

**UNIQUE IMPLEMENTATION OF ACTION PROFILES:
NECESSARY AND SUFFICIENT CONDITIONS***

BY SANDRO BRUSCO¹

Universidad Carlos III de Madrid, Spain

I study the general problem of a principal who desires to implement a given vector of actions and pay the agents according to a given compensation scheme. Previous work has provided mechanisms for implementation in various special cases. In this article, I fully characterize the set of action profiles and compensation schemes implementable in subgame-perfect equilibrium, providing necessary and sufficient conditions for implementation.

1. INTRODUCTION

The traditional implementation problem with complete information has the following structure: A group of agents observe the state of the world, all of them obtaining the same information (it is in this sense that this is a complete information problem). There is a planner who does not observe the state of the world. The planner wants to take decisions, and in principle she would like to change the decision taken depending on the state of the world. Formally, the decision rule is described by a social choice function, a mapping from the set of states of the world to the set of possible decisions. In order to implement the social choice function, the planner has to elicit the information from the agents. The problem is therefore to set up a mechanism making sure that the agents report truthfully the state of the world they observe.

When there are at least three agents, the direct mechanism has always an equilibrium in which agents tell the truth. However, there is no guarantee that the truth-telling equilibrium be unique. The problem is therefore to find a mechanism such that truth telling is the *unique* equilibrium or, more accurately, such that the outcome prescribed by the social choice function is the unique equilibrium outcome in each state of the world.

Whether or not it is actually possible to design a mechanism delivering unique implementation depends on the properties of the social welfare function and on the equilibrium notion adopted to predict the outcome of the mechanism. The literature

* Manuscript received July 1998; revised May 2000.

¹ I would like to thank seminar participants at UCSD, two anonymous referees, and the associate editor for many suggestions that helped me to improve both the content and the style of the article. Funding from Ministerio de Educación y Cultura—Dirección General de Enseñanza Superior e Investigación Científica, proyecto PB97-0084 and proyecto PB96-0118, is gratefully acknowledged. E-mail: brusco@emp.uc3m.es.

has provided complete characterizations of the set of implementable social choice functions for an array of solution concepts, such as Nash equilibrium (Moore and Repullo, 1990; Dutta and Sen, 1991), subgame-perfect equilibrium (Moore and Repullo, 1988; Abreu and Sen, 1990), and virtual implementation with iterative elimination of dominated strategies (Abreu and Matsushima, 1992).

This traditional approach can be seen as dealing with an “adverse selection,” or “hidden information,” problem. In this article I want to look at the “moral hazard,” or “hidden action,” counterpart.

This problem can be informally described as follows: There is a group of agents working for a principal. Each agent has a number of possible actions available (for example, she can work hard or shirk), and the principal wants each agent to select a particular action. The action profile taken by the agents is not observable by the principal. However, it is observable by the agents themselves; that is, each agent can observe the actions taken by her colleagues. This is the counterpart of the complete information assumption for the moral hazard problem. The action profile determines the probability distribution over some observable variable (the typical example is the revenue of the firm). While the principal cannot observe the action profile, she can make the compensation contingent on this observable variable. Furthermore, the principal can ask the agents what was the action profile actually taken, and determine the compensation scheme accordingly. The implementation problem in this case is solved if the principal can design a mechanism in which the agents take the right action profile and receive the right compensation scheme. Furthermore, we want this to be the unique equilibrium outcome of the mechanism.

Such a problem was first considered by Demski and Sappington (1984) and Mookherjee (1984). They observed that standard tournament models usually have multiple equilibria, and some of the “bad” equilibria may be Pareto superior from agents’ point of view. In particular, in direct revelation mechanisms it is often an equilibrium for all the agents to take the lowest level of effort and report unanimously that the required effort was actually undertaken. Ma et al. (1988) have shown how to get unique implementation in the Demski–Sappington model, and Ma (1988) has shown that unique implementation in subgame-perfect equilibrium is achievable in the Mookherjee model. A number of recent contributions, among them Arya and Glover (1995), Sjöström (1996), Arya et al. (1997), and Brusco (1998), have proposed mechanisms for implementation in some variant of the general model. The cited articles propose mechanisms for implementation based on some sufficient condition.

Up to now, however, there has been no attempt to analyze necessary conditions and to establish the counterpart of the results on necessary and sufficient conditions for implementation that we have for the adverse-selection problem. This is what is done in this article. I identify necessary and sufficient conditions that a pair given by an action profile and a compensation scheme has to satisfy in order to be implementable. I consider implementation in subgame-perfect equilibrium. This notion appears to be the relevant one, since the problem has a natural sequential structure: Agents first take actions and then are asked to report the information they have about the action profile observed.

My goal is to provide a full characterization of implementable pairs, obtaining for the moral-hazard case the counterpart of existing results for the adverse-selection case. This being the main scope of the article, I do not try to use simple or realistic mechanisms in my proofs. In fact, the mechanism adopted for the general proof of the sufficiency conditions is complicated and incorporates the undesirable “integer game” feature (see Jackson, 1992, for a critique of the use of this kind of “tail-chasing” devices in the implementation literature). It is important to point out, however, that in many (not so special) cases simpler mechanisms are available. As a matter of fact, most of the articles cited above identify simple mechanisms for implementation, imposing some additional assumptions on the problem. In this article, I look for necessary and sufficient conditions for implementation in the moral hazard problem without putting any constraint on the mechanisms that the planner can use. The next step in this line of research is to look for necessary and sufficient conditions when mechanisms are constrained to be “simple” in some sense (see, for example, the article by Jackson et al., 1994, for the adverse-selection problem). This is not attempted here, and it is left for future research.

The rest of the article is organized as follows: Section 2 introduces notation and describes formally the implementation problem. Section 3 provides an example showing that there are cases not covered by the standard sufficiency conditions in which implementation of action profiles is still possible. The main difference between the “hidden information” and the “hidden action” cases is that in the latter the agents report on an *endogenous* “state of nature” (the action profile). This endogeneity can be exploited to destroy some nonoptimal equilibria that would survive if the state of nature were exogenous. In Section 4, I state the result about necessary conditions. In Section 5, I show that when there are at least three agents, the necessary conditions are also sufficient, and I describe a general mechanism for implementation. Section 6 contains concluding remarks. The proofs are collected in the Appendix.

2. THE STRUCTURE OF THE PROBLEM

Let I be a finite set of agents. Each agent i can take an action a_i in the finite set A_i . Let $A \equiv \times_{i \in I} A_i$ with typical element $a \equiv (a_1, \dots, a_I)$, and $A_{-i} \equiv \times_{j \in I \setminus i} A_j$ with typical element $a_{-i} \equiv (a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_I)$. The action profile a determines a probability distribution over some variable (e.g., the revenue of a firm) v . Let V^a be the finite set of values that can be taken by v when the true action profile is a , and let $\rho(\cdot | a)$ be the probability distribution of v when the true action profile is a . Define $V \equiv \cup_{a \in A} V^a$.

An *individual compensation scheme* for agent i is a function $w_i : V \rightarrow \mathfrak{R}$; that is, w_i specifies a monetary transfer for each possible realization of v . A *compensation scheme* w is a collection of individual compensation schemes, that is, $w \equiv (w_1, \dots, w_I)$. Each agent has preferences over actions and individual compensation schemes represented by a utility function of the form

$$U^i(a, w_i) \equiv \sum_{v \in V} \rho(v | a) u^i[a, w_i(v)]$$

where the “basic” utility function $u^i : A \times \mathfrak{R} \rightarrow \mathfrak{R}$ is defined over the whole action profile a and the monetary transfer, and the probability distribution $\rho(\cdot)$ is common

to all players. I will maintain throughout the article the standard assumption that $u^i(a, w)$ is strictly increasing in w for each action profile a .

The goal of the planner is to make sure that each agent takes a given action a_i^* , so that the action profile $a^* \equiv (a_1^*, \dots, a_I^*)$ is collectively taken, and reward each agent with compensation scheme w_i^* when a^* is the true action profile. A well-known example is the case of a risk-neutral employer with risk-averse agents who desires to implement the action profile a^* , maximizing expected revenue, and pay a constant wage to each agent.² Defining $w^* \equiv (w_1^*, \dots, w_I^*)$, I will say that the problem of implementing the pair (a^*, w^*) is solved when the planner can design a mechanism such that the *unique* equilibrium outcome is that agents take the action profile a^* and are paid according to the compensation scheme w^* .

The “complete information” assumption in this model takes the form that each agent can observe the whole action profile a . This eliminates incentive compatibility considerations, since an agent can be forced to take a_i^* simply because any different action could be reported to the principal, and a stiff fine could follow. However, it leaves open the multiplicity problem.

In general, the mechanism for implementation requires that agents send some messages after the action profile has been taken, confirming that a^* was indeed the action profile. When agents confirm this, the compensation w^* is paid. The planner has to make sure that it is not an equilibrium for the agents to take a different action profile $a \neq a^*$ and then to claim that a^* has been taken.

An important difference between the traditional implementation problem and the problem of implementing action profiles is that “action taking” is given by some technological process and it cannot be manipulated by the principal. This imposes some physiological constraints on the form that the mechanism can take. In particular, any extensive form game that is adopted to implement the pair (a^*, w^*) must be made up of three parts:

- (1) A *pre-action part*, where agents send messages. This has to be intended as an extensive form at which messages are exchanged, with the final vector of messages observed by the principal and the agents. Depending on the message profile, either the game ends and payments are determined, or the game proceeds to the action stage.
- (2) An *action part*, where each agent chooses an action, possibly as a function of the outcome of the pre-action stage. Simultaneously, the agent can also issue some message.³ The set of messages available to the agents may depend on the message profile announced at the pre-action stage. The action profile is observed by the agents but not by the principal, while the message profile issued at this stage is also observed by the principal.
- (3) A *post-action part*, where messages are again issued. The particular extensive form used may depend on the messages sent at the pre-action stage and at the

² See Ma (1988) for an analysis of this case.

³ Here, “simultaneously” should be interpreted with respect to the information structure; that is, messages are sent before actions are observed, and actions are taken before messages are observed.

action stage. However, it cannot depend on the action profile a , since this is not observable by the principal.

The final vector of messages (i.e., messages issued at each stage) determines the compensation scheme selected by the principal, while the action profile implemented is simply the one chosen by the agents at the action stage. Notice that for certain mechanisms, some of these parts may be trivial. For example, we may decide that we do not need agents to send messages before actions are taken. However, when we search for necessary conditions for implementation, we have to consider all possible mechanisms, which means that we have to take into account the set of mechanisms generated by this natural structure.

Sufficient conditions in the literature on implementation of action profiles (see Ma, 1988; Brusco, 1998) take the following form: When $a \neq a^*$, then the equilibrium in which all agents take a and report that a^* is the action profile is destroyed, providing each agent with the option of “spying,” that is, announce that the true action profile is a . This leads the mechanism to a subgame in which another agent j has to choose between a compensation w_j^x and a compensation w_j^y . The two action profiles a and a^* yield different preferences over w_j^x and w_j^y , so that the choice of the agent reveals the true action profile. The principal rewards the spy when it turns out that the accusation is corroborated, while punishing her when this turns out not to be the case (so that the correct equilibrium survives).

The crucial sufficient condition is therefore that action profile $a \neq a^*$ induces different preferences between compensation schemes than a^* . More formally, if $a \neq a^*$, then we can find j , w_j^x , and w_j^y such that

$$U^j(a^*, w_j^x) \geq U^j(a^*, w_j^y) \quad U^j(a, w_j^x) < U^j(a, w_j^y)$$

Is this condition necessary as well? The answer is no. The reason is that, different from the “hidden information” case, the agents are reporting on a state of the world that is *endogenous*. Despite the fact that the switch condition is violated, it may still be the case that the agents do not want to take a . In the next section, I present an example in which this endogeneity is exploited to implement an action profile.

3. AN EXAMPLE

There are two agents, and the action set of each agent is $A_i = \{s, n, c\}$. Here s denotes “shirking,” n denotes a noncooperative effort, and c a cooperative effort. As a consequence of agents’ efforts, a pair of outcomes (v_1, v_2) is observed, where v_i can take values 0 or 1.

Production is a deterministic function of efforts. The function is summarized by Table 1.

Each entry in Table 1 indicates the pair (v_1, v_2) resulting when the corresponding pair of efforts is taken. For example, if agent 1 takes the noncooperative effort n and agent 2 takes the cooperative effort c , then only the first good is produced.

The utility function is quasilinear: $u_i(a, m) = m - \phi(a_i)$, where ϕ denotes the cost of effort. Assume

$$\phi(a_i) = \begin{cases} 0 & \text{if } a_i = s \\ 0.5 & \text{if } a_i = c \\ 1 & \text{if } a_i = n \end{cases}$$

This model can be explained as follows: Agents prefer shirking to any kind of effort, and a cooperative effort to a noncooperative one. If a planner wants to achieve $(v_1, v_2) = (1, 1)$, then the best way to do it is through the pair $(a_1, a_2) = (c, c)$. The cooperative effort is more efficient. Agent 1 can obtain $v_1 = 1$ either when she takes n , the noncooperative effort, or when agent 2 is taking c . Since the cooperative effort generates less disutility, it is better that both agents take c .

Suppose now that the planner wants to implement (c, c) and pay a fixed salary $w > 1$ to both agents. The problem for the planner is to make sure that the action profile (n, n) does not occur in equilibrium. Notice that here we cannot apply the standard sufficient conditions, since (n, n) yields exactly the same outcome as (c, c) . Therefore, in any mechanism played after actions are taken, whatever is an equilibrium after (c, c) must remain an equilibrium after (n, n) . In the hidden-information case this is also the end of the story, meaning that if the social choice function selects different outcomes under two states θ and ψ , and no reversal of preferences occurs, then the function cannot be implemented in subgame-perfect equilibrium.

However, as we have observed, the state here is endogenous. Even if we cannot avoid that *once (n, n) has been taken*, agents report (c, c) , we can still hope to be able to avoid that (n, n) be taken in the first place. This of course is not possible in the hidden-information case, where the state of the world is chosen by nature and not by the players.

In this case we can destroy the (n, n) equilibrium using the following compensation scheme:

Agent 1: w if $(1, 1)$ or $(0, 1)$ is observed, $-\epsilon$ in all other cases.

Agent 2: w if $(1, 1)$ or $(1, 0)$ is observed, $-\epsilon$ in all other cases.

Under this wage scheme, an agent can profitably deviate from the (n, n) equilibrium taking action s . This effort is less costly and produces an outcome that still guarantees that w will be paid. Notice that deviating to s is not profitable when the other agent is taking c , so that the good equilibrium is not destroyed.

TABLE 1

		Agent 2		
		s	n	c
Agent 1	s	(0, 0)	(0, 1)	(1, 0)
	n	(1, 0)	(1, 1)	(1, 0)
	c	(0, 1)	(0, 1)	(1, 1)

It can in fact be checked that (c, c) and w paid for sure is the unique equilibrium outcome of the simple mechanism in which agents just take actions and are rewarded according to the compensation scheme described above.

The example shows that the standard sufficiency conditions based on switching preferences are not the only ones that can be used to enforce an action profile. In the rest of the article I expand on this observation, providing a full characterization of implementable pairs (a^*, w^*) .

4. NECESSARY CONDITIONS

In order to better understand what type of conditions we may need, it is useful to look at the necessary conditions in the hidden-information problem. The basic idea is that if the social choice function prescribes two different allocations at state θ and at state ψ , we want to avoid that the strategy profile adopted at θ be an equilibrium when the state of the world is ψ . When implementing action profiles, the main issue becomes to make sure that an action profile $a \neq a^*$ is not taken in equilibrium. In particular, we want to avoid equilibria in which $a \neq a^*$ is the action profile taken and the agents issue the same messages as in the equilibrium in which a^* is taken, thus leading to the selection of the compensation scheme w^* . In other words, whenever an action profile $a \neq a^*$ is going to be taken, there must be a profitable deviation for some agent at some stage of the mechanism. As observed before, all mechanisms for implementation of action profiles can be partitioned in three parts, and I am now going to provide three conditions, each one corresponding to the case in which a profitable deviation can be found at a given part.

First, we have a condition making possible a deviation in the post-action part.

CONDITION 1. An action profile a satisfies the condition with respect to \bar{a} if it is possible to find an agent $j(a, \bar{a})$ (shortly, j) and a pair of individual compensation schemes $w_j^x(a, \bar{a}), w_j^y(a, \bar{a})$ (shortly, w_j^x, w_j^y) such that

$$U^j(\bar{a}, w_j^x) \geq U^j(\bar{a}, w_j^y) \quad U^j(a, w_j^x) < U^j(a, w_j^y)$$

Condition 1 permits to “test” independently a claim made by an agent that the true action profile is a rather than \bar{a} . When the action profile a is taken, any agent other than $j(a, \bar{a})$ may ask that $j(a, \bar{a})$ be forced to choose between w_j^x and w_j^y . Given the structure of preferences, the choice will reveal the true action profile. We will use the term “test agent for the pair (a, \bar{a}) ” (or just “test agent” when the pair is clear from the context) to indicate agent $j(a, \bar{a})$.

Condition 1 is the standard sufficient condition that has been used in this literature starting from Ma (1988). Brusco (1998) discusses the relationship between Condition 1 and Abreu and Sen’s condition α , a necessary condition for subgame-perfect implementation in the adverse-selection framework.

However, this condition alone cannot be also necessary for implementation. In the example presented in Section 3, the action profile (n, n) does not satisfy Condition 1 with respect to (c, c) , yet we have seen that (c, c) is implementable. To obtain a full characterization of the set of implementable pairs (a^*, w^*) , we have to turn the

attention to the action part and the pre-action part of the mechanism. The following condition refers to the action part:

CONDITION 2. An action profile a satisfies the condition with respect to (\bar{a}, w) if it is possible to find an agent $i(a, \bar{a})$ (shortly, i), an action $\tilde{a}_i(a, \bar{a})$ (shortly, \tilde{a}_i) and an individual compensation scheme $w_i^y(a, \bar{a})$ (shortly, w_i^y) such that

- (1) $U^i((\tilde{a}_i, a_{-i}), w_i^y) > U^i(a, w_i)$.
- (2) For each action \hat{a}_i , either (\tilde{a}_i, a_{-i}) satisfies Condition 1 with respect to $(\hat{a}_i, \bar{a}_{-i})$, or there exists a compensation schedule $\bar{w}(\hat{a}_i)$ such that $U^i((\tilde{a}_i, \bar{a}_{-i}), \bar{w}_i(\hat{a}_i)) \leq U^i(\bar{a}, w_i)$, and $U^i((\hat{a}_i, a_{-i}), \bar{w}_i(\hat{a}_i)) > U^i(a, w_i)$.

Condition 2 can be interpreted as follows: Suppose the planner wants to implement a pair (\bar{a}, w) . Then an action profile $a \neq \bar{a}$ that satisfies the condition can be “killed” as an equilibrium, because there is an agent i who prefers to deviate to \tilde{a}_i and ask for a different compensation scheme w_i^y . In particular, we can imagine that the mechanism gives to agent i the possibility of issuing a “warning message” like the following:

The agents are taking action profile a rather than the prescribed \bar{a} . I am therefore taking action \tilde{a}_i . The final action profile will be (\tilde{a}_i, a_{-i}) , and I ask to be compensated according to w_i^y rather than according to w_i .

By issuing the message and taking action \tilde{a}_i , agent i ends up being strictly better off (part 1 of the condition). However, we also have to make sure that by endowing the agent with the possibility of issuing the warning message, we do not jeopardize the truth-telling equilibrium, that is, the equilibrium in which the agents take action profile \bar{a} and are then compensated according to w . This is the content of part 2 of the condition.

To interpret this part of the condition, assume that agents other than i are taking the action profile \bar{a}_{-i} . Suppose that agent i takes action \hat{a}_i and then proceeds to claim that the actual action profile being taken is (\tilde{a}_i, a_{-i}) , so that she wants to be compensated according to w_i^y . We have to make sure that this deviation does not increase the utility of the agent. One simple way to achieve this is to discover that the agent lied and impose a stiff punishment. This is possible if (\tilde{a}_i, a_{-i}) satisfies Condition 1 with respect to $(\hat{a}_i, \bar{a}_{-i})$. In this case, the truth can be elicited in the subgame following the action stage (see the discussion after Condition 1). If (\tilde{a}_i, a_{-i}) does *not* satisfy Condition 1 with respect to $(\hat{a}_i, \bar{a}_{-i})$, then it may be impossible to discover that agent i lied. Nevertheless, at the post-action stage we can ask the agents what the true action profile is. If they report $(\hat{a}_i, \bar{a}_{-i})$, then we reward i according to $\bar{w}_i(\hat{a}_i)$. This makes sure that the agent is not better off deviating from the truth-telling equilibrium. At the same time we have to make sure that when agent i is right in issuing the warning message, she still ends up with higher utility. The problem here is that when the true action profile is (\tilde{a}_i, a_{-i}) and the agents announce $(\hat{a}_i, \bar{a}_{-i})$, it may be impossible to discover that they lied. The condition $U^i((\tilde{a}_i, a_{-i}), \bar{w}_i(\hat{a}_i)) > U^i(a, w_i)$ makes sure that even in the case in which $(\hat{a}_i, \bar{a}_{-i})$ is announced after the warning message by agent i , it is still profitable for agent i to deviate.

To see Condition 2 at work, let us go back to the example of Section 3. The planner wants the agents to take the action profile (c, c) and wants to pay the wage w to each agent when the outcome $(1, 1)$ is observed. The wage paid when an outcome $(v_1, v_2) \neq (1, 1)$ is observed is arbitrary. In particular, let us take the wage scheme defined in Section 3 (i.e., the wage scheme giving w to agent 1 if either $(1, 1)$ or $(0, 1)$ is observed and $-\epsilon$ otherwise, and similarly for agent 2) to show that (c, c) is in fact implementable. Let us call w^* this compensation scheme.

We have seen that the action profile (n, n) does not satisfy Condition 1 with respect to (c, c) . It can be shown, however, that (n, n) satisfies Condition 2 with respect to $((c, c), w^*)$. To see this, set $i((n, n), (c, c)) = 1$, $\tilde{a}_1((n, n), (c, c)) = s$, and $w_1^y = w_1^*$ (i.e., the wage scheme is unchanged). Part 1 of the condition is satisfied, since

$$U^1((s, n), w_1^y) = w > w - 1 = U^1((n, n), w_1^*)$$

To check that part 2 is also satisfied, observe that (s, n) satisfies condition 1 with respect to all action profiles (s, c) , (n, c) , and (c, c) . In this case, Condition 1 is somewhat trivial, since the outcomes are deterministic, and we therefore only need to check that different action profiles produce different outcomes.

Finally, we have the condition making possible a deviation in the pre-action part.

CONDITION 3. An action profile a satisfies the condition with respect to (\bar{a}, w) if there exists an agent $i(a, \bar{a})$ (shortly, i), an action profile $\tilde{a}(a, \bar{a})$ (shortly, \tilde{a}), and a compensation scheme $w^y(a, \bar{a})$ (shortly, w^y) such that

- (1) $U^i(\tilde{a}, w_i^y) > U^i(a, w_i)$.
- (2) $U^i(\tilde{a}, w_i^y) \leq U^i(\bar{a}, w_i)$.
- (3) If an action profile \hat{a} satisfies neither Condition 1 nor Condition 2 with respect to (\tilde{a}, w^y) , then $U^i(\hat{a}, w_i^y) > U^i(a, w_i)$.

To interpret the condition, assume that the planner wants to implement (\bar{a}, w) . The action profile $a \neq \bar{a}$ cannot be part of an equilibrium if a satisfies Condition 3. The reason is that before the actions are taken, agent i could ask for action profile \tilde{a} to be implemented and compensation scheme w^y be adopted. Part 1 makes sure that this is a profitable deviation, while 2 makes sure that this maneuvering does not kill the “correct” equilibrium. Part 3 guarantees that in all possible equilibria following the deviation, agent i is better off. As discussed above, if an action profile \hat{a} satisfies Condition 1 or Condition 2 with respect to (\tilde{a}, w^y) , then we can destroy the equilibrium in which agents take \hat{a} and report \tilde{a} . If \hat{a} satisfies neither of the two conditions, then we cannot avoid this equilibrium. However, part 3 states that the deviation will remain profitable for the agent.

The situations in which we need to resort to condition 3 in order to implement an action profile tend to be somewhat pathological. The following example illustrates how the condition can be used:

EXAMPLE. There are two agents. Each agent can take three actions, s, n , or c . Depending on the pair of actions taken, the outcome is either 0 or 1. The outcome is deterministic and it is related to the action profiles according to Table 2.

TABLE 2

		Agent 2		
		<i>s</i>	<i>n</i>	<i>c</i>
Agent 1	<i>s</i>	0	1	0
	<i>n</i>	1	1	0
	<i>c</i>	0	0	1

TABLE 3

		Agent 2		
		<i>s</i>	<i>n</i>	<i>c</i>
Agent 1	<i>s</i>	0, 0	1, 1	0, 1
	<i>n</i>	1, 1	1, 1	$\frac{1}{2}, 1$
	<i>c</i>	1, 0	$1, \frac{1}{2}$	0, 0

The utility function of the two agents takes the form $U^i(a, m) = m - \phi^i(a)$. The function ϕ^i gives the disutility of each action profile for agent i . The functions ϕ^1 and ϕ^2 are represented in Table 3.

In Table 3, the first entry in each box is the disutility suffered by agent 1 at that pair of actions, and the second entry is the disutility of agent 2. For example, if agent 1 takes action c and agent 2 takes action s , then the disutility of 1 is 1 and the disutility of 2 is 0.

The planner wants to implement the action pair (c, c) and a compensation scheme that pays 1 to each agent when the outcome is 1 and $-\varepsilon$ when the outcome is 0, with $\varepsilon > 0$. Let us call w^* this compensation scheme.

Consider now the action profile (n, n) . This action profile produces exactly the same outcome as (c, c) at a greater effort for the agents. Is it possible to make sure that (n, n) is not taken in equilibrium? We will see that the answer is yes, but that this requires carefully designing the pre-action stage.

We start observing that (n, n) does not satisfy Condition 1 with respect to (c, c) . It is therefore impossible to elicit the truth in the post-action stage. More precisely, the pair of equilibrium strategies adopted by the two agents in the post-action stage after (c, c) has been taken will still be an equilibrium pair when (n, n) is taken. Next, observe that no profitable deviation at the action stage is possible. Since the problem is symmetric, we can limit attention to agent 1. It is clear that the agent cannot be better off if she still takes action n and issues a new message. In order to be better off, the agent should obtain a higher salary in all the subgame-perfect equilibria following the message. But then the same deviation would be profitable also when (c, c) is the action profile, since the set of subgame-perfect equilibria is the same. The same reasoning applies when agent 1 deviates to s . Again, she needs a higher salary

to be strictly better off, and again she could ask for a higher salary when (c, c) is the action profile. Thus, any profitable deviation from (n, n) would also be a profitable deviation against (c, c) . Finally, consider a deviation to c , so that the pair (c, n) is realized. In order to be better off, the agent would have to issue a message leading to a salary of $w > 1$ being paid after the deviation in every subgame-perfect equilibrium; notice that this salary should be paid when the observed outcome is 0, since this is what occurs when the action profile (c, n) is taken. But then the agent could use the same deviation to destroy the equilibrium in which (c, c) is taken. In this case agent 1 would have to deviate to s , so that the action profile (s, c) is realized, and issue the same message. The outcome would again be 0, and in all subgame-perfect equilibria the salary $w > 1$ would be paid.

In fact, it can be checked that the second part of Condition 2 is not satisfied. Suppose, for example, that we want to use a deviation by agent 1 to s in order to break the equilibrium in which both agents take n . Using the notation of the condition, we have $a = (n, n)$, $\bar{a} = (c, c)$, $i = 1$, $\tilde{a}_1 = s$, so that $(\tilde{a}_1, a_2) = (s, n)$. To see that the second part of Condition 2 is not satisfied, consider $\hat{a}_1 = c$, so that $(\hat{a}_1, \bar{a}_2) = (c, c)$. It is immediate to see that (s, n) does not satisfy Condition 1 with respect to (c, c) . The other possibility is to find a compensation schedule $\bar{w}(\hat{a}_1)$ that satisfies the two inequalities:

$$\bar{w}(\hat{a}_1) - \phi^1(c, c) \leq 1 - \phi^1(c, c) \quad \bar{w}(\hat{a}_1) - \phi^1(s, n) > 1 - \phi^1(n, n)$$

Given the values in Table 3, the two inequalities are equivalent to $\bar{w}(\hat{a}_1) \leq 1$ and $\bar{w}(\hat{a}_1) > 1$, which are clearly incompatible.

Consider next a deviation by agent 1 to c . Now we have $a = (n, n)$, $\bar{a} = (c, c)$, $i = 1$, $\tilde{a}_1 = c$, so that $(\tilde{a}_1, a_2) = (c, n)$. To see that the second part of Condition 2 is not satisfied, consider $\hat{a}_1 = s$, so that $(\hat{a}_1, \bar{a}_2) = (s, c)$. Again, (c, n) does not satisfy Condition 1 with respect to (s, c) . When we look for a compensation schedule $\bar{w}(\hat{a}_1)$, we have the two inequalities:

$$\bar{w}(\hat{a}_1) - \phi^1(s, c) \leq 1 - \phi^1(c, c) \quad \bar{w}(\hat{a}_1) - \phi^1(c, n) > 1 - \phi^1(n, n)$$

Again this implies $\bar{w}(\hat{a}_1) \leq 1$ and $\bar{w}(\hat{a}_1) > 1$.

We have therefore established that given any mechanism having $((c, c), w^*)$ as a subgame-perfect equilibrium outcome, there is no profitable deviation at the action stage or at the post-action stage against a strategy profile that involves taking the action pair (n, n) and otherwise issuing the same messages as in the truth-telling equilibrium, so that w^* is paid. Any hope to destroy such an equilibrium must therefore lie at the pre-action stage.

In fact, we can show that it is possible to construct a mechanism yielding $((c, c), w^*)$ as the unique subgame-perfect equilibrium outcome. The mechanism is as follows:

Pre-action stage: Each agent announces either Y or N and an integer number. The outcome is as follows:

- If one agent announces $(Y, 0)$ and the other announces (Y, \cdot) , then set the compensation scheme equal to $w_i^*(1) = 1$ and $w_i^*(0) = -\epsilon$ for each agent and go to the action stage.

- If one agent announces $(Y, 0)$ and the other announces (N, \cdot) , then set the compensation scheme equal to $w_i^R(1) = -\varepsilon$ and $w_i^R(0) = 1$ for each agent and go to the action stage. We will call this “the reverse compensation scheme” and denote it by $w^R = (w_1^R, w_2^R)$.
- In all other cases, pay 10 to the agent announcing the highest number (breaking ties in favor of agent 1) and go to the action stage adopting the reverse compensation scheme w^R .

Action stage: Each agent takes the action, and no message is issued. The outcome is observed and agents are paid according to the compensation scheme decided at the pre-action stage.

We now have the following proposition:

PROPOSITION 1. *The proposed mechanism has a unique subgame-perfect equilibrium. In the equilibrium, the agents announce $(Y, 0)$ at the pre-action stage and play (c, c) at the action stage.*

The formal proof of the proposition is in the Appendix. Here I give an intuitive account.

The reason why this mechanism uniquely implements (c, c) and the correct compensation scheme is that it makes sure at the pre-action stage that an agent can always obtain a utility equal to at least 1. This is so because in the case in which the reverse compensation scheme is chosen, there is a unique equilibrium at the action stage in which agents take (s, s) , so that their utility is 1. Furthermore, at the pre-action stage the only equilibrium has both agents saying $(Y, 0)$. If not, the integer game would be triggered, and this cannot be part of any equilibrium. We conclude that in all equilibria the agents obtain a utility equal to at least 1 and they are paid according to w^* . This implies that it cannot be part of an equilibrium that (n, n) is taken at the action stage, since this would generate a utility inferior to 1 and it would induce a deviation at the pre-action stage.

We can check that in fact Condition 3 is satisfied. Let $a = (n, n)$, $\bar{a} = (c, c)$, $i(a, \bar{a}) = 1$, $\tilde{a}(a, \bar{a}) = (s, s)$, and $w^y(a, \bar{a}) = w^R$. Then,

- (1) $1 - \phi^1(s, s) > 1 - \phi^1(n, n)$.
- (2) $1 - \phi^1(s, s) \leq 1 - \phi^1(c, c)$.
- (3) The action profiles yielding 1 satisfy Condition 1 with respect to (s, s) . The remaining action profiles satisfy Condition 2 (since it is always convenient for one of the agents to switch to s).

We can now state the theorem about necessary conditions for implementation. As we observed above, if we want to implement the pair (a^*, w^*) , then any mechanism must make sure that the message profile leading to the choice of w^* is not an equilibrium when agents plan to take action profile $a \neq a^*$. This means that any action profile $a \neq a^*$ must satisfy at least one of the conditions stated above with respect to (a^*, w^*) . More formally, we have the following result:

THEOREM 1. *If the pair (a^*, w^*) is implementable, then each action profile $a \neq a^*$ satisfies at least one of the Conditions 1–3 with respect to (a^*, w^*) .*

The union of Conditions 1–3 is a necessary condition for implementation of a pair (a^*, w^*) . The role played by such condition is therefore analogous to the role played by Maskin monotonicity for Nash implementation or by condition α (see Abreu and Sen, 1990) for subgame-perfect implementation. To complete the analogy, we have to show that the condition is also sufficient when there are at least three agents. This is done in the next section.

5. SUFFICIENCY

In this section, it is shown that when each action profile $a \neq a^*$ satisfies at least one of the Conditions 1–3 with respect to a pair (a^*, w^*) and there are at least three agents, then we can build a mechanism to implement such a pair. In the following assume for simplicity that there is a sufficiently large amount of money K and a sufficiently low amount $-\xi$, such that for each pair a, \bar{a} we have $U^i(a, K) > U^i(\bar{a}, w_i(v))$ and $U^i(a, -\xi) < U^i(\bar{a}, w_i(v))$ for each agent i , for each v , and for each compensation scheme w entering in one of the Conditions 1–3. In other words, it is always possible to “reward enough” and “punish enough” an agent.

THEOREM 2. *Suppose that each action profile $a \neq a^*$ satisfies one of the Conditions 1–3 with respect to (a^*, w^*) . If there are at least three agents, then the pair (a^*, w^*) can be implemented in subgame-perfect equilibrium.*

In the remainder of this section, I describe the mechanism for implementation. In the Appendix, I show that there exists a subgame-perfect equilibrium yielding (a^*, w^*) and that no subgame-perfect equilibrium has an outcome other than (a^*, w^*) .

Define A^1 as the set of action profiles that satisfy Condition 1 with respect to a^* , A^2 the set of actions such that $a \notin A^1$ and Condition 2 is satisfied with respect to (a^*, w^*) , and A^3 the set of actions such that $a \notin A^1 \cup A^2$ and Condition 3 is satisfied with respect to (a^*, w^*) . Since (a^*, w^*) is implementable, $A^1 \cup A^2 \cup A^3 = A \setminus a^*$.

5.1. The Pre-action Stage. Each agent announces an action profile $a \in A$ and an integer number. The outcome function is as follows:

- $N - 1$ agents announce $(a^*, 0)$, and agent i announces (a, \cdot) . If $a \in A^3$ and $i = i(a, a^*)$ (i.e., i is the agent identified by Condition 3), then go to the action stage defining (\bar{a}, w^j) as the “standing message,” where $\bar{a} = \bar{a}(a, a^*)$ is the action profile identified by Condition 3 and $w^j = w^j(a, a^*)$ is the compensation scheme identified by Condition 3. Otherwise, go to the action stage defining (a^*, w^*) as the “standing message.”
- In all other cases, the agent announcing the highest integer receives a large sum K , all other agents receive $-\xi$, and the game ends. Ties are broken in favor of the lower index (this tie-breaking rule is used in all the following stages).

5.2. *The Action Stage.* Let (a, w) be the standing message. Each agent i takes an action and announces an action profile $\tilde{a} \in A$ and an integer number. The outcome function is as follows:

- $N - 1$ agents announce $(a, 0)$, and agent i announces (\hat{a}, \cdot) . Go to the post-action stage defining the standing message as follows:
 - Suppose that \hat{a} satisfies Condition 2 with respect to (a, w) and $i = i(\hat{a}, a)$, that is, i is the agent identified by the condition. In this case, let $\tilde{a}_i = \tilde{a}_i(\hat{a}, a)$ and $w_i^y = w_i^y(\hat{a}, a)$ be the action and the individual compensation scheme identified by the condition. The standing message becomes $((\tilde{a}_i, \tilde{a}_{-i}), (w_i^y, w_{-i}))$.
 - Otherwise, the standing message remains (a, w) .
- In all other cases, the agent announcing the highest integer receives a large sum K , all other agents receive $-\zeta$, and the game ends.

5.3. *The Post-action Stage.* Each agent announces $\tilde{a} \in A$ and an integer number. The outcome function differs depending on the standing message. In the sequel, when $N - 1$ agents announce an identical message $(a, 0)$, we will use the expression “the remaining agent does not challenge a ” to mean that the remaining agent k issues a message (\tilde{a}, \cdot) such that either \tilde{a} does not satisfy Condition 1 with respect to a or $k = i(\tilde{a}, a)$, that is, the remaining agent is the test agent. On the other hand, we will use the expression “ k successfully challenges a ” to mean that k issues a message (\tilde{a}, \cdot) such that \tilde{a} satisfies Condition 1 with respect to a and $k \neq i(\tilde{a}, a)$.

CASE 1. The standing message is (a, w) and it did not change at the action stage:

- $N - 1$ agents announce $(\tilde{a}, 0)$, and the remaining agent k does not challenge \tilde{a} . In this case,
 - If $\tilde{a} = a$, then pay the agents according to w .
 - If $\tilde{a} \neq a$, then all agents are paid $-\frac{3}{2}\zeta$ except the agent who announced the highest integer at the action stage,⁴ who is paid $-\zeta$.
- $N - 1$ agents announce $(\tilde{a}, 0)$, and the remaining agent k announces (\hat{a}, \cdot) and successfully challenges \tilde{a} . In this case, go to the substage $\Gamma(\hat{a}, \tilde{a})$.
- In all other cases, the agent announcing the highest integer receives a large sum K , all other agents receive $-\zeta$, and the game ends.

CASE 2. The standing message was (a, w) at the pre-action stage, and it was changed to $((\tilde{a}_i, \tilde{a}_{-i}), (w_i^y, w_{-i}))$ by agent i at the action stage:

- $N - 1$ agents announce $(\bar{a}, 0)$, and the remaining agent k does not challenge \bar{a} . In this case,

⁴ Notice that we are using the integer number *announced at the previous stage* to determine the agent paying less. This will be used in the proof to make sure that an equilibrium always exists at the post-action stage.

- If $\exists a_i''$ such that $\bar{a} = (a_i'', a_{-i})$ and $(\tilde{a}_i, \hat{a}_{-i})$ does not satisfy Condition 1 with respect to \bar{a} , then agent i is paid according to $\bar{w}_i(a_i'')$ defined by Condition 2, and all other agents are paid according to w_{-i} .
- If there is no a_i'' such that $\bar{a} = (a_i'', a_{-i})$ and $(\tilde{a}_i, \hat{a}_{-i})$ does not satisfy Condition 1 with respect to \bar{a} , then each agent is paid according to (w_i^y, w_{-i}) .
- In all other cases, all agents are paid $-\frac{3}{2}\xi$ except the agent who announced the highest integer at the action stage,⁵ who is paid $-\xi$.
- $N - 1$ agents announce $(\bar{a}, 0)$, and the remaining agent k announces (\hat{a}, \cdot) and successfully challenges \bar{a} . In this case, go to the substage $\Gamma(\hat{a}, \bar{a})$.
- In all other cases, the agent announcing the highest integer receives a large sum K , all other agents receive $-\xi$, and the game ends.

The substage $\Gamma(\hat{a}, \bar{a})$. Let j be the agent and (w_j^x, w_j^y) the two compensation schemes identified by Condition 1. Each agent announces either Y or N and an integer number. The outcome function is as follows:

- If at least $N - 1$ agents announce $(Y, 0)$ and agent j announces (Y, \cdot) , then each agent is paid -2ξ and j is paid w_j^x .
- If $N - 1$ agents announce $(Y, 0)$ and agent j announces (N, \cdot) , then each agent is paid -2ξ and j is paid w_j^y .
- If at least $N - 1$ agents announce $(N, 0)$, then agent i (the agent who “disagreed” at the post-action stage) is paid a large sum K and all other agents are paid -2ξ .
- In all other cases, the agent announcing the highest integer receives a large sum K , all other agents receive -2ξ , and the game ends.

I now provide some intuition about the structure of the mechanism and how it works.

At the pre-action stage, the agents are asked whether they are going to take action profile a^* at the action stage. If all the agents confirm that this is the case, then the game proceeds normally to the action stage. If one dissident announces that actually the agents are going to take action profile $a \neq a^*$ and this action profile satisfies Condition 3, then the rules of the game change. Agents are asked to take the action profile $\tilde{a} = \tilde{a}(a, a^*)$ (given by Condition 3) and are compensated according to $w^y = w^y(a, a^*)$ (again identified by Condition 3). In this way it is made sure that an action profile a that satisfies Condition 3 cannot be part of an equilibrium. The reason is that at this stage the only possible equilibrium is one in which all agents announce $(a^*, 0)$, since in all other cases the integer game is triggered. But if agents plan to take $a \neq a^*$ satisfying Condition 3, then agent $i(a, a^*)$ identified by the condition has a profitable deviation.

In successive stages, by and large the same reasoning applies. We now have to ensure two things. First, if everything went smoothly at the pre-action stage (i.e., all agents announced $(a^*, 0)$), then we want (a^*, w^*) to be the unique outcome. Second,

⁵ Again, notice that the integer determining the winner is the one announced at the previous stage.

if a dissident forced a change in the action profile at the pre-action stage, then we have to make sure that the deviation is profitable; that is, we have to make sure that in all equilibria following the deviation, the agent is better off (notice that this only happens out of equilibrium). This is accomplished forcing agents to announce again the action profile they are taking. When they confirm that they are taking the prescribed action profile, then the game proceeds normally to the post-action stage. If they are taking an action profile that is not the prescribed one, then a dissident can deviate and denounce the other agents. Suppose, for example, that the agents confirmed at the pre-action stage that they intend to take a^* . If they are actually planning to take a at the action stage, then an agent can denounce the deviation. If a satisfies Condition 1 with respect to a^* , then the accusation will be checked at the post-action stage. Otherwise, a satisfies Condition 2 (if a only satisfies Condition 3, then the deviation would have occurred at the pre-action stage), and there is an agent $i(a, a^*)$ who can deviate from a , ask for a different wage scheme w^v , and be better off. In equilibrium it must be the case that all agents confirm the action profile a^* , since otherwise the integer game is triggered. But there will be no deviation from an announcement of a^* only if a^* is actually the action profile taken by the agents. Thus, in equilibrium a^* is taken and the agents truthfully report the action profile they are taking. The same reasoning applies when the “standing action profile” is different from a^* , that is, when some agents forced a change in the action profile at the pre-action stage.

The post-action stage is similar to other mechanisms used for implementation in subgame-perfect equilibrium. Here agents are asked which action profile was taken at the action stage. If they unanimously report that the action profile was the prescribed one, then the game ends. If a dissident announces that a different action profile was taken, then the accusation is checked by asking a test agent to choose between two different compensations. The choice of the test agent reveals the true action profile, and the dissident is rewarded if the accusation was correct or punished if the accusation was false.

To sum up, the idea of the mechanism is to use different stages to avoid that action profiles different from a^* could be part of an equilibrium. If the agents plan to take an action profile $a \neq a^*$ that satisfies Condition 3, then a profitable deviation will be available at the pre-action stage. If a satisfies Condition 2, then the profitable deviation will be at the action stage, and if Condition 1 is satisfied, then the deviation will be at the post-action stage.

6. CONCLUSION

I have provided necessary and sufficient conditions for the implementation of an action profile a^* and a vector of compensation schemes w^* , thus obtaining a full characterization of the implementable pairs (a^*, w^*) . This constitutes the natural counterpart of the necessary and sufficient condition for implementation in the traditional (hidden-information) implementation framework. The “hidden action” case presents some interesting differences with respect to the traditional case. I have shown that mechanisms implementing action profiles have a natural sequential structure centered on the “action taking” stage, and that necessary and sufficient

conditions can be obtained looking at some reversal of preferences in different parts of the mechanism. Previous work has been more ad hoc in the sense that special pairs (a^*, w^*) were considered and a mechanism for implementation was found. In general, these mechanisms rely on some reversal of preferences at what we have called the “post-action stage.” I have shown that in fact implementation may be possible under more general conditions involving reversal of preferences at the “action stage” or at the “pre-action stage.”

APPENDIX

PROOF OF PROPOSITION 1. I first show that whenever the action stage is reached with the reverse scheme, the only equilibrium is (s, s) . Taking into account both the disutility of actions and the compensation scheme, the game to be played among the two agents will have the payoffs shown in Table A.1.

The game shown in Table A.1 has (s, s) as the unique Nash equilibrium. To see this, observe first that there is no equilibrium in which agent 2 selects n . If this were the case, agent 1 would choose c , but then n would not be a best response for agent 2. Therefore, in all equilibria agent 2 must take s or c . This in turn implies that agent 1 does not select n , which is dominated by s . But then 2 must put probability 1 on s , which strictly dominates c and n when agent 1 does not use n . The unique best response for agent 1 is therefore s , and we conclude that (s, s) is the unique equilibrium.

This implies that both agents can always attain a utility of at least 1, just by announcing $(N, 0)$. This invariably leads to a utility of 1 at the action stage, and possibly more if the integer game is triggered and won. In turn this implies that in all subgame-perfect equilibria of the mechanism, both agents must attain a utility at least equal to 1.

Next, I show that the only equilibrium is that both agents announce $(Y, 0)$, $(Y, 0)$ and play (c, c) at the action stage. There can be no equilibrium in which agent 1 announces something different from $(Y, 0)$. In this case, the best response for agent 2 would be to announce (N, k) , with k winning the integer game. But then the best response of 1 would be to announce (N, k') with $k' > k$. It is easy to see that no pair of messages can be an equilibrium, since at least one agent would want to choose a higher number. Therefore, in all equilibria agent 1 announces $(Y, 0)$ with probability 1. An identical argument establishes that the same must be true for agent 2, so that

TABLE A.1

		Agent 2		
		s	n	c
Agent 1	s	1, 1	$-\epsilon - 1, -\epsilon - 1$	1, 0
	n	$-\epsilon - 1, -\epsilon - 1$	$-\epsilon - 1, -\epsilon - 1$	$\frac{1}{2}, 0$
	c	0, 1	$0, \frac{1}{2}$	$-\epsilon, -\epsilon$

$((Y, 0), (Y, 0))$ is the only possible equilibrium announcement. This in turn implies that in all equilibria the agents are paid according to the correct compensation scheme. Furthermore, since in all equilibria all agents obtain a utility of at least 1, it must be the case that (c, c) is played at the action stage. ■

PROOF OF THEOREM 1. Suppose that there exists an extensive form mechanism implementing (a^*, w^*) in subgame-perfect equilibrium. A formal description of an extensive form mechanism is beyond the scope of this article, and I simply refer to Selten (1975). Here, I introduce only the notation that is helpful in understanding the proof.

Denote by M the set of all possible sequences of messages that can be observed, M_1 the set of possible sequences of messages that can be observed before the action stage, M_a the set of possible messages that can be observed at the action stage, and M_2 the set of possible sequences of messages that can be observed at the post-action stage. The function g denotes the mapping between the set of messages and the compensation scheme that is chosen. Thus, $g(m) = w$, with $m \in M$. For given messages $m_1 \in M_1$, $m_a \in M_a$, define $M_2(m_1, m_a)$ as the set of message profiles observable at the post-action stage and $g_2(m_2 | (m_1, m_a))$ as the outcome function for this subgame. Notice that the action profile a is not observable by the planner, so that the particular game form adopted after history (m_1, m_a, a) can depend on (m_1, m_a) but not on a . However, a is observed by the players, so that they can condition their strategies on a .

A strategy profile σ describes the behavior taken by the agents at each information set. A given σ yields a sequence of messages $m \in M$ and an action profile a . In turn, the sequence of messages yields (through the outcome function g) a compensation scheme.

Since the mechanism implements (a^*, w^*) , there exists an equilibrium strategy profile σ^* yielding this outcome. Let us call m_1^* the vector of messages sent at the pre-action stage when agents follow σ^* . Since the strategy profile yields (a^*, w^*) , we have that σ^* yields the pair (a^*, m_a^*) at the action stage (i.e., following m_1^* , the action profile a^* is taken and a message m_a^* is issued) and a message profile m_2^* at the post-action stage when (m_1^*, m_a^*, a^*) have been observed. Then $g_2(m_2^* | (m_1^*, m_a^*)) = w^*$; that is, when messages m_1^* , m_a^* , and m_2^* are issued, the selected compensation scheme is w^* .

As an observation that will be used later, notice the following fact: Consider the strategy profile σ^* . Let γ be a subgame that is reached when agent i only makes a unilateral deviation from σ^* . Call σ_γ^* the restriction of σ^* to γ , and call $U^i(\sigma_\gamma^*)$ the expected utility of player i in subgame γ when the strategy profile σ_γ^* is played. Notice that γ is not reached in equilibrium (it can only be reached when i makes a deviation). Since σ^* is a subgame-perfect equilibrium for the whole game, it must be the case that σ_γ^* is a subgame-perfect equilibrium for γ . Suppose that there is another subgame-perfect equilibrium of γ , say $\hat{\sigma}_\gamma$, such that $U^i(\hat{\sigma}_\gamma) \leq U^i(\sigma_\gamma^*)$. Then we can construct another subgame-perfect equilibrium implementing (a^*, w^*) as follows: Use strategy profile $\hat{\sigma}_\gamma$ when γ is reached, and use σ^* otherwise. This is still a subgame-perfect equilibrium, since γ can only be reached when i deviates, and this is still not profitable for i . More in general, in subgames that can only be reached

through a deviation from σ^* by a single agent, we are free to substitute to the original strategy profile, other strategy profiles that are subgame-perfect equilibria and yield a lower utility for the deviant agent.

I now show that in order to avoid that $a \neq a^*$ be the outcome of a subgame-perfect equilibrium, it is necessary that at least one of the Conditions 1–3 is satisfied.

Consider such an action profile a and the following strategy profile σ :

- At the pre-action stage, each agent follows strategy σ_1^{*i} .
- If m_1^* is observed after the pre-action stage, then each agent takes action a_i (rather than a_i^*) and issues message m_a^{*i} (i.e., the same message as in the truth-telling equilibrium). In all other cases, behave as in the original equilibrium.
- At the post-action stage, use strategy $\sigma_2^{*i}(m_1^*, m_a^*, a^*)$ if (m_1^*, m_a^*, a) has been observed. In all other cases, behave as in the original equilibrium. In other words, whenever all agents take the actions prescribed in action profile a and issue the message (m_1^*, m_a^{*i}) , then behave as if the observed action profile were a^* . Otherwise, behave as in the original equilibrium.

This strategy profile results in the action profile a and a sequence of messages (m_1^*, m_a^*, m_2^*) . Since this is the message observed in the “right” equilibrium, the compensation scheme w^* is selected. Therefore, the overall outcome is $(a, w^*) \neq (a^*, w^*)$, and the strategy profile cannot be a subgame-perfect equilibrium. We now prove that this implies that a satisfies at least one of the Conditions 1–3.

Since σ defined above is not a subgame-perfect equilibrium, there must be at least one subgame and one agent with a profitable deviation. Given the definition of the strategy profile, we observe that

- no deviation is possible in subgames following $m_1 \neq m_1^*$;
- no deviation is possible in subgames following (m_1^*, m_a) if there is an agent i such that $m_a^i \neq m_a^{*i}$;
- no deviation is possible after observing an action profile $\hat{a} \neq a$.

The reason is that in all those subgames exactly the original equilibrium strategy is followed. Deviations can therefore occur

- (1) at the post-action stage in some subgame following (m_1^*, m_a^*, a) ;
- (2) at the action stage following a message m_1^* ;
- (3) at the pre-action stage.

Suppose first that the deviation occurs at a subgame following the sequence of messages and actions (m_1^*, m_a^*, a) , and let us call j the agent with a profitable deviation. Suppose that by following strategy $\sigma_2^j(m_1^*, m_a^*, a^*)$, the outcome is w^x (notice that the subgame need not be reached in equilibrium), and that a deviation by agent j yields a compensation scheme w^y . Since the deviation is profitable, we have $U^j(a, w^y) > U^j(a, w^x)$. On the other hand, the same deviation is not profitable

in the original equilibrium, so that $U^j(a^*, w_j^x) \geq U^j(a^*, w_j^y)$. The implication is that Condition 1 is satisfied.

More in general, the reasoning above leads to the conclusion that whenever action profile a does not satisfy Condition 1 with respect to another action profile a' , then the set of subgame-perfect equilibria following (m_1, m_a, a) must contain the set of equilibria following (m_1, m_a, a') . In other words, if a strategy profile in the subgame following (m_1, m_a, a') is a subgame-perfect equilibrium, then the same strategy profile must remain a subgame-perfect equilibrium in the subgame following (m_1, m_a, a) . This observation will be used later.

Suppose next that a profitable deviation is available at the action stage. I show that in this case the action profile a must satisfy Condition 2 with respect to (a^*, w^*) . The existence of a profitable deviation implies that there exists an agent i , an action $\tilde{a}_i \neq a_i$, and a message \tilde{m}_a^i such that the agent is better off taking the action \tilde{a}_i , issuing the message \tilde{m}_a^i , and then obtaining the resulting compensation scheme. Following a deviation \tilde{a}_i , the profile of actions taken is (\tilde{a}_i, a_{-i}) . Define $E^i(m_1^*, (\tilde{m}_a^i, m_a^{*-i}), (\tilde{a}_i, a_{-i}))$ as the set of possible expected utilities that can be obtained by agent i in a subgame-perfect equilibrium following $(m_1^*, (\tilde{m}_a^i, m_a^{*-i}), (\tilde{a}_i, a_{-i}))$. Define $\underline{U}^i = \inf \{E^i(m_1^*, (\tilde{m}_a^i, m_a^{*-i}), (\tilde{a}_i, a_{-i}))\}$. We can assume without loss of generality that $\underline{U}^i \geq U^i(a, w_i^*)$. If this were not the case, then the subgame γ identified by the message $(m_1^*, (\tilde{m}_a^i, m_a^{*-i}))$ and by the action profile (\tilde{a}_i, a_{-i}) would have a subgame-perfect equilibrium $\hat{\sigma}_\gamma$ such that $U^i(\hat{\sigma}_\gamma) < U^i(a, w_i^*)$. We could then construct a new strategy profile substituting $\hat{\sigma}_\gamma$ in the relevant subgame. This would make the deviation by agent i not profitable. Since this could be done for each agent and each unilateral deviation, there would be no profitable deviation at the action stage.

I now proceed to show that the existence of a profitable deviation at the action stage implies that Condition 2 is satisfied. Define $\tilde{m}_2 = \sigma_2(m_1^*, (\tilde{m}_a^i, m_a^{*-i}), (\tilde{a}_i, a_{-i}))$ the message issued at the post-action stage by the agents in equilibrium in the subgame following the deviation, and $w^y = g_2(\tilde{m}_2 | m_1^*, (\tilde{m}_a^i, m_a^{*-i}))$ the resulting compensation scheme. Then the profitability of the deviation implies

$$(A.1) \quad U^i((\tilde{a}_i, a_{-i}), w_i^y) > U^i(a, w_i^*)$$

that is, part 1 of Condition 2 is satisfied. To see that part 2 is also satisfied, for a given \hat{a}_i define $m_2(\hat{a}_i) = \sigma_2(m_1^*, (\tilde{m}_a^i, m_a^{*-i}), (\hat{a}_i, a_{-i}^*))$ as the message resulting after action profile (\hat{a}_i, a_{-i}^*) is taken and message $(m_1^*, (\tilde{m}_a^i, m_a^{*-i}))$ is observed, and $\bar{w}(\hat{a}_i) = g_2(m_2(\hat{a}_i) | m_1^*, (\tilde{m}_a^i, m_a^{*-i}))$ the corresponding outcome. Since the deviation is not profitable in the “correct” equilibrium, we have

$$U^i((\hat{a}_i, a_{-i}^*), \bar{w}_i(\hat{a}_i)) \leq U^i(a^*, w_i^*)$$

There are two possibilities. Either the strategy profile following $(m_1^*, (\tilde{m}_a^i, m_a^{*-i}), (\hat{a}_i, a_{-i}^*))$ is a subgame-perfect equilibrium after $(m_1^*, (\tilde{m}_a^i, m_a^{*-i}), (\hat{a}_i, a_{-i}^*))$ or not. If it is not, then (\hat{a}_i, a_{-i}^*) must satisfy Condition 1 with respect to (\hat{a}_i, a_{-i}^*) . If it is, then we consider the following two cases:

- (1) $\underline{U}^i \in E^i((\tilde{m}_a^i, m_a^{*-i}), (\tilde{a}_i, a_{-i}))$. In this case, we can assume without loss of generality that in the subgame-perfect equilibrium following

$((\bar{m}_a^i, m_a^{*-i}), (\tilde{a}_i, a_{-i}))$, agent i receives an expected utility equal to \underline{U}^i ; that is, the worst possible equilibrium for agent i is played after a unilateral deviation by agent i . Then, we have

$$U^i((\tilde{a}_i, a_{-i}), \bar{w}_i(\hat{a}_i)) \geq U^i((\tilde{a}_i, a_{-i}), w_i^y)$$

since w_i^y was the worst equilibrium. Using (A.1) we conclude

$$U^i((\tilde{a}_i, a_{-i}), \bar{w}_i(\hat{a}_i)) > U^i(a, w_i^*)$$

so that the second part of Condition 2 is satisfied as well.

- (2) $\underline{U}^i \notin E^i((\bar{m}_a^i, m_a^{*-i}), (\tilde{a}_i, a_{-i}))$. In this case, for every subgame-perfect equilibrium following $((\bar{m}_a^i, m_a^{*-i}), (\tilde{a}_i, a_{-i}))$, agent i receives an expected utility strictly greater than \underline{U}^i . This implies

$$U^i((\tilde{a}_i, a_{-i}), \bar{w}_i(\hat{a}_i)) > \underline{U}^i \geq U^i(a, w_i^*)$$

so that again the second part of Condition 2 is satisfied.

At last, a profitable deviation may be available at the pre-action stage. I show that in this case Condition 3 has to be satisfied.

Let i be the agent with a profitable deviation. The profitable deviation must be along the equilibrium path and not in some unreached subgame of the pre-action stage, since in those subgames strategies are specified to be the subgame-perfect equilibria of the correct equilibrium. Let \bar{m}_1 be the message resulting at the pre-action stage as a consequence of a profitable deviation by agent i (other agents may send messages other than in the correct equilibrium because there may be substages at the pre-action stage, and agents may react to an out-of-equilibrium announcement). As a consequence of message \bar{m}_1 , the action profile \bar{a} and the message profile \bar{m}_a result at the action stage, and the message $\bar{m}_2 = \sigma_2(\bar{m}_1, \bar{m}_a, \bar{a})$ is issued at the post-action stage. These are the same action profiles and messages given by σ^* . Define $w^y = g(\bar{m}_1, \bar{m}_a, \bar{m}_2)$ and observe that, since this is a profitable deviation, we have $U^i(\bar{a}, w_i^y) > U^i(a, w_i^*)$. Since the deviation is not profitable in the truth-telling equilibrium, we have $U^i(\bar{a}, w_i^y) \leq U^i(a^*, w_i^*)$. This implies that parts 1 and 2 of Condition 3 are satisfied.

To show that part 3 is satisfied as well, consider an action profile \hat{a} such that $U^i(\hat{a}, w_i^y) \leq U^i(a, w_i^*)$. To make things simple, assume that (\bar{a}, w_i^y) is the worst subgame-perfect equilibrium outcome for agent i following the deviation at the pre-action stage (the treatment of the case in which there is no worst subgame-perfect equilibrium outcome is similar to the one discussed above for the case of deviations at the action stage, and it is omitted here). Since $U^i(\hat{a}, w_i^y) \leq U^i(a, w_i^*) < U^i(\bar{a}, w_i^y)$, the following cannot be a subgame-perfect equilibrium for the game starting at the action stage after message \bar{m}_1 :

- At the action stage, each agent takes action \hat{a}_i and issues message \bar{m}_a^i .
- After the action stage, if (\hat{a}, \bar{m}_a) is observed, then follow strategy $\sigma_2(\bar{m}_1, \bar{m}_a, \bar{a})$. Otherwise, follow the original strategy of the equilibrium yielding (\bar{a}, w^y) .

If \hat{a} does not satisfy either Condition 1 with respect to \bar{a} or Condition 2 with respect to (\bar{a}, w^y) , then this is in fact a subgame-perfect equilibrium, a contradiction. The reasoning is analogous to the one used before. If \hat{a} does not satisfy Condition 1 with respect to \bar{a} , then no profitable deviation can exist at the post-action stage, as it would also be a profitable deviation in the original equilibrium. Thus, a profitable deviation must exist at the action stage. However, if \hat{a} does not satisfy Condition 2 with respect to (\bar{a}, w^y) , then any profitable deviation against \hat{a} is also a profitable deviation against \bar{a} . ■

PROOF OF THEOREM 2. The proof is divided into two lemmas. Lemma 1 shows that there is a subgame-perfect equilibrium with outcome (a^*, w^*) . Lemma 2 shows that there is no subgame-perfect equilibrium with outcome other than (a^*, w^*) .

LEMMA 1. *There exists a subgame-perfect equilibrium with outcome (a^*, w^*) .*

PROOF. I show that the following strategy profile is a subgame-perfect equilibrium yielding (a^*, w^*) .

Pre-action stage: Each agent announces $(a^*, 0)$.

Action stage: When the standing message is (a, w) , each agent i takes action a_i and announces $(a, 0)$.

Post-action stage: If a is the true action profile observed, then announce $(a, 0)$. If a game $\Gamma(\hat{a}, \tilde{a})$ is reached, then announce $(Y, 0)$ if agent j weakly prefers w_j^x to w_j^y . Otherwise, announce $(N, 0)$.

I now check, working backward, that this is a subgame-perfect equilibrium.

At $\Gamma(\hat{a}, \tilde{a})$, the only agent who can change the outcome is j . It is clear that the prescribed strategy is optimal.

At the post-action stage, an agent i can deviate announcing a false action profile \hat{a} that satisfies Condition 1 with respect to the true action profile a and such that $j(\hat{a}, a) \neq i$. Since the other agents are announcing the true action profile, the equilibrium at $\Gamma(\hat{a}, a)$ is that everybody announces $(Y, 0)$, making the deviant agent worse off.

Consider now the action stage. Let (a, w) be the standing message. Suppose first that an agent i changes the action to $\hat{a}_i \neq a_i$ without changing the message (or issuing a message that does not change the standing message (a, w)). In this case, the mechanism moves to case 1 of the post-action stage. However, the observed action profile will be $(\hat{a}_i, a_{-i}) \neq a$. Following the equilibrium strategy, all agents other than i report the observed action profile and zero; that is the announcement of all agents other than i at the post-action stage will be $((\hat{a}_i, a_{-i}), 0)$. If agent i successfully challenges this report (that is, agent i announces an action profile that satisfies Condition 1 with respect to (\hat{a}_i, a_{-i}) and such that the test agent is not i), then the subgame Γ is reached and agent i ends up being worse off, since the true action profile is confirmed. If i does not successfully challenge the report by other agents, and since an action profile different from the standing message is reported, the best that a deviant agent can hope to obtain is $-\zeta$. This is clearly not profitable.

Suppose now that agent i changes the standing message; i.e., agent i reports (\hat{a}, \cdot) with \hat{a} satisfying Condition 2 with respect to a and $i = i(\hat{a}, a)$. Agent i may or may not also change the action. Let us call \bar{a}_i the action undertaken. Define $\tilde{a}_i = \tilde{a}_i(\hat{a}, a)$ and $w_i^y = w_i^y(\hat{a}, a)$ as identified by Condition 2. The resulting standing message is $((\tilde{a}_i, \hat{a}_{-i})(w_i^y, w_{-i}))$, and the action profile that is actually observed is (\bar{a}_i, a_{-i}) .

At the post-action stage, the mechanism will be in Case 2. Following the equilibrium strategy, all agents other than i will report the true action profile (\bar{a}_i, a_{-i}) . This implies that if agent i does not challenge the announcement of the other agents, she is paid according to $\bar{w}_i(\bar{a}_i)$ whenever $(\bar{a}_i, \hat{a}_{-i})$ does not satisfy Condition 1 with respect to (\bar{a}_i, a_{-i}) and at most $-\xi$ otherwise. By definition of $\bar{w}_i(\bar{a}_i)$, this does not make agent i better off. If agent i challenges the report, then game Γ is reached and again agent i is worse off.

At last, no deviation at the pre-action stage is profitable, by part 2 of Condition 3. ■

LEMMA 2. *There is no subgame-perfect equilibrium with an outcome other than (a^*, w^*) .*

PROOF. All equilibria must have the following structure:

- All agents announce $(a^*, 0)$ at the pre-action stage.
- If (\hat{a}, \hat{w}) is the standing message at the action stage, then all agents announce $(\hat{a}, 0)$.
- Whenever all agents announce $(\hat{a}, 0)$ at the action stage, they also announce $(\hat{a}, 0)$ at the post-action stage. If not, the integer game would be triggered at the action stage.
- When the action profile a has been taken, then all agents announce unanimously an action profile \hat{a} at the post-action stage such that a does not satisfy Condition 1 with respect to \hat{a} .
- If a substage $\Gamma(\hat{a}, \hat{a})$ is reached and agent $j(\hat{a}, \hat{a})$ strictly prefers $w_j^y(\hat{a}, \hat{a})$ to $w_j^x(\hat{a}, \hat{a})$, then the only equilibrium is that all agents announce $(N, 0)$.

The implication is that in all equilibria the action stage is reached with standing message (a^*, w^*) , and the announcement is confirmed at the action stage and then again at the post-action stage. I have only to show that when this sequence of messages occurs, no action profile $a \neq a^*$ can be part of an equilibrium.

Suppose that there is an equilibrium in which at the action stage the agents take an action profile $a \neq a^*$. It can never be the case that action a satisfies Condition 1 with respect to a^* . The reason is that at the post-action stage, one agent could announce the true action profile, thus reaching $\Gamma(a, a^*)$ and obtaining K .

If a satisfies Condition 2, then agent i identified by the condition has a profitable deviation, namely, take action \tilde{a}_i and announce (a, \cdot) . This implies that the standing message becomes $((\tilde{a}_i, a_{-i})(w_i^y, w_{-i}^*))$. After such a message profile, the only equilibria at the post-action stage are that all agents announce an action profile $(\hat{a}, 0)$ such that (\tilde{a}_i, a_{-i}) does not satisfy Condition 1 with respect to \hat{a} . Thus, either w_i^y or $w_i(\hat{a}_i)$ is implemented and agent i is better off.

Finally, suppose a satisfies Condition 3. In this case, the agent i identified by Condition 3 can profitably deviate at the pre-action stage announcing $(a, 0)$. Let $\tilde{a} = \tilde{a}(a, a^*)$ and w^y be the action profile and compensation scheme identified by Condition 3.

Consider now the subgame starting at the action stage having as standing message (\tilde{a}, w^y) . Following such an announcement, in all subgame-perfect equilibria it must be the case that all agents announce unanimously $(\tilde{a}, 0)$ at the action stage and confirm the message at the post-action stage, so that w^y is actually implemented. It remains to show that the action profile \hat{a} taken in equilibrium is such that $U^i(\hat{a}, w^y) > U^i(a, w^*)$.

Suppose not, so that $U^i(\hat{a}, w^y) \leq U^i(a, w^*)$. Then, part 3 of Condition 3 tells us that \hat{a} either satisfies Condition 1 with respect to \tilde{a} or it satisfies Condition 2 with respect to (\tilde{a}, w^y) . In the first case, a unanimous announcement of $(\tilde{a}, 0)$ cannot be an equilibrium at the post-action stage. In the second case, there is an agent with a profitable deviation at the action stage. ■

REFERENCES

- ABREU, D., AND A. SEN, "Subgame Perfect Implementation: A Necessary and Almost Sufficient Condition," *Journal of Economic Theory* 50 (1990), 285–99.
- ABREU, D., AND H. MATSUSHIMA, "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica* 60 (1992), 993–1008.
- ARYA, A., AND J. GLOVER, "A Simple Forecasting Mechanism for Moral Hazard Settings," *Journal of Economic Theory* 66 (1995), 507–21.
- , ———, AND J. HUGHES, "Implementing Coordinated Team Play," *Journal of Economic Theory* 74 (1997), 218–32.
- BRUSCO, S., "Implementing Action Profiles with Sequential Mechanism," *Review of Economic Design* 3 (1998), 271–300.
- DEMSKI, J., AND D. SAPPINGTON, "Optimal Incentive Contracts with Multiple Agents," *Journal of Economic Theory* 33 (1984), 152–71.
- DUTTA, B., AND A. SEN, "A Necessary and Sufficient Condition for Two Person Nash Implementation," *Journal of Economic Theory* 50 (1991), 285–99.
- JACKSON, M. O., "Implementation in Undominated Strategies: A Look at Bounded Mechanisms," *Review of Economic Studies* 59 (1992), 757–75.
- , T. PALFREY, AND S. SRIVASTAVA, "Undominated Nash Implementation in Bounded Mechanisms," *Games and Economic Behavior* 6 (1994), 474–501.
- MA, C., "Unique Implementation of Incentive Contracts with Many Agents," *Review of Economic Studies* 55 (1988), 555–72.
- , J. MOORE, AND S. TURNBULL, "Stopping Agents from Cheating," *Journal of Economic Theory* 46 (1988), 355–72.
- MOOKHERJEE, D., "Optimal Incentive Schemes with Many Agents," *Review of Economic Studies* 51 (1984), 433–46.
- MOORE, J., AND R. REPULLO, "Subgame Perfect Implementation," *Econometrica* 56 (1988), 1191–1220.
- , AND ———, "Nash Implementation: A Full Characterization," *Econometrica* 58 (1990), 1083–99.
- SELTEN, R., "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory* 9 (1975), 1–12.
- SJÖSTRÖM, T., "Implementation and Information in Teams," *Economic Design* 1 (1996), 327–41.