POLITICALLY REALISTIC IMPLEMENTATION WITH INSPECTION: THE EQUITY-HONESTY-WELFARE TRADE-OFF

IGNACIO ORTUÑO-ORTIN AND JOHN E. ROEMER

We study dominant strategy mechanisms where the planner knows the distribution of types and the agents are instructed to announce their types to the planner. It is assumed that the planner has access to a technology of inspection which is costly but perfect, and that he can penalize an agent who is inspected after announcements have been made if he is found to have lied about his type. It shall be shown that, in general, the welfare-maximizing mechanism that respects minimal equity will induce some agents to lie about their types.

1. A REALISTIC APPROACH TO IMPLEMENTATION

IN THEORIES of full implementation, a planner, who is ignorant of the traits of agents in the population, wishes to implement a correspondence that associates to any economic environment some allocation or set of allocations. He designs a game form, whose equilibria (whether dominant strategy, or Nash, or subgame perfect Nash, or Bayesian Nash) induce through an outcome function the desired allocation(s). The theories of implementation using Nash equilibrium as the solution concept assume that the agents have complete information of each other's types (or at least the probability distribution of types). The planner knows nothing and the players know everything. In dominant strategy implementation, the players, of course, need have no information about each other.

In the present paper, we shall study dominant strategy mechanisms where the players know nothing about each other but the planner knows the distribution of types. We shall, for the most part, limit our discussion to direct mechanisms, where the agents are instructed to announce their types to the planner. It is assumed, in addition, that the planner has access to a technology of inspection which is costly but perfect, and that he can penalize an agent who is inspected after announcements have been made if he is found to have lied about his type.

Mechanisms in this kind of environment are of the following type. The planner announces a policy that stipulates: (1) the allocation an agent shall receive if he announces that he is of type i and is not inspected; (2) the allocation an agent receives if he announces that he is of type i and, after inspection, is found to be of type j (perhaps j = i); (3) the probability that an agent announcing type i will be inspected. We shall study a sub-class of this class of mechanisms, defined by what we consider to be two political constraints that often hold in reality: (i) the planner cannot set the penalty allocations (part (2) above), but takes them as given by a legislature, and (ii) an agent who announces that he is of type i and is found to be so upon inspection shall receive the same allocation as an agent who announces he is of type i but is not inspected. This last requirement has been called horizontal equity in the literature; we prefer to call it *minimal equity*. Without this kind of horizontal equity we would allow for the possibility to penalize an honest agent who is inspected as compared with a similar agent who is not inspected.

The following points summarize the salient aspects of the planner's implementation problem.

(i) The planner knows the distribution of types in the population and each agent knows only his own type (the planner has access to survey data about the population);
(ii) each agent reports his type to the planner;
(iii) an agent's transfer is a function only of his announcement, and the results of an inspection, if any;
(iv) the solution concept is dominant strategy equilibrium;
(v) penalties (for lying to the planner) are not an instrument of the planner;
(vi) policies are restricted by a condition of minimal equity.

In particular, it follows from the above that:

(α) no agent will be penalized because of the behavior of another agent (because of (iii)); and
(β) no agent must report on the trait of any other agent.

In addition, it will follow that:

(γ) mechanisms with inspection are robust[1] in the presence of irrational behavior on the part of a small number of agents.

(α) and (β) are desirable aspects of an implementation mechanism in democratic societies, and (γ) is desirable in any society. Typically, theories of full implementation satisfy none of (α), (β), and (γ).

The technical focus of our paper is an examination of the revelation principle for the class of mechanisms defined by (i)–(vi); the qualitative point that we wish to emphasize is that social norms give rise to problems in the design of optimal economic policies that are ignored by the literature on full implementation, but must be taken into account in the design of politically acceptable mechanisms. It turns out that the revelation principle fails because of the insistence on minimal equity. That is, it is *in general false that the minimally equitable mechanism that*

[1] By robust we mean that irrational reports by a small number of agents will not render the mechanism grossly unfeasible, or change the allocation to rational agents very much.

maximizes social welfare entails honest reporting by all agents. Typically, economists are not interested in the revelation principle *per se* - that is, because of beliefs that truth is a good thing - but only because it simplifies analysis of the problem, and perhaps of computation of the mechanism. We, however, take the point of view that:

(δ) the optimal mechanism should not require rational agents to lie to the government.

We view (δ) as a political constraint. A government that asks its citizens to report a trait, penalizes them for lying, yet designs a policy the self-interested response to which is lying - "socially optimal duplicity" - will be considered by many citizens to be either inconsistent or immoral or both.[2] (One thinks of the legal prohibition of attractive nuisances.) Call a mechanism incentive compatible if the dominant strategy for every agent is to report his true type. It is an unfortunate feature of inspection mechanisms with minimal equity that incentive compatibility can only be purchased, in general, at a welfare cost.

An alternative interpretation of our results is that minimal equity and truth-telling are incompatible with the use of a randomized incentive scheme. In an equilibrium where agents do not report their true types the government may treat agents of the same type differently: those who are inspected receive a different allocation than those who are not inspected. This different treatment is, however, prohibited under a truth-telling equilibrium with the minimal equity restriction. Stiglitz (1982) and Weiss (1978) show (in a model without inspections) that even when agents are risk-averse the existence of incentive constraints creates some nonconvexities which make randomized schemes superior to deterministic schemes. Randomized schemes, however, do not satisfy minimal equity. Thus, our problem can be seen as related to the problem studied by these authors.

In the next section, we present the general model and show that, for a certain sub-class of mechanisms satisfying (i)-(vi), ones that satisfy a "Condition A", the revelation principle holds. An immediate corollary is that the revelation principle holds for all allocation problems that are one dimensional, but our sub-class is broader than that. In section 3, we give an example where "Condition A" does not hold and the optimal mechanism is not incentive compatible. Section 4 concludes.

A word should be said about the "surprise" in this paper. Some authors have proved the revelation principle for examples of implementation problems with inspection and penalties, in which minimal equity is a feature (Melumad and Mookherjee (1989), Baron and Besanko (1984), and Monkherjee and P'ng (1990)).

[2] Note that *three* conditions must hold for socially optimal duplicity to occur: that there be (1) penalties for lying to the government, but (2) lying is rational, and (3) lying is necessary for social optimality. Thus, lying to the I.R.S. in the United States, for example, is not an instance of rational duplicity, since (3) is violated.

The problems that these authors have studied are either one dimensional allocation problems, or are ones in which Condition A trivially holds. Our *Theorem 1* implies their results; although they did not explicitly stipulate our Condition A, it holds in the problems they study (usually because they involve the allocation of only one good). The inference about the universal validity of the revelation principle in problems with inspection that readers might have drawn from the existing literature is incorrect.

There are several papers in the recent literature that study models similar to ours, particularly those of Border and Sobel (1986), Mookherjee and P'ng (1989), and Melumad and Mookherjee (1989). The model here, however, differs in some important ways from those of these authors, and we insist upon those differences as elements of what politically realistic inspection involves. First of all, the above authors make the penalty vector an instrument of the planner (or the samurai, as the case may be). As a consequence, the planner chooses the harshest penalties that are feasible: for inspection is costly for him, and harsh penalties are a cheap way of inducing truth-telling. But it is clear that in modern societies penalties for mis-reporting one's income to the tax authorities, for example, are not so harsh. (They do not penalize agents their entire incomes.) There are, we believe, at least two reasons for this: (1) actual tax policies are not incentive compatible, and society does not countenance excessively severe treatment of rational liars (the attractive nuisance doctrine), and (2) even if policies were incentive compatible, society would not countenance excessively penalizing irrational liars. Under the U.S. tax code, both penalties and taxes are set by the Congress; the Internal Revenue Service sets only the vector q of inspection probabilities (and keeps it a secret, at that). We consider it an aspect of a realistic model of inspections that the penalties are either set by law, or are chosen by the planner under constraint.[3] Secondly, the above authors allow the planner to treat an honest agent of type i, who is inspected, differently from an agent who reports his income is i but is not inspected (minimal equity is not postulated). Third, the above authors study only the problem where the planner maximizes the sum of utilities in the population, while in the present work the planner need not even maximize a social welfare function.[4]

2. A GENERAL CLASS OF IMPLEMENTATION PROBLEMS WITH INSPECTIONS

There are r types of agent, 1, . . ., r, characterized each by a vector of traits a_i. Let $A=\{a_1, \ldots, a_r\}$ and $I=\{1, \ldots, r\}$. There are N^i agents of type i, and $N=\Sigma_i N^i$. All agents of type i have the *concave* von Neumann-Morgenstern utility function for z, $u^i(z)$, an increasing function in z, where $z \in D_i \subset R^n$ and D_i

[3] Baron and Besanko (1984) agree; in their model, the regulator (planner) chooses penalties from a feasible set that is specified by law, and the regulator does not reward honest firms that are inspected (p. 450).

[4] This difference, unlike the first two, is a minor one.

is a convex set. An agent's type is private information but can be discovered by an inspection that costs the planner $w_0 \in R^n$. Type i agent has an initial endowment $w_i \in D_i$. The planner knows A, N, u, D, Ω, and P where $N=(N^1, \ldots, N^r)$, $D=(D_1, \ldots, D_r)$, $\Omega=(w_1, \ldots, w_r)$, and $P \subset R^n$ is a comprehensive set, which represents the set of feasible aggregate transfers net of inspection costs.

A *mechanism* is defined by a set of messages M, a function $t:M \rightarrow D \subset R^n$, a function $q:M \rightarrow [0,1]$ and a function $h:M\text{x}I \rightarrow D$, where D is a convex set such that $D_i \subset D$, $\forall i \in I$. We write $\mu=\langle M,t,q,h\rangle$. If an agent of type i announces the message $m \in M$ to the planner, he will receive the transfer t(m) if he is not inspected, and h(m,i) if he is inspected. We call t the *pre-inspection transfer function* and h the *post-inspection transfer function*. We assume that $\forall i \in I$, there exists $m \in M$ such that $t(m)+w_i \in D_i$ (that is, i can announce a type that entails a feasible consumption for him). The function q is the *inspection function*. If the agent announces message m, he will be inspected with probability q(m). A mechanism is *feasible* (in expectation) if the sum of transfers plus costs of inspection, when every agent responds optimally, lies in P.

A mechanism is *direct* if M = I. A direct mechanism is *incentive compatible* if announcing his true type is an optimal message for each agent. Denote the optimal message for agent i to transmit, facing μ, as m(i). We assume that the planner restricts himself to the set of direct mechanisms such that each agent has at most one dishonest optimal message, and that if he has two optimal messages, he always reports the truth. For if agents of type i had two optimal dishonest messages, i_1 and i_2, and the planner had no way of knowing what fraction of agents would respond i_1, he could not propose a mechanism that was feasible in expectation. Our restriction is not a severe one. (If the planner knew, say, that 50% of type i agents would respond with announcement i_1, then our theory can be easily extended.)

One example of this class of models is an income distribution problem where an agent's income is private information but can be discovered by a costly inspection. Agents have concave von Neumann-Morgenstern utility functions of income. The planner knows the income distribution; his goal is to maximize a social welfare function which is increasing in the expected utilities of the agents. He announces a pre-inspection transfer function that stipulates the tax an agent has to pay depending on his income. The planner also announces a penalty and inspection policy. An agent who, upon inspection, is found to have misreported his income level will be penalized with higher tax. Since this problem involves the allocation of just one good, it shall follow from Corollary 1 that there exists a welfare-maximizing mechanism that is incentive compatible.

A second example is one where agents are endowed with a vector of inputs, which can either be consumed or used to produce other goods. The utility of an agent depends on his consumption of the input goods and the produced goods. The planner maximizes a social welfare function of agents' expected utilities. He must choose a feasible allocation of goods and an inspection probability function. An agent who is found, upon inspection, to have lied about his initial endowments receives a "penalty allocation" of goods. Because this is a multidimensional allocation problem, it shall be shown that, in general, the welfare-maximizing policy that respects minimal equity will induce some agents to lie about their endowments.

Let a vector of penalty transfers $t^P = (t_1^P, \ldots, t_r^P)$ be specified, $t_i^P \in R^n$. We assume that $w_i + t_i^P \in D_i$. We will be interested in the class of direct mechanisms $H(u,t^P)$ for which the following are true:

(i) $h(i,i) = t(i)$ *(minimal equity)*
(ii) $h(j,i) = t_i^P$, $j \neq i$
(iii) $u^i(w_i + t_i^P) \leqslant u^i(w_i + t(i))$.

The first condition says that the mechanism respects minimal equity; the second says that the penalty transfer for a dissembling agent of type i is t_i^P; the third says that t_i^P is, indeed, a penalty.

Condition A. $t(m(i))) \geqslant t_i^P, \quad \forall i \in I$.

Condition A states that, when an agent responds optimally to the mechanism, the transfer assigned to him by the mechanism, should he not be inspected, weakly dominates, component by component,[5] the transfer that is assigned to him after inspection. Unfortunately, to know whether or not the condition is satisfied, it is necessary to compute the optimal behavior of each agent.

There are alternative conditions to "A" that can be checked directly and are sufficient to prove our results. Thus, instead of Condition A we could impose a restriction on the utility functions to rule out the possibility of preference reversals, i.e., to rule out situations where some agents prefer alternative x to y, and at the same time other agents prefer alternative y to x. We think, however, that this alternative type of condition implies that agents have essentially similar preferences, a less appealing requirement than Condition A. It should be also mentioned that even in the case that $t_i^P = t_j^P$, i.e. the same allocation penalty for all agents, Condition A cannot be dropped from the Theorem. Thus, the difficulties are not due to the fact that the penalty allocations may be different across types.

Theorem 1 *(The Revelation Principle) Given a feasible direct mechanism $\mu = \langle l, t, q, h \rangle$ in $H(u,t^P)$ that satisfies Condition A, there exists a feasible incentive compatible direct revelation mechanism, $\mu^* = \langle l, t^*, q^*, h^* \rangle$ in $H(u,t^P)$ that weakly Pareto dominates μ.*

Remark. In the literature, the revelation principle is taken to mean that any mechanism (where M is arbitrary) can be replaced by a direct mechanism which

[5] Vector orderings: $x \geqslant y$ means $x_i \geqslant y_i$ for all i.

is incentive compatible. Furthermore, the standard proof shows that the allocation achieved by the new mechanism is identical to the one achieved by the original mechanism. *Theorem 1*, however, does not begin with a general mechanism; nor will it be the case that the allocation achieved by the incentive compatible mechanism is identical to the allocation under the original mechanism. Moreover, Theorem 1 requires concavity of the utility functions, which is not necessary for the classical revelation principle. For these reasons, the theorem is a first cousin, but not a sibling, of the classical result.

Proof: Denote type i's expected utility if he reports j under the mechanism μ by $f(j,i)$. His maximum expected utility is:

$$f(m(i),i) = (1-q(m(i)))u^i(t(m(i))+w_i) + q(m(i))u^i(h(m(i),i)+w_i)$$

Define $t^*:I \rightarrow D$ by

$$t^*(i) = (1-q(m(i)))(t(m(i))) + q(m(i))h(m(i),i) \tag{1}$$

Note that $t^*(i)+w_i \in D_i$, by the convexity[6] of D_i.
It follows from Condition A that for all j

$$t^*(j) \leqslant t(m(j)). \tag{2}$$

Define $q^*:I \rightarrow [0,1]$ by $q^*(i) = q(m(i))$, and $h^*:IxI \rightarrow D$ by:

$$h^*(j,i) = \begin{cases} t^*(i) & \text{if } j=i, \\ t_i^P & \text{if } j \neq i. \end{cases}$$

First we show that for the mechanism μ^* an optimal message for the agent is to announce his true type. If an agent of type i announces i his utility under μ^* is $f^*(i,i) = u^i(t^*(i)+w_i)$.

By concavity of u and the definition of $t^*(i)$,

$$f^*(i,i) \geqslant f(m(i),i). \tag{3}$$

If i announces type $j \neq i$ his utility is

$$\begin{aligned} f^*(j,i) &= (1-q^*(j))u^i(t^*(j)+w_i) + q^*(j)u^i(t_i^P+w_i) \\ &\leqslant (1-q(m(j))u^i(t(m(j))+w_i) + q(m(j))u^i(t_i^P+w_i) \\ &= f(m(j),i) \leqslant f(m(i),i) \quad \text{if } m(j) \neq i \\ &\leqslant f(m(j),i) \leqslant f(m(i),i) \quad \text{if } m(j) = i \end{aligned} \tag{4}$$

[6] We assume that an agent of type i only announces type m facing μ if $t(m)+w_i \in D_i$. Equivalently, we can let $u(z;a_i) = -\infty$ if z is not feasible for i.

using (2), and the definition of m(i). From (3) and (4), it follows that all agents tell the truth when facing μ^*. (We may have $z = t^*(j) + w_i \in D_i$ in which case we let $u^i(z) = -\infty$.) The total inspection costs are the same under the two mechanisms:

$$\sum_i w_0 N^i q^*(i) = \sum_i w_0 N^i q(m(i)).$$

The total transfers are the same:

$$\sum_i t^*(i) N^i = \sum_i ((1 - q(m(i)))t(m(i)) + q(m(i))h(m(i),i))N^i.$$

Therefore μ^* is feasible, and by (3), all agents are at least as well off under μ^*. Q.E.D.

The key step in the proof - the one using (2) and hence Condition A - is the circled inequality. Figure 1 illustrates how that inequality may fail without Condition A.

Corollary 1: *If $n = 1$, then for any direct mechanism μ in $H(u,t^P)$, there is an incentive compatible mechanism in $H(u,t^P)$ that Pareto dominates μ.*

Proof: We need only note that Condition A always holds for $n = 1$: for if, to the contrary, $t(m(i)) < t_i^P$, then part (iii) of the definition of $H(u,t^P)$ is violated.

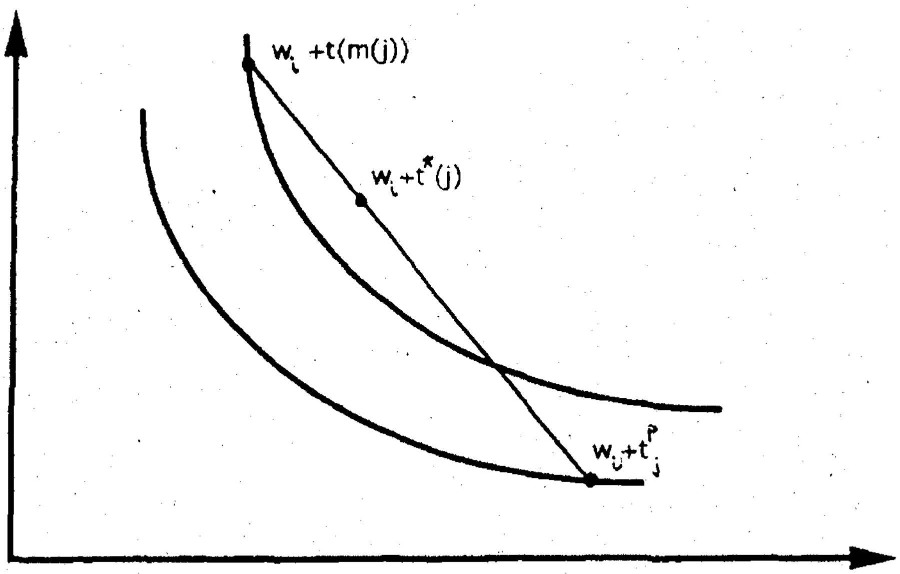

Figure 1. Possible indifference curves for u^i when $t^*(j) \nleq t(m(j))$.

We observe, this time in general, that if we do not restrict ourselves to the class of mechanisms satisfying minimal equity, then we *can* use the standard trick to prove the revelation principle, and the convexity assumptions on the utility function and the sets D_i are unnecessary. The standard trick is to set $t^*(i) = t(m(i))$, $q^*(i) = q(m(i))$, and $h^*(j,i) = h(m(j),i)$.

One weakness of the model of *Theorem 1* is that penalties must be only a function of the agent's type, rather than, more generally, a function of his type and his announced type. (We might want to penalize an agent as a function of how much he lied.) It is possible to preserve the revelation principle for a model with minimal equity and more general penalties of this type, but one cannot simultaneously take the penalty schedule as exogenous. We now state a result for any (non-direct) mechanism.

Condition A′: $t(m(i)) \geqslant h(m(i),i)$.

Theorem 2: *Given a feasible mechanism $\mu = \langle M,t,q,h \rangle$ that satisfies Condition A′, and such that for all i,j, $h(m(j),i) \geqslant h_i$, for some given set of vectors $\{h_i\}$. Then there exists a feasible incentive compatible direct mechanism $\mu^* = \langle I,t^*,q^*,h^* \rangle$ that weakly Pareto dominates μ, and in which $h^*(j,i) \geqslant h_i$ for all i,j and $h^*(i,i) = t^*(i)$, for all i.*

Proof: Appendix.

Note that in *Theorem 2* we begin with an abstract message space, in which there is no meaning to "lying" about one's type. Condition A′ says, nevertheless, that the pre-inspection transfer dominates the post-inspection transfer for the optimal announcements. Suppose we restrict ourselves to mechanisms that are limited in their severity, in the sense that the penalty transfers are bounded below by the vectors h_i. The theorem assures us that, without loss of generality, we can restrict ourselves to direct, truth-revealing mechanisms which preserve horizontal equity, and which respect the same bound on severity.[7] The planner, however, must have the freedom to assign new penalties $h^*(j,i)$ which, in general, will differ from $h(j,i)$.

Next, one might wish to relax the minimal equity requirement in a nice way, and allow the planner to *reward* agents who are inspected and have announced their true type. To this end, let $t^R = (t_1^R, \ldots, t_r^R)$, $t_i^R \in R_+^n$ be a vector of rewards given to the planner, along with t^P. The planner may reward a truth-telling agent of type i with reward t_i^R. A direct mechanism of this type is defined by the set I, the pre-inspection transfer function t, the inspection function q, and the post-inspection function h where t and q are defined as before and

[7] Some condition like the bound condition stipulated is required to make the theorem interesting; for if the planner could choose arbitrarily severe penalties, it would be easy to guarantee the theorem's conclusion.

$$h(j,i) = \begin{cases} t(i) + \partial(i) & \text{if } j = i, \\ t_i^P & \text{if } j \neq i, \end{cases}$$

where $\partial(i) \in \{0, t_i^R\}$. In general $h(i,i) \neq t(i)$, so we do not have minimal equity. Let $H(u,t^P,t^R)$ be the class of direct mechanisms so defined.

Theorem 3 *Given a feasible direct mechanism $\mu \in H(u,t^P,t^R)$ that satisfies Condition A, there exists a feasible incentive compatible direct mechanism $\mu^* \in H(u,t^P,t^R)$ that weakly Pareto dominates μ.*

Proof: Appendix.

Our proof of the result still requires Condition A. This suggests that we cannot escape the requirement of Condition A, for guaranteeing incentive compatibility, by working in the larger class of mechanisms in which rewards to honest inspected agents are allowed. Indeed, if we try to apply the standard trick for proving the revelation principle to the model in Theorem 1, we will see that the trick requires giving the planner the authority to *penalize* honest agents who are inspected. It is therefore not surprising that enlarging the space of mechanisms to allow the planner to *reward* honest agents does not seem to help.

Let us recall the interpretation of these results. We have postulated a society in which there are social norms that require any politically acceptable mechanism to be minimally equitable, and that duplicity not be socially optimal. Therefore, the government will find it necessary to restrict its policies to ones which entail truthful response from self-interested citizens. In general, this restriction requires a sacrifice in social welfare. Condition A describes a class of mechanisms in which the optimal mechanism engenders truthful response. A legislature can, in many applications, design penalties such that Condition A is likely to hold for reasonable distributions of population traits. In the next section, we provide an example where Condition A fails to hold and any truth-telling direct mechanism is Pareto dominated by a direct mechanism where some agents lie. We consider mechanisms with minimal equity, although a similar example may be constructed to show that Condition A cannot be dropped from *Theorem 3* either.

3. A PROBLEM WHERE CONDITION A DOES NOT HOLD

There are three types of agent, designated by $I = \{1,2,3\}$. There are N^i agents of type i. There are two goods, an output y and an input x. The endowments of the agents consist of the input good only, and are designated w_1, w_2, and w_3. The utility functions are:

$$u^1(y_1,x_1) = \beta y_1 + \partial x_1 u^2(y_2,x_2) = x_2 u^3(y_3,x_3) = y_3 .$$

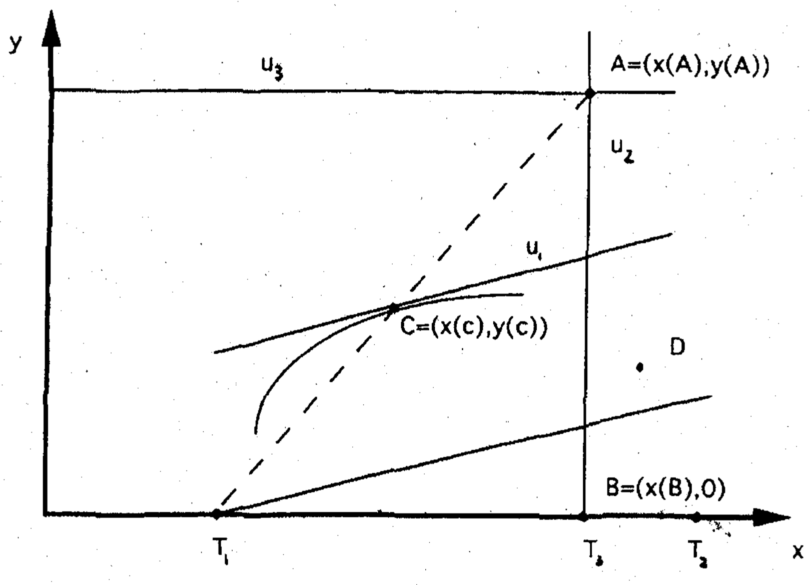

Figure 2. A mechanism violating Condition A.

We furthermore assume that type 1 agents are risk neutral (we take these special utility functions just for simplicity and a similar example may be given for the case where all agents care about the two goods). The penalty vector t^P is given by $t_i^P = (0, -x_i^P)$, where $x_i^P \leqslant w_i$, for $i = 1,3$; that is, the consumption of an agent of type i who is penalized is $(0, w_i - x_i^P)$. Suppose the planner's objective is to implement a feasible allocation that maximizes the expected utility of type 1 agents, subject to giving type 2 agents a utility level of $\mathbf{u}_2$, and type 3 agents a utility of $\mathbf{u}_3$. Examine Figure 2, in which the abscissa measures the amount of the input that an agent contributes to production (i.e., $w_i - x_i$) and the ordinate measures the amount of y that an agent consumes (or is produced). An indifference curve of a type 1 agent is the straight line through C labelled u_1; indifference curves for types 2 and 3 agents are the lines labelled u_2 and u_3, respectively, through point A. At B, type 2 agents receive exactly utility $\mathbf{u}_2$ (note it is inefficient for them to consume any output). At A, type 3 agents receive utility exactly $\mathbf{u}_3$ (they contribute their entire endowment to production, as they derive no utility from its consumption). Given that type 3 agents and type 2 agents are each contributing w_3 to production, and that type 3 agents are each consuming y(A) units of output, and that total inspection costs consume amount $q_3(N^1 + N^3)w_0$ of the x good (see below), the concave curve through C represents the output available for an agent of type 1 as a function of the amount of input he contributes to production. The production function is chosen so that this curve is tangent to the indifference curve of type 1 through C. C is clearly the allocation that maximizes type 1's consumption subject to the requirements that types 2 and 3 receive utility levels $\mathbf{u}_2$ and $\mathbf{u}_3$. C is chosen to lie on the line segment through T_1 and A: more precisely,

$$(x(C),y(C)) = q_3(w_1 - x_1^P,0) + (1-q_3)(x(A),y(A)), \text{ for some } q_3 < 1.$$

The penalty allocations for the three types are indicated by the points T_i.

Consider the mechanism that assigns the allocation A to type 3 agents, the allocation B to type 2 agents, and the allocation D to type 1 agents. Note that D is less attractive to types 1, 2 and 3 than C, B and A, respectively. Let $q(3) = q_3$. With type 2 preferences, B is preferred to A and D; with type 3 preferences, A is preferred to B and D. So under the mechanism defined, types 2 and 3 report truthfully. No one will dishonestly report that he is type 1 or 2, so set $q(1) = q(2) = 0$. Type 1, however, announces 3, and has an expected consumption of C, which dominates D. By risk neutrality of type 1, his expected utility of the lottery $(q_3,T_1;(1-q_3),A)$ is equal to his utility at the bundle C. By construction, it is feasible for type 1 agents to consume C in expectation given that types 2 and 3 consume at B and A. Call this mechanism μ. Note that the inspection costs are $q_3(N^1+N^3)w_0$, where w_0 is the cost of an inspection.

We now observe that there is no direct incentive compatible mechanism that weakly Pareto dominates the mechanism μ. The only feasible way that type 1 agents can achieve at least the expected utility they get at the lottery $(q_3,T_1;1-q_3,A)$, as an outcome of *truthful* reporting, is for them to be assigned the bundle C. But C is strictly preferred by type 2 agents to B. To dissuade type 2 agents from reporting type 1 if $t^*(1) = C$, and $t^*(2) = B$, $q^*(1)$ must be set so that the convex combination $q^*(1)T_2 + (1-q^*(1))C$ lies on or to right of the indifference curve u_2. This determines some minimum value $q^*(1)$. The inspection costs of the new mechanism will be $q^*(1)N^1w_0$. These costs will exceed the costs of μ as long as $(q^*(1)-q_3)N^1 > q_3N^3$, which can be assured as long as T_2 is chosen close to B, so that $q^*(1) > q_3$, and N^1 is chosen large compared to N^3. Then it will be the case that the allocation C is not feasible for type 1 agents, as the increased inspection costs will have lowered the production function so that C lies above it. Finally we have to check that an allocation $t^*(1) = C$ and $t^*(2) = B'$, such that $B' < B$, is not feasible. Let u_2' be type 2's indifference curve when he receives B'. In this case the input that a type 2 agent contributes, $x(B')$, is less than before. At the same time the probability of inspection $q^{*\prime}(1)$ that makes $q^{*\prime}(1)T_2 + (1-q^{*\prime}(1))C$ lie to the right of the indifference curve u_2' is less than $q^*(1)$ above. Therefore, inspection costs are less than before and C might be feasible. We always can choose the production function and agent 2's utility function such that this possibility cannot happen. Hence, for this specification of the parameters, there is no incentive compatible direct mechanism that achieves the utilities for the three types that they receive under μ.

To summarize the idea of this counterexample: a social-welfare-optimizing planner must allow type 1 agents to lie; for if he assigned them C, that would create the incentive for type 2 to lie and announce C. Therefore, the planner would have to inspect agents who announce C to dissuade type 2 from lying. Since there are type 1 agents who will truthfully announce C, and they are a large population, the inspection costs will be large.

4. CONCLUDING REMARKS

We have ignored the problem of the planner's commitment – why should he in fact inspect after announcements have been made under an incentive compatible mechanism? We have no answer for this within the model. We might mention, however, an inspection mechanism that approximately implements allocation rules at no inspection cost, and in which commitment is not a problem. Suppose the problem is incentive compatible. The planner first calculates and announces the optimal policy with inspections, and then announces that if he receives reports from the population that aggregate to the true distribution (i.e., for all i he receives N^i reports of type i), then he will implement the allocation with no inspections, and in addition each agent will receive a lump-sum equal his share of inspection costs saved. On the other hand, should he receive announcements aggregating to the wrong distribution, then he will carry out the inspections as planned. Truth-telling is a Nash equilibrium strategy under this mechanism; indeed, for all practical purposes, it is a dominant strategy.[8] But since one of our motivations in studying mechanisms with inspection is the assumption that some agents will act irrationally, we do not place much importance on this mechanism.

We have studied the possibility of implementing allocations that maximize a social welfare function when the planner knows the distribution of the unknown trait in the population, and the agents know only their own trait.[9] As two requirements of political realism, we have demanded that the mechanisms respect minimal equity, and that the penalties be specified exogenously to the planner. The first observation is that, because the revelation principle in general fails for this class of mechanisms, we may sometimes observe a government asking its citizens to report a trait, penalizing them for lying, and yet deliberately designing a policy the self-interested response to which requires some types to lie. For a general class of problems, we have proved that, if the planner restricts himself to a class of mechanisms satisfying Condition A, then he may, without welfare loss, restrict himself to mechanisms that constrain the agents to report their types truthfully. The practical import is that legislators may wish to choose penalties for which it will be the case that the optimal (welfare-maximizing) mechanism satisfies Condition A, and so the planner will never need to decide whether or not to trade-off honesty for an increase in welfare. When a trade-off exists, planners may lean towards increasing welfare, while legislators may lean towards

[8] We thank Giacomo Bonanno for pointing out this mechanism. We say that truthful revelation is almost a dominant strategy, since it will not be dominant for an agent in the very unlikely case that, should he lie in a certain way, the planner will then receive announcements aggregating to the correct distribution.

[9] More precisely, agents need only use information about their own trait.

presenting the image of good government to citizens;[10] by the judicious choice of penalties, legislatures can impose their preferences on planners. If the problem involves the allocation of only one good, then Condition A is always satisfied, and there is no conflict.

Finally, we should mention the recent work of Green and Laffont (1986), in which the authors study implementation in a situation where, instead of being able to announce any type in l, an agent of type i can only credibly announce types in a subset $l_i \subset l$. The idea is that a professor cannot credibly announce that his income is that of a janitor. (Note, however, that this must be because the planner can easily [costlessly] check that the agent in question is a professor.) Green and Laffont show that, in a setting without inspections, the revelation principle holds if and only if a certain rather strong condition, the "nested range condition", holds for the sets $l_1, \ldots, l_r$. We have assumed that the planner knows quite a bit about the population, and so this might restrict the credible reports of agents to such proper subsets of l. Suppose, to follow the above suggestion, that the type is an agent's income level, and the planner does have information on each agent's profession, if not his income. But then, we claim, types can be redefined so that professors with income i constitute one type and janitors with income j constitute another. With these new types, we have a larger set of types, call it l^*, which is now *partitioned* into subsets l_k^*, $k = 1,s$, having the property that any agent of a given type can credibly report that he is any type in only one of the subsets l_k^*. (A professor can report that he is a professor with any possible income that a professor might have; he cannot report that he is a janitor with *any* income.) For the pairwise disjoint sets l_k^* the nested range condition holds. Although we have not investigated the role of the nested range condition for implementation with inspection, this observation at least suggests that there is no reason a priori to worry that restrictions on the reports of agents, due to other facts that the planner knows about them, will further jeopardize incentive compatibility.

IGNACIO ORTUÑO-ORTIN
*Departamento de Fundamentos
del Analisis Economico,
Universidad de Alicante, Campus San Vicente,
03071-Alicánte,
Spain*

JOHN E. ROEMER
*University of California,
Davis,
U.S.A.*

REFERENCES

Baron, D. and D. Besanko, 1984, Regulation, asymmetric information, and auditing. *Rand Journal of Economics* 15, 447–470.

[10] A good government is one that does not design policies involving penalties for lying and rational duplicity.

Border, K. and J. Sobel, 1987, Samurai accounting: a theory of auditing and plunder. *Review of Economic Studies* 54, 525–540.

Green, J. and J. J. Laffont, 1986, Partially verifiable information and mechanism design. *Review of Economic Studies* 53, 447–456.

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

Mookherjee, D. and I. P'ng, 1989, Optimal auditing, insurance, and redistribution. *Quarterly Journal of Economics* 104, 399–416.

—— and ——, 1990, Enforcement costs and the optimal progressivity of income taxes. *Journal of Law, Economics, and Organization* 6, 411–431.

Stiglitz, J., 1982, Utilitarianism and horizontal equity. *Journal of Public Economics* 17, 126–140.

Weiss, L., 1978, The desirability of cheating incentives in income tax schemes. *Journal of Political Economy* 86, 210–225.

APPENDIX

Theorem 2: *Given a feasible mechanism $\mu = \langle M,t,q,h\rangle$ that satisfies Condition A', and such that for all i,j, $h(m(j),i) \geqslant h_i$, for some given set of vectors $\{h_i\}$. Then there exists a feasible incentive compatible direct mechanism $\mu^* = \langle I,t^*,q^*,h^*\rangle$ that weakly Pareto dominates μ, and in which $h^*(j,i) \geqslant h_i$ for all i,j and $h^*(i,i) = t^*(i)$, for all i.*

Proof: Denote type i's maximum expected utility under the mechanism μ by:

$$f(m(i),i) = (1 - q(m(i)))u^i(t(m(i)) + w_i) + q(m(i))u^i(h(m(i),i) + w_i)$$

Define $t^*: I \rightarrow D$ by

$$t^*(i) = (1 - q(m(i)))(t(m(i))) + q(m(i))h(m(i),i). \tag{5}$$

Note that $t^*(i) + w_i \in D_i$, by the convexity of D_i.

It follows from Condition A' that

$$t^*(i) \leqslant t(m(i)). \tag{6}$$

Define $q^*: I \rightarrow [0,1]$ by $q^*(i) = q(m(i))$, and $h^*: I\mathrm{x}I \rightarrow D$ by:

$$h^*(j,i) = \begin{cases} t^*(i) & \text{if } j = i, \\ h(m(j),i) & \text{if } j \neq i. \end{cases}$$

First we show that for the mechanism μ^* an optimal message for the agent is to announce his true type. If an agent of type i announces i his utility under μ^* is

$$\begin{aligned} f^*(i,i) &\equiv (1 - q^*(i))u^i(t^*(i) + w_i) + q^*(i)u^i(h^*(i,i) + w_i) \\ &= u^i((t^*(i) + w_i), \end{aligned}$$

since $t^*(i) = h^*(i,i)$.

By concavity of u and the definition of t*(i),

$$f^*(i,i) \geqslant f(m(i),i). \tag{7}$$

If i announces type j his utility is

$$\begin{aligned} f^*(j,i) &= (1-q^*(j))u^i(t^*(j)+w_i)+q^*(j)u^i(h^*(j,i)+w_i) \\ &= (1-q(m(j)))u^i(t^*(j)+w_i)+q(m(j))u^i(h(m(j),i)+w_i) \\ &\leqslant f(m(j),i) \leqslant f(m(i),i), \end{aligned} \tag{8}$$

where the first inequality follows from (6), the second by definition of expected utility, and the third follows by the fact that m(i) is i's optimal message at μ. From (7) and (8), it follows that all agents tell the truth when facing μ^*. (We may have $z = t^*(j)+w_i \in D_i$ in which case we let $u^i(z) = -\infty$.)

The total inspection costs remain the same under the two mechanisms:

$$\textstyle\sum_i w_0 N^i q^*(i) = \sum_i w_0 N^i q(m(i)).$$

The total transfers are the same:

$$\textstyle\sum_i t^*(i)N^i = \sum_i ((1-q(m(i)))t(m(i)) + q(m(i))h(m(i),i))N^i.$$

Therefore μ^* is feasible, and by (7), the agents are indifferent or better off under μ^*. Q.E.D.

Proof of Theorem 3:
Let $\mu = \langle I,t,q,h \rangle$. Define $q^*: I \rightarrow [0,1]$ by $q^*(i) = q(m(i))$ and $t^*: I \rightarrow D$ by

$$t^*(i) = \begin{cases} (1-q(m(i))t(m(i)) + q(m(i))t_i^P & \text{if } m(i) \neq i \\ t(i) & \text{if } m(i) = i \end{cases}$$

and $h^*: I x I \rightarrow D$ by

$$h^*(j,i) = \begin{cases} h(i,i) & \forall j,i, \quad j=i,\ m(i)=i, \\ t^*(i) & \forall j,i, \quad j=i,\ m(i) \neq i, \\ t_i^P & \text{otherwise.} \end{cases}$$

The rest of the proof parallels the proof of *Theorem 1.*

Observe that under μ^* we reward only the types that announce their true type and were rewarded under μ, and the reward vectors remain the same. Q.E.D.