CATCHING THE AGENT ON THE WRONG FOOT: EX POST CHOICE OF MONITORING

FAHAD KHALIL

Department of Economics, Box 353330

University of Washington, Seattle, WA 98195

khalil@u.washington.edu

AND

JACQUES LAWARRÉE

Department of Economics, University of Washington

and ECARES, Brussels

lawarree@u.washington.edu

September 25, 2000

Abstract:

In a principal-agent model with multiple performance measures, we show that the principal benefits by choosing ex post which variables will be monitored. If it is too costly for one type of agent to mimic all performance measures expected from another type, the principal can hope to catch the agent on the wrong foot if the agent tries to misrepresent his type. For cases of small asymmetry of information, the principal can implement the first best contract. For more serious asymmetries of information, the first best is not implementable. Then the low type may be required to overproduce, which is in contrast to the traditional result of second best contracting. We also obtain a ranking of monitoring instruments according to the frequency of their use.

JEL Classification Numbers: D23, D82, L22.

Acknowledgments: We would like to thank Y. Barzel, M. Boyer, N.Bruce, J. Crémer, T. Eicher, M. Ghatak, J. Kline, F. Laux, P.Pestieau, R Strausz, and M. Van Audenrode for valuable comments.

Corresponding author: Jacques Lawarrée, Department of Economics, Box 353330, University of Washington, Seattle, WA 98195-3330, USA, phone: 1-206-543-5632, fax: 1-206-685-7477, e-mail: lawarree@u.washington.edu

1. Introduction

In a principal agent problem under adverse selection, a principal can often find many performance measures to screen an agent and provide incentive. If the agent knows the variables that are being monitored, an agent of one type can obtain rent by mimicking the contractual obligation of another (inferior) type. In what has been touted as the biggest accounting fraud ever ($19 billion), the Securities and Exchange Commission (S.E.C.) has recently investigated a company called CUC International. It is alleged that the company fooled auditors from Ernst and Young by faking the numbers of *some* of its subsidiaries. "The S.E.C. says that the fraud was easier to pull off because CUC officials knew which subsidiaries would be audited, and therefore hid the most obvious frauds in subsidiaries that they knew the auditors would not look at." (New York Times, June 16, 2000).

It is standard in the literature to assume that the principal announces ex ante which performance measures will be used.[1] In this paper, we show that the principal could gain by choosing ex post which variables will be monitored. If it is too costly for one type of agent to mimic all performance measures expected from another type, the principal can hope to catch the agent on the wrong foot if the agent tries to misrepresent his type.

There are many examples where different measures of performance can be used to address an agency problem under adverse selection. Revisiting the standard taxation model of Mirrlees (1971), Maskin and Riley (1985) point out that a tax authority can use various instruments, such as input and output, to derive the optimal tax scheme. In regulation theory, the optimal scheme for a monopolist can also be based on various signals such as input, output, costs, etc. (Caillaud et al. (1988)). An important issue in pollution control is the choice between emission taxes and taxes on other factors of production that are correlated with emissions (see Besanko (1994), Lewis (1996), Schmutzler and Goulder, (1997)). The tax authority must therefore choose between monitoring emission directly or, alternatively, monitoring the pollution abatement technologies (e.g. end-of-pipe technology) installed by the polluting firm.

[1] See for example, Baron and Myerson (1982), Sappington (1983), and Laffont and Tirole (1993).

Analyses of multiple performance measures also appear in fields other than public economics. Lafontaine and Slade (1996, 1998) argue that, in most manufacturer-retailer relationships including franchising, the manufacturer can monitor the retailer through sales data or through more direct signals of effort, e.g., by tasting the food quality, assessing the cleanliness of the unit or by determining work hours. In a similar vein, Anderson and Oliver (1987) distinguish between behavior-based compensation and outcome-based compensation. In labor economics, the literature on piece-rate vs. wage-rate is another illustration of the availability of multiple signals (see Matutes and Régibeau (1994) for a recent contribution).[2]

Even when the principal receives several signals, if mimicking is costless, the agent simply mimics every possible performance measure so that misrepresentation is never detected. Therefore, the principal may not be able to improve the contract by increasing the number of signals in this case. However, if mimicking is costly, then increasing the potential number of signals raises the agent's cost of misrepresenting his type. This would necessarily help the principal if signals were free to observe.[3] Since performance measures are typically costly to obtain, increasing the number of signals is costly to the principal.[4] However, this is based on the implicit assumption that all the signals announced ex ante are indeed monitored eventually. Under ex post monitoring, the principal announces a large number of variables ex ante, but only decides ex post which ones he will actually monitor. This gives the principal the option of monitoring only a subset of the variables and save on monitoring cost while using the incentive power of a large number of variables.

We want to emphasize that the principal may collect or create an access to a large amount of data, even if verifying all of them is too costly. For income tax returns, the tax code

[2] There is also a vast literature on agricultural contracts where wage contracts are input (labor hours) based, and sharecropping is output based. See Singh (1989).

[3] See Holmström (1979) in the context of moral hazard and Rochet and Stole (2000) in the context of multidimensional screening.

[4] While more information is typically better, the principal might not use all the available variables. The reason could be that these variables are costly to observe or simply that it is too costly to make all of them verifiable. Such an instance would be common when the output has a quality component, and an expert witness has to convince a third party of the true value of the quality parameter. This same assumption is at the root of the incomplete contract literature (see Hart (1995)). Dewatripont and Maskin (1995) also show that observing more than one variable is not necessarily better when renegotiation is allowed after one variable has been revealed and before the other one is revealed. Crémer (1995) is another example of such an effect of revealing information too early in a repeated relationship.

requires taxpayers to report many items that can be crosschecked if the IRS wants to do so.[5] In life insurance contracts, individuals are asked to answer a large number of questions regarding their health. For compliance with the chemical weapons treaty, participating countries have to make very detailed announcements regarding production plans of every chemical plant in the country.[6] Similarly, immigration officials ask a long list of questions when foreigners apply for a US visa. In each of these examples, what is important is that the principal reserves the right to decide ex post which pieces of information to verify.

If the number of variables announced ex ante is large, mimicking every performance measure may become too costly for the agent. As a consequence, the agent trying to mimic another type has to guess which variables will actually be monitored. If he guesses incorrectly the agent will be caught on the wrong foot and penalized since his true type will then be revealed. This threat of a penalty reduces information rent, and for small cases of asymmetry of information, the principal can implement the first best contract.[7]

Our analysis may have important empirical implications in that actual rent in real world contracts may be much less than what would be inferred from a model based on one monitoring variable. In reality, organizations have access to multiple monitoring variables even though explicit incentives may primarily be based on one variable. Therefore the scope for misrepresentation may be quite limited. In their survey paper, Andreoni et al. (1998, p.821) claim that the IRS's use of "informational reporting," which requires reports on multiple items in tax returns, could partly explain why compliance levels have been so high even though penalties and probabilities of auditing are relatively low. Our results suggest policy guidelines as they establish the notion that the access to a menu of variables implies more efficient contracts than if only one variable was available.

The choice between monitoring variables has attracted attention in the recent literature on incentive problems due to hidden information. Maskin and Riley (1985) introduce the problem of input versus output monitoring in this framework, and show that output monitoring

[5] Bruce (1998, p.472) presents a nice example where the IRS caught pizza parlors committing tax fraud. The IRS auditors showed inconsistencies in the reports of input (flour) and output (sales).
[6] See the web site http://www.opcw.nl/ for more details on the treaty.
[7] In section 2, we contrast our approach with the traditional auditing model after presenting our model.

is better.[8] In a similar vein, Lewis and Sappington (1995), consider the case of pollution control, and ask if monitoring pollution or monitoring abatement equipment is more efficient. Khalil and Lawarrée (1995) extend the work of Maskin and Riley, and show that the choice of the monitoring instrument depends on whether the principal or the agent collects the output. Barzel (1997) studies a similar problem in the context of moral hazard. In a recent paper, Bontems and Bourgeon (1999) show that even the choice of the monitoring instrument can be used as a screening device. The principal lets the agent choose the instrument. They show that some types may choose input while others may choose output monitoring, and in some cases even the first best efforts can be implemented. However, this literature assumes that the agent knows the variable to be monitored before he acts, whereas we let the principal choose this variable after the productive action has occurred. The literature on multi dimensional screening (see Rochet and Stole (2000) and references therein) is also related. Whereas this literature assumes that signals are observed for free, we assume that it is costly to monitor signals. Indeed, our main point is that when signals are costly, the principal can save on monitoring cost by only observing a subset of signals while deriving benefit from all of them.

To illustrate our ideas, we use a model similar to Maskin and Riley (1985) where the principal can monitor either input or output. In addition to the result of first best contract for cases of small asymmetry of information, we find that the principal may ask a low type agent to over-produce as opposed to the traditional under-production result. Also, we obtain a ranking of instruments even when input and output monitoring are equally costly and equally accurate: the principal will monitor input more often.

The paper is organized as follows. We introduce the model in section 2. Two benchmark contracts are presented in section 3: the full information contract, and the contract under ex ante choice of monitoring. Our main results are presented in section 4, while we explore some extensions in section 5 and present our conclusions in section 6.

[8] For an earlier treatment, see Wittman (1977).

2. The Model

A risk neutral principal hires a risk neutral agent to work for him. The agent's input $e \geq 0$ along with his productivity parameter θ determines output $X=\alpha(e,\theta)$, with $\alpha(0, \theta) \geq 0$, $\alpha_e > 0$, $\alpha_\theta > 0$, $\alpha_{ee} \leq 0$, $\alpha_{e\theta} \geq 0$, $\lim_{\theta\to\infty} \alpha_e(e,\theta) = \infty$. While we refer below to input e as effort, it could also include non-effort related decisions such as the types of input purchased or the end of pipe pollution control technology adopted or the marketing strategies selected. Effort cost is inversely related to productivity.[9] The cost of effort for type θ is given by the function $\psi(e,\theta)$, with $\psi_e > 0$, $\psi_{ee} > 0$, $\psi_\theta < 0$, $\psi_{e\theta} < 0$, $\psi(0,\theta) = 0$, $\lim_{e\to\infty} \psi_e(e,\theta) = \infty$, $\lim_{e\to 0} \psi_e(e,\theta) = 0$, $\lim_{\theta\to\infty} \psi(e,\theta) = \lim_{\theta\to\infty} \psi_e(e,\theta) = 0$.

For simplicity, we assume that productivity can be either high (θ_2) or low (θ_1), $\theta_2 > \theta_1 > 0$.[10] The parameter θ is private information of the agent. The principal's subjective probability that $\theta = \theta_1$ is q. We assume that it is optimal for the principal to employ either type of agent. This implies that q is not too small given the ratio θ_2/θ_1.

A contract is a six-tuple $\{e_1, e_2, t_1, t_2, \omega_1, \omega_2\}$, where ω_i is the probability of output monitoring when the agent announces that he is of type i and t_i is the corresponding transfer. Note that the contract also implicitly specifies output levels for each type. For instance, if type θ_1 is supposed to put in effort e_1 in exchange for t_1, then the output implied by the contract is: $X_1=\alpha(e_1, \theta_1)$. The principal can either monitor the input (e) or the output (X). Monitoring is perfect but publicly reveals *only* the variable that is monitored, e.g., if output is monitored, X is perfectly known but not e nor θ. This leaves room for one type to mimic the other. If X_1 is observed, it cannot be ruled out that the high type has mimicked the low type. If e_1 is observed, it is not known whether output is $X = \alpha(e_1,\theta_2) > X_1$. Even though the principal collects the output, we assume that, under input monitoring, he cannot observe output or make it verifiable. There are many examples where output has features, such as quality, that are difficult to

[9] The previously discussed paper by Bontems and Bourgeon (2000) shows that the opposite assumption leads to the introduction of countervailing incentives.

[10] Assumptions made above regarding the functions $\alpha(.)$ and $\psi(.)$ have obvious counterparts for this discrete framework.

observe even though the beneficiary of the output is clear. When controlling pollution, the EPA, representing citizens, "consumes pollution" but does not observe its level unless it monitors it (see Swierzbinski (1994)). Healthcare authorities, like Medicare, contract with healthcare providers, but the benefit of treatment accrues to patients and is typically not observed by the authorities (see Chalkley and Malcomson (1999)). It is difficult to evaluate the work of an auto mechanic, and the military may never learn the efficacy of weapons in a nuclear war.[11]

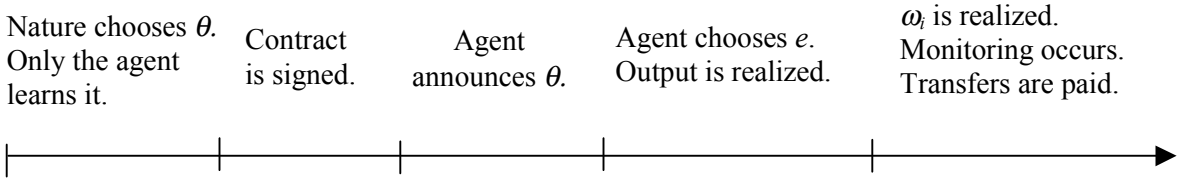
Differentials in cost and precision between input and output monitoring have straightforward consequences in our model. The principal would always be biased towards the cheaper and more precise monitoring instrument. To focus on a ranking of instruments based only upon incentives, we therefore assume that input and output monitoring are equally costly and equally precise. Monitoring is error free and it costs C to monitor either input or output. A more critical assumption is that it is not feasible to monitor both input and output. This assumption captures the fact that in general it is too costly for the principal to monitor every possible performance measure. In our model, if both input and output were observed, the first best (minus $2C$) would always be reached. We could also have allowed the principal to choose to observe neither input nor output. If nothing was observed, the transfer would be based only on the agent's announcement and the expected cost of monitoring would drop below C.[12]

We assume that the principal can commit to the probability of monitoring as part of the contract. This assumption turns out to be innocuous in our model since we assume that the principal must monitor either input or output, which can be observed with equal precision and accuracy. More generally, the commitment assumption is restrictive, but is used frequently in models with monitoring. It is typically justified using informal arguments from repeated games or delegation games. Melumad and Mookherjee (1989) model IRS audits to show that if the public can observe some aggregate variables like the IRS budget, aggregate costs and fines collected, then the government can attain the full commitment outcome even if it cannot control

[11] See Lewis-Sappington (1991), Lawarrée-Van Audenrode (1996) or Strausz (1997) for further examples.
[12] In tax-compliance models (e.g. Mookherjee and Png's (1989)) tax is only based on reported income unless there is an audit. In Swierzbinski's (1994) model of pollution control, the regulatory policy is based on the announced type unless the level of pollution is monitored. But in each case the principal announces ex ante the variable to be monitored.

audit probabilities directly.[13] In a model without commitment to the monitoring probability, the equilibrium would be in mixed strategies (see for example Khalil (1997)), and the optimal contract would be more complex to characterize. Since our goal was to demonstrate simply the benefit of basing the contract on multiple signals while monitoring only a subset, we chose the simpler model with commitment.

The principal collects the output and compensates the agent with a transfer t. The agent receives the transfer t, but bears the cost $\psi(e,\theta)$. The agent knows his productivity before signing the contract. Therefore he must receive his reservation utility (normalized to zero) in any state of the world. The agent is asked to announce his type after accepting the contract. If the monitored signal does not correspond to the announcement, shirking is detected and the agent does not receive any transfer. The penalty for shirking is therefore the uncompensated cost of effort[14]. A more stringent penalty could also be considered: besides losing his transfer the agent may be required to suffer an additional fine $F > 0$. For now, we assume that $F = 0$, and we will discuss the case of $F > 0$ later. To summarize, we present below the timing:



We have presented a model in which the principal can hope to catch the agent on the wrong foot and penalize him by choosing the monitoring variables ex post. The agent could not be penalized in a model of ex ante choice or if mimicking is costless. With ex ante choice, the agent knows which variables to mimic; with costless mimicking, the agent would mimic every variable. In either case, misrepresentation cannot be detected; it can only be deterred by giving rent, and there are no penalties on or off the equilibrium path. The monopoly regulation

[13] One could interpret the formula for computing DIF scores (see Andreoni et al. (1998)) as an attempt at coordinating actions of IRS auditors, and thus a form of commitment to auditing. Even though the formula is secret, it presumably can be inferred over time.

[14] Following Laffont-Tirole(1993), we also interpret the agent's limited liability as the principal's inability to extract money.

model of Baron and Myerson (1982) is an example with ex ante choice and costless mimicking. Sappington (1983) is another example of ex ante choice and costless mimicking, but with multiple signals. The model of input versus output monitoring used here is due to Maskin and Riley (1985), and it provides a clean way to capture the impossibility of mimicking every variable in a simple technology: when one type of agent mimics another, he can mimic another type's input or output but not both. In section 5, we extend our model to allow the agent to mimic every variable by introducing falsification cost.

Our main ideas would generalize to a model with more than two types. Consider for instance a continuum of types. The agent must still announce his type and misrepresentation will be detected in the same fashion. The only difference is that, when misrepresentation is detected, the monitored variable could correspond to the level of another existing type.[15] However, the agent still gets caught since the monitored variable does not correspond to the *equilibrium* value of the announced type.

In our model, efficiency of the optimal contract is enhanced by off-the-equilibrium path penalties. This aspect of our model is similar to auditing models where the principal can commit to audit probabilities. However, our model differs from the traditional auditing model.[16] In these models, one variable, say output, is always observed, and therefore a high type may shirk by producing the output level of a low type. In order to discover shirking, the principal has to observe a second variable (input), which is the outcome of an audit. In our model, the principal has access to two variables, but ultimately observes only one variable. If the principal randomizes between the potential monitoring variables, a shirking agent faces a positive probability of penalty even if only one variable is observed. Thus monitoring cost may be lower than under auditing since only one variable has to be monitored in equilibrium.

[15] In our model with two types, when the high type agent mimics the input of the low type he produces an output higher than the equilibrium output of the low type and lower than the equilibrium output of the high type.

[16] See for example, Baron and Besanko (1984) or Kofman and Lawarrée (1993) for auditing models with commitment, and Khalil (1997) for the case without commitment. Without commitment, there may be penalties in equilibrium.

3. The first best contract, and the contract under ex ante choice of monitoring

If the agent's type is publicly observable, the principal's problem is to choose efforts and transfers for each type of agent to maximize expected profit:

$q\,[\alpha(e_1, \theta_1) - t_1] + (1-q)\,[\alpha(e_2, \theta_2) - t_2]$

subject to the individual rationality constraints of the two types:

(IR1) $t_1 - \psi(e_1, \theta_1) \geq 0,$

(IR2) $t_2 - \psi(e_2, \theta_2) \geq 0.$

The solution is the first best contract, where marginal benefit of effort equals marginal cost, and there is no rent. For $i = 1, 2,$

$$\alpha_e(e_i^*, \theta_i) = \psi_e(e_i^*, \theta_i),$$
$$t_i^* = \psi(e_i^*, \theta_i).$$

Next, we examine the case where the agent's type is private information and the principal commits to monitor a particular variable as part of the contract (that is, $\omega_i = 0$ or 1). Under input monitoring, transfers are based on observed input, and the incentive compatibility constraints are:

(ICi1) $t_1 - \psi(e_1, \theta_1) \geq t_2 - \psi(e_2, \theta_1),$
(ICi2) $t_2 - \psi(e_2, \theta_2) \geq t_1 - \psi(e_1, \theta_2)$.

On the other hand, under output monitoring, payments are based on observed output, and the incentive compatibility constraints now become:

(ICo1) $t_1 - \psi(e_1, \theta_1) \geq t_2 - \psi(\tilde{e}_2, \theta_1),$
(ICo2) $t_2 - \psi(e_2, \theta_2) \geq t_1 - \psi(\tilde{e}_1, \theta_2)$,

where $\tilde{e}_1$ and $\tilde{e}_2$ satisfy $\alpha(\tilde{e}_1, \theta_2) = \alpha(e_1, \theta_1)$ and $\alpha(\tilde{e}_2, \theta_1) = \alpha(e_2, \theta_2)$. Indeed, when mimicking the low type, the high type agent has to produce X_1 and exert an effort $\tilde{e}_1$ smaller than e_1. The low type mimicking the high type must exert an effort $\tilde{e}_2$ higher than e_2.

The principal's problem is to maximize expected profit subject to the individual rationality constraints, and the relevant incentive compatibility constraints given the monitoring scheme. In Khalil-Lawarrée (1995), we show that input monitoring yields higher profit to the principal in this model. Intuitively, it is because the agent receives rent from two sources under output monitoring and from only one source under input monitoring. The rents for the high type agent under input and output monitoring are respectively,

$$\text{Rent}^{I} = \psi(e_1, \theta_1) - \psi(e_1, \theta_2) ,$$

$$\text{Rent}^{O} = \psi(e_1, \theta_1) - \psi(\tilde{e}_1, \theta_2) .$$

The expression for Rent^{O} shows that the high type agent receives a rent because (i) he can exert a lower level of effort ($\tilde{e}_1 < e_1$) and (ii) he has a lower cost of effort. Under input monitoring, the agent only commands a rent from a lower cost of effort. The other results are standard: the high type produces efficiently; the low type under-produces and his effort is inversely related to θ_2/θ_1; the low type does not earn rent.

4. Ex post choice of monitoring

We now return to the case where the principal does not have to decide to exclusively monitor input or output before the effort is taken. The principal can commit to a probability of output monitoring $\omega_i \in [0, 1]$. The principal's problem is to choose a contract that solves the following problem **(P)**:

$$Max\ q[\alpha(e_1,\theta_1) - t_1] + (1-q)[\alpha(e_2,\theta_2) - t_2] - C$$

s.t.

$$(IC_2) \quad t_2 - \psi(e_2,\theta_2) \geq max\ \{\omega_1 t_1 - \psi(\tilde{e}_1,\theta_2), (1-\omega_1) t_1 - \psi(e_1,\theta_2)\},$$

$$(IC_1) \quad t_1 - \psi(e_1,\theta_1) \geq max\ \{\omega_2 t_2 - \psi(\tilde{e}_2,\theta_1), (1-\omega_2) t_2 - \psi(e_2,\theta_1)\},$$

$$(IR_2) \quad t_2 - \psi(e_2,\theta_2) \geq 0,$$

$$(IR_1) \quad t_1 - \psi(e_1,\theta_1) \geq 0.$$

While the individual rationality constraints are standard, the incentive compatibility constraints require elaboration. We explain (IC_2) only since (IC_1) is analogous. A high type

agent can claim to be a low type either by mimicking input, i.e., exerting e_1, or by mimicking output, i.e., exerting $\tilde{e}_1$. If he chooses e_1, he only receives t_1 if input is monitored, which occurs with probability $(1-\omega_1)$, while he bears the cost $\psi(e_1, \theta_2)$ for certain. If he chooses $\tilde{e}_1$, he will only receive t_1 with probability ω_1, while bearing the smaller cost $\psi(\tilde{e}_1, \theta_2)$. The contract must ensure that the agent has no incentive to misrepresent his type under either option. Since the high type agent cannot simultaneously mimic both input and output of the low type, he faces a penalty with a positive probability as long as monitoring is random.

As a preliminary step, we simplify the problem (P). As is typical in these types of models, the low type will not want to claim to be of high type in equilibrium. Therefore, we now assume that the constraint (IC_1) is not binding in equilibrium, but this can be verified to be true later for appropriately chosen ω_2. We will clarify the choice of ω_2 in footnote 17 when discussing lemma 2. Then (IR_1) is binding since t_1 can be lowered without violating any constraints. We replace t_1 by $\psi(e_1, \theta_1)$ in the principal's problem and focus now on (IC_2) which is rewritten as

$$(IC^*_2) \quad t_2 - \psi(e_2, \theta_2) \geq \max \{\omega_1 \psi(e_1, \theta_1) - \psi(\tilde{e}_1, \theta_2), (1-\omega_1) \psi(e_1, \theta_1) - \psi(e_1, \theta_2)\},$$

The high type may want to claim to be a low type so that he is compensated for $\psi(e_1, \theta_1)$, whereas he has actually incurred only $\psi(\tilde{e}_1, \theta_2)$ or $\psi(e_1, \theta_2)$. This cost differential represents the benefit from shirking, and is typically the rent in models with ex ante choice of monitoring (see $Rent^I$ and $Rent^O$ in section 3). In our model, shirking has another, new consequence: there may be a penalty due to ex post choice of monitoring, which is the uncompensated cost when shirking is detected. In the standard case, there is no penalty since shirking cannot be detected; it can only be deterred. We first show that the principal will optimally use this penalty by randomizing between the two instruments.

Lemma 1. *It is optimal to monitor both input and output randomly* $(0 < \omega_1 < 1.)$.

Proof. In appendix.

The intuition is straightforward. If the principal only does input monitoring ($\omega_1=0$), then the high type agent must be given a rent since he can mimic the input of the low type agent

without any risk of being detected by the principal. If ω_1 is slightly positive, the shirking agent will be detected with positive probability and the resulting penalty implies a lower rent. Similarly, if only output monitoring occurs ($\omega_1=1$), a shirking high type agent obtains a rent from his ability to mimic the output of a low type agent without running the risk of being detected while his rent can be reduced if $\omega_1<1$. Therefore, it is never optimal for the principal to perform either input or output monitoring exclusively.

Having established that the principal will randomize, we argue next that, without loss of generality, the ω_1 will be chosen to equate the two terms on the right hand side (RHS) of the constraint (IC_2). First note that in the principal's problem the ω_1 only appears on the RHS of the incentive constraint (IC_2). On the RHS of (IC_2), the first term is increasing and the second term is decreasing in ω_1. Thus for any efforts and transfers, the principal can choose ω_1 to make the RHS as small as possible by equating the two terms. This implies that $\omega_1 < .5$ as $\psi(\tilde{e}_1,\theta_2) < \psi(e_1,\theta_2)$ in (IC_2). Mimicking output generates a larger cost differential for the agent as he saves on the cost of effort and also on the amount of effort when he mimics the low type. Therefore, to lower rent, the optimal ω_1 is biased towards input monitoring by setting $\omega_1 < 1/2$. We have proved the following lemma[17].

Lemma 2. Without loss of generality, the principal can set

$$\omega_1 = \frac{1}{2} - \frac{1}{2t_1}[\psi(e_1,\theta_2) - \psi(\tilde{e}_1,\theta_2)] < .5 \text{ and it solves}$$

$$\omega_1 t_1 - \psi(\tilde{e}_1,\theta_2) = (1-\omega_1) t_1 - \psi(e_1,\theta_2).$$

It can be readily argued, that in our simple model, the optimal ω_1 also turns out to be sequentially rational. This is because the agent does not shirk in equilibrium and monitoring input or output is equally costly. Therefore, ex post the principal will have no incentive to deviate from his pre-announced choice of ω_1 as he has to pay the transfer t_1 in either case. This

[17] Just like in the case of (IC_2) and ω_1, equating the two terms also minimizes the RHS of (IC_1) which gives ω_2 without loss of generality. Note that for some parameter values, it is possible that even if ω_2 is 0 or 1, the (IC_1) is not binding.

would not be the case if the instruments were not equally costly and precise, since then the principal would monitor the cheaper variable if he knew the agent was not shirking.

Remembering that (IR_1) is binding, and substituting ω_1 from lemma 2, (IC^*_2) can be rewritten as

$$(IC'_2) \quad t_2 - \psi(e_2,\theta_2) \geq .5[\psi(e_1,\theta_1) - \psi(\tilde{e}_1,\theta_2) - \psi(e_1,\theta_2)],$$
$$\equiv .5\, R(e_1,\theta_1,\theta_2) \qquad \text{(Rent)}$$

When the (IC_2) is binding, the RHS gives the agent's rent, which is $.5R(e_1, \theta_1, \theta_2)$. Comparing with the rent expressions in section 3, it is easily seen that the RHS of (IC'_2) is different from the standard rent under ex ante choice of monitoring. The RHS of (IC'_2) shows the benefit of mimicking the low type and saving on cost as well as the expected penalty from detection. In the standard case, there would be no penalty. Our main results will all depend on how $R(e_1,\theta_1,\theta_2)$ changes with the variables. For example, if increasing e_1 increases $R(e_1,\theta_1,\theta_2)$, there will be under-production relative to first best.

We are now ready to present the simplified problem. We use the binding IR_1 to replace t_1 and lemma 2 to replace ω_1 in problem (P). Therefore, the principal's problem is now **(SP)**

$$\text{Max } q[\alpha(e_1,\theta_1) - \psi(e_1,\theta_1)] + (1-q)[\alpha(e_2,\theta_2) - t_2] - C$$

s.t.

$$(IC'_2) \quad t_2 - \psi(e_2,\theta_2) \geq .5[\psi(e_1,\theta_1) - \psi(\tilde{e}_1,\theta_2) - \psi(e_1,\theta_2)],$$

$$(IR_2) \quad t_2 - \psi(e_2,\theta_2) \geq 0.$$

Proposition 1: *If the two agents are very similar (the ratio θ_2/θ_1 is close enough to 1), the first best contract is implementable.*

Proof. In appendix.

Thus we find that the principal may obtain the benefit of two variables when in fact he observes only one. Remember that with ex ante monitoring, both variables had to be observed to obtain the first best in this model. This result can be understood by examining why the high type agent cannot command a rent. As noted earlier, the RHS of IC'_2 represents the rent. It is

the net effect of the gain from mimicking the low type, which is the cost differential minus the expected penalty from being detected. When θ_2 is relatively small, the cost differential is small but the penalty is strong because the cost of effort for the high type is large. This penalty allows the principal to implement the first best contract.[18] With larger θ_2, the cost of (any level of) effort falls, weakening the penalty while the cost differential increases. This is why the first best contract is no longer implementable for larger values of θ_2/θ_1.

We have just argued that for high values of θ_2/θ_1, (IC′$_2$) is binding and there will be a distortion in e_1. The distortion can be explained by examining how $R(e_1, \theta_1, \theta_2)$ is affected by changes in e_1. Since a higher effort e_1 implies a higher cost of effort, and this cost is also the penalty for a shirker, the principal can strengthen the penalty by increasing the effort of the low type. We call this the penalty effect of increasing e_1 on $R(e_1, \theta_1, \theta_2)$. Increasing e_1 has an undesirable effect for the principal too: it increases the cost differential and therefore makes the gain from shirking larger. We refer to this effect as the traditional effect of increasing e_1 on $R(e_1, \theta_1, \theta_2)$. The relative strength of the two effects explains the distortion in e_1. For instance, if the penalty effect is stronger, rent decreases with e_1, and the optimal e_1 is set above its first best level. The net contribution of the two effects on rent is captured by the derivative of $R(e_1, \theta_1, \theta_2)$ with respect to e_1, which we define by $R_e(e_1, \theta_1, \theta_2)$ and is given below:

$$R_e(e_1, \theta_1, \theta_2) \equiv \psi_e(e_1, \theta_1) - \psi_e(\tilde{e}_1, \theta_2)\frac{\alpha_e(e_1, \theta_1)}{\alpha_e(\tilde{e}_1, \theta_2)} - \psi_e(e_1, \theta_2).$$

Note the difference with Proposition 1 where the levels of the penalty and cost differential were relevant, and here it is the *effect of changing* e_1 on the penalty and the cost differential.

Without the penalty effect, an increase in e_1 would simply increase the cost differential and there would be underproduction due to the traditional effect of changing e_1. In a traditional input monitoring problem, as in section 3, only two terms appear in the definition of $R^I_e(e_1, \theta_1, \theta_2)$:

$$\psi_e(e_1, \theta_1) - \psi_e(e_1, \theta_2).$$

[18] The profit is not first best because C must be deducted.

Clearly, this expression is positive and under-production always occurs. Similarly in a traditional output monitoring problem, $R^O_e(e_1,\theta_1,\theta_2)$ is:

$$\psi_e(e_1,\theta_1) - \psi_e(\tilde{e}_1,\theta_2)\frac{\alpha_e(e_1,\theta_1)}{\alpha_e(\tilde{e}_1,\theta_2)}.$$

Once again, this expression is unambiguously positive and under-production occurs.

However, when the principal does not announce whether input or output monitoring will occur, $R_e(e_1,\theta_1,\theta_2)$ has the form described above and the sign of this expression is ambiguous. In Proposition 2, we state and prove that the penalty effect prevails for relatively small values of θ_2/θ_1, i.e., the net rent of the high type $.5R(e_1,\theta_1,\theta_2)$ is decreasing in e_1 and over-production occurs. We provide more intuition about this using an example below.

Proposition 2. *When the first best contract is not implementable, the low type over-produces (with respect to the first best effort level) if θ_2/θ_1 is relatively small and under-produces if θ_2/θ_1 is relatively large.*

Proof. In appendix.

In addition, if $R_e(e_1,\theta_1,\theta_2)$ is monotonically increasing in θ_2, there exists a unique cut-off $\bar{\theta}$ separating regions of under- and over-production. In the appendix, we derive conditions under which $R_e(e_1,\theta_1,\theta_2)$ is monotonically increasing in θ_2, and we show that this occurs when $\alpha_{ee}(e,\theta)$ is not too large. For many functional forms used in the literature such as $\alpha(e,\theta)=\theta + e$ or $\alpha(e,\theta)=\theta.e$, we have $\alpha_{ee}(e,\theta)=0$, and a unique cut-off $\bar{\theta}$ is found such that over-production occurs for $\theta_2 < \bar{\theta}$ and under-production for $\theta_2 > \bar{\theta}$.

An Example: $\alpha(e,\theta) = e + \theta$, $\psi(e,\theta) = e^2/2\theta^2$, $\theta_1=1$, $q = .5$

We clarify our analysis further by considering an example with $\alpha_{ee}(e,\theta) = 0$, so that $R_e(e_1,\theta_1,\theta_2)$ is unambiguously increasing in θ_2. The optimal e_1 is graphed for various values of θ_2 in figure 1. For θ_2 close to θ_1, the penalty is larger than the cost differential, and we have the first best. For larger values of θ_2, the penalty falls with θ_2 and the cost differential rises with the difference in productivity. For $\theta_2 > 1.38\theta_1$, $R(e_1,\theta_1,\theta_2)$ is positive at the first best e_1 and IC'_2 is violated under the first best contract. To satisfy IC'_2, the principal can increase t_2 and give rent

or distort e_1 and lower $R(e_1,\theta_1,\theta_2)$. Since there is no first order loss from distorting e_1 at this point, he distorts e_1 first. For further increases in θ_2, rent is provided.

To see why overproduction occurs first (as opposed to underproduction), we have to consider how the penalty effect and the traditional effect are influenced by changes in θ_2. For small θ_2, the penalty effect is larger than the traditional effect and there is overproduction. Since the marginal cost falls and the marginal cost differential increases with θ_2, the penalty effect decreases and the traditional effect increases with θ_2. When the traditional effect dominates (for $\theta_2 > 1.5\theta_1$), we finally obtain underproduction.

At $\theta_2 = 1.38\theta_1$, when the IC_2 is violated under the first best incentive scheme, the penalty effect is still stronger ($R_e(e_1,\theta_1,\theta_2) < 0$): an increase in e_1 above the first best level increases the penalty more than the cost differential and decreases $R(e_1, \theta_1, \theta_2)$. This maintains the agent's rent at zero for θ_2 above (but close to) $1.38\theta_1$. As θ_2 becomes larger, increasing e_1 keeps rent at zero, but at some point ($1.4\theta_1$) this distortion in e_1 is too costly, and it is better to yield rent. When rent is provided, the distortion in e_1 is reduced. See figure 1.

5. Extensions

It is interesting to mention what happens if the principal can impose a fine $F > 0$ besides withholding the transfer to the shirking agent. As expected, if F becomes very large, the principal can secure the first best contract. More importantly, notice that F affects the penalty (see proposition 1), but not the penalty effect (see proposition 2). So as long as the first best is not implementable, the result of overproduction survives even if there is an additional penalty F.

Note that if the penalty was transfer independent and only consisted of a fixed fine F, no over production would occur in equilibrium. However, a transfer dependent penalty, as we have assumed, by itself does not generate over production[19]. It is the simultaneous presence of ex post choice of monitoring and a transfer dependent penalty that lies behind the result of over production in our model.

We can also extend our model to allow the agent to mimic both variables but at some cost. For instance, when he mimics the input, the agent can also at some cost $A(.)$ falsify the output. In our model, $A(.)$ would represent the cost of destroying the extra output $(X_2\text{-}X_1)$. The other option for the agent, i.e., mimicking the output, would lead him to falsify the input. This could be achieved by making observed input unproductive (e.g., employees sitting at the computer but playing *solitaire*). By analogy we call it the cost of destroying input: $B(.)$.

We model the two functions as $A(X_2\text{-}X_1)$ and $B(X_2\text{-}X_1)$. They are assumed to be increasing and convex with respect to $(X_2\text{-}X_1)$. [20]

The (IC_2) now becomes

$$(IC_2^{+})\ t_2 - \psi(e_2,\theta_2) \geq \text{Max}\ \{\ .5[t_1 - \psi(\tilde{e}_1,\theta_2) - \psi(e_1,\theta_2)];\ t_1 - \psi(e_1,\theta_2) - A(.);\ t_1 - \psi(\tilde{e}_1,\theta_2) - B(.)\}$$

[19] Khalil (1997) shows that in a standard auditing model with commitment to auditing, but with transfer dependent penalty, over production does not occur.

[20] For simplicity we assume that the functions $A(.)$ and $B(.)$ are independent of θ. An alternative way to model mimicking costs can be found in the literature on costly state falsification. (See Crocker and Morgan (1998), Maggi and Rodríguez-Clare (1995) and references therein.) However, it would complicate the analysis by introducing countervailing incentives as in Lewis and Sappington (1989). As long as $A(.)$and $B(.)$ are not increasing in θ, countervailing incentives would not be present.

The agent has now four options besides telling the truth: he can mimic (i) input or (ii) output as before; (iii) he can mimic input and falsify output; (iv) he can mimic output and falsify input.

If *A(.)* and *B(.)* are so large that the second and third terms of the RHS of (IC_2^+) are negative, our model applies unchanged. When it is not the case, over production can still occur. Indeed, the functions *A(.)* and *B(.)* can play a role similar to the penalty term in (IC'_2). Therefore the RHS of (IC_2^+) has once again three terms vs. two in the ex ante model of monitoring. Consider, for example, the situation where the agent mimics the input and falsifies the output. The (IC_2) is now:

$$t_2 - \psi(e_2,\theta_2) \geq t_1 - \psi(e_1,\theta_2) - A(X_2 - X_1)$$

With (IR_1) binding, the rent R is equal to $\psi(e_1,\theta_1)-\psi(e_1,\theta_2)-A(X_2-X_1)$ and its derivative with respect to e is $R_e = \psi_e(e_1,\theta_1) - \psi_e(e_1,\theta_2) - A'(.)[\alpha_e(e_1,\theta_2)-\alpha_e(e_1,\theta_1)]$. The objective function can be written as $q[\alpha(e_1,\theta_1)-\psi(e_1,\theta_1)] + (1-q)[\alpha(e_2,\theta_2)-\psi(e_2,\theta_2) - R]$. According to the first order conditions, $e_2=e_2^{FB}$ and e_1 is such that $q[\alpha_e(e_1,\theta_1) - \psi_e(e_1,\theta_1)] - (1-q) R_e = 0$. So if $R_e < 0$ we have over-production.

As before, $\psi_e(e_1,\theta_1) - \psi_e(e_1,\theta_2) > 0$, and once again over-production can arise because $\{-A'(.)[\alpha_e(e_1,\theta_2)-\alpha_e(e_1,\theta_1)]\} < 0$., i.e., R_e can be negative. So what is crucial to get over-production is the existence of another term in R_e (the derivative of the rent) that has a negative sign. Ex post monitoring produces this extra term even when we explicitly model falsification costs.

Whether we continue to obtain the result that the first best is achieved for small asymmetry of information depends on the properties of the specific falsification cost function chosen. For instance, if there are fixed costs involved in falsification, the first best will still be reached for small asymmetry of information since the falsification cost does not disappear as the asymmetry of information reduces.

Therefore, our results generalize to the case where the agent can mimic all variables. Also, as the number of potential signals increase, so does falsification cost, and the principal benefits.

6. Conclusions

When multiple screening variables are available, we show that the principal can use the agent's fear of getting caught on the wrong foot by choosing the monitoring variable ex post. If the agent's types are similar (if θ_2/θ_1 is small in our model), this strategy of the principal has strong incentive effects: it yields the first best. For more serious situations of asymmetry (larger θ_2/θ_1), the first best is no longer implementable. We characterize the optimal contract and show that the traditional result of second best contracting no longer holds. Indeed, the principal might find it desirable to require the low type to overproduce.

For hidden information problems, our analysis provides a ranking of signals in terms of the likelihood of their use as monitoring instruments. Aside from issues of cost and accuracy, the probability of use is driven by the rent generated. In a contract with ex ante choice of instrument, if the agent can command more rent under one signal, then that variable will be monitored less often under ex post choice of instrument. We illustrate this by showing how input monitoring is used more frequently than output monitoring since there is more rent under output monitoring under the ex ante contract.

In real world applications, it may not be possible for the principal to hide for long the variable to be monitored. This is true, for example, if the chosen variable must be monitored as soon as the contract takes effect and this cannot be hidden. However, there are many cases where the evaluation takes place once the agent has performed his contractual obligations (see the introduction for examples),[21] and our analysis becomes relevant.

In many contractual relationships, the principal has the possibility to observe several variables but seems to observe only one of them most of the time. Our model stressed the role of rent, but we need to remember that we have abstracted from differences in cost or accuracy between the monitoring instruments. Often, a particular monitoring instrument reveals itself as more efficient.[22] The principal should therefore use that instrument more often. However, our model shows that having the opportunity to monitor an alternative variable has important

[21] This does not preclude early monitoring as long as the agent is not aware of it.
[22] This could be because it is more accurate and/or less expensive.

incentive effects. If this alternative variable is much less accurate and/or much more costly, the principal should use it with a very small, but positive probability. Also, in reality, different variables have different mimicking costs, which we have abstracted from. The principal will take these costs into account too when determining the frequency of use of monitoring instruments. But our main message will survive, as the principal will benefit from an increase in the dimensionality of the admissible signaling space while monitoring only a subset of signals.

Appendix

Proof of lemma 1

Replacing (IC^*_2) in the principal's problem (P), we see that the constraint (IC^*_2) is binding if $\omega_1 \in \{0, 1\}$. If $\omega_1 = 0$, the principal commits to monitor input, and the high-type's rent, which is $t_2 - \psi(e_2, \theta_2)$, equals $\psi(e_1, \theta_1) - \psi(e_1, \theta_2) > 0$. By choosing an ω_1 slightly positive, the principal can make (IC^*_2) slack and decrease t_2. A similar argument can be made for decreasing ω_1 from $\omega_1 = 1$.

Proof of Proposition 1

We show that there exists a cut-off $\underline{\theta} > 0$, such that if $\theta_2/\theta_1 < \underline{\theta}$ then (IC'_2) is slack under the first best contract and the first best contract is implementable, and that if $\theta_2/\theta_1 > \underline{\theta}$ then IC'_2 is binding and the first best contract is not implementable.

If $R(e_1^*, \theta_1, \theta_2)$ is negative, IC'_2 is slack under the first best contract. By the definition of $\alpha(e, \theta)$, $\tilde{e}_1(e_1^*, \theta_1, \theta_2)$ is continuous, and $\lim_{\theta_2 \to \theta_1} \tilde{e}_1(e_1^*, \theta_1, \theta_2) = e_1^*$. Hence,

$$\lim_{\theta_2 \to \theta_1} R(e_1^*, \theta_1, \theta_2) = -\psi(e_1^*, \theta_1) < 0.$$

On the other hand, $\lim_{\theta_2 \to \infty} \psi(e_1, \theta_2) = 0$ and $\lim_{\theta_2 \to \infty} \psi(\tilde{e}_1, \theta_2) = 0$. Hence,

$$\lim_{\theta_2 \to \infty} R(e_1^*, \theta_1, \theta_2) = \psi(e_1^*, \theta_1) > 0,$$

and the first best is no longer implementable for θ_2 high enough.

We complete the proof by showing that $R(e_1, \theta_1, \theta_2)$ is monotonically increasing in θ_2:

$$\frac{\partial R(.)}{\partial \theta_2} = [-\psi_\theta(e_1, \theta_2) + \psi_e(\tilde{e}_1, \theta_2) \frac{\alpha_\theta(\tilde{e}_1, \theta_1)}{\alpha_e(\tilde{e}_1, \theta_2)} - \psi_\theta(\tilde{e}_1, \theta_2)] > 0.$$

Proof of proposition 2

The Lagrangian is:

$$L = q[\alpha(e_1, \theta_1) - \psi(e_1, \theta_1)] + (1-q)[\alpha(e_2, \theta_2) - t_2] - C$$

$$+ \lambda_1[\, t_2 - \psi(e_2,\theta_2) - .5\{\psi(e_1,\theta_1) - \psi(\tilde{e}_1,\theta_2) - \psi(e_1,\theta_2)\}] + \lambda_2[t_2 - \psi(e_2,\theta_2)].$$

Consider the case where IC′$_2$ is binding, i.e., $\lambda_1 > 0$. It is easily checked that the optimal e_1 must satisfy

$$q[\alpha_e(e_1,\theta_1) - \psi_e(e_1,\theta_1)] = \lambda_1\left[\psi_e(e_1,\theta_1) - \psi_e(\tilde{e}_1,\theta_2)\frac{\alpha_e(e_1,\theta_1)}{\alpha_e(\tilde{e}_1,\theta_2)} - \psi_e(e_1,\theta_2)\right],$$
$$\equiv \lambda_1 R_e(e_1,\theta_1,\theta_2),$$

where $R_e(e_1, \theta_1, \theta_2) \equiv \psi_e(e_1,\theta_1) - \psi_e(\tilde{e}_1,\theta_2)\frac{\alpha_e(e_1,\theta_1)}{\alpha_e(\tilde{e}_1,\theta_2)} - \psi_e(e_1,\theta_2)$.

If $R_e(e_1, \theta_1, \theta_2)$ is positive (resp. negative), then under-production (resp. over-production) will result. The existence of a solution is demonstrated using an example in section 4, and here it is sufficient to show that for θ_2 close to θ_1, $R_e(e_1, \theta_1, \theta_2) < 0$ and for large θ_2, $R_e(e_1, \theta_1, \theta_2) > 0$.

$\lim_{\theta_2 \to \theta_1} R_e(e_1,\theta_1,\theta_2) = -\psi_e(e_1,\theta_1) < 0$ since $\lim_{\theta_2 \to \theta_1} \tilde{e}_1 = e_1$.

$\lim_{\theta_2 \to \infty} R_e(e_1,\theta_1,\theta_2) = \psi_e(e_1,\theta_1) > 0$ since $\lim_{\theta \to \infty} \alpha_e(e,\theta) = \infty$, and $\lim_{\theta \to \infty} \psi_e(e,\theta) = 0$.

Conditions for monotonicity of R(e_1, θ_1, θ_2)

$$\frac{\partial R_e(e_1,\theta_1,\theta_2)}{\partial \theta_2} = -\frac{\partial \psi_e(\tilde{e}_1,\theta_2)}{\partial \theta_2}\frac{\alpha_e(e_1,\theta_1)}{\alpha_e(\tilde{e}_1,\theta_2)} - \psi_e(\tilde{e}_1,\theta_2)\frac{\partial \frac{\alpha_e(e_1,\theta_1)}{\alpha_e(\tilde{e}_1,\theta_2)}}{\partial \theta_2} - \frac{\partial \psi_e(e_1,\theta_2)}{\partial \theta_2}.$$

From the definitions of $\psi(e, \theta)$ and $\tilde{e}_1(e_1, \theta_1, \theta_2)$, we know that both $\psi_e(e_1, \theta_2)$ and $\tilde{e}_1(e_1, \theta_1, \theta_2)$ are decreasing in θ_2. Therefore, $R_e(e_1, \theta_1, \theta_2)$ is monotonically increasing in θ_2 if $\frac{\partial \frac{\alpha_e(e_1,\theta_1)}{\alpha_e(\tilde{e}_1,\theta_2)}}{\partial \theta_2} \le 0$. One can check that this occurs when $\alpha_{ee}(e,\theta)$ is not too large. A sufficient condition is $\alpha_{ee}(e, \theta) = 0$, while the necessary condition is

$$\alpha_{e\theta}(\tilde{e}_1,\theta_2) - \alpha_{e_e}(\tilde{e}_1,\theta_2)\frac{\alpha_e(e_1,\theta_1)}{\alpha_e(\tilde{e}_1,\theta_2)} \ge 0.$$

Bibliography

Anderson, E., Oliver, R.L., 1987. Perspectives on behavior-based versus outcome-based sales force control systems. Journal of Marketing 51, 76-88.

Baron, D., Besanko, D., 1984. Regulation, asymmetric information, and auditing. RAND Journal of Economics 15, 267-302.

Baron, D., Myerson, R., 1982. Regulating a monopolist with unknown cost. Econometrica 50, 911-930.

Barzel, Y., 1997. Economic Analysis of Property Rights. Second ed. Cambridge University Press, Cambridge, UK.

Besanko, D., 1987. Performance versus design standards in the regulation of pollution. Journal of Public Economics 43, 19-44.

Bontems, P., Bourgeon, and J-M., 2000. Creating countervailing incentives through the choice of instruments. Journal of Public Economics 76, 181-202.

Bruce, N., 1998. Public Finance and the American Economy. Addison Wesley.

Caillaud, B., Guesnerie, R., Rey, P., Tirole, J., 1988. Government intervention in production and incentive theory: a review of recent contributions. RAND Journal of Economics 19, 1-26.

Chalkley, M., Malcomson, J., 1999. Government purchasing of health services. In: Culyer, A, Newhouse, J. (Eds), Handbook of Health Economics. North Holland.

Crémer, J., 1995. Arm's length relationships. Quarterly Journal of Economics 110, 275-295.

Crocker, K., Morgan, J., 1998. Is honesty the best policy? Curtailing insurance fraud through optimal incentive contracts. Journal of Political Economy 106, 355-375.

Dewatripont, M., Maskin, E.,1995. Contractual contingencies and renegotiation. Rand Journal of Economics 26(4), 704-719.

Hart, O., 1995. Firms, Contracts and Financial Structure. Oxford University Press.

Holmstrom, B., 1979. Moral hazard and observability. Bell Journal of Economics 10(1), 74-91.

Khalil, F., 1997. Auditing without commitment. Rand Journal of Economics 28(4), 629-640.

Khalil, F., Lawarrée, J., 1995. Input versus output monitoring: who is the residual claimant? Journal of Economic Theory 66 (1), 139-157.

Kofman, F., Lawarrée, J., 1993. Collusion in hierarchical agency. Econometrica 61, 629-656.

Laffont, J-J., Tirole, J., 1993. A Theory of Incentive in Procurement and Regulation. MIT Press.

Lafontaine, F., Slade, M., 1996. Retail contracting and costly monitoring: theory and evidence. European Economic Review 40, 923-932.

----------, 1998. Incentive contracting and the franchise decision. Forthcoming in Chatterjee, K., Samuelson, W. (Eds), Advances in Business Applications of Game Theory, Kluwer Academic Press, pp. ???-???.

Lawarrée, J., Van Audenrode, M., 1996. Optimal contract, imperfect output observation and limited liability. Journal of Economic Theory 71(2), 514-531

Lewis, T., 1996. Protecting the environment when costs and benefits are privately known. RAND Journal of Economics 27(4), 819-847.

Lewis, T., Sappington, D., 1989. Countervailing incentives in agency problems. Journal of Economic Theory 49, 294-313.

----------, 1991. Incentives for monitoring quality. RAND Journal of Economics 22 (3), 370-384

----------, 1995. Using markets to allocate pollution permits and other scarce resources rights under limited information. Journal of Public Economics 57, 431-435.

Maskin, E., Riley, J., 1985. Input versus output based incentive schemes. Journal of Public Economics 28, 1-23.

Maggi, G., Rodrígues-Clare, A., 1995. Costly distortion of information in agency problems. RAND Journal of Economics 26, 675-689.

Matutes, C., Régibeau, P., 1994. Compensation schemes and labor market competition: Piece rate versus wage rate. Journal of Economics and Management Strategy 3 (2), 325-53.

Melumad, N., Mookherjee, D., 1989. Delegation as commitment: the case of income tax audits. RAND Journal of Economics 20, 139-163.

Mirrlees, J, 1971. An exploration in the theory of optimum income taxation. Review of Economic Studies 38 (114), 175-208.

Mookherjee, D., Png, I., 1989. Optimal auditing, insurance and redistribution. Quarterly Journal of Economics CIV, 399-416.

New York Times, 2000, June 16. Asleep at the Books: A Fraud that went On and On and On. Section C, page 1.

Rochet, J.-C., Stole, L., 2000. The economics of multidimensional screening. Working paper University of Chicago.

Sappington, D., 1983. Optimal regulation of a multiproduct monopoly with unknown technological capabilities. *Bell Journal of Economics* 14, 453-463.

Schmutzler, A., Goulder, L., 1997. The choice between emission taxes and output taxes under imperfect monitoring. *Journal of Environmental Economics and Management* 32 (1), 51-64.

Singh, N., 1989. Theories of sharecropping. In: Bardhan, P. (Ed.), *The Economic Theory of Agrarian Institutions*. Clarendon, Oxford.

Strausz, R., 1997. Delegation of monitoring in a principal-agent relationship. *Review of Economic Studies* 64 (3), 337-357.

Swierzbinski, J., 1994. Guilty until proven innocent – regulation with costly and limited enforcement. *Journal of Environmental Economics and Management*, 27 (2), 127-146.

Wittman, D., 1977. Prior regulation versus post liability: The choice between input and output monitoring. *Journal of Legal Studies* VI (1), 193-211.

Figure 1: Effort of the Low Type as a Function of θ_2/θ_1