Partially Verifiable Information and Mechanism Design

Author(s): Jerry R. Green and Jean-Jacques Laffont

Source: *The Review of Economic Studies*, Jul., 1986, Vol. 53, No. 3 (Jul., 1986), pp. 447-456

Published by: Oxford University Press

Stable URL: https://www.jstor.org/stable/2297639

Review of Economic Studies (1986) LIII, 447-456
0034-6527/86/00300447$02.00

Partially Verifiable Information and Mechanism Design

JERRY R. GREEN
Harvard University

and

JEAN-JACQUES LAFFONT
Université des Sciences Sociales de Toulouse and EHESS

In a principal-agent model with adverse selection, we study the implementation of social choice functions when the agent's message space is a correspondence which depends on this true characteristic. We characterize such correspondence for which the Revelation Principle is valid.

1. INTRODUCTION

Organizations function by giving their members some discretionary power. As a check on their freedom of choice, each agent is often held accountable for his decisions. He is required to have a credible explanation for the action he has chosen. Formally speaking, the set of allowable choices for the agent varies with his actual state of knowledge of the system.

Examples of this type of situation abound. Financial reports and tax returns may be distorted slightly, but auditing precludes excessive misreporting. Business decisions may be made to suit objectives other than that of the organization as a whole, but gross neglect of responsibility to the organization can be detected and punished. Income tax returns must not contradict easily observable elements of the taxpayer's lifestyle.

In this paper we present a model of two-person organizations in which these considerations are important. *We ask whether the restrictions on fallacious statements inherent in the system are sufficient to achieve the goals of the organization. We show that the optimal way of eliciting the agent's response may not be to establish incentives for truthtelling in all instances.* In our model there is one player, the "agent", who observes the state of the economic system. Another player, the "principal", takes an action based on the information presented to him by the agent. This class of models is often referred to as the principal-agent problem. In the extensive and growing literature on this problem, it has always been assumed that the agent could lie to the principal, and therefore that he had to be properly motivated to act at least partially in the principal's interest.

Our formal way of modelling the partial verifiability of information is to introduce the restriction that the agent's responses must lie in a set $M(\theta)$ that varies with the true state θ. It is assumed that this set always admits the possibility of responding with the true state. The way in which the allowable response set varies with the true state is known to the principal. In other respects we retain the usual principal-agent formalization. In particular, the principal is the Stackelberg leader in this game. He can commit himself to choose a collective action as a function of the message transmitted by the agent.

Whether the variation of the message space with the true observation is purely technological, or whether it is induced by the severity of potential actions of the principal, is largely a matter of interpretation. The important part is that the implementability of any collective decision rule is determined by the interaction of the allowable messages and the associated actions. If the messages are restricted in some circumstances, there is an enhanced potential to implement mutually beneficial courses of action. The goal of this paper is to develop a theory of incentive compatibility and implementability for the variable message space case. It is grounded in the observation that although lying is a problem within most organizations, there is only a limited range of distortions against which the system must guard itself.

In Section 2 we describe formally the model and define our notion of implementability. Section 3 is devoted to the proof of a characterization theorem giving a necessary and sufficient condition for a form of the Revelation Principle to hold in our framework. The notion of implementation we propose is then discussed. Section 4 presents some comparative statics results of the implementable set of social choice functions with respect to the correspondence $M(\cdot)$. Examples of the usefulness of the mechanisms with partial verification of information are given in Section 5.

2. THE MODEL

We consider a principal–agent problem. The agent's utility function depends on a parameter, or characteristic, $\theta \in \Theta$ and on a decision $x \in X$. We denote this utility by $u(x, \theta)$. He observes θ and then transmits information to the principal in a manner described below. The principal chooses an action as a function of this information.

Because the principal acts as a Stackelberg leader, fixing the dependence of x on the transmission in an immutable way before θ is observed, the principal's problem can be expressed equivalently as a delegation problem. The agent is offered the choice of actions from a limited menu. The principal's role is thus rather passive: the selection of this menu.

The crucial feature of our model is that the space of possible messages for an agent with characteristic θ is a subset of Θ which varies with θ. Let $M : \Theta \to \Theta$ be the correspondence determining the admissible messages. That is, for each θ, $M(\theta) \subseteq \Theta$ is the set of messages to which his transmission is restricted and we always assume that $\theta \in M(\theta)$ for all $\theta \in \Theta$. Clearly, the analysis is therefore restricted to direct mechanisms (see, however, Section 4 for an extension).

Definition 1. A mechanism $(M(\cdot), g)$ consists of a correspondence $M : \Theta \to \Theta$ such that $\theta \in M(\theta)$ for all $\theta \in \Theta$, and an outcome function $g : \Theta \to X$.

Given the correspondence $M(\cdot)$, the outcome function g induces *a*[1] *response rule* $\phi_g : \Theta \to \Theta$ defined by

$$\phi_g(\theta) \in \arg\max_{m \in M(\theta)} u(g(m), \theta).$$

Our goal is to study the class of social choice functions f from Θ into X that can be achieved despite the asymmetry of information between the two players.

Definition 2. A social choice function $f : \Theta \to X$ is $M(\cdot)$-*implementable* iff there exists an outcome function $g : \Theta \to X$ such that:

$$g(\phi_g(\theta)) = f(\theta) \quad \text{for any } \theta \text{ in } \Theta$$

where $\phi_g(\cdot)$ is an induced response rule.

Definition 3. A social choice function $f:\Theta\to X$ is *truthfully* $M(\cdot)$*-implementable* iff there exists an outcome function $g^*:\Theta\to X$ such that, for any θ in Θ,

$$g^*(\phi_{g^*}(\theta))=f(\theta)$$

and

$$\phi_{g^*}(\theta)=\theta$$

where ϕ_{g^*} is an induced response rule.

In the traditional principal-agent literature, $M(\theta)=\Theta$ for any θ in Θ. Then, a straightforward but important result of incentives theory, known as the revelation principle,[2] states that any M-implementable social choice function is truthfully implementable.

In the next section we characterize the message space correspondences $M(\cdot)$ for which the revelation principle is valid. We will then discuss the interpretation to be given to $M(\cdot)$-implementability.

3. A CHARACTERIZATION THEOREM

We establish that the following condition is both necessary and sufficient for the equivalence of the set of $M(\cdot)$-implementable social choice functions to the set of truthfully $M(\cdot)$-implementable social choice functions. In the presence of this condition, the principal's problem can be written as a constrained optimization problem in which the truth-telling restrictions operate as constraints. This is the usual form in which this problem has been solved in the principal-agent literature. When this condition does not hold, its necessity implies that the solution to such a constrained optimization may fail to identify the true solution to the principal's problem.

Nested Range Condition (NRC). For any three distinct elements θ_1, θ_2, θ_3 in Θ, if $\theta_2\in M(\theta_1)$ and $\theta_3\in M(\theta_2)$, then $\theta_3\in M(\theta_1)$.

The nested range condition states that for any θ_2 such that $\theta_2\in M(\theta_1)$, the entire image $M(\theta_2)$ is contained in $M(\theta_1)$. If $\theta_2\notin M(\theta_1)$, then there is no implied relationship between the images.

Theorem 1. (1) *If* $M(\cdot)$ *satisfies* NRC *then for any* X *and* $u:X\times\Theta\to R$, *the set of implementable social choice functions coincides with the set of truthfully implementable social choice functions.*

(2) *If* $M(\cdot)$ *violates* NRC *then there exists* X, $u:X\times\Theta\to R$, *and an* $M(\cdot)$*-implementable social choice function* f *such that* f *is not truthfully* $M(\cdot)$*-implementable.*

Proof.

(1) Suppose that $(M(\cdot),g)$ implements f and $\bar{g}$, defined by $\bar{g}(\theta)=f(\theta)$, is not $M(\cdot)$-truthfully implementable. We will show that NRC is violated.

Without loss of generality, let $\theta_1\in\Theta$ be such that $u(\bar{g}(\theta_1),\theta_1)<u(\bar{g}(\theta_2),\theta_1)$ for some $\theta_2\in M(\theta_1)$. As g was able to $M(\cdot)$-implement f, we must have $\bar{g}(\theta_2)\neq g(\theta)$ for all $\theta\in M(\theta_1)$. In particular, $\bar{g}(\theta_2)\neq g(\theta_1)$ and $\bar{g}(\theta_2)\neq g(\theta_2)$, as we know that θ_1 and θ_2 are in $M(\theta_1)$. But $\bar{g}(\theta_2)=g(\theta)$ for some $\theta\in M(\theta_2)$, since $\bar{g}(\theta_2)=g(\phi_g(\theta_2))$ and $\phi_g(\theta_2)\in M(\theta_2)$. Without loss of generality, let $\theta_3=\phi_g(\theta_2)$. Then $\theta_3\notin M(\theta_1)$ because of the M-implementability of f as shown above.

Collecting these results, we have $\theta_2\in M(\theta_1)$, $\theta_3\in M(\theta_2)$ and $\theta_3\notin M(\theta_1)$, violating NRC.

(2) Consider the following example where NRC is not satisfied: $\Theta=\{\theta_1,\theta_2,\theta_3\}$; $M(\theta_1)=\{\theta_1,\theta_2\}$; $M(\theta_2)=\{\theta_2,\theta_3\}$; $M(\theta_3)=\{\theta_3\}$ and the social choice function $f(\cdot)$ such

that $f(\theta_1)=x_1$; $f(\theta_2)=x_2$; $f(\theta_3)=x_2$ with the agent's utility function such that

$$\begin{array}{lll} u(x_1,\theta_1)=10 & u(x_2,\theta_1)=15 & u(x_3,\theta_1)=20 \\ u(x_1,\theta_2)=5 & u(x_2,\theta_2)=10 & u(x_3,\theta_2)=0 \\ u(x_1,\theta_3)=10 & u(x_2,\theta_3)=15 & u(x_3,\theta_3)=20. \end{array}$$

Choose g such that $g(\theta_1)=g(\theta_2)=x_1$ and $g(\theta_3)=x_2$.

Observe that $g(\phi_g(\theta_i))\equiv f(\theta_i)$ for $i=1,2,3$. Therefore the desired SCF is implemented; but $\phi_g(\theta_2)=\theta_3$ and this implementation is not truthful.

However, $f(\cdot)$ cannot be truthfully implemented. If g^* were to truthfully implement $f(\cdot)$, then $g^*(\theta_1)=x_1$, $g^*(\theta_2)=x_2$, $g^*(\theta_3)=x_2$. In response to this g^*, we have $\phi_{g^*}(\theta_1)=\theta_2$ because $u(x_2,\theta_1)>u(x_1,\theta_1)$ and $\theta_2\in M(\theta_1)$. Thus $g^*(\phi_{g^*}(\theta_1))=x_2\neq f(\theta_1)$.

For more general Θ, there must exist, if NRC is not satisfied, $(\theta_1,\theta_2,\theta_3)$ such that $M(\theta_1)=\{\theta_1,\theta_2\}\cup A$, $M(\theta_2)=\{\theta_2,\theta_3\}\cup B$, $M(\theta_3)=\{\theta_3\}\cup C$, with $A\cap\{\theta_1,\theta_2,\theta_3\}=\varnothing$, $B\cap\{\theta_1,\theta_2,\theta_3\}=\varnothing$, $C\cap\{\theta_1,\theta_2,\theta_3\}=\varnothing$. Then, the same argument as above applies. ‖

In order to find all the implementable SCF's for a given u, $M(\cdot)$ pair, an appeal to the NRC condition may provide a constructive approach. First, NRC must be verified. If it holds, these SCF's are precisely those implementable via truthtelling. This is constructive insofar as a constructive method is used to check the transitivity of the relation induced by $M(\cdot)$.

When NRC fails, there might still be a constructive approach if the set of $M(\cdot)$ implementable SCF's were identical to the $M'(\cdot)$-implementable SCF's for some correspondence M' that is directly derivable from M and satisfies NRC. For example, one possibility which does not work would be to take the transitive closure of the binary relation induced by M and consider its graph as the graph of M'. The failure of this method can be observed by considering the f in part (2) of the above theorem. It is not $M'(\cdot)$-implementable for $M'(\theta_1)=\{\theta_1,\theta_2,\theta_3\}$, $M'(\theta_2)=\{\theta_2,\theta_3\}$, $M'(\theta_3)=\{\theta_3\}$, which is the transitive closure of the M in the theorem.

We now investigate whether any method of this sort can be given. Specifically, we ask whether, for each M, there is some M' satisfying NRC whose implementable set is the same as that for M. The answer we arrive at is negative. Take u and M as in part (2) of the theorem. The following SCF's are implementable:

$$\begin{array}{lll} f_1(\theta_1)=x_1 & f_1(\theta_2)=x_1 & f_1(\theta_3)=x_3 \\ f_2(\theta_1)=x_1 & f_2(\theta_2)=x_2 & f_2(\theta_3)=x_2 \\ f_3(\theta_1)=x_2 & f_3(\theta_2)=x_1 & f_3(\theta_3)=x_1 \\ f_4(\theta_1)=x_3 & f_4(\theta_2)=x_2 & f_4(\theta_3)=x_1. \end{array}$$

(Among these, only f_2 is not truthfully $M(\cdot)$-implementable. It can be implemented via $g(\theta_1)=x_1$, $g(\theta_2)=x_1$, $g(\theta_3)=x_2$.)

Note that the following SCF's are not $M(\cdot)$-implementable:

$$\begin{array}{lll} f_5(\theta_1)=x_1 & f_5(\theta_2)=x_1 & f_5(\theta_3)=x_2 \\ f_6(\theta_1)=x_1 & f_6(\theta_2)=x_2 & f_6(\theta_3)=x_3. \end{array}$$

Now suppose that M' is a correspondence satisfying NRC with the same implementable set. Specifically f_1–f_4 are $M'(\cdot)$-implementable and f_5 and f_6 are not.

Since f_5 is not, either $\theta_3\in M'(\theta_2)$ or $\theta_3\in M'(\theta_1)$. But as f_1 is implementable, $\theta_3\notin M'(\theta_1)$, leaving $\theta_3\in M'(\theta_2)$ as the remaining possibility. From the implementability of f_2 and f_3, we have $\theta_2 \notin M'(\theta_1)$, $\theta_3 \notin M'(\theta_1)$, $\theta_1 \notin M'(\theta_2)$ and $\theta_1 \notin M'(\theta_3)$. Finally, from the implementability of f_4, $\theta_2 \notin M'(\theta_3)$. Using the fact that M' is assumed to satisfy NRC it is determined by these relations to be

$$M'(\theta_1) = \{\theta_1\} \qquad M'(\theta_2) = \{\theta_2, \theta_3\} \qquad M'(\theta_3) = \{\theta_3\}.$$

But then f_6 would be $M'(\cdot)$-implementable. Hence, as claimed, there is no such correspondence M'.

At this point we return to the conceptual and motivational questions behind the definition of $M(\cdot)$-implementability, as promised at the end of Section 2. Up to now we have simply viewed the points in $M(\theta)$ as possible "statements" that the agent could make or "messages" that he could send. Without further elaboration this seems to be a very simplistic and structureless idea. Why should some mere statements be feasible while others are not?

To give the model some real content, the proper interpretation of $M(\cdot)$ is that the principal also has some information, and that the principal can act on this information, to inflict severe punishment on the agent in some circumstances. It is important to be very clear about the source and accuracy of the principal's information. It is not the case that the principal has an independent observation on θ. Rather, the principal can observe a binary variable whose value is (non-stochastically) jointly determined by the truth θ and the message θ', sent by the agent. Its value indicates whether or not $\theta' \in M(\theta)$.

The payoffs can be constructed to be very bad whenever the principal learns $\theta' \notin M(\theta)$. Thus, no agent will ever send such a message, and the mechanism will implement his chosen payoff $g(\theta')$ for $\theta' \in M(\theta)$.

This way of interpreting $M(\theta)$ is different from allowing the principal to see the entire set $M(\theta)$ of permissible messages, for then the principal would be able to infer the true θ itself, except where $M(\theta)$ is invariant to θ. It is also different from permitting the principal to observe a sample from $\{\tilde{\theta} \mid \tilde{\theta} \notin M(\theta)\}$ and to punish the agent whenever $\tilde{\theta} = \theta'$. In the real world, our notion of $M(\theta)$ corresponds to a situation where the points in Θ have a natural meaning in some language. They are not just abstract messages. Points θ' that are not in $M(\theta)$ cause an observable signal and are thus detectable as lies.

An example might make this interpretation clearer. Let θ be income and let the principal be the income tax authority. An individual can understate his income somewhat but, if he does so to too great an extent, he becomes very nervous in an income tax audit. This nervousness is observable to the auditor, who can then discover the source of the understatement and implement a severe punishment. Moderate understatements will not result in nervousness, and will go undetected. In the resulting implementation, no one ever exhibits nervousness. Therefore the decision rule is a function of only reported income, and reported income is chosen within the range that will not result in nervousness $(M(\theta))$, so as to minimize tax liability.

4. COMPARATIVE STATICS OF THE IMPLEMENTABLE SET OF SCF's WITH RESPECT TO THE CORRESPONDENCE $M(\cdot)$

As the set of possible messages expands, there are two possible effects on the implementability of SCF. The possible statement of messages not previously in $M(\cdot)$ means that the SCF may now fail to be implementable because some additional "incentive-compatibility" constraints are present. On the other hand, the additional flexibility in assigning actions to messages afforded by the larger $M(\cdot)$ could expand the implementable set of SCF's to include some not previously feasible. In this section we will use Theorem 1 to discuss these two effects. We will consider two cases: an expansion of $M(\theta)$ to include messages not corresponding to any of the truthful states θ, and the special case of this in which these auxiliary messages are independent of the state.

Theorem 1 provides a complete characterization of the restrictions on messages that remain compatible with the revelation principle. We have identified the points in $M(\theta)$ with points in Θ, the interpretation being that these possible observations could be transmitted. We now expand $M(\theta)$ as follows:

The set of all messages is described as the union of Θ, the "true" messages, and N, a set of messages other than those in Θ. The allowable message correspondence

$$M:\Theta\to\Theta\cup N$$

is defined by the union of two correspondences

$$M(\theta)=T(\theta)\cup N(\theta)$$

with $T(\theta)\subseteq\Theta$ and $N(\theta)\subseteq N$. We continue to assume that $\theta\in T(\theta)$ for all $\theta\in\Theta$.

The condition corresponding to NRC is:

Condition NRC'. For any three distinct elements θ_1, θ_2, m, with θ_1, $\theta_2\in\Theta$ and $m\in\Theta\cup N$, if $\theta_2\in M(\theta_1)$ and $m\in M(\theta_2)$, then $m\in M(\theta_1)$.

Theorem 2. *Condition* NRC' *is necessary and sufficient for the set of* $M(\cdot)$*-implementable* SCF's *to coincide with the set of truthfully* $M(\cdot)$*-implementable* SCF'*s.*

The proof of Theorem 2 is a straightforward extension of that of Theorem 1.

An important special case in which there are messages that can be sent other than the observations arises when a fixed set of messages, K, is possible for every $\theta\in\Theta$ in addition to $M(\theta)$. The "right to remain silent" is of this nature.

We already know from Theorem 1 that if NRC is satisfied by $M(\theta)$, then the addition of a set of common messages, K, will not enlarge the set of implementable plans. This follows since NRC' will continue to hold after K is added, and hence, with or without K, truthful implementation can be used. With fewer possible messages there are fewer incentive compatibility constraints to be satisfied. The feasible SCF's may actually contract because of the necessity to assign *some* outcome to messages in K, and these may create adverse incentive effects. In the next theorem we show that this situation is representative even of cases in which NRC or its generalization NRC' does not hold initially.

Theorem 3. *Assume that* $M(\theta)$ *is enlarged by the addition of a fixed set of messages* K *that is distinct from* $T(\theta)\cup N(\theta)$ *for all* $\theta\in\Theta$. *Then if* f *is an implementable* SCF, *it must also be an implementable* SCF *without the messages in* K.

Proof. Suppose that $\phi_g(\theta)\in K$ for some θ, where g implements f in the presence of K. Let L be the set of all $\theta\in\Theta$ for which $\phi_g(\theta)\in K$. Consider each θ in L and define

$$\begin{aligned}\tilde{g}(\theta)&=g(\phi_g(\theta)) \qquad &\text{if } \theta\in L\\ &=g(\theta) &\text{if } \theta\notin L.\end{aligned}$$

Since the outcomes defined by $g(\phi_g(\theta))$, $\theta\in L$, were available to an agent with *any* characteristic θ in the absence of K, no agent with characteristic $\theta\notin L$ is induced to change his message and clearly any agent with characteristic $\theta\in L$ is induced to tell the truth. Therefore f is implementable in the absence of K. ||

The addition of a set of common messages K may, *for a given set of social states*, decrease the set of implementable functions. Take the following example: $\Theta=\{\theta_1,\theta_2,\theta_3\}$;

$M(\theta_1)=\{\theta_1\}$; $M(\theta_2)=\{\theta_2\}$; $M(\theta_3)=\{\theta_3\}$ and $X=\{x_1, x_2, x_3\}$. $f(\theta_1)=x_1$; $f(\theta_2)=x_2$; $f(\theta_3)=x_3$ is clearly implementable. But assume that

$$u(x_2, \theta_1)>u(x_1, \theta_1) \qquad u(x_3, \theta_2)>u(x_2, \theta_2) \qquad u(x_1, \theta_3)>u(x_3, \theta_3).$$

Suppose we add a strategy $\boldsymbol{n}$ to each set $M(\theta)$. Some social state must be associated with $\boldsymbol{n}$. Let it be x_1 (without loss of generality because of the symmetry of the problem). Then an agent with characteristic θ_3 will now obtain x_1 by announcing $\boldsymbol{n}$ and $f(\cdot)$ is not implementable.

5. EXAMPLES

a. *Conditions implying condition NRC*

(1) *The common intersection condition*

For all pairs $(\theta^i, \theta^j)\in\Theta\times\Theta$, such that $\theta^i\neq\theta^j$,

$$M(\theta^i)\cap M(\theta^j)=\bigcap_{\theta\in\Theta} M(\theta).$$

Under this condition, NRC must be satisfied since for any distinct θ_1, θ_2, θ_3 we have

$$M(\theta_1)\cap M(\theta_2)=M(\theta_2)\cap M(\theta_3).$$

This common intersection condition is also clearly sufficient when $M(\theta)$ is enlarged by a common set K of messages not in M.

There is then an interesting special case of a game, which we call the *silence game*, that satisfies this condition. In the silence game the agent's choice is simply to tell the truth or to remain silent. Thus for all $\theta\in\Theta$

$$M(\theta)=\{\theta, \theta_0\}$$

where θ_0 is another point adjoined to the space of messages. (In the notation of condition NRC′, $\{\theta_0\}=N$.)

A related game is the *no-evidence game*. Here, the points in Θ are of two kinds. Evidence, if it exists, is one of the points $\theta_1, \dots, \theta_N$. The agent may have seen some evidence, or he may have no-evidence, denoted θ_0. Thus,

$$\Theta=\{\theta_0, \theta_1, \dots, \theta_N\}$$
$$M(\theta_i)=\{\theta_i, \theta_0\} \qquad \text{for all } i=1, \dots, N$$
$$M(\theta_0)=\{\theta_0\}.$$

Of course, the possibility of partial verification of transmitted information improves the power of the principal. In an example of the silence game, this type of partial verifiability may allow the first best to be implemented when it would not be implementable if $M(\theta)=\{\theta_0, \dots, \theta_N\}$ for all θ.

(2) *Unidirectional distortions in cases with an ordered space* Θ.

These are some cases in which the points in Θ stand in a natural ordered relation to each other.

Let $\gtrsim$ be an ordering of Θ.

$$M(\theta)=\{\tilde{\theta}, \tilde{\theta}\in\Theta, \tilde{\theta}\gtrsim\theta\}.$$

Observe that condition NRC is satisfied, because, for any three different values θ_1, θ_2, θ_3 such that $\theta_2\in M(\theta_1)$, and $\theta_3\in M(\theta_2)$, we must have $\theta_2>\theta_1$ and $\theta_3>\theta_2$, which imply that $\theta_3\in M(\theta_1)$.

We call this game the *overstatement* (or understatement) game. Overstatement games arise, for example, when the agent must support his claim that he has a given level of wealth or income by a demonstration that could not be duplicated by those in a lower wealth or income class.

b. *Examples not satisfying condition NRC*

(1) *Interval of non-verifiability*

Let $\gtrsim$ be an ordering of θ, taken to be a closed interval $[\underline{\theta}, \bar{\theta}]$ of $\mathbb{R}$, and let I be another closed interval containing zero,

$$M(\theta) = \{\tilde{\theta}, \tilde{\theta} \in [\underline{\theta}, \bar{\theta}] \cap (\theta + I)\}.$$

This is the *partial-overstating game* where an agent can lie upward only by a certain amount. As an illustration, take $\Theta = \{\theta_1, \theta_2, \theta_3\}$ and

$$M(\theta_1) = \{\theta_1, \theta_2\} \qquad M(\theta_2) = \{\theta_2, \theta_3\} \qquad M(\theta_3) = \{\theta_3\}.$$

This is the $M(\cdot)$ correspondence used in part (2) of Theorem (1).

In this environment we know that some implementable rules may require a non-truthtelling response by the agent. Not only is there no need to know the truth in order to implement a given SCF, insisting on the truth may make the SCF non-implementable.

Another example where the revelation principle does not hold arises when the agent can partially overstate or understate the value of θ. Using the same notation as above, let I, I' be two intervals containing zero, and

$$M(\theta) = \{\tilde{\theta}, \tilde{\theta} \in [\underline{\theta}, \bar{\theta}] \cap \{(\theta + I) \cup (\theta - I')\}.$$

(2) *Two-dimensional action spaces*

A great deal of work in incentive theory has been done in the case where the space of actions is two-dimensional, one of the dimensions being interpreted as an allocation decision, the other dimension being interpreted as a compensatory transfer. Suppose for simplicity that each agent's utility function is additively separable

$$u(z, \theta) + t$$

where z is the allocation decision and t the transfer and that u is strictly increasing in z and continuously differentiable of class C^2. Restricting the analysis to piecewise C^1 allocation functions, it can be shown (see Guesnerie and Laffont (1984)) that truthful implementation requires that the transfers take the form

$$t(\theta) = -\int_0^\theta \frac{\partial u}{\partial z}(z(s), s)\frac{dz}{ds}\, ds + \text{constant} \tag{5.1}$$

and that the necessary local condition for truthtelling imposes

$$\frac{\partial^2 u}{\partial z\, \partial \theta}(z(\theta), \theta)\frac{dz}{d\theta}(\theta) \geqq 0. \tag{5.2}$$

If, in addition, $\partial^2 u/\partial z\, \partial \theta$ is positive (negative) over all its domain, condition (5.1) plus the associated monotonicity condition (5.2) are jointly sufficient for implementability. However, when $\partial^2 u/\partial z\, \partial \theta$ changes sign, functions $z(\cdot)$ satisfying (5.1) and (5.2) are generally not implementable. To see this consider the following example: Suppose that

$$\begin{aligned} \frac{\partial^2 u}{\partial z\, \partial \theta}(z, \theta) &> 0 \qquad \text{if } \theta < \theta_0, \quad \theta \in \Theta \\ &\leqq 0 \qquad \theta \geqq \theta_0, \quad \theta \in \Theta. \end{aligned}$$

By (5.2) $z(\cdot)$ must change sign at θ_0, and there will be ranges of values of θ, θ_1 and θ_2 such that $z(\theta_1)=z(\theta_2)$. From (5.1) we have

$$t(\theta_1)=-\int_0^{\theta_1}\frac{\partial u}{\partial z}(z(s),s)\frac{dz}{ds}(s)\,ds+C$$

$$t(\theta_2)=-\int_0^{\theta_2}\frac{\partial u}{\partial z}(z(s),s)\frac{dz}{ds}(s)ds+C.$$

But

$$u(z(\theta_2),\theta_2)+t(\theta_2)\geqq u(z(\theta_1),\theta_2)+t(\theta_1)$$

$$u(z(\theta_1),\theta_1)+t(\theta_1)\geqq u(z(\theta_2),\theta_1)+t(\theta_2)$$

implies

$$t(\theta_1)=t(\theta_2). \tag{5.3}$$

Clearly (5.3) will almost never hold with an arbitrary $z(\cdot)$ satisfying (5.2).

Let us now expand the set of implementable functions by introducing the more restrictive message correspondence

$$M(\theta)=\{\tilde{\theta}: |\theta-\tilde{\theta}|\leqq\Sigma\}.$$

The following functions are truthfully $M(\cdot)$-implementable. Take $z(\cdot)$ such that

$z(\cdot)$ is increasing if $\theta<\theta_0-\Sigma$

$z(\cdot)$ constant if $\theta_0-\Sigma\leqq\theta\leqq\theta_0+\Sigma$

$z(\cdot)$ is decreasing if $\theta_0+\Sigma<\theta$

and take $t(\cdot)$ satisfying (5.1).

Moreover, since $M(\cdot)$ does not satisfy NRC, there are $M(\cdot)$-implementable social choice functions which are not truthfully $M(\cdot)$-implementable. An example of such a function can be constructed as follows:

Take $k(\theta)$ increasing for $\theta\leqq\theta_0$ and decreasing for $\theta\geqq\theta_0$. Consider the following outcome function.

If $\theta\neq\theta_0$

$$z(\theta)=k(\theta)$$

$$t(\theta)=-\int_0^{\theta}\frac{\partial u}{\partial z}(z(s),s)\frac{dz}{ds}(s)ds+\tilde{k}$$

where $\tilde{k}$ is chosen so that $u(z(\theta),\theta)+t(\theta)\leqq\bar{u}$ for all $\theta\in\Theta$.

$$z(\theta_0)=k$$

$$t(\theta_0)=0 \qquad \text{with } u(k,\theta)>\bar{u} \text{ for any } \theta\in[\theta_0-\Sigma,\theta_0+\Sigma].$$

The optimal response rule is

$$\begin{aligned}\phi(\theta)&=\theta && \text{for } \theta<\theta_0-\Sigma \text{ and } \theta>\theta_0+\Sigma\\ &=\theta_0 && \text{for } \theta\in[\theta_0-\Sigma,\theta_0+\Sigma].\end{aligned}$$

The implemented social choice function is

$$\begin{aligned}&(k(\theta),t(\theta)) && \text{for } \theta\notin[\theta_0-\Sigma,\theta_0+\Sigma]\\ &(k,0) && \text{for } \theta\in[\theta_0-\Sigma,\theta_0+\Sigma].\end{aligned}$$

6. CONCLUSION

Until very recently, incentive theory has neglected the observability of features related to the characteristics of the agents. Outcome functions can depend both on the agents' answers and on these observations which are connected to the true characteristics.

In this paper we have studied a particular type of monitoring technology which makes some answers of the agent noncredible. In addition to enlarging the class of implementable social choice functions, this monitoring technology invalidates the revelation principle. The idea of the revelation principle is to construct an outcome function which, for each value of the characteristics, θ, mimics the optimal response of the agent faced with the original mechanism. This is an abstract construction which may not be feasible for the monitoring technology which is only defined in terms of the original language.

First version received October 1983; final version accepted January 1986 (Eds.).

This research was supported by Commissariat du Plan, Paris, and NSF grants IST-8310118 and SES-8207338 to Harvard University. Helpful comments from J. Demski, B. Holmström and B. Nalebuff are gratefully acknowledged. An earlier version was presented at the Center for Economic Policy Research conference at Stanford, April 1983. We also benefited from the referees' and the editor's comments.

NOTES

[1] If $\operatorname{argmax}_{m\in M(\theta)} u(g(m), \theta)$ is not a singleton, any selection can be made and is an induced *response rule.*

[2] See Gibbard (1973), Green and Laffont (1979), Dasgupta, Hammond and Maskin (1979).

REFERENCES

DASGUPTA, P., HAMMOND, P., and MASKIN, E. (1979), "Implementation of Social Choice Rules: Some General Results on Incentive Compatibility",' *Review of Economic Studies*, **46**, 185-216.

GIBBARD, A. (1973), "Manipulation of Voting Schemes: A General Result", *Econometrica*, **41**, 587-601.

GREEN, J. and LAFFONT, J. J. (1979) *Incentives in Public Decision Making* (Amsterdam: North-Holland).

GUESNERIE, R. and LAFFONT, J. J. (1984), "Control of Public Firms Under Incomplete Information", *Journal of Public Economics*, **25**, 329-369.