C(X,\theta)}{\partial X \partial \theta} \leq 0$. Hence, from the Theorem1, the optimal solution $X^S(\theta)$ is nondecreasing in the parameter p. Particularly, $X^S(\theta)$ is strictly increasing in p for $k \in [0,1)$ from Theorem 1. The latter part can also be proved in the same way: The derivative $J_X(X,\theta)$ is nonincreasing in the parameter k for $p \in [0,1]$, and so the optimal solution $X^S(\theta)$ is nonincreasing in the parameter k. ■

This result could be said to demonstrate the advantage of our approach, because the extensions of the Tirole (1986) model, such as Laffont and Tirole (1991), Kofman and Lawarree (1993), Laffont and Martimort (1997), and Suzuki (1999), often have the complicated structure of a Kuhn-Tucker problem with many IC and IR constraints, and so the global characterization of the optimal solutions as well as the robust comparative statics are difficult to obtain, and only a local characterization of the solution and comparative statics is possible in the above collusion literature, while on the other hand, we can readily perform a robust (monotone) comparative statics, and the rationale of the results is clear and intuitive.

We present economic insight on corporate governance. Under collusive supervision (auditing), that is, $p \downarrow$ and $k \uparrow$, the optimal collusion-proof solution (output) $X^S(\theta)$ by the agent (manager) becomes lower, as does the principal's (shareholder's) payoff. Such lower performance firms should move to some organizational form achieving $p \uparrow$ and $k \downarrow$. Hence, a company with committees could be said to be one of the desirable forms, in that it tightens the monitoring of the agent (manager)

$p \uparrow$[11] and ensures the independence of supervision $k \downarrow$ by employing outside directors as a majority of committee members.

4 A Problem from Lack of Commitment

So far, we have considered a situation where the principal can *commit* to the collusion-proof contract. That is, 'full commitment'. Here, we examine more explicitly the timing of the game. The principal has access to the supervisor, who chooses a report r $\in \{\phi, \theta\}$, where ϕ means that he did not obtain any information. If the principal receives the message from the supervisor that the type information is θ, the principal will have an incentive to modify the original contract. The principal can raise her payoff by *eliminating the downward distortions in all other types* than the highest one θ. Namely, instead of $\{X(\theta), W(\theta)\}$, she will offer the efficient (first best) contract $\left\{X^{FB}(\theta), W^{FB}(\theta)\right\}$, and the information rent $U(\theta)$ will be exploited by the principal. In summary, the principal commits herself to the reward scheme for the supervisor, but does not commit to the one for the agent. Thus, she is tempted to modify the initial contract (or the outcome $\{X(\theta), W(\theta)\}$) unilaterally, using the information revealed by the supervisor.[12]

If the agent of type θ anticipates this modification, since he can benefit from a failure by the supervisor to report his type θ truthfully, he will offer the supervisor the transfer (side payment) $t = U(\theta)$,the amount equivalent to his information rent, of which the supervisor benefits up to kt, where $k \in [0, 1]$. Thus, the principal must pay $W_s(\theta) = kU(\theta)$ to the supervisor in opposition to the collusive offer by the agent, in order to elicit true information. In summary, the principal can strictly improve his payoff ex-post by changing $X(\theta)$ into $X^{FB}(\theta)$, but must bear the ex-ante incentive cost $kU(\theta)$. This is the trade-off for the principal when the supervisor obtains the proof of true information, with probability p.

Only when the supervisor cannot obtain any information for θ with probability $1-p$, does the principal commit herself to the initial scheme $\{X(\theta), W(\theta)\}\,\forall\theta$, and the same trade-off between the total surplus and the information rent emerges.

The expected total surplus minus the information rent for type θ in this regime is written as

$$(1-p)\left[X(\theta) - C\left(X(\theta), \theta\right)\right] + \underbrace{p}_{\substack{\theta \text{ is} \\ \text{revealed}}} \times \left[\underbrace{X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)}_{\text{(Ex post) First Best Allocative Efficiency}}\right] - \left[(1-p) + pk\right] U(\theta)$$

Eventually, in this regime, the principal maximizes the *virtual surplus* $J(X, \theta)$,

$$\max_{X \in \mathrm{X}} J(X, \theta) = (1-p)\left[X(\theta) - C\left(X(\theta), \theta\right)\right] + \frac{\left[(1-p) + pk\right]}{h(\theta)} \frac{\partial C(X, \theta)}{\partial \theta}$$

The first order condition for the optimum is,

$$\frac{\partial J(X, \theta)}{\partial X} = (1-p)\left[1 - \frac{\partial C(X, \theta)}{\partial X}\right] + \frac{\left[(1-p) + pk\right]}{h(\theta)} \frac{\partial^2 C(X, \theta)}{\partial X \partial \theta} = 0$$

$$\Leftrightarrow \quad \underbrace{1 - \frac{\partial C(X, \theta)}{\partial X}}_{\text{Marginal Total Surplus}} + \underbrace{\frac{\left[1 + \frac{p}{1-p}k\right]}{h(\theta)} \frac{\partial^2 C(X, \theta)}{\partial X \partial \theta}}_{\text{Marginal Information Rent}} = 0$$

[11] Of course, free-riding among committee members would lead to a lower monitoring, which corresponds to a lower p. Then, needless to say, the performance of a company with committees would be deteriorated.

[12] This idea is similar to the ratchet effect and the renegotiation problem caused by lack of a principal's commitment in the dynamics of incentive contracts, which were studied early by Laffont-Tirole (1988), and Dewatripont (1988) etc.

Noting that the marginal information rent for each $\theta \in \left[\underline{\theta}, \bar{\theta}\right)$ becomes larger than any other former regimes, we have the following proposition on the comparison of equilibrium incentives.

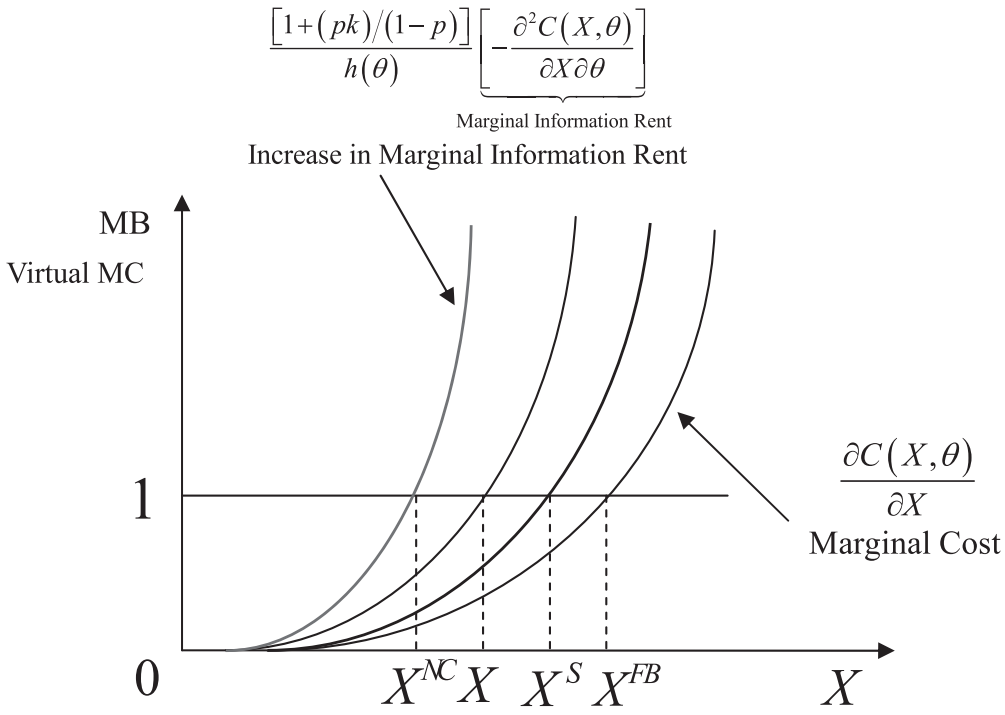
Proposition 4 *Supposing that $X^{NC}(\theta)$ is the solution (in the no-information phase $\emptyset$) of this 'No-Commitment' regime, we obtain:*

$$X^{NC}(\theta) \leq X(\theta) \leq X^{S}(\theta) \leq X^{FB}(\theta) \text{ for all } \theta \in \left[\underline{\theta}, \bar{\theta}\right]$$

Graphical Explanation

$X^{NC}(\theta) \leq X(\theta)$ in Proposition 4 comes from the increase in the virtual cost, i.e., the total and marginal information rents in this regime. Virtual marginal cost increases by $pk/(1-p)$, compared with the standard no-supervisor case. The below figure2 clearly shows this point.

Figure 2



Now, we can perform a comparative statics on the optimal solution $X^{NC}(\theta)$.

Proposition 5 *Comparative statics on $X^{NC}(\theta)$*

The optimal output $X^{NC}(\theta)$ in this no commitment/renegotiation regime is nonincreasing in the parameter p (in opposition to Proposition 3), and nonincreasing in the parameter k (in parallel to Proposition3).

Proof: The coefficient of the marginal information rent $1 + (pk)/(1-p)$ increases as the parameter p increases. Hence, the marginal information rent (and so the marginal virtual cost) $-\frac{[1+(pk)/(1-p)]}{h(\theta)}\frac{\partial^2 C(X,\theta)}{\partial X \partial \theta}$ increases as p increases. This brings about the decrease in the optimal output $X^{NC}(\theta) \downarrow$. Similarly, the coefficient of the marginal information rent $1+(pk)/(1-p)$ increases as the parameter k increases. Hence, the marginal information rent (and so the marginal virtual cost) increases as k increases. This brings about the decrease in the optimal output $X^{NC}(\theta) \downarrow$. ■

5 Payoff Comparison between Three Regimes

We compare the payoffs between three regimes, that is, 'No-Commitment' regime (NC), 'Principal-Supervisor-Agent' regime (S) with full commitment, and 'No-Supervisor' second best regime (NS).

The expected payoff for the principal in the 'No-Commitment' regime (NC) is

$$(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{NC}(\theta)-C\left(X^{NC}(\theta),\theta\right)\right]f(\theta)\,d\theta+p\times\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta$$
$$+\left[(1-p)+pk\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$

The expected payoff for the principal in the 'Principal-Supervisor-Agent' regime (S) with full commitment is

$$(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)\right]f(\theta)\,d\theta+p\times\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)\right]f(\theta)\,d\theta$$
$$+\left[(1-p)+pk\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X^{S}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$

The expected payoff for the principal in No-Supervisor regime (NS), which is the standard second best regime and corresponds to $p=0$, $k=0$ in the Principal-Supervisor-Agent regime (S), is

$$\int_{\underline{\theta}}^{\bar{\theta}}\left[X(\theta)-C\left(X(\theta),\theta\right)+\frac{1}{h(\theta)}\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}\right]f(\theta)\,d\theta$$

We consider the comparison between the three regimes under $Z=0$, that is, the cost of introducing the supervisor (a transaction cost) is zero.

Step1

First, we compare the equilibrium payoffs between the 'No-Commitment' (NC) regime and the 'Principal-Supervisor-Agent **Collusion-proof, Commitment**' (S) regime.
By definition, $X^{NC}(\theta)$ is the optimal decision over the problem

$$\max_{X(\cdot)}(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X(\theta)-C\left(X(\theta),\theta\right)\right]f(\theta)\,d\theta$$
$$+\left[(1-p)+pk\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$

Then, from the *revealed preference* argument, the following holds.

$$(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{NC}(\theta)-C\left(X^{NC}(\theta),\theta\right)\right]f(\theta)\,d\theta$$
$$+\left[(1-p)+pk\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$
$$\geq(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)\right]f(\theta)\,d\theta$$
$$+\left[(1-p)+pk\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X^{S}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$

The following inequality holds by the same *revealed preference* argument.

$$\underbrace{p\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\text{First Best Expected Total Surplus}}\geq p\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)\right]f(\theta)\,d\theta$$

Hence, we have the following inequality.

$$\begin{aligned}
&(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{NC}(\theta)-C\left(X^{NC}(\theta),\theta\right)\right]f(\theta)\,d\theta+\underbrace{p\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\text{First Best Expected Total Surplus}}\\
&\qquad+\left[(1-p)+pk\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta\\
&\geq(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)\right]f(\theta)\,d\theta+p\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)\right]f(\theta)\,d\theta\\
&\qquad+\left[(1-p)+pk\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X^{S}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta\\
&=\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)\right]f(\theta)\,d\theta+\left[(1-p)+pk\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X^{S}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta
\end{aligned}$$

Thus, 'No Commitment' regime (NC) is payoff dominant over the 'Principal-Supervisor-Agent Collusion-proof, Commitment' (S) regime for the principal.

Step2

Next, we compare the equilibrium payoffs between the 'Principal-Supervisor-Agent **Collusion-proof, Commitment**' regime (S) and 'No Supervisor' regime (NS).
By definition, $X(\theta)$ is the optimal decision over the problem

$$\max_{X(\cdot)}\int_{\underline{\theta}}^{\bar{\theta}}\left[X(\theta)-C\left(X(\theta),\theta\right)+\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}\frac{1}{h(\theta)}\right]f(\theta)\,d\theta$$

By definition, $X^{S}(\theta)$ is the optimal decision over the problem

$$\max_{X(\cdot)}\int_{\underline{\theta}}^{\bar{\theta}}\left[X(\theta)-C\left(X(\theta),\theta\right)+\left[(1-p)+pk\right]\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}\frac{1}{h(\theta)}\right]f(\theta)\,d\theta$$

Then, from the *revealed preference* argument, the following holds.

$$\begin{aligned}
&\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)+\left[(1-p)+pk\right]\frac{\partial C\left(X^{S}(\theta),\theta\right)}{\partial\theta}\frac{1}{h(\theta)}\right]f(\theta)\,d\theta\\
&\geq\int_{\underline{\theta}}^{\bar{\theta}}\left[X(\theta)-C\left(X(\theta),\theta\right)+\left[(1-p)+pk\right]\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}\frac{1}{h(\theta)}\right]f(\theta)\,d\theta\\
&\qquad+\underbrace{p(1-k)}_{\geq0}\underbrace{\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta}_{-}\\
&=\int_{\underline{\theta}}^{\bar{\theta}}\left[X(\theta)-C\left(X(\theta),\theta\right)+\frac{1}{h(\theta)}\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}\right]f(\theta)\,d\theta
\end{aligned}$$

Thus, 'Principal-Supervisor-Agent **Collusion-proof, Commitment**' regime (S) is payoff dominant over the 'No Supervisor' regime (NS) when $Z = 0$

Combining the results of these two steps, we obtain

$$(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{NC}(\theta)-C\left(X^{NC}(\theta),\theta\right)\right]f(\theta)\,d\theta+p\underbrace{\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\text{First Best Expected Total Surplus}}$$

$$+\left[(1-p)+pk\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$

$$\geq\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)+\left[(1-p)+pk\right]\frac{\partial C\left(X^{S}(\theta),\theta\right)}{\partial\theta}\frac{1}{h(\theta)}\right]f(\theta)\,d\theta$$

$$\geq\int_{\underline{\theta}}^{\bar{\theta}}\left[X(\theta)-C\left(X(\theta),\theta\right)+\frac{1}{h(\theta)}\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}\right]f(\theta)\,d\theta$$

Proposition 6 *Suppose that $Z = 0$. Then,*
1. The principal prefers the 'No Commitment' regime (NC) to the Principal-Supervisor-Agent Collusion-proof, Commitment regime (S) in terms of her expected payoff.
2. The principal prefers the Principal-Supervisor-Agent Collusion-proof, Commitment regime (S) to the 'No Supervisor' regimes (NS) in terms of her expected payoff.
3. Hence, the principal prefers the 'No Commitment' regime (NC) to the 'No Supervisor' regimes (NS) in terms of her expected payoff.

Rationale

In the step1 where the principal compares the 'No Commitment' regime (NC) with the Principal-Supervisor-Agent Collusion-proof, Commitment regime (S), the principal designs a "more state-contingent" contract through more efficient use of supervisor's report $r \in \{\theta, \phi\}$, that is, she sets $X^{FB}(\theta)$ for the states $\{\theta, s = \theta\}$ where the agent type is θ and the supervisor's signal is $s = \theta$, and sets $X^{NC}(\theta)$ for the states $\{\theta, s = \phi\}$ where the agent type is θ and the supervisor's signal is $s = \phi$. On the other hand, in the Principal-Supervisor-Agent Collusion-proof, Commitment regime (S), the principal does not use the supervisor's report $r \in \{\theta, \phi\}$ in a state-dependent way, but unanimously imposes the pooling solution $X^S(\theta)$ for both states $\{\theta, s = \theta\}$ and $\{\theta, s = \phi\}$, which would not be efficient.

If we use the terminology in Weitzman's paper (1974) "Prices vs. Quantities", the "Commitment" regime (S) is the regime where the principal adjusts only the price rule $W(\theta)$ under the commitment to the output (quantity) rule $X^S(\theta)$, in the form that she does not pay the information rent $U(\theta)$ to the agent of type θ when the supervisor's report is $r = \theta$. On the other hand, "No Commitment" regime (NC) is the regime where the principal cannot commit herself not to adjust the output (quantity) rule $X(\theta)$ as well as the price rule $W(\theta)$, that is, the principal optimally adjusts both of them under the supervisor's report $r = \theta$.

In the step2 where the principal compares the Principal-Supervisor-Agent Collusion-proof, Commitment regime (S) with the 'No Supervisor regimes, the virtual surplus for type θ is more increased in the former regime through the reduction of information rent due to $(1-p)+pk \leq 1$, that is,

$$X(\theta)-C\left(X(\theta),\theta\right)+\underbrace{\left[(1-p)+pk\right]}_{\leq 1}\underbrace{\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}}_{-}\frac{1}{h(\theta)}$$

$$\geq X(\theta)-C\left(X(\theta),\theta\right)+\underbrace{\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}}_{-}\frac{1}{h(\theta)}$$

Hence, the principal prefers the former regime to the latter one.

An Interpretation in Top Management Organization in Corporate Governance

The advantage in the ex-post expected total surplus becomes bigger in the "state-contingent, No-Commitment" regime. This may be consistent with a situation, where in companies with committees, the committee (the supervisor in our model) accurately grasps the state (type informationθ) of the agent (operating officer) with a high probability p and the first best scheme $X^{FB}(\theta)$ is imposed for the agent.

The Choice of Organization Structure

Now define $Z^*(p,k)$ be the payoff difference between the 'No Commitment' regime (NC) and the 'No Supervisor' regime (NS). That is,

$$\begin{aligned} Z^*(p,k) := & \left\{ (1-p)\int_{\underline{\theta}}^{\bar{\theta}} \left[X^{NC}(\theta) - C\left(X^{NC}(\theta),\theta\right)\right] f(\theta)\, d\theta + p\int_{\underline{\theta}}^{\bar{\theta}} \left[X^{FB}(\theta) - C\left(X^{FB}(\theta),\theta\right)\right] f(\theta)\, d\theta \right. \\ & \left. + [(1-p)+pk]\int_{\underline{\theta}}^{\bar{\theta}} \frac{1}{h(\theta)} \frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial \theta} f(\theta)\, d\theta \right\} \\ & - \int_{\underline{\theta}}^{\bar{\theta}} \left[X(\theta) - C(X(\theta),\theta) + \frac{\partial C(X(\theta),\theta)}{\partial \theta}\frac{1}{h(\theta)}\right] f(\theta)\, d\theta \end{aligned}$$

Then, the optimal organizational structure $R^*$is determined based on the following rule.

$$R^*(p,k,Z) = \begin{cases} NC & if\ Z \le Z^*(p,k) \\ NS & if\ Z > Z^*(p,k) \end{cases}$$

That is to say, the principal's optimal strategy is to choose the three-tier structure with supervision (NC) if $Z \le Z^*(p,k)$, and to choose the two-tier structure with no supervision (NS) if $Z > Z^*(p,k)$, for $0 \le p,k \le 1$.

From the simple comparative statics, we have

$$\begin{aligned} \frac{\partial Z^*(p,k)}{\partial p} = & \int_{\underline{\theta}}^{\bar{\theta}} \underbrace{\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta),\theta\right)\right] - \left[X^{NC}(\theta) - C\left(X^{NC}(\theta),\theta\right)\right]\right\}}_{\ge 0} f(\theta)\, d\theta \\ & + \underbrace{(k-1)}_{\le 0}\int_{\underline{\theta}}^{\bar{\theta}} \frac{1}{h(\theta)} \underbrace{\frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial \theta}}_{<0} f(\theta)\, d\theta \ge 0 \ \forall (p,k) \in [0,1]^2 \\ \frac{\partial Z^*(p,k)}{\partial k} = & \, p\int_{\underline{\theta}}^{\bar{\theta}} \frac{1}{h(\theta)} \frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial \theta} f(\theta)\, d\theta \le 0 \end{aligned}$$

As p (the accuracy of supervision/monitoring) increases,the relative importance[13] $Z^*(p,k)$ of the three-tier structure with supervision increases. As for k(the efficiency of collusion) increases, it decreases.

[13]Relative importance could be rephrased as Comparative (Relative) advantage of the three-tier structure over the two-tier structure.

6 Behavioral Organization Design

6.1 Psychological Cost Model

In this section, we incorporate behavioral elements ala Fehr and Schmidt (1999) into the model.[14] Concretely, we introduce the behavioral assumption that the supervisor feels "psychological benefits and/or costs" from reporting activity, which is non-monetary and non-transferable, and then consider the optimal organization design problem by the (rational, payoff maximizing) principal.

We assume that the supervisor may bear a psychological per unit cost β (≥ 0) expressed in monetary terms when he discloses the type information θ about the agent, which reflects the deterioration of his relationship with the agent. On the other hand, he may incur a psychological per unit cost γ (≥ 0) of lying to the principal. The supervisor may be convinced that loyally reporting any information is valuable for his future career because it benefits the shareholders (principal), or he, as an auditor, may suffer from a guilty conscience for remaining silent $r = \phi$ against an efficient use of ability θ within the firm, or he may just be an honest outside auditor. Thus, γ is a parameter similar to β.

Whenever the supervisor observes the information θ (with probability p), he must decide whether to report this information or not. The payoff for him is $W_s(\theta) - \beta U(\theta)$ if he observes θ and reports it truthfully, and $kU(\theta) - \gamma\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - \left[X(\theta) - C(X(\theta), \theta) - U(\theta)\right]\right\}$ if he observes θ and does not report this evidence in collusion with the agent. Here, $\beta U(\theta)$ implies that the supervisor may feel bad about extracting the information rent $U(\theta)$ from the agent of type θ by reporting the information truthfully $r = \theta$, and then the psychological cost is β times $U(\theta)$.

Next, $\gamma\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - \left[X(\theta) - C(X(\theta), \theta) - U(\theta)\right]\right\}$ implies that the supervisor may feel bad about preventing the principal from attaining the first best efficiency by telling a lie $r = \phi$, given the agent type θ, and then the psychological cost is γ times the following value: $\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - \left[X(\theta) - C(X(\theta), \theta) - U(\theta)\right]\right\}$.

Hence, the supervisor will report the information θ truthfully only when the following incentive constraint holds, where the psychological costs are taken into consideration.

$$\underbrace{W_s(\theta)}_{\text{Wage Payment}} - \underbrace{\beta U(\theta)}_{\substack{\text{Psychological Cost} \\ \text{for supervisor}}} \geq \underbrace{kU(\theta)}_{\substack{\text{Side Payment from} \\ \text{agent to supervisor}}} - \underbrace{\gamma\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - \left[X(\theta) - C(X(\theta), \theta) - U(\theta)\right]\right\}}_{\substack{\text{Psychological Cost} \\ \text{for supervisor}}}$$

where $U(\theta) = U\left(\underline{\theta}\right) - \int_{\underline{\theta}}^{\theta} \frac{\partial C(X(\tau),\tau)}{\partial \tau} d\tau$ is the information rent for the agent given the output rule $X(\theta)$.Otherwise they will collude to keep the information secret.

In order to avoid collusion and induce the (behavioral) supervisor to tell the truth, the principal must offer the supervisor a monetary reward $W_s(\theta)$ for providing θ, such that the following *modified coalition incentive compatibility constraint* is satisfied:

$$W_s(\theta) \geq kU(\theta) + (\beta - \gamma) U(\theta) - \gamma\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - \left[X(\theta) - C(X(\theta), \theta)\right]\right\}$$

Therefore, at the optimum, the reward is

$$W_s(\theta) = kU(\theta) + (\beta - \gamma) U(\theta) - \gamma\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - \left[X(\theta) - C(X(\theta), \theta)\right]\right\}, \forall \theta$$

[14] Fehr and Schmidt (1999) give a logical approach to behavioral economics and explain multitude of evidence. Suzuki (2007) considers a setting where the existence of behavioral elements with a zero-sum structure leads to a strong incentive for vertical collusion in the principal-supervisor-two agent hierarchy, and analyzes the optimal (incomplete) contract design problem.

if (hard) information concerning θ is provided. Note that $\beta - \gamma \geq 0$ implies that the supervisor is rather agent-oriented and so more reward is required for him to tell the truth, and vice versa.

The principal will reflect this psychological cost in her optimal contract design. We expect that the characteristic of the supervisor (agent-oriented $\beta - \gamma \geq 0$ or principal-oriented $\beta - \gamma \leq 0$) will have an important impact on the optimal solution through the increase or decrease in the information rent as the reward for inducing truth telling.

Note that the above formulation is related to the "other-regarding preferences" in behavioral economics, which take many related forms. For example, "reciprocal altruism" means "I am made better off when someone else who has tried to help me is made better off". In our context, the three players (the principal, the supervisor, and the agent) are connected with one another. Then, when the supervisor reports the truth θ, he obtains the reward $W_s(\theta)$, but feels worse off by $\beta U(\theta)$, since the agent who is connected with him will be worse off by $U(\theta)$ due to the loss of information rent. Similarly, when the supervisor colludes with the agent and hides the truth θ, he obtains the side payment $kU(\theta)$ from the agent, but feels worse off by γ times the principal's payoff loss, since the principal who employs him will be worse off by

$$-\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - \left[X(\theta) - C\left(X(\theta), \theta\right) - U(\theta)\right]\right\}$$

due to the failure to exploit the information rent from the agent and at the same time to achieve the first best optimum $X^{FB}(\theta)$.

6.2 Shading Model: Observable Collusion

Next, let us compare the above formulation with the supervisor's psychological cost with the formulation based on the "shading" model[15] by Hart and Moore (2008), which introduced the new idea that a contract provides a reference point for parties' feelings of entitlement. A party who felt aggrieved in terms of his entitlement shades the party who aggrieved him to the point where his payoff falls by a constant multiplied by the aggrievement level, that is, the former punishes the latter by a constant times the aggrievement level.

In our model, the agent of type θ feels entitled to the information rent (indirect utility) $U(\theta)$ indicated by the initial contract. Nevertheless, the supervisor reported $r = \theta$ and aggrieved (disappointed) the agent by exploiting the information rent $U(\theta)$. Then, the agent shades (punishes) the supervisor by $\beta U(\theta)$. So, the net payoff of the supervisor when he reports the truth $r = \theta$ is

$$\underbrace{W_s(\theta)}_{\text{Wage Payment}} - \underbrace{\beta U(\theta)}_{\text{shading loss}}.$$

As for the principal's shading, there exists a subtle informational point. Our model is basically a hidden information model and the supervisor's signal θ is not observed by the principal. Otherwise (if the principal directly observed θ), she would not need the supervisor. We have already assumed that the supervisor, with probability p, obtains a proof (evidence) that the agent type is θ. Now suppose that the principal can know that the above state (of probability p) has happened, i.e., the supervisor has observed *some* signal θ. But suppose that she cannot know the *exact* value of θ, and also cannot verify that the supervisor has observed some signal θ. Then, if the supervisor provides no proof (evidence), the principal knows that the collusion has occurred (a side contract has been signed) between the agent of some type and the supervisor, though it is *not verifiable*. Only when the principal commits herself to the initial scheme $\{X(\theta), W(\theta)\}$ and enforces $X(\theta)$ for the agent's report $\hat{\theta} = \theta$, she can know the *exact* value of θ, and understand how much she has been aggrieved by the supervisor. Then, she can shade the supervisor. In summary, this information structure means that collusion (side contracting) between the supervisor and the type θ agent is *observable ex post* but *unverifiable*.

[15] It is related to negative reciprocity in the behavioral economics literature, that is, "I am better off when someone who has tried to hurt me is hurt".

Formally, then, the principal would feel that she had been entitled to $X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)$ since the type information was θ. Nonetheless, she could only attain the payoff under an asymmetric information regime between the principal and the agent θ, $X(\theta) - C(X(\theta), \theta) - U(\theta)$, since the supervisor colluded with the agent and hid the information θ. In summary, she was aggrieved by $\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - \left[X(\theta) - C(X(\theta), \theta) - U(\theta)\right]\right\}$ and so shades (punishes) the supervisor by a constant times the aggrievement level

$$\gamma\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - \left[X(\theta) - C(X(\theta), \theta) - U(\theta)\right]\right\}$$

Thus, we obtain the supervisor's incentive constraint with behavioral assumptions

$$\underbrace{W_s(\theta)}_{\text{wage payment}} - \underbrace{\beta U(\theta)}_{\text{shading loss}} \geq \underbrace{kU(\theta)}_{\text{side payment}} - \underbrace{\gamma\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - \left[X(\theta) - C(X(\theta), \theta) - U(\theta)\right]\right\}}_{\text{shading loss}}$$

$$\Leftrightarrow \underbrace{W_s(\theta)}_{\text{wage payment}} \geq \underbrace{kU(\theta)}_{\text{side payment}} + \underbrace{(\beta - \gamma) U(\theta)}_{\text{shading loss by agent}} - \underbrace{\gamma\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - \left[X(\theta) - C(X(\theta), \theta)\right]\right\}}_{\text{shading loss by principal}}$$

This is consistent with the "psychological cost" formulation, which we solve below.

Substituting $W(\theta) = C(X(\theta), \theta) + U(\theta)$ and

$$W_s(\theta) = kU(\theta) + (\beta - \gamma) U(\theta) - \gamma\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - \left[X(\theta) - C(X(\theta), \theta)\right]\right\}$$

into the principal's objective function, we have the formulation of virtual surplus for type θ

$$\begin{aligned}
&p\left(X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right) - W_s(\theta)\right) + (1-p)\left(X(\theta) - C(X(\theta), \theta) - U(\theta)\right) \\
&= (1 - (1+\gamma)p)\left(X(\theta) - C(X(\theta), \theta)\right) \\
&- \left[(1-p) + p\left\{k + (\beta - \gamma)\right\}\right] \underbrace{U(\theta)}_{\substack{\text{Information} \\ \text{Rent}}} + (1+\gamma) p\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right]
\end{aligned}$$

Hence, the program of designing the optimal collusion-proof contract *with a behavioral supervisor* can be rewritten as

$$\begin{aligned}
&\max_{X(.)} \int_{\underline{\theta}}^{\bar{\theta}} \left[\begin{array}{l}(1 - (1+\gamma)p)\left(X(\theta) - C(X(\theta), \theta)\right) \\ - \left[(1-p) + p\left\{k + (\beta - \gamma)\right\}\right] U(\theta) + (1+\gamma) p\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right]\end{array}\right] f(\theta)\, d\theta - z \\
&\text{s.t.} \qquad dX(\theta)/d\theta \geq 0 \quad (M) \quad \forall \theta
\end{aligned}$$

From lemma 3, the program becomes

$$\begin{aligned}
&\max_{X(.)} \int_{\underline{\theta}}^{\bar{\theta}} \left[(1 - (1+\gamma)p)\left(X(\theta) - C(X(\theta), \theta)\right) + \left[(1-p) + p\left\{k + (\beta - \gamma)\right\}\right] \frac{\partial C(X(\theta), \theta)}{\partial \theta} \frac{1 - F(\theta)}{f(\theta)}\right] f(\theta)\, d\theta \\
&\qquad + (1+\gamma) p\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - z \\
&\text{s.t.} \qquad dX(\theta)/d\theta \geq 0 \quad (M) \quad \forall \theta
\end{aligned}$$

We ignore the Monotonicity Constraint (M) and solve the *relaxed program*. The principal maximizes the expected value of the *modified virtual surplus*, denoted by $J^B(X, \theta)$. This expected value is maximized by simultaneously maximizing the modified virtual surplus for (almost) every type θ, i.e.

$$X^B(\theta) \in \arg\max_{X(\cdot)} \underbrace{(1 - (1+\gamma)p)\left(X(\theta) - C(X(\theta), \theta)\right) + \left[(1-p) + p\left\{k + (\beta - \gamma)\right\}\right] \frac{\partial C(X(\theta), \theta)}{\partial \theta} \frac{1 - F(\theta)}{f(\theta)}}_{J^B(X, \theta)}$$

This defines the optimal output rule $X^B(\cdot)$ for the program.[16]

The principal's program can then be rewritten as

$$\max_{X\in \mathrm{X}} J^B(X,\theta) = (1-(1+\gamma)p)(X(\theta)-C(X(\theta),\theta)) + \frac{[(1-p)+p\{k+(\beta-\gamma)\}]}{h(\theta)}\frac{\partial C(X(\theta),\theta)}{\partial \theta}$$

where $h(\theta) = f(\theta)/(1-F(\theta))$ is the hazard rate.

We take the derivative:

$$\frac{\partial J^B(X,\theta)}{\partial X} = [1-(1+\gamma)p]\left[1-\frac{\partial C(X,\theta)}{\partial X}\right] + \frac{[(1-p)+p\{k+(\beta-\gamma)\}]}{h(\theta)}\frac{\partial^2 C(X,\theta)}{\partial X \partial \theta} \quad (**)$$

Proposition 7 *The optimal solution $X^B(\theta)$ with behavioral elements is smaller than the solution $X^{NC}(\theta)$ with no behavioral elements, that is, $X^B(\theta) \leq X^{NC}(\theta)$*

Proof: Since

$$\frac{\partial J^{NC}(X,\theta)}{\partial X} = (1-p)\left(1-\frac{\partial C(X,\theta)}{\partial X}\right) + \frac{[(1-p)+pk]}{h(\theta)}\frac{\partial^2 C(X,\theta)}{\partial X\partial\theta} = 0 \text{ at } X = X^{NC}(\theta),$$
$$\Leftrightarrow \frac{1}{h(\theta)}\frac{\partial^2 C\left(X^{NC}(\theta),\theta\right)}{\partial X\partial\theta} = -\frac{(1-p)}{[(1-p)+pk]}\left(1-\frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial X}\right)$$

we find from the above $(**)$ that

$$\begin{aligned}\frac{\partial J^B(X,\theta)}{\partial X} &= -p\gamma\left(1-\frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial X}\right) + \frac{p(\beta-\gamma)}{h(\theta)}\frac{\partial^2 C\left(X^{NC}(\theta),\theta\right)}{\partial X\partial\theta} \text{ at } X = X^{NC}(\theta)\\ &= -p\gamma\left(1-\frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial X}\right) - \frac{p(\beta-\gamma)(1-p)}{[(1-p)+pk]}\left(1-\frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial X}\right)\\ &= -p\left(\gamma-\frac{(\gamma-\beta)(1-p)}{[(1-p)+pk]}\right)\left(1-\frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial X}\right)\end{aligned}$$

Since $1-\frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial X} \geq 0$ for $X^{NC}(\theta) \leq X^{FB}(\theta)$, the sign of $\frac{\partial J^B\left(X^{NC}(\theta),\theta\right)}{\partial X}$ depends on $-p\left(\gamma-\frac{(\gamma-\beta)(1-p)}{[(1-p)+pk]}\right)$. We easily see that $\gamma \geq \frac{(\gamma-\beta)(1-p)}{[(1-p)+pk]} \Leftrightarrow \frac{(1-p)+pk}{(1-p)} \geq \frac{\gamma-\beta}{\gamma} \Leftrightarrow 1+\frac{pk}{1-p} \geq 1-\frac{\beta}{\gamma}$ holds for any $0 \leq p, k \leq 1$, and $\beta, \gamma \geq 0$. Then, since $\frac{\partial J^B\left(X^{NC},\theta\right)}{\partial X} \leq 0$ evaluated at $X = X^{NC}(\theta)$, $X^{NC}(\theta)$ cannot be optimal for the behavioral regimes. A marginal decrease in $X(\theta)$ from $X^{NC}(\theta)$ would increase the virtual surplus $J^B\left(X^{NC},\theta\right)$ of the behavioral regime. Thus, we have $X^B(\theta) \leq X^{NC}(\theta)$[17] ■

[16]The principal can design the optimal output rule $X^B(\cdot)$ to modify shading behaviors by controlling the potential for aggrievement, e.g. information rent $U(\theta)$. In that sense, our framework of shading model is similar to the idea of efficient organization design which counters "influence activities" by Milgrom (1988). A difference is that influence activities are made *before* an important decision making, while shading behaviors are made *after* an important and aggrieving decision making.

[17]This result can be obtained also from the comparison in virtual marginal cost between two regimes: No Commitment (NC) and Behavioral (B) regimes.

Theoretical Intuition

The supervisor's reward is

$$W_s(\theta) = kU(\theta) + (\beta - \gamma) U(\theta) - \gamma \left\{ \left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta \right) \right] - \left[X(\theta) - C(X(\theta), \theta) \right] \right\}$$

First, when the output $X(\theta)$ increases marginally, the information rent $U(\theta)$ goes up.

Next, since $X(\theta) - C(X(\theta), \theta)$ goes up for $X(\theta) \leq X^{FB}(\theta)$, the potential aggrievement $\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta \right) \right] - \left[X(\theta) - C(X(\theta), \theta) \right]$ decreases, and the shading threat goes down $\gamma \left\{ \left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta \right) \right] - \left[X(\theta) - C(X(\theta), \theta) \right] \right\} \Downarrow$. These two effects will increase the supervisor's wage $W_s(\theta)$ discretely, which generates **a first-order loss**. Though the increase in $X(\theta)$ generates **a second-order gain** through the change of optimal solution, the principal's profit will go down totally(due to the **first-order loss vs. second-order gain**). Thus, the optimal solution with the behavioral supervisor $X^B(\theta)$ will fall below the No-commitment (no behavioral supervisor) solution $X^{NC}(\theta)$.

Now, we can perform a comparative statics on the optimal solution on the optimal solution $X^B(\theta)$

Corollary 7.1 *The optimal solution with behavioral supervisor $X^B(\theta)$ is nonincreasing in both parameter β (the degree of shading strength by the agent), and γ (the degree of shading strength by the principal)*[18]

Proof: From $(**)$, the derivative $J^B_X(X, \theta)$ is nonincreasing in β(behavioral elements). That is,

$$J^B_{X\beta}(X, \theta) = \frac{p}{h(\theta)} \frac{\partial^2 C(X, \theta)}{\partial X \partial \theta} \leq 0$$

Hence, the optimal solution with behavioral supervisor $X^B(\theta)$ is nonincreasing in β. Further,

$$J^B_{X\gamma}(X, \theta) = -p \left\{ \left[1 - \frac{\partial C(X, \theta)}{\partial X} \right] + \frac{1}{h(\theta)} \frac{\partial^2 C(X, \theta)}{\partial X \partial \theta} \right\}$$

We already know that $\frac{\partial J(X, \theta)}{\partial X} = \left[1 - \frac{\partial C(X, \theta)}{\partial X} \right] + \frac{1}{h(\theta)} \frac{\partial^2 C(X, \theta)}{\partial X \partial \theta} = 0$ at $X = X(\theta)$

Then, since $X^B(\theta) \leq X(\theta)$ from proposition 7, we have $\frac{\partial J(X, \theta)}{\partial X} \geq 0$ at $X = X^B(\theta)$. Hence, we have $J^B_{X\gamma}(X, \theta) \leq 0$, which means that the optimal solution $X^B(\theta)$ is nonincreasing in γ. ■

Theoretical Intuition

The intuition is very close to the former argument. The supervisor's reward is

$$W_s(\theta) = kU(\theta) + (\beta - \gamma) U(\theta) - \gamma \left\{ \left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta \right) \right] - \left[X(\theta) - C(X(\theta), \theta) \right] \right\}$$

Now, suppose that γ increases. If the optimal solution $X(\theta)$ decreases marginally, the information rent $U(\theta)$ goes down, and $X(\theta) - C(X(\theta), \theta)$ also goes down for $X(\theta) \leq X^{FB}(\theta)$. Therefore, the shading threat $\gamma \left\{ \left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta \right) \right] - \left[X(\theta) - C(X(\theta), \theta) \right] \right\}$ goes up. These two effects will decrease the supervisor's wage $W_s(\theta)$ discretely, which generates **a first-order gain**. Though the decrease in $X(\theta)$ generates **a second-order loss** through the change of optimal solution, the principal's profit will goes up totally(due to the **first-order gain vs. second-order loss**). Thus, the optimal solution with the behavioral supervisor $X^B(\theta)$ will decrease as the shading parameters β, γ increase.

[18] Comparative statics on p, k is essentially the same as proposition 5.

Proposition 8.1 *The principal's equilibrium payoff can increase more likely in the regime (B) with behavioral supervisor, in comparison with the regime (NC) without behavioral supervisor, when the shading strength γ by the principal is greater than the shading strength β by the agent, i.e. $\gamma \geq \beta$.*

Proof:

First, the principal's virtual surplus for type θ in the regime (NC) is

$$\begin{aligned}
&p\left[X^{FB}(\theta) - C\left(X^{FB}(\theta),\theta\right) - kU(\theta)\right] + (1-p)\left(X(\theta) - C\left(X(\theta),\theta\right) - U(\theta)\right) \\
&= (1-p)\left(X(\theta) - C\left(X(\theta),\theta\right)\right) + p\left[X^{FB}(\theta) - C\left(X^{FB}(\theta),\theta\right)\right] - \left[(1-p) + pk\right]U(\theta)
\end{aligned}$$

Hence the maximized expected virtual surplus in the regime (NC) is, by using the lemma 3,

$$\begin{aligned}
(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{NC}(\theta) - C\left(X^{NC}(\theta),\theta\right)\right]f(\theta)\,d\theta + p\underbrace{\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta) - C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\text{First Best Expected Total Surplus}} \\
+\left[(1-p) + pk\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta
\end{aligned}$$

Next, the principal's virtual surplus for type θ in the regime (B) is

$$p\left[X^{FB}(\theta) - C\left(X^{FB}(\theta),\theta\right) - W_S(\theta)\right] + (1-p)\left(X(\theta) - C\left(X(\theta),\theta\right) - U(\theta)\right)$$

By remembering the following coalition-proof constraint with behavioral supervisor

$$W_S(\theta) - \underbrace{\beta U(\theta)}_{\substack{\text{Shading Loss}\\ \text{by the Agent}}} \geq kU(\theta) - \underbrace{\gamma\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta),\theta\right)\right] - \left(X(\theta) - C\left(X(\theta),\theta\right) - U(\theta)\right)\right\}}_{\text{Shading Loss by the Principal}}$$

the virtual surplus for type θ in the regime (B) is transformed as follows.

$$\begin{aligned}
&(1-(1+\gamma)p)\left(X(\theta) - C\left(X(\theta),\theta\right)\right) - \left[(1-p) + p\left\{k + (\beta-\gamma)\right\}\right]U(\theta) + (1+\gamma)p\left[X^{FB}(\theta) - C\left(X^{FB}(\theta),\theta\right)\right] \\
&= (1-p)\left(X(\theta) - C\left(X(\theta),\theta\right)\right) - \left[(1-p) + pk\right]U(\theta) + p\left[X^{FB}(\theta) - C\left(X^{FB}(\theta),\theta\right)\right] \\
&\quad + p\gamma\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta),\theta\right)\right] - \left[X(\theta) - C\left(X(\theta),\theta\right)\right]\right\} - p(\beta-\gamma)U(\theta)
\end{aligned}$$

Now, the expected virtual surplus is as follows by the lemma 3.

$$\begin{aligned}
(1-(1+\gamma)p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X(\theta) - C\left(X(\theta),\theta\right)\right]f(\theta)\,d\theta + (1+\gamma)p\underbrace{\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta) - C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\text{First Best Expected Total Surplus}} \\
+\left[(1-p) + p\left(k + (\beta-\gamma)\right)\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}
\end{aligned}$$

Let $X^B_{CP}(\theta)$ be the optimal output rule over the maximization problem

$$\begin{aligned}
\max_{X(\cdot)}(1-(1+\gamma)p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X(\theta) - C\left(X(\theta),\theta\right)\right]f(\theta)\,d\theta + (1+\gamma)p\underbrace{\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta) - C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\text{First Best Expected Total Surplus}} \\
+\left[(1-p) + p\left(k + (\beta-\gamma)\right)\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}
\end{aligned}$$

Then, the maximized expected virtual surplus in the regime (B) is transfotormed as follows.

$$(1-(1+\gamma)p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X_{CP}^{B}(\theta)-C\left(X_{CP}^{B}(\theta),\theta\right)\right]f(\theta)\,d\theta+(1+\gamma)p\underbrace{\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\text{First Best Expected Total Surplus}}$$
$$+\left[(1-p)+p(k+(\beta-\gamma))\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X_{CP}^{B}(\theta),\theta\right)}{\partial\theta}$$

$$=(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X_{CP}^{B}(\theta)-C\left(X_{CP}^{B}(\theta),\theta\right)\right]f(\theta)\,d\theta+p\underbrace{\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\text{First Best Expected Total Surplus}}$$
$$+\left[(1-p)+pk\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X_{CP}^{B}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$
$$+p\gamma\left\{\underbrace{\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\text{First Best Expected Total Surplus}}-\int_{\underline{\theta}}^{\bar{\theta}}\left[X_{CP}^{B}(\theta)-C\left(X_{CP}^{B}(\theta),\theta\right)\right]f(\theta)\,d\theta\right\}$$
$$+p(\beta-\gamma)\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X_{CP}^{B}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$

Hence, the condition for the principal's equilibrium profit to increase in the regime (B) relative to the regime (NC) is as follows

$$p\gamma\left\{\underbrace{\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\text{First Best Expected Total Surplus}}-\int_{\underline{\theta}}^{\bar{\theta}}\left[X_{CP}^{B}(\theta)-C\left(X_{CP}^{B}(\theta),\theta\right)\right]f(\theta)\,d\theta\right\}$$
$$+p(\beta-\gamma)\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\underbrace{\frac{\partial C\left(X_{CP}^{B}(\theta),\theta\right)}{\partial\theta}}_{-}f(\theta)\,d\theta$$
$$\geq(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left\{\left[X^{NC}(\theta)-C\left(X^{NC}(\theta),\theta\right)\right]-\left[X_{CP}^{B}(\theta)-C\left(X_{CP}^{B}(\theta),\theta\right)\right]\right\}f(\theta)\,d\theta$$
$$+\left[(1-p)+pk\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\left\{\frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial\theta}-\frac{\partial C\left(X_{CP}^{B}(\theta),\theta\right)}{\partial\theta}\right\}f(\theta)\,d\theta$$

The RHS of the inequality is the payoff difference between $X^{NC}(\theta)$ and $X_{CP}^{B}(\theta)$ coming from the following revealed preference relation:

$$(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{NC}(\theta)-C\left(X^{NC}(\theta),\theta\right)\right]f(\theta)\,d\theta+\left[(1-p)+pk\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X^{NC}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$
$$\geq(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X_{CP}^{B}(\theta)-C\left(X_{CP}^{B}(\theta),\theta\right)\right]f(\theta)\,d\theta+\left[(1-p)+pk\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X_{CP}^{B}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$

The LHS of the inequality is totally the principal's payoff increase through *discretely relaxing* the coalition incentive constraint by the principal's shading threat $\gamma\geq\beta$. That is, the principal can reduce the reward to the supervisor discretely through her shading threat (γ times aggrievement)

to the supervisor, thereby increasing her profit.[19] ■

This proposition implies that under the information structure where collusion (side contracting) between supervisor and agent is observable ex post for the principal but unverifiable, the introduction of behavioral supervisor, together with the fear of being "shaded" by the principal, can *relax* the supervisor's incentive constraint (coalition incentive constraint), thereby can increase the principal's equilibrium profit. This exercise can be viewed as the first one made in the three-tier agency hierarchy framework.

Remark

The supervisor's equilibrium payoff under shading is

$$W_s(\theta) - \beta U(\theta) = kU(\theta) - \gamma\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - \left[X(\theta) - C(X(\theta), \theta) - U(\theta)\right]\right\}$$

Thus, the condition for the supervisor's IR constraint to be satisfied is

$$kU(\theta) - \underbrace{\gamma\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - \left[X(\theta) - C(X(\theta), \theta) - U(\theta)\right]\right\}}_{\text{Shading Loss}} \geq 0, \forall\theta \in \left[\underline{\theta}, \overline{\theta}\right]$$

This requires that the shading by the principal is not too strong. Hence, a necessary condition under which (1) the principal's equilibrium profit more likely increases by the introduction of the behavioral supervisor and (2) his IR constraint also holds is $\beta \leq \gamma \leq \frac{U(\theta)}{(\text{FBprofit})-(\text{SBprofit})}k$, more concretely,

$$\beta \leq \gamma \leq \min_{\theta} \frac{U(\theta)}{\underbrace{\left\{\left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] - \left[X(\theta) - C(X(\theta), \theta) - U(\theta)\right]\right\}}_{\text{Aggrievement}}}k, \forall\theta \in \left[\underline{\theta}, \overline{\theta}\right]$$

Suzuki (2012) derives the fact that the condition under which the principal's equilibrium profit increases by the introduction of the behavioral supervisor and his IR constraint also holds is $\beta \leq \gamma \leq k$ in the two-type model. Hence, we find that the corresponding condition for (β, γ) becomes severer in the continuous-type model.

proposition 8.2 *The principal's equilibrium payoff tends to decrease in the regime (B) with behavioral supervisor, in comparison with the regime (NC) without behavioral supervisor when the shading strength β by the agent is greater than the shading strength γ by the principal, i.e., $\beta \geq \gamma$ That is particularly so when p, γ are smaller.*

Proof: When $\beta \geq \gamma$, the second term of the LHS of the corresponding inequality in Proposition 8.1

$p\underbrace{(\beta - \gamma)}_{\geq 0}\int_{\underline{\theta}}^{\bar{\theta}} \frac{1}{h(\theta)} \underbrace{\frac{\partial C\left(X_{CP}^{B}(\theta), \theta\right)}{\partial\theta}}_{-} f(\theta)\, d\theta \leq 0$. That is, when $\beta \geq \gamma$, the net positive shading cost by the agent must be compensated for the supervisor by the principal. Only the first term of the LHS is positive, which becomes smaller when p, γ are smaller. This makes the inequality more difficult to hold. ■

[19] As an analogy for the moral hazard model with risk averse agent, we can say that the principal can decrease the risk cost (risk compensation) discretely, where the risk cost (risk compensation) corresponds to the shading cost in our paper. The point is that the principal ultimately bears the shading cost for the supervisor in order to satisfy his IR constraint.

6.3 Shading Model: Unobservable Collusion

Now, suppose that the supervisor's signal $s \in \{\theta, \phi\}$ is not observed at all by the principal ex post, that is, the principal cannot know at all ex post whether the supervisor obtained the informative signal (evidence, proof on θ) or not (ϕ), as well as which state θ has occurred. Then, the principal cannot distinguish whether she was aggrieved or whether the supervisor just obtained no informative signal (ϕ). Hence, the principal cannot shade the supervisor. This information structure means that collusion (side contracting) between supervisor and agent is *unobservable*, and thus the shading loss by the principal would be zero due to $\gamma = 0$.

Then, the supervisor's incentive constraint (coalition incentive constraint) is reduced to

$$\underbrace{W_s(\theta)}_{\text{wage payment}} - \underbrace{\beta U(\theta)}_{\text{shading loss}} \geq \underbrace{kU(\theta)}_{\text{side payment}} \Leftrightarrow \underbrace{W_s(\theta)}_{\text{wage payment}} \geq \underbrace{kU(\theta)}_{\text{side payment}} + \underbrace{\beta U(\theta)}_{\text{shading loss}}$$

Hence, shading only by the agent $\beta > 0$ tightens the supervisor's incentive constraint (coalition incentive constraint), and makes it more likely that the supervisor will collude with the agent.

Proposition 9 *Suppose that collusion (side contracting) between supervisor and agent is* ***unobservable*** *ex post for the principal. Then, only agent can shade the supervisor, which corresponds to $\beta > 0, \gamma = 0$. Then, the principal's equilibrium payoff is reduced in the regime with behavioral supervisor, in comparison with that without behavioral supervisor $\beta = \gamma = 0$. That is,* ***"shading" becomes detrimental*** *to organization design.*

Proof: The principal's virtual surplus is written as follows.

$$J^B(X,\theta) = \underbrace{(1-p)(X(\theta) - C(X(\theta),\theta)) + \frac{[(1-p)+pk]}{h(\theta)}\frac{\partial C(X(\theta),\theta)}{\partial \theta}}_{\text{(Standard) Virtual Surplus Surplus}} + \underbrace{\frac{p\beta}{h(\theta)}}_{\geq 0}\underbrace{\frac{\partial C(X(\theta),\theta)}{\partial \theta}}_{-}$$

where $\underbrace{\frac{p\beta}{h(\theta)}}_{\geq 0}\underbrace{\frac{\partial C(X(\theta),\theta)}{\partial \theta}}_{-}$ is the increase in Dead Weight Loss (Information Rent) through shading by the type θ agent, which decreases the principal's virtual surplus. This just completes the proof. ■

6.3.1 Collusion-proof Regime vs. Equilibrium Collusion Regime

Now, the principal has two options, one of which is the Collusion-proof Regime, where the principal deters the collusion between the agent θ and the supervisor through the collusion-proof constraint and induces the supervisor's truth telling $r = \theta$, and the other of which is the Equilibrium Collusion Regime, where the principal allows the collusion between them in equilibrium and induces the truthful information from the agent by himself, while the supervisor reports $r = \phi$. Which regime the principal chooses between the Collusion-proof regime and the Equilibrium Collusion regime depends on the condition, which will be analyzed below.

Collusion-proof Regime (CP)

In order to satisfy the collusion-proof constraint, the principal must set the reward for the supervisor

$$\underbrace{W_s(\theta)}_{\text{wage payment}} = \underbrace{kU(\theta)}_{\text{side payment}} + \underbrace{\beta U(\theta)}_{\text{shading loss by agent}} = (k+\beta)U(\theta)$$

Then, the virtual surplus for type θ is

$$(1-p)\left[X(\theta)-C(X(\theta),\theta)-U(\theta)\right]+p\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)-(k+\beta)U(\theta)\right]$$

Hence, the expected virtual surplus for the principal is, due to Lemma3,

$$\begin{aligned}
&\int_{\underline{\theta}}^{\bar{\theta}}\left\{(1-p)\left[X(\theta)-C(X(\theta),\theta)-U(\theta)\right]+p\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)-(k+\beta)U(\theta)\right]\right\}f(\theta)\,d\theta\\
&=\int_{\underline{\theta}}^{\bar{\theta}}\left\{(1-p)\left[X(\theta)-C(X(\theta),\theta)\right]+\left[(1-p)+p(k+\beta)\right]\frac{\partial C(X(\theta),\theta)}{\partial\theta}\frac{1-F(\theta)}{f(\theta)}\right\}f(\theta)\,d\theta\\
&\qquad-\left[(1-p)+p(k+\beta)\right]U(\underline{\theta})+\int_{\underline{\theta}}^{\bar{\theta}}p\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta
\end{aligned}$$

The principal simultaneously maximizes the modified virtual surplus for (almost) every type θ, i.e.

$$\underbrace{(1-p)\left(X(\theta)-C(X(\theta),\theta)\right)+\left[(1-p)+p(k+\beta)\right]\frac{\partial C(X(\theta),\theta)}{\partial\theta}\frac{1-F(\theta)}{f(\theta)}}_{J^{B}_{CP}(X,\theta)}$$

$$\text{Or }\underbrace{\left(X(\theta)-C(X(\theta),\theta)\right)}_{\text{Total Surplus}}+\underbrace{\left[1+\frac{pk}{1-p}\right]\frac{\partial C(X(\theta),\theta)}{\partial\theta}\frac{1}{h(\theta)}}_{\text{Information Rent}}\quad\underbrace{+\frac{p\beta}{1-p}\frac{\partial C(X(\theta),\theta)}{\partial\theta}\frac{1}{h(\theta)}}_{\text{Change in Dead Weight Loss through Shading}}$$

First order condition for the optimality is

$$\frac{\partial J^{B}_{CP}(X,\theta)}{\partial X}=\underbrace{[1-p]\left[1-\frac{\partial C(X,\theta)}{\partial X}\right]+\frac{[(1-p)+pk]}{h(\theta)}\frac{\partial^2 C(X,\theta)}{\partial X\partial\theta}}_{\text{Marginal Virtual Surplus}}+\underbrace{\frac{p\beta}{h(\theta)}\underbrace{\frac{\partial^2 C(X,\theta)}{\partial X\partial\theta}}_{-}}_{\text{Marginal Shading Cost}}=0$$

$$\Leftrightarrow\underbrace{\left[1-\frac{\partial C(X,\theta)}{\partial X}\right]+\frac{\left[1+\frac{pk}{1-p}\right]}{h(\theta)}\frac{\partial^2 C(X,\theta)}{\partial X\partial\theta}}_{\text{Marginal Virtual Surplus}}+\underbrace{\frac{p}{1-p}\frac{1}{h(\theta)}\beta\underbrace{\frac{\partial^2 C(X,\theta)}{\partial X\partial\theta}}_{-}}_{\text{Marginal Shading Cost}}=0$$

Proposition 10 *The optimal solution $X^{B}_{CP}(\theta)$ with behavioral elements under the collusion-proof regime is smaller than the optimal solution $X^{NC}(\theta)$ with no behavioral elements, that is, $X^{B}_{CP}(\theta)\leq X^{NC}(\theta)$*

Proof: See the proof of Proposition7. This is the case where $\beta>0,\gamma=0$

Equilibrium Collusion Regime (EC)

In this regime, when the supervisor obtains the proof on θ with probability p, the principal allows the collusion between the agent θ and the supervisor in equilibrium, which means that the supervisor reports $r=\phi$ and the agent θ self-selects $\{X(\theta),W(\theta)\}$ and obtains the information rent $U(\theta)$. Then, the principal pays the information rent $U(\theta)$ to the agent θ at the unit transfer price 1. Now, the virtual surplus for type θ is

$$\begin{aligned}
&(1-p)\left[X(\theta)-C(X(\theta),\theta)-U(\theta)\right]+p\left[X(\theta)-C(X(\theta),\theta)-U(\theta)\right]\\
&=X(\theta)-C(X(\theta),\theta)-U(\theta)
\end{aligned}$$

Hence, the expected virtual surplus for the principal is, due to Lemma3,

$$\int_{\underline{\theta}}^{\bar{\theta}} \{X(\theta) - C(X(\theta),\theta) - U(\theta)\} f(\theta)\, d\theta$$
$$= \int_{\underline{\theta}}^{\bar{\theta}} \left\{ X(\theta) - C(X(\theta),\theta) + \frac{\partial C(X(\theta),\theta)}{\partial \theta} \frac{1 - F(\theta)}{f(\theta)} \right\} f(\theta)\, d\theta - U(\underline{\theta})$$

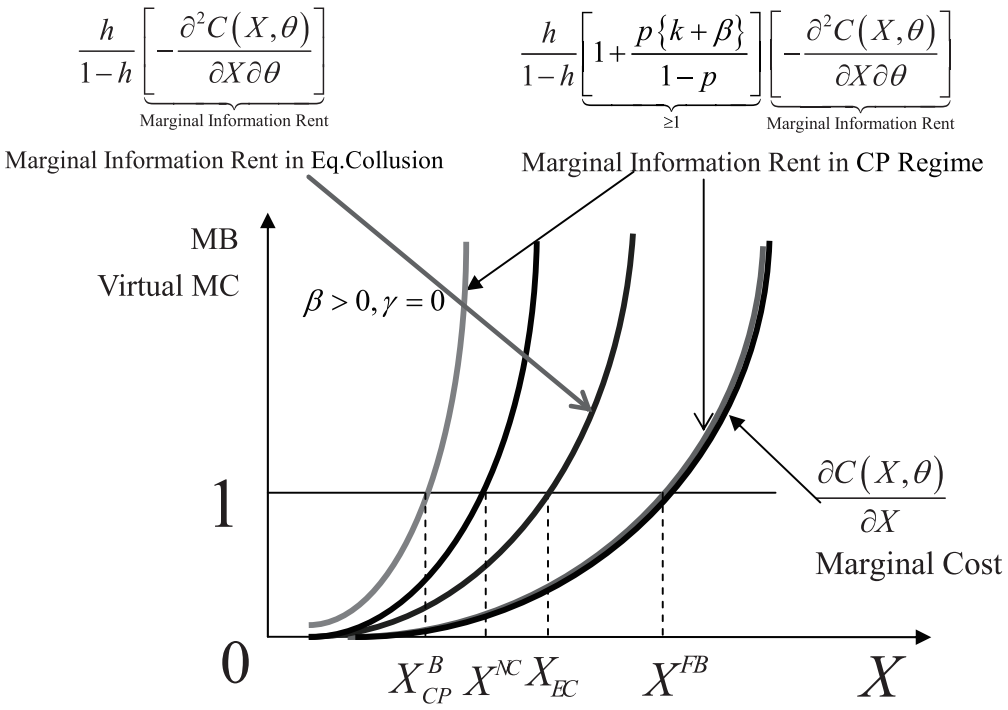
Then, the principal simultaneously maximizes the modified virtual surplus for (almost) every type θ, i.e.

$$\underbrace{(X(\theta) - C(X(\theta),\theta)) + \frac{\partial C(X(\theta),\theta)}{\partial \theta} \frac{1}{h(\theta)}}_{J^B_{EC}(X,\theta)}$$

First order condition for the optimality is

$$\left(1 - \frac{\partial C(X(\theta),\theta)}{\partial X(\theta)}\right) + \frac{\partial^2 C(X(\theta),\theta)}{\partial X \partial \theta} \frac{1}{h(\theta)} = 0$$

Comparing First Order Conditions on marginal incentives in the two regimes (CP and EC), we find that the coefficient of the marginal information rent $1 + \frac{pk}{1-p}$ in the collusion-proof regime (CP) is greater than that 1 in the equilibrium collusion regime (EC), that is, $1 + \frac{p(k+\beta)}{1-p} \geq 1$ for $\forall p, k, \beta \geq 0$. Hence, we have $X^B_{CP}(\theta) \leq X(\theta) = X_{EC}(\theta)$. The below figure represents the determination of equilibrium incentives.



6.3.2 Payoff Comparison between Collusion-proof and Equilibrium Collusion Regimes

We analyze which regime the principal chooses between the Collusion-proof regime (CP) and the Equilibrium Collusion Regime (EC). We compare the payoffs between Collusion-proof vs. Equilibrium Collusion Regimes.

The expected payoff for the principal in the 'Collusion-proof' regime (CP) is

$$(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X_{CP}^{B}(\theta)-C\left(X_{CP}^{B}(\theta),\theta\right)\right]f(\theta)\,d\theta+p\times\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta$$
$$+\left[(1-p)+p(k+\beta)\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X_{CP}^{B}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$

The expected payoff for the principal in the 'Commitment, Pooling' regime (S) is

$$\int_{\theta}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)+\left[(1-p)+p(k+\beta)\right]\frac{\partial C\left(X^{S}(\theta),\theta\right)}{\partial\theta}\frac{1}{h(\theta)}\right]f(\theta)\,d\theta$$

The expected payoff for the principal in the Equilibrium Collusion Regime (EC) is

$$\int_{\underline{\theta}}^{\bar{\theta}}\left[X(\theta)-C\left(X(\theta),\theta\right)+\frac{1}{h(\theta)}\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}\right]f(\theta)\,d\theta$$

We consider the comparison between the three regimes under $Z=0$, that is, the cost of introducing the supervisor (a transaction cost) is zero.

Step1

First, we compare the equilibrium payoffs between the 'Collusion-proof' (CP) regime $\left[X_{CP}^{B}(\theta)\ \text{w.p}\ 1-p, X^{FB}(\theta)\ \text{w.p}\ p\right]$ and the '**Commitment**, Pooling' (S) regime $\left[X^{S}(\theta)\ \text{w.p}\ 1\right]$ By definition, $X_{CP}^{B}(\theta)$ is the optimal output rule over the problem

$$\max_{X(\cdot)}(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X(\theta)-C\left(X(\theta),\theta\right)\right]f(\theta)\,d\theta$$
$$+\left[(1-p)+p(k+\beta)\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$

Then, from the *revealed preference* argument, the following holds.

$$(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X_{CP}^{B}(\theta)-C\left(X_{CP}^{B}(\theta),\theta\right)\right]f(\theta)\,d\theta$$
$$+\left[(1-p)+p(k+\beta)\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X_{CP}^{B}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$
$$\geq(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)\right]f(\theta)\,d\theta$$
$$+\left[(1-p)+p(k+\beta)\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X^{S}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$

The following inequality holds by the same *revealed preference* argument.

$$\underbrace{p\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\text{First Best Expected Total Surplus}}\geq p\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)\right]f(\theta)\,d\theta$$

Hence, we have the following inequality.

$$(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X_{CP}^{B}(\theta)-C\left(X_{CP}^{B}(\theta),\theta\right)\right]f(\theta)\,d\theta+p\underbrace{\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\text{First Best Expected Total Surplus}}$$

$$+\left[(1-p)+p(k+\beta)\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X_{CP}^{B}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$

$$\geq(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)\right]f(\theta)\,d\theta+p\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)\right]f(\theta)\,d\theta$$

$$+\left[(1-p)+p(k+\beta)\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X^{S}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$

$$=\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)\right]f(\theta)\,d\theta+\left[(1-p)+p(k+\beta)\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X^{S}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$

Thus, 'Collusion-proof' regime (CP) $\left[X_{CP}^{B}(\theta)\ \text{w.p}\ 1-p, X^{FB}(\theta)\ \text{w.p}\ p\right]$ is payoff dominant over the 'Pooling, Commitment' (S) regime $\left[X^{S}(\theta)\ \text{w.p}\ 1\right]$ for the principal.

Step2

Next, we compare the equilibrium payoffs between the 'the 'Commitment, Pooling' regime (S) $X^{S}(\theta)$ and Equilibrium Collusion Regime (EC) $X_{EC}(\theta)$
By definition, $X_{EC}(\theta)$ is the optimal output rule over the problem

$$\max_{X(\cdot)}\int_{\underline{\theta}}^{\bar{\theta}}\left[X(\theta)-C\left(X(\theta),\theta\right)+\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}\frac{1}{h(\theta)}\right]f(\theta)\,d\theta$$

By definition, $X^{S}(\theta)$ is the optimal output rule over the problem

$$\max_{X(\cdot)}\int_{\underline{\theta}}^{\bar{\theta}}\left[X(\theta)-C\left(X(\theta),\theta\right)+\left[(1-p)+p(k+\beta)\right]\frac{\partial C\left(X(\theta),\theta\right)}{\partial\theta}\frac{1}{h(\theta)}\right]f(\theta)\,d\theta$$

Then, from the *revealed preference* argument, the following holds.

$$\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)+\left[(1-p)+p(k+\beta)\right]\frac{\partial C\left(X^{S}(\theta),\theta\right)}{\partial\theta}\frac{1}{h(\theta)}\right]f(\theta)\,d\theta$$

$$\begin{matrix}\geq\\ \leq\end{matrix}\int_{\underline{\theta}}^{\bar{\theta}}\left[X_{EC}(\theta)-C\left(X_{EC}(\theta),\theta\right)+\frac{\partial C\left(X_{EC}(\theta),\theta\right)}{\partial\theta}\frac{1}{h(\theta)}\right]f(\theta)\,d\theta\ :\text{Equilibrium Collusion Payoff}$$

We have the following result on the marginal incentives (outputs), by comparing the coefficients of the information rents between Two Regimes

$$X_{EC}(\theta)=X(\theta)\leq X^{S}(\theta)\ \text{ if }\ (1-p)+p(k+\beta)\leq 1\Leftrightarrow\beta\leq 1-k$$
$$X_{EC}(\theta)=X(\theta)\geq X^{S}(\theta)\ \text{ if }\ (1-p)+p(k+\beta)\geq 1\Leftrightarrow\beta\geq 1-k$$

1. When $X_{EC}(\theta)=X(\theta)\leq X^{S}(\theta)\ \text{ if }\ (1-p)+p(k+\beta)\leq 1\Leftrightarrow(0\leq)\,\beta\leq 1-k$

Combining the results of the above two steps, we obtain

$$(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X_{CP}^{B}(\theta)-C\left(X_{CP}^{B}(\theta),\theta\right)\right]f(\theta)\,d\theta+p\underbrace{\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\text{First Best Expected Total Surplus}}$$

$$+\left[(1-p)+p(k+\beta)\right]\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X_{CP}^{B}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta$$

$$\geq\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{S}(\theta)-C\left(X^{S}(\theta),\theta\right)+\left[(1-p)+p(k+\beta)\right]\frac{\partial C\left(X^{S}(\theta),\theta\right)}{\partial\theta}\frac{1}{h(\theta)}\right]f(\theta)\,d\theta$$

$$\geq\int_{\underline{\theta}}^{\bar{\theta}}\left[X_{EC}(\theta)-C\left(X_{EC}(\theta),\theta\right)+\frac{1}{h(\theta)}\frac{\partial C\left(X_{EC}(\theta),\theta\right)}{\partial\theta}\right]f(\theta)\,d\theta\quad\text{Eq.Collusion Payoff}$$

The principal prefers the Collusion-proof regime (CP) to the Equilibrium Collusion regimes (EC) in terms of her expected payoff when the shading parameter $\beta\leq 1-k$, which is a sufficient condition for the Collusion-proof regime (CP) to be optimal. In this case, the **"collusion-proof principle"** still holds.

2. When $X_{EC}(\theta)=X(\theta)\geq X^{S}(\theta)$ if $(1-p)+p(k+\beta)\geq 1\Leftrightarrow\beta\geq 1-k$
The optimal solution $X_{CP}^{B}(\theta)$ is determined by

$$\frac{\partial J_{CP}^{B}(X,\theta)}{\partial X}=\underbrace{[1-p]\left[1-\frac{\partial C(X,\theta)}{\partial X}\right]+\frac{[(1-p)+pk]}{h(\theta)}\frac{\partial^{2}C(X,\theta)}{\partial X\partial\theta}}_{\text{Marginal Virtual Surplus}}+\underbrace{\frac{p\beta}{h(\theta)}\underbrace{\frac{\partial^{2}C(X,\theta)}{\partial X\partial\theta}}_{-}}_{\text{Marginal Shading Cost}}=0$$

$$\underbrace{\left[1-\frac{\partial C(X,\theta)}{\partial X}\right]+\frac{\left[1+\frac{pk}{1-p}\right]}{h(\theta)}\frac{\partial^{2}C(X,\theta)}{\partial X\partial\theta}}_{\text{Marginal Virtual Surplus}}+\underbrace{\frac{p}{1-p}\frac{1}{h(\theta)}\beta\underbrace{\frac{\partial^{2}C(X,\theta)}{\partial X\partial\theta}}_{-}}_{\text{Marginal Shading Cost}}=0$$

Then, as the shading parameter β becomes larger (as $\beta\to+\infty$), the optimal output rule goes to zero, $X_{CP}^{B}(\theta)\to 0$ for all $\theta\in\left[\underline{\theta},\bar{\theta}\right]$[20] Then, the potential aggrievement (information rent) for the agent also goes to zero, $U(\theta)\to 0$ for all $\theta\in\left[\underline{\theta},\bar{\theta}\right]$ Hence, the equilibrium payoff of the Collusion-proof regime with Behavioral supervisor goes to

$$\underbrace{(1-p)\int_{\underline{\theta}}^{\bar{\theta}}\left[X_{CP}^{B}(\theta)-C\left(X_{CP}^{B}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\to 0}+\underbrace{p\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\text{First Best Expected Total Surplus}}$$

$$+\left[(1-p)+p(k+\beta)\right]\underbrace{\int_{\underline{\theta}}^{\bar{\theta}}\frac{1}{h(\theta)}\frac{\partial C\left(X_{CP}^{B}(\theta),\theta\right)}{\partial\theta}f(\theta)\,d\theta}_{\to 0}$$

$$\to p\underbrace{\int_{\underline{\theta}}^{\bar{\theta}}\left[X^{FB}(\theta)-C\left(X^{FB}(\theta),\theta\right)\right]f(\theta)\,d\theta}_{\text{First Best Expected Total Surplus}}$$

[20] An example of cost function is $C(X,\theta)=\left(\frac{X}{\theta}\right)^{\alpha}$, $\frac{\partial C}{\partial X}=\frac{\alpha X^{\alpha-1}}{\theta^{\alpha}}$, $\alpha\geq 2$

On the other hand, the payoff of the Equilibrium Collusion regime is independent of β, p

$$\int_{\underline{\theta}}^{\bar{\theta}} \left[X_{EC}(\theta) - C\left(X_{EC}(\theta), \theta\right) - U(\theta)\right] f(\theta)\, d\theta$$
$$= \int_{\underline{\theta}}^{\bar{\theta}} \left[X_{EC}(\theta) - C\left(X_{EC}(\theta), \theta\right) + \frac{\partial C\left(X_{EC}(\theta), \theta\right)}{\partial \theta} \frac{1}{h(\theta)}\right] f(\theta)\, d\theta - U(\underline{\theta})$$

Therefore, which payoff is greater between (CP) and (EC) at $\beta \to +\infty$ depends on the relative size of

$$\underbrace{p \int_{\underline{\theta}}^{\bar{\theta}} \left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] f(\theta)\, d\theta}_{\text{First Best Expected Total Surplus}} \begin{matrix} \geq \\ \leq \end{matrix} \underbrace{\int_{\underline{\theta}}^{\bar{\theta}} \left[X_{EC}(\theta) - C\left(X_{EC}(\theta), \theta\right) - U(\theta)\right] f(\theta)\, d\theta}_{\text{Eq.Collusion Payoff}}$$

Case1

$$\text{If } \underbrace{p \int_{\underline{\theta}}^{\bar{\theta}} \left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] f(\theta)\, d\theta}_{\text{First Best Expected Total Surplus}} \leq \underbrace{\int_{\underline{\theta}}^{\bar{\theta}} \left[X_{EC}(\theta) - C\left(X_{EC}(\theta), \theta\right) - U(\theta)\right] f(\theta)\, d\theta}_{\text{Eq.Collusion Payoff}}$$

$$\Leftrightarrow p \leq \frac{\underbrace{\int_{\underline{\theta}}^{\bar{\theta}} \left[X_{EC}(\theta) - C\left(X_{EC}(\theta), \theta\right) - U(\theta)\right] f(\theta)\, d\theta}_{\text{Eq.Collusion Payoff=Second Best Payoff}}}{\underbrace{\int_{\underline{\theta}}^{\bar{\theta}} \left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] f(\theta)\, d\theta}_{\text{First Best Expected Total Surplus}}} = p^*$$

Then, there exists aβ^*such that for $\beta \geq \beta^* (> 1 - k)$ "Equilibrium Collusion" Payoff dominates "Collusion-proof" payoff, that is, Equilibrium Collusion is optimally chosen by the principal.

Case2

$$\text{If } \underbrace{p \int_{\underline{\theta}}^{\bar{\theta}} \left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] f(\theta)\, d\theta}_{\text{First Best Expected Total Surplus}} \geq \underbrace{\int_{\underline{\theta}}^{\bar{\theta}} \left[X_{EC}(\theta) - C\left(X_{EC}(\theta), \theta\right) - U(\theta)\right] f(\theta)\, d\theta}_{\text{Eq.Collusion Payoff}}$$

$$\Leftrightarrow p \geq \frac{\underbrace{\int_{\underline{\theta}}^{\bar{\theta}} \left[X_{EC}(\theta) - C\left(X_{EC}(\theta), \theta\right) - U(\theta)\right] f(\theta)\, d\theta}_{\text{Eq.Collusion Payoff=Second Best Payoff}}}{\underbrace{\int_{\underline{\theta}}^{\bar{\theta}} \left[X^{FB}(\theta) - C\left(X^{FB}(\theta), \theta\right)\right] f(\theta)\, d\theta}_{\text{First Best Expected Total Surplus}}} = p^*$$

In this case, Equilibrium Collusion is not optimal even for $\beta \to +\infty$, but Collusion-proof regime is optimally chosen. (A clear example is $p \to 1$). The point is that Shut-down is endogenously chosen in the states of (θ, ϕ), that is, the optimal output rule goes to zero, $X_{CP}^{B}(\theta) \to 0$ for all $\theta \in \left[\underline{\theta}, \bar{\theta}\right]$, and the potential aggrievement (information rent) also goes to zero, $U(\theta) \to 0$ for all $\theta \in \left[\underline{\theta}, \bar{\theta}\right]$.

As p becomes smaller, the states of (θ, ϕ) with probability $1-p$ increase. Then the principal cannot neglect her decision $X_{CP}^{B}(\theta)$ any more in the supervisory no information state ϕ, in the form of $X_{CP}^{B}(\theta) \to 0$. Nonetheless, the cost of collusion-proof constraint, or the shading cost which the principal will eventually bear becomes very large. Since it is too costly for the principal, the principal optimally switches to the Equilibrium Collusion Regime which induces $X_{EC}(\theta)$ in both states. Thus, we have the proposition.

Proposition 11 ***Collusion-proof vs. Equilibrium Collusion***

1. The principal prefers the Collusion-proof regime (CP) to the Equilibrium Collusion regime (EC) in terms of her expected payoff when $\beta \leq 1-k$.

2. The principal prefers the Collusion-proof regime (CP) to the Equilibrium Collusion regime (EC) for all $\beta \geq 0$ when $p \geq p^*$. Especially, as $\beta \to +\infty$, the optimal Collusion-proof contract has a property of "Shut Down" in all states of supervisory no information (θ, ϕ)

3. The principal prefers the Equilibrium Collusion regime (EC) to the Collusion-proof regime (CP) when $\beta > \beta^*$ and $p < p^*$

Rationale

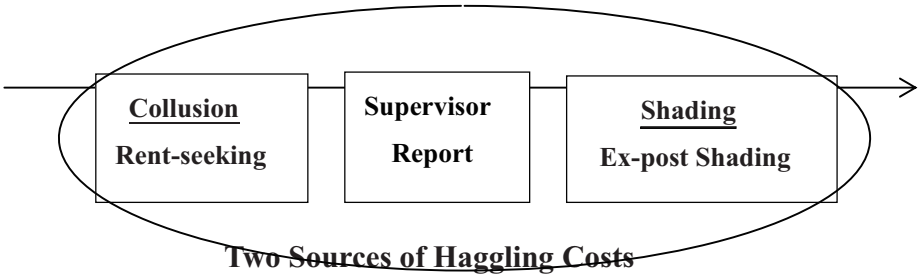
As the degree of shading β ("threat" by the agent) increases, the incentive for collusion between the agent of type θ and the supervisor increases. Thereby, it becomes more costly for the principal to impose collusion-proof schemes and deter collusion, and to induce the truth telling from the supervisor. Theoretically, this implies that as the set of collusion-proof, Incentive compatible schemes becomes smaller, the attainable efficiency becomes lower.

Then, it may be better for the principal to allow collusion between the agent of type θ and the supervisor, and then attain the higher efficiency through discretely reducing the ex-post aggrievement and shading by the agent of type θ.

This is a new idea in the Collusion literature a la Tirole (1986, 1990) in that the increase in shading pressure (**behavioral element**) strengthens the incentive for collusion, thereby makes it difficult to implement the collusion-proof (Supervisor's truth telling) incentive schemes, which leads to the Equilibrium Collusion. The principal allows collusion between the high productivity agent and the supervisor in equilibrium, and the supervisor reports $r = \phi$ ("I did not observe any information") and the high productivity type reveals his type information θ by self-selecting $\{X(\theta), W(\theta)\}$ and obtains the information rent $U(\theta)$

Interpretation of the Result

We can interpret the results from the viewpoint of Transaction Cost Economics a la Coase (1937) and Williamson (1975). Let us assume that "Haggling Cost" in Transaction Cost Economics has two sources: Cost of Rent-seeking or Influence activity which accompanies Ex-ante Collusion *before* the supervisor's decision making (report), and Cost of Ex-post Shading which results from Ex-post aggrievement and shading behavior *after* the supervisor's decision making (report), as the below figure suggests.



Two Sources of Haggling Costs

(CP) Collusion –Proof but Ex-post Shading

(EC) Equilibrium Collusion but Ex-post No Shading

In the Collusion-proof regime, the principal deters collusion through collusion-proof schemes, and thus no ex-ante collusion occurs. But, ex-post shading by the agent of type θ occurs, since the agent of type θ expected to obtain the best reward for him, that is, the information rent $U(\theta)$, but was aggrieved to have lost it due to the supervisory report $r = \theta$. Therefore, the agent of type θ shades the supervisor by the shading parameter β times the aggrievement level $U(\theta)$. In this case, we have ex-ante no collusion costs but ex-post shading costs.

On the other hand, in the Equilibrium Collusion Regime, the principal allows ex-ante collusion between the agent of type θ and the supervisor, which may be costly by itself but does not generate any aggrievement for the agent of type θ, since he can indeed obtain the information rent $U(\theta)$(as his "entitlement"). Hence, he does not shade the supervisor ex-post. In this case, we have ex-ante collusion costs but ex-post no shading costs.

As the degree of shading β increases, the incentive for collusion between the agent of type θ and the supervisor increases. Thereby, it becomes more costly for the principal to impose collusion-proof schemes and deter collusion, and to induce the truth telling from the supervisor. Then, it can be better for the principal to let them collude in equilibrium, and attain the higher efficiency through reducing discretely the ex-post aggrievement and shading by the agent of type θ.

We believe that this is not only a new idea in the Collusion literature a la Tirole (1986, 1990) in that the increase in shading pressure (behavioral element) strengthens the incentive for collusion, thereby makes it difficult to implement the collusion-proof (Supervisor's truth telling) incentive schemes, which leads to the Equilibrium Collusion, but also gives a micro-foundation (an explicit modeling) for the "Ex-post Haggling Cost" in Transaction Cost Economics a la Williamson (1975).

7 Conclusion

Recently, auditing to meet the needs of corporate governance has rapidly been increasing in importance in Japan, as well as in the U.S. and other Western countries. Given this trend, we were motivated to build a theoretical model to examine how supervision (auditing) could be utilized in order to enhance the effectiveness of corporate governance and to deter collusive supervision (auditing). We introduced the outcomes of "Monotone Comparative Statics" à la Topkis (1978) and Edlin and Shannon (1998), and Milgrom and Segal (2002)'s generalized envelope theorem into a familiar screening (self selection) model with a continuum of types, and constructed a three-tier agency model with a mathematically tractable structure. This should be an advantage in modeling in comparison with the collusion literature e.g., Kofman and Lawarree (1993)'s auditing application of the three-tier agency model à la Tirole (1986, 1992). The basic trade-off involved in adding the auditor (supervisor) into the hierarchy is the benefit obtained by the discrete reduction in information rent and the improvement of marginal incentives (outputs) versus the resource cost of the auditor (supervisor). This bottom line was consistently preserved through the model.

Throughout the basic model of the paper we considered a situation where the principal can *commit* to a collusion-proof contract, that is, 'full commitment'. We used the revelation principle,

solving programs in which the principal always prevents collusion between the auditor (supervisor) and the manager (agent). In the optimal contract, nobody colludes: this is called the collusion-proof principle. However, this does not imply an obvious inconsistency with reality, where collusive supervision (auditing) often makes headlines, as stated in the introduction. The revelation principle and the collusion-proof principle are *solution techniques* which facilitate characterization of the optimal contract.[21]

We then showed as an extension what happens when the principal cannot fully commit to the mechanism and the renegotiation is unavoidable.When the principal commits herself to the reward scheme for the supervisor, but does not commit to the one for the agent, she is tempted to modify the initial contract (or the outcome) unilaterally, using the information revealed by the supervisor. The situation is similar to the ratchet problem and the renegotiation problem caused by lack of the principal's commitment in the dynamics of incentive contracts, studied early by Laffont-Tirole(1988), and Dewatripont (1988) etc. If the agent anticipates such a modification, since he can benefit from a failure by the supervisor to report his type truthfully, he will offer the supervisor the transfer (side payment)equivalent to his information rent. Thus, the principal must pay the supervisor in opposition to the collusive offer by the agent.Thus, the principal can strictly improve his payoff ex-post, but must bear the ex-ante incentive cost.

We compared the payoffs for the principal between three regimes, that is, the 'No-Commitment' regime (NC), the 'Principal-Supervisor-Agent' Collusion-proof, Commitment regime (S), and the 'No-Supervisor' (standard second best) regime (NS). Under the assumption that the cost of introducing the supervisor (a transaction cost) is zero, the principal prefers the 'No Commitment' regime (NC) most in terms of her expected payoff. Intuitively, since the principal does not commit herself not to adjust the output (quantity) rule as well as the price rule in the "No Commitment" regime, she optimally adjusts both of them and tries to design a "more state-contingent" contract through more efficient use of supervisor's report, which is more efficient than the pooling output (quantity) rule in the "Collusion-proof, Commitment" regime. This may be consistent with a situation in the top management organization in corporate governance, where in companies with committees, the committee (the supervisor in our model) accurately grasps the state (type information) of the agent (operating officer) with a high probability and imposes the first best scheme for the agent. Under the positive cost of introducing the supervisor (a transaction cost), which regime the principal prefers most depends on whether the comparative (relative) advantage of the three-tier structure ('No Commitment' regime (NC)) over the two-tier structure ('No Supervisor regimes (NS)) is greater or not than the cost of introducing the supervisor.

Then, we incorporated behavioral elements ala Fahr and Schmidt (1999) into the model, and examined their effects on the optimal solution in the principal-supervisor-agent hidden information model with collusion. We found that these behavioral elements could change the monetary reward for inducing the true information, and so the virtual surplus for each type was also altered through the change in the information rent (an incentive cost for inducing a truthful information revelation). Thus, the optimal solution with behavioral elements could be different from the one with no behavioral elements. More concretely, we introduced the recent behavioral contract theory idea, "shading" (Hart and Moore (2008)) into the collusion model a la Tirole (1986, 1992). By combining the two ideas, i.e., *collusion* and *shading*, we enriched the existing collusion model and obtained a new result on *Collusion-proof* vs. *Equilibrium Collusion* in that the increase in shading pressure (behavioral element) strengthened the incentive for collusion, thereby made it difficult to implement the collusion-proof (Supervisor's truth telling) incentive schemes, which led to the Equilibrium Collusion. Then, the collusion-proof principle does not hold any more. Further, by considering shading as a component of ex-post haggling (addressed by Coase (1937) and Williamson (1975), more generally, Transaction Cost Economics (TCE)), we could give a micro foundation (an

[21]Indeed, if we consider an *incomplete* grand contract situation like Tirole (1992), Laffont and Tirole (1991), and Suzuki (2007), *equilibrium collusion* can improve efficiency. Such models indeed could be usefully applied, in such fields as political economy, regulation, and authority delegation in organizations.

explicit modeling) to ex-post adaptation costs, where we viewed rent-seeking associated with collusive behavior and ex-post haggling generated from aggrievement and shading as the two sources of the costs.

In summary, we applied the Monotone Comparative Statics method and the First Order (Mirrlees) Approach to the continuous-type, three-tier agency model with hidden information and collusion a la Tirole (1986,1992), thereby providing a framework that can address the issues treated in the existing literature *in a much simpler fashion*. We characterized the nature of equilibrium contract that can be implemented under the possibility of collusion, and obtained a general comparison result on the organization structures. Then, we introduced the recent behavioral contract theory idea, "shading" into the model. By combining the two ideas, i.e., *collusion* and *shading*, we could not only enrich the existing collusion model, including a new result on the choice of Collusion-proof vs. Equilibrium Collusion regimes, but also give a micro foundation to ex-post haggling costs, addressed by Transaction Cost Economics. We believe that our model can help a deep understanding of resource allocation and decision process in internal organizations of large firms.

APPENDICES

Appendix 1 Proof of Lemma 2

Proof: The "⇒" part was established above. It remains to show that local IC and monotonicity imply that $U\left(\widehat{\theta},\theta\right) \leq U(\theta)$ for all $\widehat{\theta}$, θ. For $\widehat{\theta} > \theta$, we can write

$$\begin{aligned}
U\left(\widehat{\theta},\theta\right) - U(\theta) &= W\left(\widehat{\theta}\right) - C\left(X\left(\widehat{\theta}\right),\theta\right) - U(\theta) \\
&= U\left(\widehat{\theta}\right) + C\left(X\left(\widehat{\theta}\right),\widehat{\theta}\right) - C\left(X\left(\widehat{\theta}\right),\theta\right) - U(\theta) \\
&= \left[C\left(X\left(\widehat{\theta}\right),\widehat{\theta}\right) - C\left(X\left(\widehat{\theta}\right),\theta\right)\right] + \left[U\left(\widehat{\theta}\right) - U(\theta)\right] \\
&= \int_{\theta}^{\widehat{\theta}} \frac{\partial C\left(X\left(\widehat{\theta}\right),\tau\right)}{\partial \tau} d\tau + \int_{\theta}^{\widehat{\theta}} \left[-\frac{\partial C\left(X(\tau),\tau\right)}{\partial \tau}\right] d\tau \qquad (*) \\
&= \int_{\theta}^{\widehat{\theta}} \left[\frac{\partial C\left(X\left(\widehat{\theta}\right),\tau\right)}{\partial \tau} - \frac{\partial C\left(X(\tau),\tau\right)}{\partial \tau}\right] d\tau \leq 0 \qquad (**)
\end{aligned}$$

In $(*)$, we used the following fact by **(ICFOC)** and Envelope theorem

$$U\left(\widehat{\theta}\right) - U(\theta) = \int_{\theta}^{\widehat{\theta}} \frac{dU}{d\tau}(\tau)\, d\tau = \int_{\theta}^{\widehat{\theta}} \frac{\partial C\left(X(\tau),\tau\right)}{\partial \tau} d\tau$$

In $(**)$, the last inequality is obtained by **SCP** and the fact that $X\left(\widehat{\theta}\right) \geq X(\theta)$ by **(M)**. As explained just below the Definition 1, **SCP** implies that the marginal cost of output $\frac{\partial C(X,\theta)}{\partial X}$ is decreasing in type θ in our model. That is $\frac{\partial^2 C(X,\theta)}{\partial X \partial \theta} < 0$. This condition implies that $\frac{\partial C\left(X\left(\widehat{\theta}\right),\theta\right)}{\partial \theta} - \frac{\partial C\left(X(\theta),\theta\right)}{\partial \theta} \leq 0$ for $X\left(\widehat{\theta}\right) \geq X(\theta)$ due to **(M)**. So, we obtain the last inequality. The proof for $\theta > \widehat{\theta}$ is similar. **Q.E.D**

Appendix 2 Proof of Lemma 3

Proof: We transform the *expected information rents* by exploiting "Integration by Parts".

Now, remember that

$$\int_{\underline{\theta}}^{\bar{\theta}} \left[U\left(\underline{\theta}\right) - \int_{\underline{\theta}}^{\theta} \frac{\partial C\left(X\left(\tau\right),\tau\right)}{\partial \tau} d\tau \right] f\left(\theta\right) d\theta = \int_{\underline{\theta}}^{\bar{\theta}} U\left(\theta\right) f\left(\theta\right) d\theta$$

Because $\left[U\left(\theta\right)F\left(\theta\right)\right]' = U\left(\theta\right)f\left(\theta\right) + \underbrace{\frac{dU\left(\theta\right)}{d\theta}}F\left(\theta\right) = U\left(\theta\right)f\left(\theta\right) \quad \underbrace{-\frac{\partial C\left(X\left(\theta\right),\theta\right)}{\partial \theta}}_{\text{(Due to Envelope Theorem)}} \quad F\left(\theta\right)$, and so

$U\left(\theta\right)f\left(\theta\right) = \left[U\left(\theta\right)F\left(\theta\right)\right]' + \frac{\partial C\left(X\left(\theta\right),\theta\right)}{\partial \theta}F\left(\theta\right)$, we have

$$\int_{\underline{\theta}}^{\bar{\theta}} U\left(\theta\right)f\left(\theta\right)d\theta = \left[U\left(\theta\right)F\left(\theta\right)\right]_{\underline{\theta}}^{\bar{\theta}} + \int_{\underline{\theta}}^{\bar{\theta}} \frac{\partial C\left(X\left(\theta\right),\theta\right)}{\partial \theta}F\left(\theta\right)d\theta$$

$$= U\left(\bar{\theta}\right) + \int_{\underline{\theta}}^{\bar{\theta}} \frac{\partial C\left(X\left(\theta\right),\theta\right)}{\partial \theta}F\left(\theta\right)d\theta = U\left(\underline{\theta}\right) - \int_{\underline{\theta}}^{\bar{\theta}} \frac{\partial C\left(X\left(\theta\right),\theta\right)}{\partial \theta}d\theta + \int_{\underline{\theta}}^{\bar{\theta}} \frac{\partial C\left(X\left(\theta\right),\theta\right)}{\partial \theta}F\left(\theta\right)d\theta$$

$$\left(\because U\left(\bar{\theta}\right) = U\left(\underline{\theta}\right) - \int_{\underline{\theta}}^{\bar{\theta}} \frac{\partial C\left(X\left(\theta\right),\theta\right)}{\partial \theta}d\theta\right)$$

$$= U\left(\underline{\theta}\right) - \int_{\underline{\theta}}^{\bar{\theta}} \frac{\partial C\left(X\left(\theta\right),\theta\right)}{\partial \theta}\left(1 - F\left(\theta\right)\right)d\theta = U\left(\underline{\theta}\right) - \int_{\underline{\theta}}^{\bar{\theta}} \frac{\partial C\left(X\left(\theta\right),\theta\right)}{\partial \theta}\frac{1 - F\left(\theta\right)}{f\left(\theta\right)}f\left(\theta\right)d\theta \ \textbf{Q.E.D}$$

REFERENCES

1. Bolton, P and M. Dewatripont (2005) *Contract Theory* MIT Press
2. Coase, R. (1937). "The Nature of the Firm", *Economica*, N.S., 4(16), pp. 386-405.
3. Dewatripont, M (1989) "Renegotiation and Information Revelation over Time: The Case of Optimal Labor Contracts", *Quarterly Journal of Economics*, Vol. 104. No3, August, 589-619.
4. Edlin, A and Shannon, C (1998) "Strict Monotonicity in Comparative Statics." *Journal of Economic Theory*, 81, July, 201-219.
5. Fehr, E. and K. Schmidt (1999) "A Theory of Fairness, Competition and Cooperation" *Quarterly Journal of Economics*, Vol.114, No3, 817-868.
6. Fehr.E, Hart.O,and C. Zehnder, (2011). "Contracts as Reference Points–Experimental Evidence," *American Economic Review*, vol. 101(2), pp. 493-525, April.
7. Fudenberg, D and J. Tirole (1991) *Game Theory* MIT Press
8. Hart, O. and Holmstrom, B. (2010) "A Theory of Firm Scope", *Quarterly Journal of Economics*, 125 (2): 483-513.
9. Hart, O. and J. Moore (2008) "Contracts as Reference Points," *Quarterly Journal of Economics*, vol. 123(1), pp. 1-48, 02.
10. Kofman, F, and Lawarree, J, (1993) "Collusion in Hierarchical Agency", *Econometrica*, Vo1. 61, No3, May, 629-656.
11. Laffont, J and D. Martimort (1997) "Collusion under Asymmetric Information" *Econometrica*, Vol. 65, No4, 875-911.

12. Laffont, J-J. and J. Tirole (1988) "The Dynamics of Incentive Contracts," *Econometrica*, Vol. 56 No. 5, pp. 1153-75, September

13. Laffont, J. and J. Tirole (1991) "The Politics of Government Decision-Making: A Theory of Regulatory Capture," *Quarterly Journal of Economics*. 106, 1089-1127.

14. Milgrom, P. (1988) "Employment Contracts, Influence Activities and Efficient Organization Design," *Journal of Political Economy*, 96, 42-60.

15. Milgrom, P. (2004): Putting Auction Theory to Work, Cambridge University Press: Cambridge.

16. Milgrom, P, and Segal, I (2002) "Envelope Theorems for Arbitrary Choice Sets," *Econometrica* 70 (2), March, 583-601

17. Mirrlees, J. A. (1971) "An Exploration in the Theory of Optimum Income Taxation" *Review of Economic Studies*, 38, 175-208.

18. Myerson, R. (1981) "Optimal auction design", *Mathematics of Operations Research* vol. 6, 58-73.

19. Suzuki, Y., (2007) "Collusion in Organizations and Management of Conflicts through Job Design and Authority Delegation", *Journal of Economic Research* 12. pp. 203-241

20. Suzuki, Y. (2007) "Mechanism Design with Collusive Auditing: A Three-Tier Agency Model with "Monotone Comparative Statics" and an Implication for Corporate Governance", Institute of Comparative Economic Studies, Hosei University, Working Paper No. 128.

21. Suzuki, Y., (2008) "Mechanism Design with Collusive Supervision: A Three-tier Agency Model with a Continuum of Types," *Economics Bulletin*, Vol. 4 no. 12 pp. 1-10

22. Suzuki, Y., (2012) "Collusive Supervision, Shading, and Efficient Organization Design", mimeo-graphed.

23. Tirole, J. (1986) "Hierarchies and Bureaucracies: On the role of Collusion in Organizations". *Journal of Law, Economics and Organization*. 2. 181-214.

24. Tirole, J. (1992) "Collusion and the Theory of Organizations" in *Advances in Economic Theory: The Sixth World Congress* edited by J. J. Laffont. Cambridge: Cambridge University Press.

25. Topkis, D (1978) "Minimizing a submodular function on a lattice", *Operations Research*, 26 (2), 255-321.

26. Weitzman, M. (1974) "Prices vs. quantities", *Review of Economic Studies* 41(4), 477-491, October.

27. Williamson, O. (1971) "The Vertical Integration of Production: Market Failure Considerations." *American Economic Review* 61: 112-23

28. Williamson, O. (1975) Markets and Hierarchies: Analysis and Antitrust Implications. New York, NY: Free Press.