ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 25/26 — 1992

Threat-Based Implementation of Incentive Compatible Mechanisms

Liang ZOU *

ABSTRACT. – We investigate necessary and sufficient conditions for threat-based incentive mechanisms (TBIMs), an extension of the Mirrlees's schemes, to approximately eliminate moral hazard in a principal-agent relationship with both problems of moral hazard and adverse selection. Assuming normal distributions of output, a necessary condition is that the agent's type and effort do not affect the variance of the distribution. Sufficient conditions on the likelihood ratio of the distributions appear to be stronger than those used in the pure adverse-selection models, and have their particular implications.

Implémentation de mécanismes incitatifs par des menaces

RÉSUMÉ. – On examine, dans l'esprit de schémas introduits par Mirrlees, des conditions nécessaires et suffisantes pour que, dans une relation principal-agent comportant aléa moral et selection contraire, l'aléa moral soit éliminé approximativement par le recours à des mécanismes incitatifs fondés sur des menaces. Dans le cas d'un output stochastique distribué normalement, une condition nécessaire est que le type et l'effort de l'agent n'affectent pas la variance de la distribution. Des conditions suffisantes portent sur le rapport de vraisemblance; elles sont plus fortes que les conditions introduites traditionnellement pour résoudre des modèles de sélection contraire pure et comportent des conséquences spécifiques.

---

* L. Zou: Department of Finance, Faculty of Economics and Business Administration, University of Limburg, P.O. Box 616, MD Maastricht, The Netherlands. I would like to thank Claude d'Aspremont for helpful comments and suggestions. Of course, remaining errors are all mine.

1 Introduction

Recent research in the principal-agent literature has been successful in extending the solutions to the standard problems involving either pure adverse selection (AS) or pure moral hazard (MH) to the framework where both types of these problems (MH-AS) are present. [1] One of the interesting observations is that when the agent is income risk neutral, the AS solutions can in most cases be implemented via a family of linear (or quadratic) incentive contracts. That is, moral hazard does not really cause a problem even in the face of adverse selection. [2]

However, in the MH-AS models where the agent is risk averse, linear incentive mechanisms can no longer costlessly implement the AS-solutions – just as the case where, in the pure MH models with risk-averse agents, linear incentive contracts cannot be used to achieve a first-best solution (*e.g.* HOLMSTRÖM [1979] and SHAVELL [1979]). In the light of MIRRLEES's observation [1974], a natural question thus arises as to whether AS-solutions could still be implemented via other kinds of mechanisms if reasonably low costs of implementation are acceptable. The answer to this question has been widely believed to be positive (*e.g.* GUESNERIE *et al.* [1989]), and has indeed been confirmed for a class of problems in which the agent's output follows a special normal distribution with constant variance (ZOU [1989 *a*]). [3] Of course, this is a stronger condition than what one might expect as an extension of the MH situation in which MIRRLEES's schemes work. [4]

The difficulty of this extension arises from the fact that when the agent possesses private information, he has freedom to misreport the information and influence the principal's belief about the distribution of the output. This problem becomes extremely serious when the likelihood ratio of the output distribution is sensitive to the agent's hidden information or hidden effort. The purpose of this paper is to investigate some necessary and sufficient conditions for TBIMs to implement an AS-solution with negligible costs.

---

1. Amongst others, *see* MYERSON [1982], LAFFONT and TIROLE [1986], PICARD [1987], BARON and BESANKO [1987], PAGE [1989], and GUESNERIE *et al.* [1989].
2. For the case with a single agent, *see* LAFFONT and TIROLE [1986], BARON and BESANKO [1987], PICARD [1987], MCAFEE and MCMILLAN [1987], CAILLAUD *et al.* [1989]. For the case with multiple agents, *see* ZOU [1989 *b*], PICARD and REY [1990], and MCAFEE and MCMILLAN [1990].
3. Some trivial cases where the output distributions move with the agent's effort have also been shown to be free from moral hazard problem (*e.g.* GUESNERIE *et al.* [1989] and ZOU [1989 *a*]).
4. Regarding the distribution of output, MIRRLEES's schemes only require that its likelihood ratio is unbounded from below as the output approaches the lower bound of its support. See MIRRLEES [1974] and OSBAND [1987].

In the next section we present the model and define the relevant concepts. In Section 3 we show that under the assumption of normal distribution of output with mean equal to the agent's private effort, a necessary condition for asymptotically optimal TBIMs to exist is that the variance of the distribution be constant. In Section 4 we identify some conditions regarding the likelihood ratio of the distribution of output, and show that they are sufficient for TBIMs to work effectively in approximately eliminating moral hazard. The possible implications of these conditions are discussed and a number of examples satisfying these conditions are given. In Section 5 we summarize the paper, discuss possible justifications for our analysis, and sketch some further lines of research.

2 The Model

The setting is a principal-agent relationship, where the agent possesses two-dimensional private information represented by $\theta \in \Theta = [\underline{\theta}, \bar{\theta}]$ and $e \in A = [0, L]$; θ is the agent's "type", e is "effort", and Θ and A are the set of all possible types and effort levels respectively. The agent is a utility maximizer, whose utility is assumed to be $U(s) - V(e, \theta)$ ($U' > 0$, $U'' \leqslant 0$, $V_e > 0$, $V_{ee} > 0$, $V_\theta < 0$, and $V_{e\theta} < 0$), where s denotes some money equivalent. That is, for given s, e and θ, $U(s)$ denote the agent's utility level of income s, and $V(e, \theta)$ denote the cost (or disutility) of effort e for the agent having type θ. Assume that there is a publicly observable variable $x \in X$ which is a performance indicator of e, *i.e.* x conveys information about the actual level of e made by the agent. For interpretational convenience we call x the output.

Let $G(x|e, \theta)$ and $g(x|e, \theta)$ denote respectively the cumulative distribution and the density function of x, conditional on the level of effort and the type of the agent. The functional forms of $G(.)$ and $g(.)$ are assumed to be common knowledge.

By the "revelation principle" (*e.g.* MYERSON [1982]), we can limit our attention without loss of generality to the class of incentive compatible direct mechanisms in which the agent reports directly to the principal his true type and chooses the effort level recommended by the principal.

DEFINITION 1: When e is verifiable, an individually rational revelation mechanism (IRRM) is the class of vector functions $\{e^A(.), S^A(.)\}$, where $e^A(.) : \Theta \mapsto A$ is the effort level that the agent will be required to make if he reports θ and $S^A(.) : \Theta \mapsto R$ is an incentive contract specifying the payment to the agent as a function of the reported type, satisfying

$$\theta = arg \max_{\hat{\theta} \in \Theta} U(S^A(\hat{\theta})) - V(e^A(\hat{\theta}), \theta)), \qquad \forall \theta \in \Theta \tag{1}$$

$$U(S^A(\theta)) - V(e^A(\theta), \theta) \geqslant 0, \qquad \forall \theta \in \Theta \tag{2}$$

We shall limit our attention to implementing (continuously) differentiable incentive compatible mechanisms with $e^{A'}(\theta) > 0$ on $(\underline{\theta}, \bar{\theta})$. [5] From (1), such a mechanism should satisfy the first-order condition

$$U'(S^A(\theta))\,S^{A'}(\theta) - V_e(e^A(\theta), \theta)\,e^{A'}(\theta) = 0 \tag{3}$$

We also assume that $U(s) \to -\infty$ as $s \to s^*$ where $s^* \geq -\infty$. What is important here is not the actual level of s^*, but the existence of such an s^*.

DEFINITION 2: When x instead of e is verifiable, an incentive compatible direct mechanism (ICDM) is the class of vector functions $\{e(.), S(.,.)\}$ where $e(.) : \Theta \mapsto A$ is the effort level that the agent will be recommended to make if he reports θ, and $S(.,.) : \Theta \times X \mapsto R$ is an incentive contract specifying the payment to the agent as a function of the reported type and the observed output, satisfying

$$(\theta, e^A(\theta)) \in arg \max_{\substack{\hat{\theta} \in \Theta, \\ e \in A}} E_{x \mid (e, \theta)}[U(S(\hat{\theta}, x)) - V(e, \theta)], \qquad \forall \theta \in \Theta \tag{4}$$

$$E_{x \mid (e^A(\theta), \theta)}[U(S^A(\theta, x)) - V(e^A(\theta), \theta)] \geq 0. \qquad \forall \theta \in \Theta \tag{5}$$

where $E_{x \mid (e, \theta)}$ is the expectation operator over X, conditional on (e, θ).

Constraints (4) and (5) are the standard incentive compatibility and individual rationality conditions respectively. Note the difference between an IRRM and an ICDM. In the former mechanism the payment function depends only on the reported message because the agent's effort is perfectly observable, and the effort allocation $e(\theta)$ can be enforced given any reported θ. In the latter mechanism $e(\theta)$ is not enforceable because e is unobservable, and the realized output x plus report θ may not allow the principal to infer the actual level of effort. Therefore ICDMs have to deal with a compound incentive problem, concerning both the revelation of private information and the implementation of the effort allocations.

DEFINITION 3: A threat-based incentive mechanism (TBIM) is an ICDM $\tau = (e(\theta), S(\theta, x))$ in which the incentive contract $S(\theta, x)$ takes the form

$$\begin{aligned} S^T(\theta, x) &= R(\theta) \qquad x \geq T(\theta) \\ &= M(\theta) \qquad x < T(\theta) \end{aligned}$$

where $R(\theta) > M(\theta)$. We call such contracts the threat-based incentive contracts (TBICs).

Note that this is a natural extension of the MIRRLEES's schemes that are defined with (T, R, M) being constant variables (*see* MIRRLEES [1974]). To denote a particular TBIM, we shall often use the triple notation $(R(.), M(.), T(.)) : \Theta \mapsto R^3$, where the first, second and third element denote the reward payment, the penalty payment, and the minimal required

5. In fact, in our context, e^A being non-decreasing in θ is a necessary condition for incentive compatibility (*see* GUESNERIE and LAFFONT [1984]). Notice further that $e^{A'} > 0$ implies $S^{A'} > 0$ from the following first-order condition.

output (abbreviated as target) respectively. [6] Given a TBIM, the agent selects $\theta \in \Theta$ to settle a TBIC. When a particular θ is reported, $T(\theta)$ serves as the minimal target. Success in reaching $T(\theta)$ guarantees the agent a fixed reward $R(\theta)$, and failure in meeting $T(\theta)$ causes him to suffer a loss of $R(\theta) - M(\theta)$.

Given a TBIM τ, let π^{τ} denote the expected utility of agent θ reporting $\hat{\theta}$ and making effort e:

$$\pi^{\tau}(e, \hat{\theta}, \theta) = E_{x \mid (e, \theta)}[U(S(\hat{\theta}, x)) - V(e, \theta)]$$

DEFINITION 4 : An IRRM $(e^{A}(.), S^{A}(.))$ is said to be asymptotically implementable via ICDMs if for any $\delta > 0$, there exists an ICDM $(e^{A}(.), S(.,.))$ such that

$$\pi^{\tau}(e^{A}(\theta), \theta, \theta) \geq \pi^{\tau}(e, \hat{\theta}, \theta), \qquad \forall e \in A, \forall \theta, \hat{\theta} \in \Theta; \tag{6}$$

$$\pi^{\tau}(e^{A}(\theta), \theta, \theta) \geq U(S^{A}(\theta)) - V(e^{A}(\theta), \theta), \qquad \forall \theta \in \Theta; \tag{7}$$

$$E_{x \mid (e^{A}(\theta), \theta)} S(\theta, x) \leq S^{A}(\theta) + \delta, \qquad \forall \theta \in \Theta \tag{8}$$

Condition (6) is the incentive compatibility condition; (7) says that the agent is at least weakly better-off in $(e^{A}(.), S(.,.))$ than in $(e^{A}(.), S^{A}(.))$; and (8) requires that the principal's cost of implementing $(e^{A}(.), S^{A}(.))$ is no more than δ.

3 Using TBIMs to Implement an IRRM: Necessary Conditions

In this section, the two propositions are meant to illustrate some difficulties and limitations of using the incentive structure as specified in a TBIM for solving agency problems. Contrary to what one might think from MIRRLEES [1974], high penalties applied with small probabilities may not work as effectively in the MH-AS situations as in the MH situations, even when the output is normally distributed. Of course, our purpose is not to repudiate the use of TBIMs, since we also show in the next section that there exist typical cases where the TBIMs are effective.

PROPOSITION 1: Suppose that the agent is strictly risk averse, that $G(x \mid e, \theta)$ is normal with mean e and continuously differentiable variance $\sigma^{2}(\theta)$ with $\sigma > 0$ on Θ, and that there exists $\theta \in \Theta$ such that $\sigma'(\theta) \neq 0$.

6. More precisely, a TBIM should also contain a specification of the effort function. We let it drop out for notational convenience.

Then no IRRM$(e^A(.), S^A(.))$ is asymptotically implementable via TBIMs. [7]

Proof: Given a TBIM $\tau=((R(.), M(.), T(.))$, if the agent θ reports $\hat{\theta}$ and exerts effort e, his expected utility is

$$(9)\quad \pi^{\tau}(e, \hat{\theta}, \theta)=G(T(\hat{\theta})\,|\,e, \theta)\,U(M(\hat{\theta})) + (1-G(T(\hat{\theta})\,|\,e, \theta))\,U(R(\hat{\theta}))-V(e, \theta)$$

We need to verify that conditions (6), (7) and (8) are satisfied. First notice that for any $\delta>0$, from $U''<0$ and (7) and (8) we have

$$U(S^A+\delta)\geqslant U(GM+(1-G)R)>GU(M)+(1-G)U(R)\geqslant U(S^A),$$

which implies $U(GM+(1-G)R)-[GU(M)+(1-G)U(R)]\leqslant U'(S^A)\delta$. Thus as δ tends to zero, $U(GM+(1-G)R)-[GU(M)+(1-G)U(R)]$ must also tend to zero. Under strict risk aversion, this can happen only when G goes to 0 or 1. We shall focus on the case where G goes to zero. The other case corresponds to a reward-based incentive mechanism and may be analyzed similarly, although under risk aversion it is generally more costly to reward then to penalize (*see* OSBAND [1987]). Also notice that as δ tends to 0, to maintain conditions (7) and (8) $R(.)$ must tend to $S^A(.)$.

Condition (6) holds only if the following first and second order conditions are satisfied at $\hat{\theta}=\theta$ and $e=e^A(\theta)$ for all $\theta\in\Theta$:

$$(10)\quad \pi^{\tau}_e(e, \hat{\theta}, \theta)\Big|_{\substack{\hat{\theta}=\theta,\\ e=e^A(\theta)}} = G_e(T(\theta)\,|\,e^A(\theta), \theta)\,[U(M(\theta))-U(R(\theta))]-V_e(e^A(\theta), \theta)=0$$

$$(11)\quad \pi^{\tau}_{\theta}(e, \hat{\theta}, \theta)\Big|_{\substack{\hat{\theta}=\theta,\\ e=e^A(\theta)}} = g(T(\theta)\,|\,e^A(\theta), \theta)\,T'(\theta)\,(U(M(\theta))-U(R(\theta))) + G(T(\theta)\,|\,e^A(\theta), \theta)\,U'(M(\theta))\,M'(\theta) + (1-G(T(\theta)\,|\,e^A(\theta), \theta))\,U'(R)\,R'(\theta)=0$$

We show that there exist no solutions $(R(.), M(.), T(.))$ to these two equations for sufficiently low initial values of T. Differentiating (10) w.r.t. θ yields

$$(12)\quad U'(M)\,M'-U'(R)\,R'=\frac{(V_{ee}\,e^{A'}+V_{e\theta})\,G_e-V_e\,(G_{ee}\,e^{A'}+g_e\,T'+G_{e\theta})}{G_e^2}$$

Inserting (12) into (11) gives

$$(13)\quad T'(\theta)=\frac{U'(R)\,R'\,G_e^2+GG_e\,(V_{ee}\,e^{A'}+V_{e\theta})-V_e\,GG_{ee}\,e^{A'}-V_e\,GG_{e\theta}}{V_e\,[G\,g_e-g\,G_e]}$$

---

7. The differentiability assumptions are adopted for ease of analysis. Although exceptions might exist, we surmise that similar results hold for non-differentiable distribution functions.

For T negatively large, it is easy to verify that the terms in the right-hand side of equation (13) satisfy [8]

$$\left|\frac{GG_e(V_{ee}e^{A'}+V_{e\theta})}{V_e[Gg_e-gG_e]}\right|=O(|T|),$$

$$\left|\frac{U'(R)R'G_e^2}{V_e[Gg_e-gG_e]}\right|=O(T^2),$$

$$\left|\frac{V_eGG_{ee}e^{A'}}{V_e[Gg_e-gG_e]}\right|=O(T^2),$$

and

$$\left|\frac{-V_eGG_{e\theta}}{V_e[Gg_e-gG_e]}\right|=O(|T|^3)$$

as $T\to-\infty$. For instance, consider the last term in (13). By L'Hopital's rule, as $T\to-\infty$,

$$\frac{-V_eGG_{e\theta}}{V_e[Gg_e-gG_e](T-e)^3}=\frac{-G\left[\frac{(T-e)^2}{\sigma}-1\right]\frac{\sigma'}{\sigma}}{\left[\frac{T-e}{\sigma^2}G+g\right](T-e)^3}$$

$$\to\frac{G\frac{\sigma'}{\sigma^2}}{\left[\frac{T-e}{\sigma^2}G+g\right](T-e)}$$

$$\to\frac{g\frac{\sigma'}{\sigma^2}}{2\frac{T-e}{\sigma^2}G+g}$$

$$\to-\frac{\sigma'}{\sigma^2}$$

By the assumption that σ' is continuous and $\sigma'(.)\neq0$ for some values of θ, there exists an interval $[\theta_1,\theta_2]$ on which $\sigma'>0$ or on which $\sigma'<0$.

Consider first the case where $\sigma'>0$ on $[\theta_1,\theta_2]$. From the above analysis, there exists T such that for all solutions to (13) that satisfy $T(.)<T$ on Θ, the following inequality must hold

$$T'(\theta)\geqslant\frac{-\sigma'(\theta)}{2\sigma^2(\theta)}T^3(\theta)>0,\qquad\forall\theta\in[\theta_1,\theta_2] \tag{14}$$

---

8. As usual, a function $F(x)$ satisfies the expression $|F(x)|=O(x)$ as $x\to\infty$ if and only if $|F(x)|/x\to c\in(0,\infty)$ as $x\to\infty$, where c is a constant (independant of x).

Choose the initial value $T(\theta_2)=T_0<T$. Multiply both sides in (14) by $-2/T^3(\theta)$ and integrate over $[\theta, \theta_2]$ we derive

$$\frac{1}{T^2(\theta)} \leqslant \frac{1}{T_0^2} - \int_{\theta}^{\theta_2} \frac{\sigma'(\theta)}{\sigma^2(\theta)} d\theta \tag{15}$$

Obviously, for T_0 negatively large such that $1/T_0^2 < \int_{\theta_1}^{\theta_2} \frac{\sigma'(\theta)}{\sigma^2(\theta)} d\theta$, the inequality in (15) cannot hold for all $\theta \in [\theta_1, \theta_2]$. Thus no solution to (13) exists on the whole range of Θ when T is negatively large. The case where $\sigma'<0$ can be proved similarly by choosing $T(\theta_1)=T_0$. □

A similar result holds when the agent's effort affects the variance of the distribution.

PROPOSITION 2: Suppose the agent is strictly risk averse, and $G(x|e)$ is normal with mean e and continuously differentiable variance $\sigma^2(e)$ with $\sigma>0$ and $\sigma'(e)\neq 0$ on [0, L], then no IRRM $(e^A(.), S^A(.))$ is asymptotically implementable via TBIMs.

Proof: We only need to replace $G_{e\theta}=G_{e\sigma}\sigma'(\theta)$ in the proof of Proposition 1 by $G_{e\theta}=G_{e\sigma}\sigma'(e^A(\theta))e^{A'}(\theta)$, and the rest of the argument will go through exactly the same way as in the proof of Proposition 1. □

From these two propositions we see that it is necessary to have a constant variance, when the distribution of output is normal, in order for TBIMs to approximately eliminate moral hazard under adverse selection. The insight here is that when the agent has private information that influences, either directly (as in Proposition 1) or indirectly (as in Proposition 2), the principal's belief about the distribution function of the output, a TBIM may have to be extremely discriminating to the different types of agents. For the normal distributions where the variance depends on hidden information or hidden effort, the minimum target $T(\theta)$ has to dramatically vary with θ (recollect that $|T'|=O(|T^3|)$). As a result, there exists no feasible target function $T(.)$ satisfying low enough initial conditions that can be defined on the whole interval of $[\underline{\theta}, \bar{\theta}]$.

Indeed, this difficulty would not arise when the agent is risk neutral, in which case, if linear incentive schemes are used, the variance of the distribution will have no effect on the expected utilities of the players. Of course, global risk neutrality in income without institutional contraints (such as limited liability) is a very strong assumption, and has been ruled out in our model. [9]

---

9. An unverified conjecture is that under limited liability and risk neutral preferences, the agent can be better motivated through a family of target incentive systems (*see* ZOU [1991]), which is a more general form of TBIMs with the reward or penalty levels dependent on x as well.

4 Using TBIMs to Implement an IRRM: Sufficient Conditions

The impossibility result in the preceding section, of course, does not in general invalidate the use of threat-based incentive mechanisms; it only serves to strike out a class of situations in which the use of threats does not work as effectively as we might wish. In this section we investigate sufficient conditions that ensure that TBIMs can be used to asymptotically implement an IRRM. Later on we will prove a positive result, but first let us discuss the conditions that are posed in the following assumptions.

ASSUMPTIONS: Suppose the distribution of x can be written as $G(x|e)$, *i.e.* it is not directly dependent on θ. Further, for $G(T|e)$ (viewing T as a variable), the function satisfies [10]

$$\text{(16)} \qquad G_e/G \to -\infty \quad \text{and} \quad G_{ee} \geqslant 0 \quad \text{as} \quad T \to -\infty;$$

$$\text{(17)} \qquad \frac{\left|\frac{d}{de}\left(\frac{G_e}{G}\right)\right|}{\left|\frac{d}{dT}\left(\frac{G_e}{G}\right)\right|} \leqslant O(|T|) \quad \text{as} \quad T \to -\infty;$$

$$\text{(18)} \qquad \left|\frac{d}{dT}\ln\left(\frac{-G_e}{G}\right)\right| \geqslant O\left(\frac{1}{|T|}\right) \quad \text{as} \quad T \to -\infty;$$

In the pure MH context, only Assumption (16) on the output distribution is needed for a TBIC (or MIRRLEES's scheme) to asymptotically implement a first-best contract, whereas here stronger conditions are required. Clearly, all the above three conditions are about the likelihood ratio G_e/G of the distribution function, where e is viewed as an unknown parameter. Since this ratio is the derivative of the likelihood function ln G w.r.t. e, the larger $|G_e/G|$ is, the better the principal is able to discern whether the right effort is not made (*see* HOLMSTRÖM [1979]). Assumption (16) is thus seen as allowing the principal to arbitrarily increase his ability to discern shirking by choosing an arbitrarily low target.

A complicating factor here, however, is that in a TBIM, both T and e are functions of the agent's report θ and, hence, subject to the agent's manipulation. Assumptions (17) and (18) might thus be viewed as restrictions that limit the agent's ability to manipulate the likelihood ratio. Because of the complicated expressions, we find it hard to provide a satisfactory interpretation of the assumptions (17) and (18), although a large class of distribution functions do meet all these assumptions. The following are some examples.

10. Here, a function $F(x)$ satisfies the expression $|F(x)| \geqslant (\leqslant) O(x)$ as $x \to \infty$ if and only if $|F(x)|/x \to c \in (0, \infty] (c \in [0, \infty))$ as $x \to \infty$, where c is a constant independent of x.

EXAMPLE 1: The normal distributions with mean e and with a constant variance satisfy the assumptions (16)-(18) (*see* ZOU [1989*a*]). [11]

EXAMPLE 2: Consider the "reversed" exponential distribution

$$G(T|e)=\exp\left\{\frac{T-b}{b-e}\right\}=(b-e)g(T|e), \qquad T\in(-\infty, b].$$

We have $G_e=\dfrac{T-b}{b-e}g(T|e)$, $g_e=\dfrac{g}{b-e}\left(1+\dfrac{T-b}{b-e}\right)$. As $T\to-\infty$,

$$\frac{G_e}{G}=\frac{T-b}{(b-e)^2}\to-\infty, \qquad \frac{G_{ee}}{G_e}=\frac{1}{b-e}+\frac{g_e}{g}\to-\infty,$$

$$\frac{|(d/de)(G_e/G)|}{|(d/dT)(G_e/G)|}=\frac{2(b-T)}{b-e}=O(|T|), \qquad \left|\frac{d}{dT}\ln\left(\frac{-G_e}{G}\right)\right|=\frac{1}{b-e}>O\left(\frac{1}{|T|}\right).$$

It thus meets the assumptions (16)-(18).

EXAMPLE 3: Suppose

$$G(T|e)=\exp\left\{-\frac{(T-b)^2}{2(b-e)}\right\}=\frac{b-e}{b-T}g(T|e), \qquad T\in(-\infty, b].$$

We have $G_e=\dfrac{T-b}{2(b-e)}g(T|e)$, $g_e=\dfrac{g}{b-e}\left(1-\dfrac{(T-e)^2}{2(b-e)}\right)$. As $T\to-\infty$,

$$\frac{G_e}{G}=-\frac{(T-b)^2}{2(b-e)^2}\to-\infty, \qquad \frac{G_{ee}}{G_e}=\frac{2}{b-e}+\frac{G_e}{G}\to-\infty,$$

$$\frac{|(d/de)(G_e/G)|}{|(d/dT)(G_e/G)|}=\frac{(b-T)}{b-e}=O(|T|), \qquad \left|\frac{d}{dT}\ln\left(\frac{-G_e}{G}\right)\right|=\frac{1}{b-e}>O\left(\frac{1}{|T|}\right).$$

Thus the assumptions (16)-(18) are satisfied for this distribution function.

EXAMPLE 4: Now let the right bound of the distribution be influenced by the agent, in a way such that $G(T|e)=\exp\left\{-\dfrac{(T-e)^2}{2}\right\}=\dfrac{1}{e-T}g(T|e)$, $T\in(-\infty, e]$. We have $G_e=-g(T|e)$, $g_e=\dfrac{[1-(T-e)^2]}{e-T}g$. As $T\to-\infty$,

$$\frac{G_e}{G}=T-e\to-\infty, \qquad \frac{G_{ee}}{G_e}=\frac{g_e}{g}\to-\infty, \qquad \frac{|(d/de)(G_e/G)|}{|(d/dT)(G_e/G)|}=1<O(|T|),$$

$\left|\dfrac{d}{dT}\ln\left(\dfrac{-G_e}{G}\right)\right|=\dfrac{1}{e-T}=O\left(\dfrac{1}{|T|}\right)$. Thus this distribution function also meets the conditions in the assumptions (16)-(18).

---

11. It is easy to verify that the normal distribution with mean e and with standard deviation $\sigma(e)$, $\sigma'(e)\neq 0$, does not satisfy condition (18).

PROPOSITION 4: Suppose the distribution of x satisfies the assumptions (16)-(18). And suppose the given IRRM (e^A, S^A) satisfies $V_{ee}e^{A'} + V_{e\theta} > 0$. Then (e^A, S^A) is asymptotically implementable via TBIMs.

Proof: [12] Let an IRRM (e^A, S^A) be given. We need to show that for any $\delta > 0$ there exists a TBIM $\tau = ((R(.), M(.), T(.))$ that satisfies conditions (6)-(8). Again, let $\pi^\tau(e, \hat{\theta}, \theta)$ denote the θ agent's utility who reports $\hat{\theta}$ and exerts effort e.

Condition (6) holds if the first- and second-order conditions are satisfied at $\hat{\theta} = \theta$ and $e = e^A(\theta)$ for all $\theta \in \Theta$:

$$(19)\quad \pi^\tau_e(e, \hat{\theta}, \theta)\Big|_{\substack{\hat{\theta}=\theta,\\ e=e^A(\theta)}} = G_e(T(\theta)\,|\,e^A(\theta))[U(M(\theta)) - U(R(\theta))] - V_e(e^A(\theta), \theta) = 0,$$

$$(20)\quad \pi^\tau_{\hat{\theta}}(e, \hat{\theta}, \theta)\Big|_{\substack{\hat{\theta}=\theta,\\ e=e^A(\theta)}} = g(T(\theta)\,|\,e^A(\theta))\,T'(\theta)\,(U(M(\theta)) - U(R(\theta)))) + G(T(\theta)\,|\,e^A(\theta))\,U'(M(\theta))\,M'(\theta) + (1 - G(T(\theta)\,|\,e^A(\theta)))\,U'(R)\,R'(\theta) = 0,$$

and

$$(21)\quad \pi^\tau_{ee} < 0,\ \pi^\tau_{\theta\theta} < 0 \qquad \text{and} \qquad \pi^\tau_{ee}\pi^\tau_{\theta\theta} - \pi^{\tau 2}_{e\theta} > 0.$$

By differentiating $\pi^\tau_e(e^A(\theta), \theta, \theta) = 0$ and $\pi^\tau_{\hat{\theta}}(e^A(\theta), \theta, \theta) = 0$ as an identity in θ, we have $\pi^\tau_{ee}e^{A'} + \pi^\tau_{e\hat{\theta}} - V_{e\theta} = 0$ ($V_{e\theta} = -\pi^\tau_{e\theta}$) and $\pi^\tau_{\hat{\theta}e}e^{A'} + \pi^\tau_{\hat{\theta}\hat{\theta}} = 0$ ($\pi^\tau_{\hat{\theta}\hat{\theta}} = 0$), which allow us to write, for T sufficiently low,

$$(22)\quad \pi^\tau_{ee} = G_{ee}(U(M) - U(R)) - V_{ee} = G_{ee}V_e/G_e - V_{ee} < 0$$

$$(23)\quad \pi^\tau_{e\hat{\theta}} = V_{e\theta} - \pi^\tau_{ee}e^{A'} > 0$$

$$(24)\quad \pi^\tau_{\hat{\theta}\hat{\theta}} = -\pi^\tau_{e\hat{\theta}}e^{A'} < 0$$

and

$$(25)\quad \pi^\tau_{ee}\pi^\tau_{\hat{\theta}\hat{\theta}} - \pi^{\tau 2}_{e\hat{\theta}} = -V_{e\theta}\pi^\tau_{e\hat{\theta}} > 0$$

Next consider (7). If we set $R(\theta) = S^A(\theta) + \delta$, as we do, using (19) we have

$$(26)\quad \begin{aligned}\pi^\tau(e^A(\theta), \theta, \theta) &= U(S^A(\theta) + \delta) - V(e^A(\theta), \theta) \\ &\quad + V_e(e^A(\theta), \theta)\,G(T\,|\,e^A(\theta))/G_e(T\,|\,e^A(\theta)) \\ &\geqslant U(S^A(\theta)) - V(e_A(\theta), \theta) + r\delta \\ &\quad + V_e(e_A(\theta), \theta)\,G(T\,|\,e_A(\theta))/G_e(T\,|\,e_A(\theta))\end{aligned}$$

where $r = U'(S^A(\bar{\theta}) + \delta)$. The last term on the right-hand side V_eG/G_e is negative and tends to zero as T tends to $-\infty$. Therefore for whatever

12. Most of this proof is adapted from ZOU [1989*a*], in which only the normal distribution is considered.

small $\delta > 0$, it is always possible to choose a number T such that

$$r\delta + V_e G(T \mid e_A(\theta)/G_e(T \mid e^A(\theta)) \geq 0, \tag{27}$$

i.e. $\pi^\tau(e^A(\theta), \theta, \theta)) \geq U(S^A(\theta)) - V(e^A(\theta), \theta)$ for all $T(.)$ and $\theta \in \Theta$ such that $T(\theta) \leq T$.

Last consider (8). Given $R(\theta) = S^A(\theta) + \delta$ and $(T(\theta), M(\theta))$ satisfying (19)-(27), the expected payment by the principal

$$E_{x \mid e^A(\theta)} S^T(\theta, x) = GM(\theta) + (1-G) S^A(\theta) + \delta) \leq S^A(\theta) + \delta \tag{28}$$

since $M(\theta) < S^A(\theta) + \delta$. Thus (8) is also satisfied.

All that remains now is to check the existence of such $(T(\theta), M(\theta))$ which should satisfy (19)-(27) for all $\theta \in \Theta$. First note that M can be solved as a function of T and θ from (19):

$$M(\theta, T(\theta)) = U^{-1}[V_e(e^A(\theta), \theta)/G_e(T(\theta) \mid e^A(\theta)) + U(S^A(\theta) + \delta)] \tag{29}$$

Differentiating (29) w.r.t. θ yields

$$U'(M) M' = \frac{(V_{ee} e^{A'} + V_{e\theta}) G_e - V_e G_{ee} e^{A'} - V_e g_e T'}{G_e^2} + U'(S^A + \delta) S^{A'} \tag{30}$$

Inserting (30) into (20) gives

$$T'(\theta) = \frac{U'(S^A + \delta) S^{A'} G_e^2 + GG_e(V_{ee} e^{A'} + V_{e\theta}) - V_e GG_{ee} e^{A'}}{V_e[G g_e - g G_e]} \tag{31}$$

Denote the right-hand side of (31) by $\Psi(\theta, T)$, which is continuously differentiable in T and in θ. We can rewrite

$$\Psi(\theta, T) = \frac{-R_A(\xi) U'(\xi) S^{A'} G_e^2 \delta + GG_e(V_{ee} S^{A'} + V_{e\theta}) + V_e(G_e^2 - GG_{ee}) e^{A'}}{V_e[G g_e - g G_e]}$$

where $S^A < \xi < S^A + \delta$, $R_A(\xi) = -U''(\xi)/U'(\xi)$. The above equality results from the Mean-Value Theorem and the fact that $U'(S^A) S^{A'} = V_e e^{A'}$ (*see* (3)).

Note that condition (27) implies $-G_e \delta \geq V_e G/r$. It follows that

$$\Psi(\theta, T) \geq \frac{G_e G[R_A(\xi) U'(\xi) S^{A'} V_e/r + V_{ee} e^{A'} + V_{e\theta}]}{V_e[G g_e - g G_e]} - \left| \frac{(G_e^2 - GG_{ee}) e^{A'}}{G g_e - g G_e} \right|$$

$$\geq \frac{G_e G}{G g_e - g G_e} B_1 - \left| \frac{G_e^2 - GG_{ee}}{G g_e - g G_e} \right| B_1$$

where

$$B_1 = \sup_{\theta \in \Theta} \sup_{\xi \in [S^A(\theta), S^A(\theta) + \delta]} \left(\frac{|(R_A(\xi) U'(\xi) S^{A'} V_e/r + V_{ee} e^{A'} + V_{e\theta})|}{V_e} + |e^A(\theta)| \right) < \infty$$

Moreover, we have

$$\Psi(\theta, T) \leqslant -\frac{G_e G}{G g_e - g G_e}\left|\frac{V_{ee} e^{A'} + V_{e\theta}}{V_e}\right| + \left|\frac{(G_e^2 - G G_{ee}) e^{A'}}{G g_e - g G_e}\right|$$

$$\leqslant -\frac{G_e G}{G g_e - g G_e} \cdot B_2 + \left|\frac{G_e^2 - G G_{ee}}{G g_e - g G_e}\right| B_2$$

where

$$B_2 = \sup_{\theta \in \Theta}\left(\frac{|(V_{ee} e^{A'} + V_{e\theta})|}{V_e} + |e^A(\theta)|\right) < \infty$$

Finally, by the assumptions (17) and (18),

$$\left|\frac{G_e^2 - G G_{ee}}{G g_e - g G_e}\right| = \frac{\left|\frac{d}{de}\left(\frac{G_e}{G}\right)\right|}{\left|\frac{d}{dT}\left(\frac{G_e}{G}\right)\right|} \leqslant O(|T|)$$

$$\left|\frac{G_e G}{G g_e - g G_e}\right| = \frac{1}{\left|\frac{d}{dT} \ln\left(\frac{-G_e}{G}\right)\right|} \leqslant O(|T|)$$

Thus, for $-T$ large enough, we must have

$$BT(\theta) \leqslant T'(\theta) \leqslant -BT(\theta), \qquad \forall T(\theta) \leqslant T, \qquad \forall \theta \in \Theta,$$

for some $B > 0$. The solution to the differential equations $T' = BT$ and $T' = -BT$ with the initial condition $T(\underline{\theta}) = T_0$ are $T(\theta) = T_0 \exp\{B(\theta - \underline{\theta})\}$ and $T(\theta) = T_0 \exp\{-B(\theta - \underline{\theta})\}$ respectively. Consequently, the solution to (31) is bounded between $T_0 \exp\{B(\theta - \underline{\theta})\}$ and $T_0 \exp\{-B(\theta - \underline{\theta})\}$, and therefore must be defined on the whole interval $[\underline{\theta}, \bar{\theta}]$ (*see* HUREWICZ, Theorem 12). Finally, given $T(\theta)$, $M(\theta)$ is determined from (29). □

It is worth remarking that the assumption on the IRRM (e^A, S^A), $V_{ee} e^{A'} + V_{e\theta} > 0$, is only needed for the second-order conditions (22)-(25), it can be replaced by a condition $G_{ee}/G_e \geqslant 0$ for low enough T. In the above examples, we have shown that all the illustrated distribution functions satisfy this condition (in fact, $G_{ee}/G_e \to -\infty$ as $T \to -\infty$).

5 Concluding Remarks

We have examined some necessary and sufficient conditions for using TBIMs to approximately implement an IRRM in a principal-agent relationship involving both hidden information and hidden effort, and risk aversion

of the agent. The assumptions regarding the likelihood ratio of the distribution function of output, under which MIRRLEES's schemes work effectively in the pure moral hazard settings, are not sufficient to ensure that TBIMs would asymptotically eliminate moral hazard (Propositions 1 and 2). Under the assumption that the distribution of output is normal, it is necessary that the agent's private information and effort do not affect the variance of the distribution. The reason is that when the agent's information or effort affects the variance, the likelihood ratio of the distribution can become extremely sensitive to the agent's report for low output levels. Consequently, the penalty payment level and the minimum target have to be set in a way that they are also sensitive enough to the agent's report, in order to induce truthful revelation. But this cannot in general be done because the derivative of these variables w.r.t. the type reports quickly diverges to infinity and, thus, they cannot be defined on the whole interval of the type parameters.

However, there are also situations where TBIMs do work effectively. In Proposition 3, a set of conditions is identified under which an IRRM can be asymptotically implemented via a family of TBIMs. These conditions, though cumbersome to allow straightforward interpretations, do provide some insight into where and how the agent's capability of manipulating the principal's belief does not entail significant moral hazard costs. Broadly speaking, apart from the standard assumption that the likelihood ratio goes to minus infinity as output approaches to its lower bound, we need to strengthen the principal's ability to manipulate this ratio at the left tail of the distribution function [condition (18)], and limit the agent's ability to do so as compared with the principal's [condition (17)].

Our focus in this paper has been on the implementation of incentive compatible mechanisms, and not on the design of optimal mechanisms. But it has been observed that if the agent's type-parameter θ does not directly influence the distribution function of output, an ICDM that asymptotically implements an optimal IRRM is also approximately optimal (*see* ZOU [1989*a*]). Thus, Proposition 3 also implies that the TBIMs can be asymptotically optimal under the specified assumptions.

If the output distribution depends also on θ, the principal might be able to use the observation of output to infer not only information about e, but also information about θ. In this case, the observation of e is not necessarily superior to the observation of x, thus even if an ICDM implements the optimal IRRM derived under effort verifiability, it might still be suboptimal. [13] However, there are also possible situations in which implementation of IRRMs may be justified even if the distribution of output depends directly on θ. For example,

– The given IRRM may already be the first-best. Think of Groves mechanisms, for instance, where a benevolent principal maximizes the sum of many (independent) agents' utilities under effort observation (*see* ZOU [1989*b*]).

---

13. *See* a similar remark in GUESNERIE *et al.* [1989].

– The principal may be just an intermediary whose job is to see that an incentive compatible mechanism is implemented at the lowest possible cost. He has no direct personal interests in the performance of the agent.

– The principal may be primarily concerned in having a decision carried out in each corresponding state of nature, here the effort function $e^A(\theta)$. At the same time he is under a tight budget constraint.

– The principal may be a planner who coordinates different activities. An IRRM may be globally desirable even if it is not locally optimal as to the specific relationship between the principal and the agent of concern.

– Understanding the problem of implementation of IRRMs is important for designing optimal incentive mechanisms in more complicated situations, such as in the dynamic situations where mechanisms used in earlier periods are not necessarily optimal in the short-run.

One might be tempted as well to examine how **reward**-based incentive mechanisms (RBIMs) would work at the other extreme. Namely to offer the agent an astronomically high price for fulfilment of a terrific target and pay him a normal constant salary otherwise. It is usually more difficult to use RBIMs to implement an IRRM because when the agent is risk averse, it is more costly to motivate than to menace (*see* OSBAND [1987). In fact, there are more problems with RBIMs because, to assign a small probability to the event of rewarding, one must look at the right-tail of the output distribution. The necessary second-order conditions are easily violated for most commonly used distribution functions.

The impossibility result in Propositions 1 and 2 also motivate us to think of a related topic for further research, that is, to what extent the incentive power of a TBIM can be extended if the levels of rewards or penalties are allowed to depend on output as well. This would be of particular interest when the players' utilities are bounded because of institutional contraints or of the limited liabilities.

• References

BARON, D. P. and BESANKO, D. (1987). – "Monitoring, Moral Hazard, Asymmetric Information, and Risk Sharing in Procurement Contracting", Discussion Paper, Stanford University.

CAILLAUD, B., GUESNERIE, R. and REY, P. (1987). – "Noisy Observation in Adverse Selection Models", Document de Travail No. 8714, C.E.R.A.S., Paris.

GUESNERIE, R. and LAFFONT, J.-J. (1984). – "A Complete Solution to a Class of Principal-Agent Problems with an Application to the Control of a Self-Managed Firm", *Journal of Public Economics*, 25, pp. 329-369.

GUESNERIE, R., PICARD, P. and REY, P. (1989). – "Adverse Selection and Moral Hazard with Risk Neutral Agents", *European Economic Review*, 33, pp. 807-823.

HOLMSTRÖM, B. (1979). – "Moral Hazard and Observability", *Bell Journal of Economics*, 10, pp. 74-91.

HUREWICZ, W. (1970). – "Lectures on Ordinary Differential Equations", M.I.T. Press, 2nd Edition, March.

LAFFONT, J.-J. and TIROLE, J. (1986). – "Using Cost Observation to Regulate Firms", *Journal of Political Economy*, Vol. 94, pp. 614-641.

MCAFEE, R. P. and MCMILLAN, J. (1987). – "Competition for Agency Contracts", *Rand Journal of Economics*, Vol. 18, No. 2, pp. 287-307.

MCAFEE, R. P. and MCMILLAN, J. (1990). – "Optimal Contracts for Teams", Working Paper, University of Western Ontario and University of California.

MIRRLEES, J. A. (1974). – "Notes on Welfare Economics, Information, and Uncertainty", in BALCH, MCFADDEN and WU, eds., *Essays on Economic Behaviour under Uncertainty*, Amsterdam: North Holland Publishing Co.

MYERSON, R. B. (1982). – "Optimal Coordination Mechanisms in Generalized Principal-Agent Problems", *Journal of Mathematical Economics*, 10, pp. 67-81.

OSBAND, K. (1987). – "Speak Softly, but Carry a Big Stick: On Optimal Targets under Moral Hazard", *Journal of comparative Economics*, 11, pp. 584-595.

PAGE Jr., F. H. (1989). – "Mechanism Design for General Screening Problems with Moral Hazard", Working Paper, Department of Finance, Indiana University.

PICARD, P. (1987). – "On the Design of Incentive Schemes under Moral Hazard and Adverse Selection", *Journal of Public Economics*, 33, pp. 305-331.

PICARD, P. and REY, P. (1990).– "Incentives in Cooperative Research and Development, in P. CHAMPSAUR *and Ali*, ed.: *Essays in Honor of Edmond Malinvaud*, (MIT Press).

SHAVELL, S. (1979). – "Risk Sharing and Incentives in the Principal and Agent Relationship", *Bell Journal of Economics*, 10, pp. 55-73.

ZOU, L. (1989 *a*). – "Threat-Based Incentive Mechanisms under Moral Hazard and Adverse Selection", CORE Discussion Paper 8909, forthcoming in the March 1992 issue in *Journal of Comparative Economics*.

ZOU, L. (1989 *b*). – "Ownership Structure and Efficiency: An Incentive Mechanism Approach", CentER Discussion Paper No. 8955, Tilburg University, The Netherlands.

ZOU, L. (1991). – "The Target-Incentive System vs. the Price-Incentive System under Adverse Selection and the Ratchet Effect", *Journal of Public Economics*, 46, pp. 51-89.