Commitment and Robustness in Mechanisms with Evidence[1]

Elchanan Ben-Porath [2] Eddie Dekel [3] Barton L. Lipman[4]

First Draft
June 2016

[1]We thank the National Science Foundation, grant SES–0820333 (Dekel), and the US–Israel Binational Science Foundation for support for this research.

[2]Department of Economics and Center for Rationality, Hebrew University. Email: benporat@math.huji.ac.il.

[3]Economics Department, Northwestern University, and School of Economics, Tel Aviv University. Email: dekel@northwestern.edu.

[4]Department of Economics, Boston University. Email: blipman@bu.edu.

Abstract

We show that in a class of I–agent mechanism design problems with evidence, commitment has no value for the principal, randomization has no value for the principal, and robust incentive compatibility has no cost. In particular, there is an equilibrium with a relatively simple structure in which the principal obtains the same outcome without commitment as the best he can achieve with commitment.

1 Introduction

We show that in a class of I–agent mechanism design problems with evidence, randomization has no value for the principal and robust incentive compatibility has no cost. Also, commitment has no value for the principal in the sense that there is an equilibrium of the game when the principal is not committed to the mechanism with the same outcome as in the optimal mechanism with commitment. We also show that this equilibrium has a relatively simple structure.

To understand the class of mechanism design problems our result applies to, consider the following examples.

Example 1. The simple allocation problem. The principal has a single unit of an indivisible good which he can allocate to one of a set of I agents. Each agent has a type which affects the value to the principal of allocating the good to that agent. Each agent prefers getting the good to not getting it, regardless of her type. Types are independent across agents and monetary transfers are not possible. Each agent may have concrete evidence which proves to the principal some facts about her type. For example, the principal may be a state government which needs to choose a city in which to locate a public hospital. Each city wants the hospital. The state wants to place the hospital where it will be most efficiently utilized, but each city has private information on local needs. The state could ask the city to bear the cost of the hospital, but that would imply diverting the city's funds from other projects that the goverement considers important.

Example 2. The public goods problem. The principal has to choose whether or not to provide a public good which affects the utility of I agents. If the principal provides the good, the cost must be evenly divided among the agents. Each agent has a type which determines her willingness to pay for the good. If the willingness to pay exceeds her share of the cost, she wants the good provided and otherwise prefers that it not be provided. Types are independent across agents and monetary transfers other than the cost sharing are not possible. Each agent may have evidence which enables her to prove some facts to the principal about the value of the public good to her. The principal wishes to maximize the sum of the agents' utilities. For example, the principal may be a government agency deciding whether or not to build a hospital in a particular city and the agents may be residents of that city who will be taxed to pay for the hospital if it is built. Then an agent might show documentation of a health condition or past emergency room visits to prove to the principal that she has a high value for a nearby hospital.

Example 3. The allocation problem with externalities. This problem is the same as Example 1, but now agent i cares about which other agent receives the good if she does not. For example, suppose agent i's most preferred outcome is to receive the good, her second–best option is that agent $i+1$ receives it (mod I), and all other allocations

are tied for worst. In Example 3a, the payoff to the principal to allocating the good to agent i is still just a function of i's type. Alternatively, in Example 3b, we assume that the principal's payoff is a weighted sum of the payoffs to the agents where the weight on agent i's utility (which could be negative) depends on i's type.

Example 4. The public goods problem with interdependence. This problem is the same as Example 2 except now the payoff to agent i from the provision of the public good depends on both her type and the types of the other agents.

We will show that optimal mechanisms for Examples 1, 2, and 3b share several significant features. First, there is no value to commitment. In other words, if the principal is not committed to the mechanism, there is still an equilibrium of the game with the same outcome as in the optimal mechanism. Second, the optimal mechanism is deterministic — the principal does not need to randomize. Third, the optimal mechanism is not just incentive compatible but is also what we will call *robustly incentive compatible*. We define this precisely later, but for now simply note that it is a strengthening of dominant strategy incentive compatibility. Thus the robustness of dominant strategy incentive compatibility comes at no cost to the principal. This strong robustness of the mechanism in turn implies strong robustness properties of the equilibrium which achieves the same outcome. By contrast, the optimal mechanism for Examples 3a and 4 satisfies *none* of these properties, in general.

One useful implication of this result is that we can compute optimal mechanisms by considering equilibria. In particular, we give a relatively simple characterization of the optimal equilibrium for the principal which does not rely on much information regarding the principal's preferences or the structure of the set of actions. This makes determining the optimal mechanism straightforward in some cases. To illustrate, we consider optimal mechanisms when the evidence technology is the one originally proposed by Dye (1985). In Dye's model, each agent has some probability of having evidence that would enable her to exactly prove her true type and otherwise has no evidence at all.

When we apply this approach to the simple allocation problem described in Example 1 above, we find that the optimal mechanism has a favored–agent structure reminiscent of Ben-Porath, Dekel, and Lipman (2014), henceforth BDL. BDL considered the simple allocation problem described above, but with one difference. BDL did not consider evidence but instead considered what in some respects is the opposite assumption, where the principal can verify claims by the agents at a cost in a manner that the agents cannot control. BDL showed that the optimal mechanism was a favored–agent mechanism. More precisely, a favored–agent mechanism specifies an agent, say i^*, and a threshold net value, say v^*. Call the *net value* of a type of an agent the value to the principal of giving her the good when she has that type minus the cost of verifying this claim. In a favored–agent mechanism, if every agent other than the favored agent claims a net value below v^*, then

the principal gives the good to the favored agent i^* and does not verify any reports. Otherwise, he verifies the report of the agent who makes the highest net value claim and, if the report is truthful (as it will be in equilibrium), he gives the good to that agent.

Turning to the model of the current paper, we define a favored–agent mechanism to specify a favored agent i^* and a threshold value v^* such that if every agent other than the favored agent either provides no evidence or proves a value below v^*, then the favored agent i^* receives the good. Otherwise, the good goes to the agent who proves the highest value. The fact that the principal does not need commitment makes this result quite straightforward to prove, by contrast to the complex proof in BDL.[1] We also show that the optimal mechanism for the public good problem is very similar to the optimal mechanism for the same problem with costly verification as derived by Erlanson and Kleiner (2015).

The paper is organized as follows. Section 2 states the model formally. In Section 3, we show the main results sketched above, including the characterization of the best equilibrium for the principal. The proof of this theorem is sketched in Section 5. In Section 4, we specialize to the Dye (1985) evidence structure and provide a characterization of optimal mechanisms in this setting. We then use this characterization to give optimal mechanisms for a variety of more specific settings including the simple allocation problem and the public goods problem. We offer concluding remarks in Section **??**. Proofs not contained in the text are in the Appendix.

Related literature. There are several literatures related to this paper. First, there is a literature on mechanism design with evidence — see, for example, Green and Laffont (1986), Bull and Watson (2007), Deneckere and Severinov (2008), Ben-Porath and Lipman (2012), Kartik and Tercieux (2012), and Sher and Vohra (2015).

A particularly relevant subset of this literature is a set of papers on one–agent mechanism design problems which show that, under certain conditions, the principal does not need commitment to obtain the same outcome as under the optimal mechanism. This was first shown by Glazer and Rubinstein (2004, 2006) for the case where the principal has two actions and was extended by Sher (2011) to concave utility functions and by Hart, Kremer, and Perry (2015) to single–peaked utility. Our results extend these in two ways. First and most importantly, we consider multi–agent problems rather than single–agent settings. Of course, our result on robust incentive compatibility cannot have any analog in the one–agent setting. Second, unlike these previous results, we allow the preferences of an agent to vary with her type, albeit in a specific fashion. In particular, in the two–action case considered by Glazer and Rubinstein, we can allow arbitrary dependence of an agent's preference on her type. We discuss the connections to these papers in more detail in Section 4.

[1] Though see the simpler proof of that result in Lipman (2015) or Erlanson and Kleiner (2016.

Second, our result showing that commitment is not valuable can be thought of as a characterization of equilibria in games with evidence. Hence our work is also related to the literature on communication games with evidence. See, for example, Dye (1985), Jung and Kwon (1988), Shin (1994, 2003), Lipman and Seppi (1995), and Guttman, Kremer, and Skrzypacz (2014).

As noted above, our work is also related to our previous work on mechanism design with costly verification (Ben Porath, Dekel, and Lipman, 2014) and to Erlanson and Kleiner (2015). We discuss this connection in Section 3.

2 Model

The set of agents is $\mathcal{I} = \{1, \ldots, I\}$ where $I \geq 1$. The principal has a finite set of actions A and can randomize over these.

Each agent i has private information in the form of a type t_i where types are distributed independently across agents. The set of types of i is denoted T_i and the (full support) prior is denoted ρ_i. T_i is finite for all i.

Given action a by the principal and type profile t, agent i's utility is $u_i(a, t_i)$, independent of t_{-i}. We add significantly more structure on the agents' utility functions below.

The principal's utility is

$$v(a, t) = u_0(a) + \sum_i u_i(a, t_i)v_i(t_i).$$

In what follows, we often write this simply as $\sum_i u_i(a)v_i(t_i)$ with the convention that the sum runs from $i = 0$ to I and $v_0(t_0) \equiv 1$.

There are two ways to interpret the principal's utility function. The most obvious is a social welfare interpretation where the principal maximizes a weighted sum of the agent's utilities and t_i determines how much he "cares" about agent i's utility. On the other hand, this utility function does not require the principal to care about the agents at all. A different interpretation is to think of $v_i(t_i)$ as measuring the extent to which the principal's interests are aligned with those of agent i. That is, a high value of $v_i(t_i)$ doesn't mean that the principal likes agent i but means that the principal likes what agent i likes.

For some settings, both interpretations seem natural. For example, consider the simple allocation problem, Example 1 in Section 1. Here the principal has a single unit

of an indivisible good that he can either keep or allocate to one of the agents. Thus we can write the set of actions as $\{0, 1, \ldots, I\}$ where 0 is interpreted as the principal keeping the good and $i \neq 0$ is interpreted as allocating it to agent i. Assuming every type of every agent prefers having the good to not having it, we take the utility function for agent i to be

$$u_i(a_i, t_i) = \begin{cases} 1, & \text{if } a = i; \\ 0, & \text{otherwise.} \end{cases}$$

Letting $v_i(t_i)$ denote the value to the principal of allocating the good to agent i when she is type t_i and letting $u_0(0)$ denote the value to the principal of keeping it, we obtain the utility function for the principal of $u_0(a) + \sum_i u_i(a, t_i)v_i(t_i)$. Thus this formulation is consistent with the second interpretation, but, of course, it is also consistent with an interpretation that $v_i(t_i)$ measures how much the principal "cares about" type t_i's utility.

For other problems, the social welfare interpretation is more natural. For example, suppose the principal has a set of objects to allocate. If each agent wants at most one object and is indifferent across objects, then, just as above, we can interpret $v_i(t_i)$ as the value to the principal of giving an object to t_i. On the other hand, if agents may want multiple objects and have nontrivial preferences regarding bundles of such objects, it seems most natural to interpret the principal's utility function as a social welfare function.

Another important issue for interpretation is that we cannot entirely separate assumptions about the principal's utility function and the agents' utility functions. For example, suppose $v_i(t_i) > 0$ for all t_i and all i. Then consider changing agent i's utility function from $u_i(a, t_i)$ to $\hat{u}_i(a, t_i) = u_i(a, t_i)v_i(t_i)$ and changing the principal's utility function to $\sum_i \hat{u}_i(a, t_i)$. Becaue $\hat{u}_i(a, t_i)$ is a positive affine transformation of $u_i(a, t_i)$, we haven't changed best responses for the agents. Clearly, the principal's preferences have not changed since this is simply a different way of writing the same function. Hence we cannot separate the extent to which $v_i(t_i)$ is part of the principal's utility function or a "scaling factor" for agent i's utility function.

Finally, note that $v_i(t_i)$ is allowed to be zero or negative. Thus the principal's interests can be in conflict with those of some or all agents in a way which can depend on the agents' types.

Each agent may have evidence which would prove some claims about her type. To model evidence, we assume that for every i, there is a function $\mathcal{E}_i : T_i \to 2^{2^{T_i}}$. In other words, $\mathcal{E}_i(t_i)$ is a collection of subsets of T_i, interpreted as the set of events that t_i can prove. The idea is that if $E_i \in \mathcal{E}_i(t_i)$, then type t_i has some set of documents or other tangible evidence which she can present to the principal which demonstrates conclusively that her type is in the set E_i. We require the following properties. First, proof is true. Formally, $E_i \in \mathcal{E}_i(t_i)$ implies $t_i \in E_i$. Second, proof is consistent in the sense that

$s_i \in E_i \in \mathcal{E}(t_i)$ implies $E_i \in \mathcal{E}_i(s_i)$. In other words, if there is a piece of evidence that some type can present which does not rule out s_i, then it must be true that s_i could present that evidence. Clearly, if s_i could not present it, the evidence actually refutes the possibility of s_i. Putting these two properties together, we have $t_i \in E_i$ if and only if $E_i \in \mathcal{E}_i(t_i)$.

The last property we assume is not necessary for the model to be internally consistent but is an additional restriction used in much of the literature. This property is called *normality* by Bull and Watson (2007) and the *full reports condition* by Lipman and Seppi (1995). The condition says that there is one event that t_i can present which summarizes all the evidence she has available. Intuitively, this condition means that there are no time or other restrictions on the evidence an agent can present, so that she can present everything she has. Formally, the statement is that for every t_i, we have

$$\bigcap_{E_i \in \mathcal{E}_i(t_i)} E_i \in \mathcal{E}_i(t_i).$$

That is, the event proved by showing all of t_i's evidence is itself an event that t_i can prove. Henceforth, we let $M_i(t_i)$ denote the *maximally informative event* t_i can prove. I.e., we define

$$M_i(t_i) = \bigcap_{E_i \in \mathcal{E}_i(t_i)} E_i.$$

We sometimes refer to t_i presenting $M_i(t_i)$ as presenting *maximal evidence*.

Before formally defining a mechanism, we note that given our assumptions, it is without loss of generality to focus on mechanisms where the agents simultaneously make cheap talk reports of types and present evidence and where each agent truthfully reveals her type and presents maximal evidence. This is not the standard Revelation Principle but has been shown by, among others, Bull and Watson (2007) and Deneckere and Severinov (2008). Formally, let $\mathcal{E}_i = \cup_{t_i \in T_i} \mathcal{E}_i(t_i)$ and $\mathcal{E} = \prod_i \mathcal{E}_i$. A *mechanism* is then a function $P : T \times \mathcal{E} \to \Delta(A)$.

For notational brevity, given a mechanism P, $t_i \in T_i$, $(s_i, e_i) \in T_i \times \mathcal{E}_i(t_i)$, and $(t_{-i}, e_{-i}) \in T_{-i} \times \mathcal{E}_{-i}$, let

$$\tilde{u}_i(s_i, e_i, t_{-i}, e_{-i} \mid t_i, P) = \sum_a P(a \mid s_i, e_i, t_{-i}, e_{-i}) u_i(a, t_i)$$

and

$$\hat{u}_i(s_i, e_i \mid t_i, P) = \mathrm{E}_{t_{-i}} \tilde{u}_i(s_i, e_i, t_{-i}, M_{-i}(t_{-i}) \mid t_i, P).$$

In words, $\tilde{u}_i(s_i, e_i, t_{-i}, e_{-i} \mid t_i, P)$ is agent i's expected utility under mechanism P when her type is t_i but she reports s_i, presents evidence e_i, and expects all other agents to claim types t_{-i} and report evidence e_{-i}. Then $\hat{u}_i(s_i, e_i \mid t_i, P)$ is i's expected utility from

reporting (s_i, e_i) when her type is t_i and she expects the other agents to report their types truthfully and to provide maximal evidence.

A mechanism P is *incentive compatible* if for every agent i,

$$\hat{u}_i(t_i, M_i(t_i) \mid t_i, P) \geq \hat{u}_i(s_i, e_i \mid t_i, P),$$

for all $s_i, t_i \in T_i$ and all $e_i \in \mathcal{E}_i(t_i)$. The principal's expected payoff from an incentive compatible mechanism P is

$$\mathrm{E}_t \sum_a P(a \mid t, M(t)) v(a, t).$$

Our main result is that if the type dependence of the agents' utility is sufficiently simple, then for the principal, there is no value to commitment, no cost to robust incentive compatibility, and no need to randomize. We now make this statement more precise.

Before defining our notion of robust incentive compatibility, we begin with more standard notions. A mechanism is *ex post incentive compatible* if for every agent i,

$$\tilde{u}_i(t_i, M_i(t_i), t_{-i}, M_{-i}(t_{-i}) \mid t_i, P) \geq \tilde{u}_i(s_i, e_i, t_{-i}, M_{-i}(t_{-i}) \mid t_i, P),$$

for all $s_i, t_i \in T_i$, all $t_{-i} \in T_{-i}$, and all $e_i \in \mathcal{E}_i(t_i)$. In other words, a mechanism is ex post incentive compatible if each agent i has an incentive to report honestly and present maximal evidence even if she knows all the other agents' types and that they are reporting truthfully.

Say that a reporting strategy $\sigma_j : T_j \to T_j \times \mathcal{E}_j$ is *feasible* if whenever $\sigma_j(t_j) = (s_j, e_j)$, we have $e_j \in \mathcal{E}_j(t_j)$. A mechanism is *dominant strategy incentive compatible* if for every agent i,

$$\mathrm{E}_{t_{-i}} \tilde{u}_i(t_i, M_i(t_i), \sigma_{-i}(t_{-i}) \mid t_i, P) \geq \mathrm{E}_{t_{-i}} \tilde{u}_i(s_i, e_i, \sigma_{-i}(t_{-i}) \mid t_i, P)$$

for all $s_i, t_i \in T_i$, all feasible $\sigma_{-i} : T_{-i} \to T_{-i} \times \mathcal{E}_{-i}$, and all $e_i \in \mathcal{E}_i(t_i)$.

In mechanisms with evidence, neither of these notions of incentive compatibility implies the other. A mechanism could be ex post incentive compatible, but an agent might want to deviate if she knew another agent were going to report (s_i, e_i) where $e_i \neq M_i(s_i)$. That is, an agent might want to deviate from truth telling and maximal evidence if she knew another agent was going to deviate from truth telling and maximal evidence in a detectable way. Similarly, a mechanism could be dominant strategy incentive compatible but an agent could wish to deviate if she knew the specific types of her opponents. The robustness notion we will use combines both the ex post and dominant strategy features of the above definitions.

We say that a mechanism is *robustly incentive compatible* if for every agent i,

$$\tilde{u}_i(t_i, M_i(t_i), t_{-i}, e_{-i} \mid t_i, P) \geq \tilde{u}_i(s_i, e_i, t_{-i}, e_{-i} \mid t_i, P),$$

for all $s_i, t_i \in T_i$, all $(t_{-i}, e_{-i}) \in T_{-i} \times \mathcal{E}_{-i}$, and all $e_i \in \mathcal{E}_i(t_i)$. In other words, even if i knew the exact type and evidence reports of all other agents, it would be optimal to report truthfully and provide maximal evidence regardless of what those reports are. In mechanisms without evidence but with independent private values, robust incentive compatibility, ex post incentive compatibility, and dominant strategy incentive compatibility are all equivalent. While we have independent private values (i.e., the t_i's are independent across i and u_i is not a function of t_{-i}), these concepts are not equivalent here because of the evidence structure.

Obviously, robust incentive compatibility implies incentive compatibility, but the converse is not true. Hence the best robustly incentive compatible mechanism for the principal yields her a weakly lower expected payoff than the best incentive compatible mechanism, typically strictly lower. Our result will show that under our assumptions, the best incentive compatible mechanism for the principal is always robustly incentive compatible.

We say a mechanism P is *deterministic* if for every $(t, e) \in T \times \mathcal{E}$, $P(t, e)$ is a degenerate distribution. In other words, for every report and presentation of evidence, whether or not it involves truth telling and maximal evidence, the principal chooses an $a \in A$ without randomizing. Of course, randomization is an important feature of optimal mechanisms in some settings. We will show that under our assumptions, there is an optimal mechanism which is deterministic.

Finally, to state what it means that there is no value to commitment, we must define what the principal can accomplish in the absence of commitment. Without commitment, we assume that there is a game in which, just as in the revelation mechanism, agents simultaneously make type reports and present evidence, perhaps with randomization. The principal observes these choices and then chooses some allocation a, again perhaps with randomization. More formally, the set of strategies for agent i, Σ_i, is the set of functions $\sigma_i : T_i \to \Delta(T_i \times \mathcal{E}_i)$ such that $\sigma(s_i, e_i \mid t_i) > 0$ implies $e_i \in \mathcal{E}_i(t_i)$. That is, if agent i is type t_i and puts positive probability on providing evidence e_i, then this evidence must be feasible for t_i in the sense that $e_i \in \mathcal{E}_i(t_i)$. The principal's set of feasible strategies, Σ_P, is the set of functions $\sigma_P : T \times \mathcal{E} \to \Delta(A)$. We consider the set of perfect Bayesian equilibria of this game. More precisely, a belief by the principal is a function $\mu : T \times \mathcal{E} \to \Delta(T)$, giving the principal's beliefs about t given a profile of reports and evidence presentation. For notational convenience, given $\sigma_{-i} \in \Sigma_{-i}$, $\sigma_P \in \Sigma_P$, $a \in A$, and $(s_i, e_i) \in T_i \times \mathcal{E}_i$, let

$$Q_i(a \mid s_i, e_i, \sigma_{-i}, \sigma_P) = \mathrm{E}_{t_{-i}} \sum_{(s_{-i}, e_{-i})} \sigma_P(a \mid s, e) \prod_{j \neq i} \sigma_j(s_j, e_j \mid t_j).$$

In other words, this is the probability the principal chooses allocation a given that she uses strategy σ_P, agents other than i use strategies σ_j, $j \neq i$, and agent i reports s_i and presents evidence e_i.

We say that $(\sigma_1, \ldots, \sigma_I, \sigma_P, \mu)$ is a perfect Bayesian equilibrium[2] if the following conditions hold. First, for every i and every $t_i \in T_i$, $\sigma_i(s_i, e_i \mid t_i) > 0$ implies

$$(s_i, e_i) \in \arg \max_{s_i' \in T_i, e_i' \in \mathcal{E}_i(t_i)} \sum_{a \in A} Q_i(a \mid s_i', e_i', \sigma_{-i}, \sigma_P) u_i(a, t_i).$$

Second, for every $(s, e) \in T \times \mathcal{E}$, $\sigma_P(a \mid s, e) > 0$ implies

$$a \in \arg \max_{a' \in A} \sum_{t \in T} \mu(t \mid s, e) v(a', t).$$

Third, for every (s, e), $\mu(\cdot \mid s, e)$ respects independence across agents. That is, i's report (s_i, e_i) only affects the principal's beliefs about t_i and his beliefs about t_i and t_j respect independence for all $i \neq j$. Formally, we have functions $\mu_i : T_i \times \mathcal{E}_i \to \Delta(T_i)$ such that for all $t \in T$ and all $(s, e) \in T \times \mathcal{E}$,

$$\mu(t \mid s, e) = \prod_i \mu_i(t_i \mid s_i, e_i).$$

Fourth, for all (s, e), $\mu(\cdot \mid s, e)$ respects feasibility. That is, the principal's beliefs must put zero probability on any type which is infeasible given (s, e). Formally, for every $t_i \in T_i$ and $(s_i, e_i) \in T_i \times \mathcal{E}_i$, we have $\mu_i(t_i \mid s_i, e_i) = 0$ if $e_i \notin \mathcal{E}_i(t_i)$.

Finally, the principal's beliefs are consistent with Bayes' rule whenever possible in the sense that for every $(s_i, e_i) \in T_i \times \mathcal{E}_i$ such that there exists t_i with $\sigma_i(s_i, e_i \mid t_i) > 0$, we have

$$\mu_i(t_i \mid s_i, e_i) = \frac{\sigma_i(s_i, e_i \mid t_i)\rho_i(t_i)}{\sum_{t_i' \in T_i} \sigma_i(s_i, e_i \mid t_i')\rho_i(t_i')}.$$

(Recall that ρ_i is the principal's prior over t_i.)

The equilibria which will give the principal the same payoff as in the optimal mechanism will satisfy a certain robustness property that, for lack of a better phrase, we simply call *robustness*. Specifically, a perfect Bayesian equilibrium (σ, μ) is *robust* if for every i and every $t_i \in T_i$, $\sigma_i(s_i, e_i \mid t_i) > 0$ implies

$$(s_i, e_i) \in \arg \max_{s_i' \in T_i, e_i' \in \mathcal{E}_i(t_i)} \sum_{a \in A} \sigma_P(a \mid s_i', e_i', s_{-i}, e_{-i}) u_i(a, t_i), \ \forall (s_{-i}, e_{-i}) \in T_{-i} \times \mathcal{E}_{-i}.$$

[2]Our definition is the natural adaptation of Fudenberg and Tirole's (1991) definition of perfect Bayesian equilibrium for games with observed actions and independent types to allow type–dependent sets of feasible actions.

In other words, $\sigma_i(t_i)$ is optimal for t_i regardless of the actions played by the other agents, given the strategy of the principal.

Given a perfect Bayesian equilibrium (σ, μ), the principal's expected utility is

$$\mathrm{E}_t \sum_a \prod_i \sigma_i(s_i, e_i \mid t_i)\sigma_P(a \mid s, e)v(a, t).$$

We will show that there is a robust perfect Bayesian equilibrium of this game which gives the principal the same expected utility as the optimal mechanism. In this sense, the principal does not need the commitment assumed in characterizing the optimal mechanism.

3 Commitment, Determinism, and Robust Incentive Compatibility

Our result assumes that the type dependence of the agents' utility functions takes a particularly simple form. Formally, we say that $u_i(a, t_i)$ satisfies *simple type dependence* if there exist functions $u_i : A \to \mathbf{R}$ and $\beta_i : T_i \to \mathbf{R}$ such that $u_i(a, t_i) = u_i(a)\beta_i(t_i)$ where $\beta_i(t_i) \neq 0$ for all $t_i \in T_i$.[3]

This multiplicative separability is more restrictive than it may appear. Under simple type dependence, we effectively have a model where all types have the same indifference curves in $\Delta(A)$ space, those defined by utility function $u_i(a)$, but the direction of improvement may vary across types. Obviously, this is quite restrictive in general. To see this, renormalize $u_i(a, t_i)$ by dividing through by $|\beta_i(t_i)|$. Hence it is strategically equivalent to define

$$\bar{u}_i(a, t_i) = \frac{u_i(a, t_i)}{|\beta_i(t_i)|}.$$

With this renormalization, we have

$$\bar{u}_i(a, t_i) = \begin{cases} u_i(a), & \text{if } t_i \in T_i^+; \\ -u_i(a), & \text{if } t_i \in T_i^- \end{cases}$$

where

$$T_i^+ = \{t_i \in T_i \mid \beta_i(t_i) > 0\}$$

and $T_i^- = T_i \setminus T_i^+$. We can also rewrite the principal's utility function in terms of the

[3]If $\beta_i(t_i) = 0$ for some t_i, then that type is indifferent over all actions by the principal and so will always truthfully reveal. Hence we may as well disregard such types.

$u_i(a)$'s. Specifically, note that

$$\begin{aligned} v(a,t) &= u_0(a) + \sum_i v_i(t_i)u_i(a,t_i) \\ &= u_0(a) + \sum_i v_i(t_i)\beta_i(t_i)u_i(a) \\ &= u_0(a) + \sum_i \bar{v}_i(t_i)u_i(a), \end{aligned}$$

where $\bar{v}_i(t_i) = v_i(t_i)\beta_i(t_i)$. With some abuse of notation, we can then redefine v_i and write the principal's utility function as

$$v(a,t) = u_0(a) + \sum_{i=1}^{I} u_i(a)v_i(t_i) = \sum_{i=0}^{I} u_i(a)v_i(t_i).$$

Because it is more convenient for analysis, we will write the principal and agents' utility functions in the form above hereafter, referring to T_i^+ as the *positive types* of i and T_i^- as the *negative types*.

While this assumption is restrictive in general, there is one case where it is not restrictive at all, namely, where each agent has only two type–independent indifference curves over A. For example, this obviously must hold in the case when the principal has only two pure actions, the case originally considered by Glazer and Rubinstein (2004, 2006). Similarly, consider a type–dependent version of the simple allocation problem where each agent cares only about whether she receives the good or not, but some types prefer to get the good and other prefer not to. Here the principal has as many actions as there are agents (more if she can keep the good), but each agent has only two indifference curves over A. In this case, there are only two (nontrivial) preferences over $\Delta(A)$, so this formulation is not restrictive in that context.

Of course, simple type dependence also nests type independence where $\beta_i(t_i) = 1$ for all t_i or, equivalently, $T_i^+ = T_i$ and $T_i^- = \emptyset$. This is the case considered in most of the literature on mechanism design with evidence, particularly the papers on the value of commitment. For example, it is easy to see that the allocation problems Example 1 and Example 3b discussed in Section 1 satisfy simple type dependence because of the type independence assumed.

For an example which satisfies simple type dependence but not type independence, consider the public goods problem, Example 2, discussed in the Introduction. Suppose $A = \{0, 1\}$ where 0 corresponds to not providing the public good and 1 corresponds to provision. Let $u_i(0, t_i) = 0$ and $u_i(1, t_i) = \beta_i(t_i)$. The interpretation is that $\beta_i(t_i)$ is the value of the public good to type t_i net of i's share of the cost of provision. For types who do not value the public good very much, we will have $\beta_i(t_i) < 0$, while there may be other types who value it substantially and therefore have $\beta_i(t_i) > 0$. Assume the principal's

utility function is the sum of the agent's utilities. Letting T_i^+ denote the set of types with $\beta_i(t_i) > 0$, we can rewrite this model in the form of simple type dependence with

$$u_i(a) = \begin{cases} 0, & \text{if } a = 0; \\ 1, & \text{if } a = 1 \end{cases}$$

and $v_i(t_i) = |\beta_i(t_i)|$ for all $t_i \in T_i$ and all i.

When we show that commitment has no value, we will construct an equilibrium with the same outcome as in the optimal mechanism. The equilibrium we construct is particularly simple in that it can be constructed from a set of I one–agent games which do not depend on A or preferences over A.

Specifically, we define the *artificial game for agent i* as follows. This is a game with two players, the principal and agent i. Agent i has type set T_i. Type t_i has action set $T_i \times \mathcal{E}_i(t_i)$. The principal has action set $X \subseteq \mathbf{R}$ where X is the compact interval $[\min_j \min_{t_j \in T_j} v_j(t_j), \max_j \max_{t_j \in T_j} v_j(t_j)]$. Agent i's payoff as a function of t_i and the principal's choice of x is

$$\begin{cases} x, & \text{if } t_i \in T_i^+; \\ -x, & \text{otherwise.} \end{cases}$$

The principal's utility in this situation is $-(x - v_i(t_i))^2$. In other words, the artifical game is a persuasion game where positive types want the principal to believe that $v_i(t_i)$ is large and negative types want him to believe it is small. The structure of A and $u_i(a)$ play no role. As in the real game defined earlier, a strategy for agent i is a function $\sigma_i : T_i \to \Delta(T_i \times \mathcal{E}_i)$ with the property that $\sigma_i(s_i, e_i \mid t_i) > 0$ implies $e_i \in \mathcal{E}_i(t_i)$. We denote a strategy for the principal as $X_i : T_i \times \mathcal{E}_i \to X$.

Theorem 1. *If every u_i exhibits simple type dependence, then commitment and randomization have no value for the principal, while robust incentive compatibility has no cost. That is, there is an optimal incentive compatible mechanism for the principal which is deterministic and robustly incentive compatible. In addition, there is a robust perfect Bayesian equilibrium with the same outcome as in this optimal mechanism. In this equilibrium, agent i's strategy is a perfect Bayesian equilibrium strategy in the artifical game for agent i.*

As mentioned in Section 1, there are earlier results for the one–agent setting showing that commitment is not valuable to the principal. Our result extends these in several ways. First, we consider multiple agents. Second, because we have multiple agents, we can consider robust incentive compatibility — that is, the question of robustness with respect to agents' beliefs about other agents, an issue absent in the one–agent setting. Third, our characterization of these equilibrium strategies is novel.

Even when we restrict our analysis to the case of $I = 1$ so that we also only have one agent, our results are not nested by the previous literature. For the remainder of this

section, we discuss the one–agent case, so t refers to the type of the single agent, T her set of types, and u her utility function. The first papers to show no value to commitment in a mechanism design problem with evidence were Glazer and Rubinstein (2004, 2006). These papers used weaker assumptions on evidence than we use as they do not require normality. However, they assumed that the principal only had two actions available and the agent's preference over these actions was independent of her type. Our assumptions on preferences, restricted to the one–agent case where the principal has only two actions, are completely general, though. Unlike our model, randomness may be important for optimal mechanism in Glazer–Rubinstein because they allow for violations of normality.

Sher (2011) generalizes the Glazer–Rubinstein result via a concavity assumption. In our notation, his assumptions are as follows. First, the utility function of the agent u is independent of her type t. Sher assumes that the agent's utility is strictly increasing with respect to an order over the principal's actions. The concavity assumption is that for every type t, there exists a concave function φ_t such that $v(a,t) = \varphi_t(u(a))$. That is, given the agent's type, the principal's utility function over A is a concave transformation of the agent's utility function. It is not hard to see that this implies that the principal does not need to randomize in the optimal mechanism.

In the one–agent version of our model, the principal's utility function is $v(a,t) = u_0(a) + v(t)u(a)$. Since the principal's utility over A given t depends on more than just the agent's utility, this is not nested by Sher's assumptions, even in the type–independent version of our model.

Finally, Hart, Kremer, and Perry (2015) give a version of the Glazer–Rubinstein result which, like our result, assumes normality of evidence. Like Glazer–Rubinstein and Sher, they assume that the agent's utility function is independent of her type. Like Sher, they assume that the agent's utility is increasing in a. In addition, they discuss two other preference assumptions. First, they give a no–value–to–commitment result which assumes that the principal *cannot* randomize. For this result, they assume that for each $t \in T$, the utility function of the principal over A can be written as $v(a,t) = \varphi_t(u(a))$ where given any $\mu \in \Delta(T)$, $\sum_t \mu(t)\varphi_t$ is a single–peaked function. That is, there exists u^*_μ such that $\sum_t \mu(t)\varphi_t(u)$ is strictly increasing in u for $u < u^*_\mu$ and strictly decreasing in u for $u > u^*_\mu$. Again, because we allow the principal's utility to depend on a directly as well as through $u(a)$ in the form $v(a,t) = u_0(a) + u(a)v(t)$, our model violates this assumption in general, even in the type–independent version of our model. Also, our assumptions *imply* that the principal does not need to randomize, while Hart, Kremer, and Perry's first theorem *assumes* that he cannot.

Hart, Kremer, and Perry's second no–value–to–commitment result allows the principal to randomize. Here they use an assumption called PUB or Principal's Uniform Best. To state this, say that $\bar{u} \in \mathbf{R}$ is a feasible utility level for the agent if there exists

$p \in \Delta(A)$ such that $\sum_a p(a)u(a) = \bar{u}$. Given a feasible $\bar{u}$, let $P(\bar{u})$ denote the set of $p \in \Delta(A)$ such that $\sum_a p(a)u(a) = \bar{u}$. The assumption then is that for every feasible $\bar{u}$, there exists $p \in P(\bar{u})$ such that $\sum_a p(a)v(a,t) \geq \sum_a \hat{p}(a)v(a,t)$ for every $\hat{p} \in P(\bar{u})$ for every $t \in T$. In other words, if the principal is constrained to give the agent utility $\bar{u}$, then there is a utility–maximizing way for him to do this which is independent of the agent's type. In the one–agent case, our model does satisfy this assumption. Clearly, if $v(a,t) = u_0(a) + u(a)v(t)$, then for any t, the principal's preferred $p \in P(\bar{u})$ is any p maximizing $\sum_a p(a)u_0(a)$. So our model satisfies their PUB assumption. Thus the type–independent one–agent version of our model is nested in their second result.

4 Optimal Mechanisms with Dye Evidence

In light of Theorem 1, we can compute the outcomes of optimal mechanisms by identifying the best robust equilibrium for the principal. In particular, we can compute these equilibria by considering the artificial game for each agent i. In some cases, these equilibria are very easy to characterize. In this section, we illustrate by considering optimal mechanisms with a particular evidence structure introduced by Dye (1985) and studied extensively in both the economics and accounting literatures.

We say that the model has Dye evidence if for every i, for all $t_i \in T_i$, either $\mathcal{E}_i(t_i) = \{T_i\}$ or $\mathcal{E}_i(t_i) = \{\{t_i\}, T_i\}$. In other words, any given type either has no evidence in the sense that she can only prove the trivial event T_i or can choose between proving nothing and proving exactly her type. Let T_i^0 denote the set of $t_i \in T_i$ with $\mathcal{E}_i(t_i) = \{T_i\}$.

Our artificial games differ in one respect from the usual persuasion games in the literature. In our artificial game, agent i both presents evidence and makes a cheap talk claim regarding her type. Of course, if these cheap talk claims convey information, we can always permute agent i's use of these claims and the principal's interpretation of them to obtain another equilibrium.

There is also another form of multiplicity which is more standard in the literature on games with evidence. In some cases, we may have an equilibrium where the principal has the same beliefs about the agent whether she presents evidence e or evidence e'. In these cases, we can construct an equilibrium where the agent presents evidence e and another where she presents evidence e'.

Note that in both of these cases, the principal's beliefs about the agent along the equilibrium path are the same across these various equilibria. That is, if the agent is type t, the belief the principal will have about t is the same across these equilibria. With this issue in mind, we say that an equilibrium in the artificial game for agent i is

essentially unique if all equilibria have the same outcome in this sense.

To be precise, given equilibria (σ_i^*, x_i^*) and $(\hat{\sigma}_i^*, \hat{x}_i^*)$ of the artificial game for i, we say these equilibria are *essentially equivalent* if for every $x \in X$ and every $t_i \in T_i$, we have

$$\begin{aligned}&\sigma_i^*\left(\{(s_i, e_i) \in T_i \times \mathcal{E}_i(t_i) \mid x_i^*(s_i, e_i) = x\} \mid t_i\right)\\ &\quad = \hat{\sigma}_i^*\left(\{(s_i, e_i) \in T_i \times \mathcal{E}_i(t_i) \mid \hat{x}_i^*(s_i, e_i) = x\} \mid t_i\right).\end{aligned}$$

If there is an equilibrium with the property that every other equilibrium is essentially equivalent to it, we say the equilibrium is *essentially unique*.

The simplest case to consider with Dye evidence is where the utility functions are not type dependent at all. We say that the model exhibits *type–independent utility* if $u_i(a, t_i)$ is independent of t_i for all i and a. In this case, we abuse notation and write $u_i(a, t_i) = u_i(a)$. Note that this is equivalent to assuming $T_i^- = \emptyset$ (or redefining u_i and taking $T_i^+ = \emptyset$).

The following results build on well–known characterizations of equilibria in evidence games using the Dye evidence structure.

Theorem 2. *In any model with Dye evidence, for every i, there exists a unique v_i^* such that*

$$v_i^* = \mathrm{E}\left[v_i(t_i) \mid t_i \in T_i^0 \text{ or } v_i(t_i) \leq v_i^*\right].$$

If $T_i^- = \emptyset$, the essentially unique equilibrium in the artificial game for i is a pure strategy equilibrium where every type makes the same cheap talk claim, say s_i^*, and only types with evidence with $v_i(t_i) \geq v_i^*$ present (nontrivial) evidence. That is, type t_i sends $(s_i^*, e_i^*(t_i))$ with probability 1 where

$$e_i^*(t_i) = \begin{cases} T_i, & \text{if } t_i \in T_i^0 \text{ or } v_i(t_i) < v_i^*; \\ \{t_i\}, & \text{otherwise.} \end{cases}$$

To see the intuition, note first that cheap talk cannot be credible in this game since every type wants the principal to believe that v_i is large. So if i has no evidence (i.e., can only prove the trivial event T_i), then she has no ability to convey any information to the principal — she can only send an uninformative cheap talk message and prove nothing. If i can prove her type is t_i, she wants to do so only if $v_i(t_i)$ is at least as large as what the principal would believe if she showed no evidence. Thus types with evidence but lower values of $v_i(t_i)$ will pool with the types who have no evidence, leading to an expectation of $v_i(t_i)$ equal to v_i^*.

In this equilibrium, the principal's expectation of $v_i(t_i)$ will be v_i^* given a type with no evidence or with $v_i(t_i) < v_i^*$ and will equal the true value otherwise. More formally,

let

$$\hat{v}_i(t_i) = \begin{cases} v_i^*, & \text{if } t_i \in T_i^0 \text{ or } v_i(t_i) < v_i^*; \\ v_i(t_i), & \text{otherwise.} \end{cases}$$

For every $\hat{v} = (\hat{v}_1, \ldots, \hat{v}_I)$, let $\hat{p}(\cdot \mid \hat{v})$ denote any $p \in \Delta(A)$ maximizing

$$\sum_{a \in A} p(a) \left[u_0(a) + \sum_i u_i(a) \hat{v}_i \right].$$

The following is a corollary to Theorems 1 and 2.

Corollary 1. *In any model with type–independent utility and Dye evidence, there is an optimal mechanism P with $P(\cdot \mid t, M(t)) = \hat{p}(\cdot \mid \hat{v}(t))$. In other words, the outcome selected by the principal when the profile of types is t is $\hat{p}(\cdot \mid \hat{v}(t))$.*

We can use Corollary 1 to give simple characterizations of optimal mechanisms in many cases of interest.

Example 1. The simple allocation problem with Dye evidence. In this case, $\hat{p}(i \mid t) > 0$ iff $\hat{v}_i(t_i) = \max_j \hat{v}_j(t_j)$. We can break indifferences in a particularly simple way and recast this characterization in the form of a *favored–agent mechanism*. More specifically, say that P is a favored–agent mechanism if there is a *threshold* $v^* \in \mathbf{R}$ and an agent i, the *favored agent*, such that the following holds. First, if no agent $j \neq i$ proves that $v_j(t_j) > v^*$, then i receives the good. Second, if some agent $j \neq i$ does prove that $v_j(t_j) > v^*$, then the good is given to the agent who proves the highest $v_j(t_j)$ (where this may be agent i).

More specifically, a favored–agent mechanism where the favored agent is any i satisfying $v_i^* = \max_j v_j^*$ and the threshold v^* is given by v_i^* is an optimal mechanism. To see this, fix any t. By definition, $\hat{v}_j(t_j) \geq v_j^*$ for all j. Hence if $v_i^* \geq v_j^*$ for all j, then $\hat{v}_i(t_i) \geq v_j^*$ for all j. Hence for any j such that $\mathcal{E}_j(t_j) = \{T_j\}$ or $v_j(t_j) < v_j^*$, we have $\hat{v}_i(t_i) \geq v_i^* \geq v_j^* = \hat{v}_j(t_j)$. So if every $j \neq i$ satisfies this, it is optimal for the principal to give the good to i. Otherwise, it is optimal for him to give it to any agent who proves the highest value.

As noted in Section 1, this mechanism is very similar to the favored–agent mechanism discussed by BDL for the allocation problem with costly verification, a point we return to below.

Example 2. The multi–unit allocation problem with Dye evidence. It is not hard to extend the above analysis to the case where the principal has multiple identical units of the good to allocate. Suppose he has $K < I$ units and, for simplicity, assume he must allocate all of them. Suppose each agent can only have either 0 or 1 unit. Then

the principal's action can be thought of as selecting a subset of $\{1, \ldots, I\}$ of cardinality K. The principal's utility given the set $\hat{\mathcal{I}}$ is $\sum_{i \in \hat{\mathcal{I}}} v_i(t_i)$. As before, agent i's utility is 0 if she does not get a unit, 1 if she does.

In this case, it is easy to see that the principal allocates units to the K agents with the highest values of $\hat{v}_i(t_i)$. This can be computed recursively as a kind of favored–agent mechanism. In other words, we allocate the first unit to the agent with the highest value of v_i^* if no other agent proves a higher value and to the agent with the highest proven value otherwise. Once removing this agent and unit, we follow the same procedure for the second unit, and so on. It is easy to see that the agent with the highest value of v_i^* is the most favored agent in the sense that at least K agents must prove a value above her v_i^* for her to not get a unit. Similarly, the agent with the second–highest value of v_i^* is the second–most favored agent in the sense that at least $K-1$ of the "lower ranked" agents must prove a value above her v_i^* for her not to get a unit, etc.

Example 3. Allocating a "bad." Another setting of interest is where the principal has to choose one agent to carry out an unpleasant task (e.g., serve as department chair). It is easy to see that this problem is effectively identical to having $I-1$ goods to allocate since not receiving the assignment is the same as receiving a good. Thus we can treat the principal's set of feasible actions as the set of subsets of $\{1, \ldots, I\}$ of cardinality $I-1$, interpreted as the set of agents who are *not* assigned the task. The one aspect of this example that may seem odd is that the principal's utility if he assigns the task to agent i is then $\sum_{j \neq i} v_j(t_j)$. On the other hand, it is an innocuous renormalization of the principal's utility function to subtract the allocation–independent term $\sum_j v_j(t_j)$ from her utility. In this case, we see that the principal's payoff to assigning the task to agent i is $-v_i(t_i)$, so $v_i(t_i)$ is naturally interpreted as t_i's level of *in*competence in carrying out the task. One can apply the analysis of the previous example for the special case of $K = I - 1$ to characterize the optimal mechanism for this example.

While the case of type–independent utility with Dye evidence is particularly tractable, the case of simple type dependence is not much more difficult. To see the intuition, again consider the artificial game for i where some types wish to persuade the principal that $v_i(t_i)$ is large and other types want to convince him $v_i(t_i)$ is small. Suppose that when the agent doesn't prove her type, she makes a cheap talk claim regarding whether her type is positive (i.e., she wants the principal to think $v_i(t_i)$ is large) or negative (i.e., the reverse). Let v_i^+ denote the principal's belief about v_i if i does not prove her type but says it is positive and let v_i^- be the analog for the case where i claims her type is negative. If $v_i^+ > v_i^-$, then every positive type without evidence prefers to truthfully report that her type is positive, while every negative type without evidence will honestly reveal that her type is negative since this leads to the best possible belief from i's point of view. If i is a positive type with evidence, she will want to prove her type only if $v_i(t_i) > v_i^+$, while a negative type with evidence will prove her type only if $v_i(t_i) < v_i^-$. Hence for this to

be an equilibrium, we must have

$$v_i^+ = \mathrm{E}\left[v_i(t_i) \mid (t_i \in T_i^+ \cap T_i^0) \text{ or } (t_i \in T_i^+ \setminus T_i^0 \text{ and } v_i(t_i) \leq v_i^+)\right]$$

and

$$v_i^- = \mathrm{E}\left[v_i(t_i) \mid (t_i \in T_i^- \cap T_i^0) \text{ or } (t_i \in T_i^- \setminus T_i^0 \text{ and } v_i(t_i) \geq v_i^-)\right].$$

Suppose this gives a unique value for v_i^+ and v_i^-. If these values do not satisfy $v_i^+ \geq v_i^-$, then this doesn't work as the positive types will prefer to act like negative types and vice versa. In this case, we must pool all types.

This motivates the following result.

Theorem 3. *In any model with Dye evidence, for every i, there exists a unique triple v_i^+, v_i^-, and v_i^* such that*

$$v_i^+ = \mathrm{E}\left[v_i(t_i) \mid (t_i \in T_i^+ \cap T_i^0) \text{ or } (t_i \in T_i^+ \setminus T_i^0 \text{ and } v_i(t_i) \leq v_i^+)\right],$$

$$v_i^- = \mathrm{E}\left[v_i(t_i) \mid (t_i \in T_i^- \cap T_i^0) \text{ or } (t_i \in T_i^- \setminus T_i^0 \text{ and } v_i(t_i) \geq v_i^-)\right],$$

and

$$v_i^* = \mathrm{E}\left[v_i(t_i) \mid (t_i \in T_i^0) \text{ or } (t_i \in T_i^- \setminus T_i^0 \text{ and } v_i(t_i) \geq v_i^*) \text{ or } (t_i \in T_i^+ \setminus T_i^0 \text{ and } v_i(t_i) \leq v_i^*)\right].$$

If $v_i^+ \leq v_i^-$, then there is an essentially unique equilibrium in the artificial game for i. In this pure strategy equilibrium, there is a fixed type $\hat{s}_i$ such that t_i reports $(\hat{s}_i, e_i^*(t_i))$ where

$$e_i^*(t_i) = \begin{cases} T_i, & \text{if } t_i \in T_i^0 \text{ or } (t_i \in T_i^+ \text{and } v_i(t_i) < v_i^*) \text{ or } (t_i \in T_i^- \text{and } v_i(t_i) > v_i^*); \\ \{t_i\}, & \text{otherwise.} \end{cases}$$

If $v_i^+ > v_i^-$, there are two equilibria that are not essentially equivalent to one another and every other equilibrium is essentially equivalent to one of the two. One of these is the equilibrium described above. The other is another equilibrium in pure strategies. In this second equilibrium, there are types $\hat{s}_i^+$ and $\hat{s}_i^-$ with $\hat{s}_i^+ \neq \hat{s}_i^-$ such that $t_i \in T_i^k$ sends $(\hat{s}_i^k, e_i^k(t_i))$, $k = -, +$, where

$$e_i^+(t_i) = \begin{cases} T_i, & \text{if } t_i \in T_i^0 \text{ or } v_i(t_i) < v_i^+; \\ \{t_i\}, & \text{otherwise,} \end{cases}$$

and

$$e_i^-(t_i) = \begin{cases} T_i, & \text{if } t_i \in T_i^0 \text{ or } v_i(t_i) > v_i^-; \\ \{t_i\}, & \text{otherwise.} \end{cases}$$

If $v_i^+ > v_i^-$, then there are (essentially) two equilibria in the artificial game. As the result below will show, we can always compare these equilibria for the principal and the better one is the one which separates the positive and negative types. Thus this is the

equilibrium that corresponds to the optimal mechanism. With this in mind, now define $\hat{v}_i(t_i)$ as follows. If $v_i^+ > v_i^-$, we let

$$\hat{v}_i(t_i) = \begin{cases} v_i^+, & \text{if } t_i \in T_i^0 \cap T_i^+ \text{ or } t_i \in T_i^+ \setminus T_i^0 \text{ and } v_i(t_i) < v_i^*; \\ v_i^-, & \text{if } t_i \in T_i^0 \cap T_i^- \text{ or } t_i \in T_i^- \setminus T_i^0 \text{ and } v_i(t_i) > v_i^*; \\ v_i(t_i), & \text{otherwise.} \end{cases}$$

If $v_i^+ \leq v_i^-$, let

$$\hat{v}_i(t_i) = \begin{cases} v_i(t_i), & \text{if } (t_i \in T_i^+ \setminus T_i^0 \text{ and } v_i(t_i) > v_i^*) \text{ or } (t_i \in T_i^- \setminus T_i^0 \text{ and } v_i(t_i) < v_i^*); \\ v_i^*, & \text{otherwise.} \end{cases} \tag{1}$$

For each $t \in T$, let $\hat{p}(\cdot \mid \hat{v})$ denote any $p \in \Delta(A)$ maximizing

$$\sum_{a \in A} p(a) \left[u_0(a) + \sum_i u_i(a, t_i) \hat{v}_i \right].$$

The following result is a corollary to Theorems 1 and 3.

Corollary 2. *In any model with simple type dependence and Dye evidence, there is an optimal mechanism P with $P(\cdot \mid t, M(t)) = \hat{p}(\cdot \mid \hat{v}(t))$. In other words, the outcome selected by the principal when the profile of types is t is $\hat{p}(\cdot \mid \hat{v}(t))$.*

The only part of this result that requires proof is the claim above that when $v_i^+ > v_i^-$, the equilibrium that is better for the principal is the one that separates the positive and negative types. This is shown in Appendix D.

Example 4. The public goods problem. As an application, consider the public goods model discussed in Section 3. For simplicity, we write out the optimal mechanism only for the case where $v_i^+ > v_i^-$ for all i, but similar comments apply more generally. We know that in equilibrium, given a profile of types t, the principal's expectation of v_i will be given by $\hat{v}_i(t_i)$ defined in equation (1) above. Then the principal will provide the public good iff $\sum_i \hat{v}_i(t_i) > 0$. This is the analog of the optimal mechanism for costly verification identified by Erlanson and Kleiner (2015). Specifically, in their model, we compute what they call an "adjusted report" for each agent i given t_i. Exactly as above for the types with evidence, the adjusted report for a positive type with evidence is $\max\{v_i^+, v_i(t_i)\}$, while the adjusted report for a negative type with evidence is $\min\{v_i^-, v_i(t_i)\}$ for certain cutoffs v_i^+ and v_i^-. These reports are adjusted by the verification cost and then summed to determine the optimal action by the principal.

There is a simple reason why the optimal mechanism here for the simple allocation problem parallels the optimal mechanism under costly verification identifed by BDL and why the optimal mechanism here for the public goods problem parallels the optimal

mechanism under costly verification identified by Erlanson and Kleiner. To explain, first note that in both of the costly verification models, the assumption is that the principal can pay a cost c_i to learn the realization of agent i's type, t_i. The agent cannot affect this verification process.

So consider the following "hybrid" model. Assume the Dye structure for evidence as above. However, change our assumptions so that when an agent provides evidence, this does not immediately prove anything to the principal. Instead, this is a *verifiable* message, but one that is costly to verify. Equivalently, the principal has to pay to "read" the evidence to see that it does indeed prove what the agent has said that it proves. This adds verification costs to the current model and adds nonverifiable types to the earlier costly verification models, where the principal cannot verify the validity (or lack thereof) of a claim to be an unverifiable type. It is not hard to show that this latter change alters the optimal mechanism in BDL only in small and relatively obvious ways. We conjecture that a similar result occurs in the Erlanson and Kleiner model.

As we take the verification costs in those models to zero, the optimal mechanisms converge continuously to the optimal mechanisms in the model of this paper. Thus it is not surprising that for *small* verification costs, the optimal mechanisms look similar. The more intriguing observation is that this is true for large costs as well.

While the optimal mechanisms look similar in the two models, this does not mean that all properties are the same. While BDL note that their mechanism is ex post and dominant strategy incentive compatible, the optimal mechanism in Erlanson and Kleiner is ex post incentive compatible but not dominant strategy incentive compatible. While these results have been shown only for the model without unverifiable types, it is easy to show the same holds for the hybrid version of BDL and we conjecture it holds for the extension of Erlanson–Kleiner. Also, we believe that commitment is valuable for the principal in both models under costly verification, whether or not there are unverifiable types.

5 Proof Sketch

In this section, we sketch the proof of Theorem 1. For simplicity, we sketch the proof in the context of a special case, namely, the simple allocation problem. So assume for this section that the principal has one unit of an indivisible good to allocate to some agent. All agents desire the good and the principal must give it to one of the agents. So $A = \{1, \ldots, I\}$ where $a = i$ means that the principal allocates the good to agent i. The

utility functions of the agents are type independent with

$$u_i(a) = \begin{cases} 1, & \text{if } a = i \\ 0, & \text{otherwise.} \end{cases}$$

The payoff to the principal to allocating the good to agent i given type profile t is $v_i(t_i)$ which we assume is strictly positive for all i and all t_i.

One convenient simplification in the type independent case is that we can write a mechanism as a function only from type reports into choices by the principal, where it is understood that if i claims type t_i, she also reports maximal evidence for t_i, namely, $M_i(t_i)$. This works in the type independent case because if i claims type t_i but does not show evidence $M_i(t_i)$, the principal knows how to punish i — namely, he can give i the good with zero probability, the worst possible outcome for i. This will deter any such "obvious" deviation. Of course, the mechanism must still deter the more subtle deviations to reporting some $s_i \neq t_i$ and providing evidence $M_i(s_i)$. So for this proof sketch, we will write a mechanism as a function $P : T \to \Delta(A)$.

Fix an optimal mechanism P. Given this mechanism, we can construct the probability that any given type of any given agent receives the good. Let

$$\hat{p}_i(t_i) = \mathrm{E}_{t_{-i}} P(i \mid t_i, t_{-i}).$$

This is type t_i's probability of being allocated the good in mechanism P. Partition each T_i according to equality under $\hat{p}_i$. In other words, for each $\alpha \in [0, 1]$, let

$$T_i^\alpha = \{t_i \in T_i \mid \hat{p}_i(t_i) = \alpha\}.$$

Of course, since T_i is finite, there are only finitely many values of α such that $T_i^\alpha \neq \emptyset$. Unless stated otherwise, any reference below to a T_i^α set assumes that this set is nonempty. Let $\mathcal{T}_i$ denote the partition of T_i so defined and $\mathcal{T}$ the induced (product) partition of T.

It is easy to see that incentive compatibility is equivalent to the statement that $M_i(s_i) \in \mathcal{E}_i(t_i)$ implies $\hat{p}_i(t_i) \geq \hat{p}_i(s_i)$. In other words, if t_i can report s_i credibly in the sense that t_i has available the maximal evidence of s_i, then the mechanism must give the good to t_i at least as often as s_i.

The first key observation is that without loss of generality, we can take the mechanism to be measurable with respect to $\mathcal{T}$. While this property may seem technical, it is the key property behind our results and is not generally true for models with more type dependence.

To see why this property holds in the simple allocation problem, suppose it is violated. In other words, suppose we have some pair of types $s_i, s_i' \in T_i$ such that $\hat{p}_i(s_i) = \hat{p}_i(s_i')$ but

P is not measurable with respect to $\{s_i, s_i'\}$ in the sense that there is some $t_{-i} \in T_{-i}$ with $P(\cdot \mid s_i, t_{-i}) \neq P(\cdot \mid s_i', t_{-i})$. Consider the alternative mechanism P^* which is identical to P unless i's report is either s_i or s_i'. For either of these actions by i, P^* specifies the *expected* allocation generated by P. In other words, if q is the probability of type s_i conditional on $\{s_i, s_i'\}$, then for every $a \in A$ and every $t_{-i} \in T_{-i}$, we set

$$P^*(a \mid s_i, t_{-i}) = P^*(a \mid s_i', t_{-i}) = qP(a \mid s_i, t_{-i}) + (1-q)P(a \mid s_i', t_{-i}).$$

It is easy to see that the incentives of agents $j \neq i$ are completely unaffected. The payoffs to these agents don't depend on i's type directly — they are only affected by i's type through its effect on the outcome chosen by the principal. Since this change in the mechanism preserves the probability distribution over outcomes from the point of view of these agents, their incentives are unaffected.

Also, the incentives of agent i are not affected. The payoff to i from reporting anything other than s_i or s_i' are not changed. The expected payoff to i from reporting s_i was $\hat{p}_i(s_i)$ in the original mechanism, while the expected payoff from reporting s_i' was $\hat{p}_i(s_i')$. In the new mechanism, we have "averaged" these two types together, so that in the new mechanism, the probability i receives the good if she reports s_i is now $q\hat{p}_i(s_i)+(1-q)\hat{p}_i(s_i')$. But since $\hat{p}_i(s_i) = \hat{p}_i(s_i')$, this means that the probability that i receives the good if she reports s_i does not change and similarly for s_i'. Hence the expected payoff to i from every action is the same under P and P^*, so P^* must be incentive compatible.

To see that this change does not affect the principal's payoff, recall that the principal's utility function is

$$v(a,t) = \sum_j u_j(a)v_j(t_j).$$

For the same reason that agents $j \neq i$ are unaffected by the proposed change in the mechanism, the expectation of $\sum_{j\neq i} u_j(a)v_j(t_j)$ is unaffected as well. Hence if the principal's utility changes, it is because the expectation of $u_i(a)v_i(t_i)$ changes. Note, though, that under the original mechanism, this expectation is

$$\begin{aligned} \mathrm{E}_t \sum_a P(a \mid t)u_i(a)v_i(t_i) &= \mathrm{E}_{t_i}\left[\mathrm{E}_{t_{-i}} \sum_a P(a \mid t)u_i(a)\right] v_i(t_i) \\ &= \mathrm{E}_{t_i}\hat{p}_i(t_i)v_i(t_i). \end{aligned}$$

As noted above, the expected probability of receiving the good from any report by i is unchanged in the new mechanism, so this expectation is unchanged as well. Hence the principal receives the same payoff in both mechanisms. So without loss of generality, we can change from P to P^*. Repeating as needed, we construct an alternative optimal mechanism which is measurable with respect to $\mathcal{T}$.

To see why this property is critical, fix any event $\hat{T}$ in $\mathcal{T}$ and suppose that the principal learns *only* that the type profile t is contained in $\hat{T}$. Since the mechanism is measurable

with respect to $\mathcal{T}$, it specifies the same response for every $t \in \hat{T}$. Suppose, though, that this response is not sequentially rational for the principal in the sense that learning that $t \in \hat{T}$ would lead him to strictly prefer some different response from what the mechanism specifies so that the commitment to the mechanism is crucial. If this is true, we can consider the following alternative mechanism. If $t \notin \hat{T}$, the new mechanism is the same as the original one. If $t \in \hat{T}$, then the new mechanism chooses the same as the old one with probability $1 - \varepsilon$ and chooses the strictly better response for the principal with probability ε for some very small $\varepsilon > 0$. It is easy to see that this alternative mechanism must give the principal a strictly higher expected payoff than the original mechanism. Hence if the new mechanism is incentive compatible, we have a contradiction to the optimality of the original mechanism.

To see that the new mechanism must be incentive compatible, fix any t_i and any $s_i \neq t_i$ with $M_i(s_i) \in \mathcal{E}_i(t_i)$. Since the original mechanism is incentive compatible, we know that $\hat{p}_i(t_i) \geq \hat{p}_i(s_i)$. Suppose in the original mechanism, we have $\hat{p}_i(t_i) = \hat{p}_i(s_i)$. Since the new mechanism is measurable with respect to $\mathcal{T}$, we must have this same equality in the new mechanism. Hence in this case, t_i has no incentive to claim to be s_i. So suppose in the original mechanism, we had $\hat{p}_i(t_i) > \hat{p}_i(s_i)$. Then if we choose ε sufficiently small, the interim probabilities must still satisfy this condition. Hence the new mechanism must be incentive compatible, a contradiction.

This result is the key to Theorem 1. While it strongly suggests that the principal does not need commitment, it does not establish this entirely. The argument above only says that if the principal learns *only* the event of the mechanism partition containing t, then he wants to follow the mechanism. But the equilibrium strategies of the game will determine what information the principal receives. Intuitively, if we can construct an equilibrium where the only information the principal receives from the agents is the event of the mechanism partition containing t, then the result above shows that the principal will find it optimal to follow the mechanism, at least on the equilibrium path. We explain the equilibrium construction further below.

The result above also tells us a great deal about the structure of the optimal mechanism. In particular, for any event $\hat{T}_i$ in the partition of T_i, let $\bar{v}_i(\hat{T}_i) = \mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i]$. Given any event $\hat{T} = \prod_j \hat{T}_j$, the optimal mechanism must give the good to some agent $i \in \arg\max_j \bar{v}_j(\hat{T}_j)$. Ignoring ties to keep the discussion simple, we see that this pins down the optimal mechanism given the partition $\mathcal{T}$ and that the optimal mechanism is deterministic.

This also implies that the mechanism is robustly incentive compatible. To see this, simply note that incentive compatibility implies that every agent i must maximize $\bar{v}_i(\hat{T}_i)$ by her honest report. Hence whatever i thinks the other agents are doing, she could not do better by misreporting.

We conclude this proof sketch by explaining how we construct equilibrium strategies to obtain the same outcome as in the optimal mechanism. As noted above, the key is to construct an equilibrium of the game where the information revealed to the principal by the equilibrium strategies is effectively the same as the partition $\mathcal{T}$. More precisely, the key is that *at least* this much information must be revealed to the principal. If less information is revealed, then it's impossible for him to implement the outcome from the optimal mechanism. If more than this information is revealed, then he can't gain by using the extra information. If he would be better off using the extra information, then we could use this equilibrium to construct a superior mechanism, contradicting the hypothesis that we have the optimal mechanism.

We construct the equilibrium in two steps. First, we consider equilibria in what we call the *restricted artificial game for* i in which type t_i of agent i is restricted to messages in the same event of the partition as t_i and where agent i simply wants to make the principal believe that $v_i(t_i)$ is as high as possible. In other words, we can let the principal's action in this game be the choice of a number x where his payoff is $-(x - v_i(t_i))^2$ and where agent i's utility is x. Thus x is, in effect, the principal's estimate of $v_i(t_i)$. In this game, the principal *must* learn at least that $t_i \in T_i^\alpha$ since the agent cannot claim types outside this set. Second, we show that by only adjusting off equilibrium path beliefs, we can turn this into an equilibrium of the *unrestricted artificial game for* i, where t_i can claim any type for which she has the needed evidence. Intuitively, this works because the original mechanism is incentive compatible. Thus no agent can "trick" the principal into too high a belief about v_i relative to what the mechanism is based on. Finally, we show that these strategies by the agents together with an appropriate response by the principal form an equilibrium in the original game. Since the principal responds to higher beliefs about v_i in ways that are preferred by agent i, agent i's incentives in both the original game and the artificial game are simply to induce as high an expectation of v_i as possible.

A natural question to ask at this point is why similar results do not hold in "typical" mechanism design problems in the literature. Of course, in a mechanism design problem that does not involve allocation of a good, there is no obvious analog of the probability that i receives the good. In the more general proof (and as the sketch above may suggest), the key is whether indifference can be equated with equality of outcome. To be specific, fix an optimal mechanism and suppose there is a type s_i who is indifferent between truthful revelation and reporting some other type, say s_i'. In our setting, this implies that s_i' is also indifferent between truthful revelation and reporting s_i. Note that this symmetry holds even when we allow type dependent preferences as long as the type dependence is simple in the sense we have defined. With simple type dependence, indifference by one type implies indifference by all types.

Of course, one reason these types may be indifferent between telling the truth versus claiming to be the other type could be that the two types are treated exactly the same

way — that is, that they receive exactly the same outcome. The key property here is that if they are not treated exactly the same way, we can find another optimal mechanism in which they are. In this sense, indifference can be equated with equality of outcomes. Once we have this property, we immediately obtain the sequential rationality property above and the rest of the argument follows.

Note, though, that indifference matching equality of outcomes is *not* a property that holds in typical mechanism design problems. For example, consider a standard adverse selection problem where there is a "high" type who is more productive than a "low" type and the principal must choose effort and wages for the two types. The standard result is that the low type is indifferent between her allocation and that of the high type. But, of course, there is no optimal mechanism where this indifference is turned into equality of outcome — that is, where the high and low type have the same effort and wage.

Appendix

A Proof of Theorem 1

Throughout the Appendix, we assume that each u_i satisfies simple type dependence and consequently write the agents' utility functions as

$$u_i(a_i, t_i) = \begin{cases} u_i(a), & \text{if } t_i \in T_i^+; \\ -u_i(a), & \text{if } t_i \in T_i^-, \end{cases}$$

where $T_i^+ \cup T_i^- = T_i$. We write the principal's utility function as

$$v(a, t) = u_0(a) + \sum_i u_i(a) v_i(t_i).$$

To simplify notation, we define $T_0 = \{t_0\}$ and $v_0(t_0) = 1$. With this notation, we can write the principal's utility function as $\sum_{i=0}^{I} u_i(a) v_i(t_i)$.

For each i, let $R_i \equiv T_i \times \mathcal{E}_i$. Given a mechanism P and $r_i \in R_i$, let

$$\hat{u}_i(r_i \mid P) = \mathrm{E}_{t_{-i}} \sum_a P(a \mid r_i, t_{-i}, M_{-i}(t_{-i})) u_i(a).$$

Recall that the Revelation Principle for this class of problems says that we can restrict attention to equilibria where t_i honestly states her type and provides maximal evidence. Hence $\hat{u}_i(r_i \mid P)$ will be the expected utility of t_i from report r_i in the mechanism.

Fix an optimal mechanism P. For each $\alpha \in \mathbf{R}$, let

$$R_i^\alpha = \{r_i \in R_i \mid \hat{u}_i(r_i \mid P) = \alpha\}.$$

Finiteness of T_i implies that $\mathcal{E}_i$ is finite. Given this, "most" R_i^α will be empty. When we refer to one of these sets below, we often take as given that it is nonempty. Note that the nonempty R_i^α's form a partition of R_i. In what follows, we refer to this partition as the *mechanism partition for* i, denoted $\{R_i^\alpha\}$ and refer to the product partition of R formed by the cells $\prod_i R_i^{\alpha_i}$ simply as the *mechanism partition*, denoted $\{\prod_i R_i^{\alpha_i}\}$. It will also be useful to define

$$T_i^\alpha = \{t_i \in T_i \mid \hat{u}_i(t_i, M_i(t_i) \mid P) = \alpha\} = \{t_i \in T_i \mid (t_i, M_i(t_i)) \in R_i^\alpha\}.$$

Note that we could have some values of α with $\hat{u}_i(t_i, e_i \mid P) = \alpha$ only for $e_i \neq M_i(t_i)$ in which case $R_i^\alpha \neq \emptyset$ but $T_i^\alpha = \emptyset$.

Lemma 1. *Fix $(s_i, e_i) \in R_i^\alpha$. For any $t_i \in T_i^+$, if $(t_i, M_i(t_i)) \in R_i^\beta$ with $\alpha > \beta$, then we have $e_i \notin \mathcal{E}_i(t_i)$. For any $t_i \in T_i^-$, if $(t_i, M_i(t_i)) \in R_i^\beta$ with $\alpha < \beta$, we have $e_i \notin \mathcal{E}_i(t_i)$.*

Proof. Follows from incentive compatibility. ▌

Lemma 2. *Without loss of generality, we can assume the optimal mechanism P has the property that for all i and all α, if $R_i^\alpha \neq \emptyset$, then $T_i^\alpha \neq \emptyset$.*

Proof. Suppose the optimal mechanism does not have this property. Fix any $R_i^\alpha \neq \emptyset$ such that $T_i^\alpha = \emptyset$. By the Revelation Principle, the $r_i \in R_i^\alpha$ are not used in equilibrium since $T_i^\alpha = \emptyset$ implies they are all of the form (t_i, e_i) where $e_i \neq M_i(t_i)$. Intuitively, then, we can change the outcome from these off equilibrium reports so that they remain off equilibrium without changing the principal's payoff.

More specifically, choose any β such that $T_i^\beta \neq \emptyset$ and there does not exist $\gamma \in (\alpha, \beta) \cup (\beta, \alpha)$ with $T_i^\gamma \neq \emptyset$. It is easy to see that such a β must exist. First, there must be some β with $T_i^\beta \neq \emptyset$ since the nonempty T_i^β sets partition T_i. So we can simply choose the smallest $\beta > \alpha$ such that $T_i^\beta \neq \emptyset$ if such a β exists and the largest $\beta < \alpha$ with $T_i^\beta \neq \emptyset$ otherwise.

Fix any $(\hat{t}_i, \hat{e}_i) \in T_i^\beta$ and consider the mechanism P^* given by

$$P^*(\cdot \mid t, e) = \begin{cases} P(\cdot \mid t, e), & \text{if } (t_i, e_i) \notin R_i^\alpha; \\ P(\cdot \mid \hat{t}_i, \hat{e}_i, t_{-i}, e_{-i}), & \text{otherwise.} \end{cases}$$

Note that we have only changed the mechanism for reports by i which are in R_i^α and hence are *not* of the form $(t_i, M_i(t_i))$. Hence the incentive compatibility of P for $j \neq i$ implies incentive compatibility of P^* for $j \neq i$. Similarly, the principal's payoff from P^* is the same as his payoff from P.

To see that P^* is incentive compatible for i, fix any t_i and any (s_i, e_i) such that $e_i \in \mathcal{E}_i(t_i)$. Clearly, $\hat{u}_i(t_i, M_i(t_i) \mid P^*) = \hat{u}_i(t_i, M_i(t_i) \mid P)$. If $(s_i, e_i) \notin R_i^\alpha$, then $\hat{u}_i(s_i, e_i \mid P^*) = \hat{u}_i(s_i, e_i \mid P)$, so the fact that t_i prefers reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i) in P implies the same is true for P^*.

So suppose $(s_i, e_i) \in R_i^\alpha$. In this case, $\hat{u}_i(s_i, e_i \mid P^*) = \hat{u}_i(\hat{t}_i, \hat{e}_i \mid P) = \beta$, while $\hat{u}_i(s_i, e_i \mid P) = \alpha$. For concreteness, suppose $\beta > \alpha$ (the case where $\beta < \alpha$ is analogous). From the way we chose β, we cannot have $\alpha \leq \hat{u}_i(t_i, M_i(t_i) \mid P) < \beta$. So either

$$\hat{u}_i(t_i, M_i(t_i) \mid P) < \alpha = \hat{u}_i(s_i, e_i \mid P) < \beta = \hat{u}_i(s_i, e_i \mid P^*)$$

or

$$\hat{u}_i(s_i, e_i \mid P) = \alpha < \beta = \hat{u}_i(s_i, e_i \mid P^*) \leq \hat{u}_i(t_i, M_i(t_i) \mid P).$$

Recall that $e_i \in \mathcal{E}_i(t_i)$ by assumption. Hence in the former case, incentive compatibility implies $t_i \in T_i^-$ and therefore t_i prefers reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i) in P^*. In the latter case, incentive compatibility implies $t_i \in T_i^+$ and therefore t_i (weakly) prefers

reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i) in P^*. Either way, P^* is incentive compatible and is also an optimal mechanism. By repeating this argument, we construct an optimal mechanism with the desired property. ▌

Lemma 3. *Without loss of generality, we can take the mechanism P to be measurable with respect to the mechanism partition for each i, $\{R_i^\alpha\}$, in the sense that if $(s_i, e_i), (t_i, e_i')$, then $P(\cdot \mid s_i, e_i, t_{-i}, e_{-i}) = P(\cdot \mid t_i, e_i', t_{-i}, e_{-i})$ for all $(t_{-i}, e_{-i}) \in R_{-i}$. Hence we can take P to be measurable with respect to the mechanism partition $\{\prod_i R_i^{\alpha_i}\}$ in the sense that $P(\cdot \mid s, e) = P(\cdot \mid s', e')$ if $(s, e), (s', e') \in \prod_i R_i^{\alpha_i}$.*

Proof. Fix an optimal mechanism P which is not measurable in this sense. We construct an alternative mechanism which is measurable, is incentive compatible, and has the same payoff for the principal as P. Fix any i and any α such that $R_i^\alpha \neq \emptyset$. By Lemma 2, $T_i^\alpha \neq \emptyset$.

Define a mechanism P^* by $P^*(\cdot \mid s_i, e_i, t_{-i}, e_{-i}) = P(\cdot \mid s_i, e_i, t_{-i}, e_{-i})$ if $(s_i, e_i) \notin R_i^\alpha$ and otherwise

$$P^*(a \mid s_i, e_i, t_{-i}, e_{-i}) = \mathrm{E}_{t_i}(P(a \mid t_i, M_i(t_i), t_{-i}, e_{-i}) \mid (t_i, M_i(t_i)) \in R_i^\alpha),$$

for all $a \in A$ and all $(t_{-i}, e_{-i}) \in R_{-i}$.

For any agent $j \neq i$, the expected payoff under the mechanism, both from honest reporting with maximal evidence and from any deviation, is unaffected. Hence we have incentive compatibility of P^* for all $j \neq i$ from incentive compatibility of P.

For agent i for $(s_i, e_i) \in R_i^\alpha$, we have

$$\begin{aligned}
\hat{u}_i(s_i, e_i \mid P^*) &= \mathrm{E}_{t_{-i}}\left[\sum_a P^*(a \mid s_i, e_i, t_{-i}, M_{-i}(t_{-i}))u_i(a)\right] \\
&= \mathrm{E}_{t_{-i}}\left[\sum_a \mathrm{E}_{t_i}(P(a \mid t_i, M_i(t_i), t_{-i}, M_{-i}(t_{-i}))) \mid (t_i, M_i(t_i)) \in R_i^\alpha)u_i(a)\right] \\
&= \mathrm{E}_{t_i}\left[\mathrm{E}_{t_{-i}}\sum_a (P(a \mid t_i, M_i(t_i), t_{-i}, M_{-i}(t_{-i}))u_i(a) \mid (t_i, M_i(t_i)) \in R_i^\alpha)\right] \\
&= \mathrm{E}_{t_i}\left[\alpha \mid (t_i, M_i(t_i)) \in R_i^\alpha\right] \\
&= \alpha = \hat{u}_i(s_i, e_i \mid P).
\end{aligned}$$

So every type of agent i receives the same expected payoff under the new mechanism for every report as she did in the original mechanism. Hence the incentive compatibility of P implies incentive compatibility of P^*. It is easy to see that P^* gives the principal the same expected payoff as P. Hence P^* is an optimal mechanism as well. Iterating this construction, we construct an optimal mechanism which is measurable with respect to the mechanism partition.

To see that this implies measurability with respect to the mechanism partition, apply the argument above iteratively over i. ▌

Added the "Part B" as you guys suggested.

In light of Lemma 3, we henceforth assume P is measurable with respect to the mechanism partition.

Given any partition $\mathcal{R}$ of $T \times \mathcal{E}$, we say that a mechanism $\tilde{P}$ (not necessarily the optimal mechanism) is *sequentially rational given* $\mathcal{R}$ if the following is true. First, $\tilde{P}$ is measurable with respect to $\mathcal{R}$. Second, for every event E of the partition, for every $(t, M(t)) \in E$, $\tilde{P}(\cdot \mid t, M(t))$ is some $p \in \Delta(A)$ which maximizes

$$\sum_a p(a) \mathrm{E}_t \left[v(a,t) \mid (t, M(t)) \in E \right].$$

In other words, the mechanism is optimal for the principal given the information contained the partition $\mathcal{R}$.

Lemma 4. *P is sequentially rational given the mechanism partition.*

Proof. Suppose not. Fix any $(t, M(t))$ and let $P(\cdot \mid t, M(t)) = \bar{p}$. Let $\hat{R} = \prod_j R_j^{\alpha_j}$ denote the event of the mechanism partition containing $(t, M(t))$ and suppose

$$\sum_a \tilde{p}(a) \mathrm{E}[v(a,t) \mid (t, M(t)) \in \hat{R}] > \sum_a \bar{p}(a) \mathrm{E}[v(a,t) \mid (t, M(t)) \in \hat{R}].$$

We construct a new mechanism P^* as follows. For any $(t, e) \notin \hat{R}$, $P^*(\cdot \mid t, e) = P(\cdot \mid t, e)$. For $(t, e) \in \hat{R}$,

$$P^*(\cdot \mid t, e) = (1 - \varepsilon)\bar{p} + \varepsilon \tilde{p}$$

for some small $\varepsilon > 0$. Clearly, for any $\varepsilon \in (0, 1)$, P^* yields a strictly higher payoff for the principal than P.

We now show that that for ε sufficiently small, P^* is incentive compatible. To see this, fix any $(t_i, M_i(t_i))$ and any (s_i, e_i) with $e_i \in \mathcal{E}_i(t_i)$. Suppose that under P, t_i strictly preferred reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i). Then for ε sufficiently small, this must still be true. So suppose t_i was indifferent between $(t_i, M_i(t_i))$ and (s_i, e_i). That is, $\hat{u}_i(t_i, M_i(t_i) \mid P) = \hat{u}_i(s_i, e_i \mid P)$. But then by measurability of the new mechanism P^* with respect to the mechanism partition of the original mechanism P, t_i must still be indifferent between these reports in the new mechanism, so it remains incentive compatible. This contradicts the optimality of P. ▌

In what follows, for any α such that $T_i^\alpha \neq \emptyset$, let

$$\bar{v}_i(\alpha) = \mathrm{E}[v_i(t_i) \mid (t_i, M_i(t_i)) \in R_i^\alpha].$$

The following lemma will be useful.

Lemma 5. *Let*

$$\mathcal{U} = \{(\bar{u}_0, \bar{u}_1, \ldots, \bar{u}_I) \in \mathbf{R}^{I+1} \mid \exists p \in \Delta(A) \text{ with } \sum_a p(a)u_i(a) = \bar{u}_i, \ \forall i\}.$$

Given any belief of the principal over each t_i, let $\hat{v}_i$ denote the expectation of $v_i(t_i)$ under the belief over t_i and let $\hat{v} = (1, \hat{v}_1, \ldots, \hat{v}_I)$. Let $\mathcal{U}^*(\hat{v})$ denote the set of $u \in \mathcal{U}$ maximizing $\hat{v} \cdot u$. Fix any i, v, and v' such that $v_i > v'_i$ and $v'_j = v_j$ for $j \neq i$. Fix any $u \in \mathcal{U}^*(v)$ and any $u' \in \mathcal{U}^*(v')$. Then $u_i \geq u'_i$.

Proof. This result is standard, but we include a proof for completeness. Clearly, we must have

$$v \cdot u \geq v \cdot u'$$

$$v' \cdot u' \geq v' \cdot u$$

implying

$$(v - v') \cdot (u - u') \geq 0.$$

But this is $(v_i - v'_i)(u_i - u'_i)$. Since $v_i > v'_i$, we must have $u_i \geq u'_i$. ∎

Lemma 6. *For all $\alpha > \beta$, we have $\bar{v}_i(\alpha) \geq \bar{v}_i(\beta)$. In other words, "more valuable" sets for the principal receive higher utilities.*

Proof. Fix $\alpha > \beta$. Since $\alpha > \beta$, there must exist events $T_j^{\alpha_j}$ for $j \neq i$ such that

$$\bar{u}_i(\alpha) \equiv \sum_a P(a \mid t, M(t))u_i(a) > \sum_a P(a \mid t'_i, M_i(t'_i), t_{-i}, M_{-i}(t_{-i}))u_i(a) \equiv \bar{u}_i(\beta),$$

where t_i is an arbitrary element of T_i^α, t'_i is an arbitrary element of T_i^β, and $t_j \in T_j^{\alpha_j}$ for each $j \neq i$. Let $\hat{T}_{-i} = \prod_{j \neq i} T_j^{\alpha_j}$. Let $\bar{v}_j = \bar{v}_j(\alpha_j)$ for $j \neq i$, let $\bar{v}^\alpha$ denote the vector $(\bar{v}_i(\alpha), \bar{v}_{-i})$, and define $\bar{v}^\beta$ analogously. For each $j \neq i$, let

$$\bar{u}_j(\alpha) = \sum_a P(a \mid t, M(t))u_j(a)$$

for any $t \in T_i^\alpha \times \hat{T}_{-i}$ and define $\bar{u}_j(\beta)$ analogously using any $t \in T_i^\beta \times \hat{T}_{-i}$. Finally, let $\bar{u}^\alpha$ denote the vector $(\bar{u}_i(\alpha), \bar{u}_{-i}(\alpha))$ and define $\bar{u}^\beta$ analogously. From Lemma 4, we know that $\bar{u}^\alpha$ maximizes $\bar{v}^\alpha \cdot u$ over u that can be generated by some $p \in \Delta(A)$. Similarly, $\bar{u}^\beta$ maximizes $\bar{v}^\beta \cdot u$. Since $\bar{u}_i^\alpha > \bar{u}_i^\beta$, Lemma 5 implies that $\bar{v}_i(\alpha) \geq \bar{v}_i(\beta)$. ∎

Lemma 7. *Without loss of generality, we can take the mechanism P to be measurable with respect to $\bar{v}_i$ in the sense that if $\bar{v}_i(\alpha) = \bar{v}_i(\beta)$ and $(s_i, e_i) \in R_i^\alpha$, $(t_i, e'_i) \in R_i^\beta$, then $P(\cdot \mid s_i, e_i, t_{-i}, e_{-i}) = P(\cdot \mid t_i, e'_i, t_{-i}, e_{-i})$ for all $(t_{-i}, e_{-i}) \in R_{-i}$. In other words, we can take the mechanism to have the property that $\alpha \neq \beta$ implies $\bar{v}_i(\alpha) \neq \bar{v}_i(\beta)$.*

Proof. Fix an optimal mechanism p which does not satisfy this property. Fix the relevant α and let $\mathcal{A} = \{\beta \mid R_i^\beta \neq \emptyset \text{ and } \bar{v}_i(\beta) = \bar{v}_i(\alpha)\}$. By assumption, there exists at least one $\beta \neq \alpha$ with $\beta \in \mathcal{A}$.

By Lemma 6, we have that $\alpha' > \beta'$ implies $\bar{v}_i(\alpha') \geq \bar{v}_i(\beta')$. Hence for any $\gamma \notin \mathcal{A}$, either γ is strictly smaller than every $\beta \in \mathcal{A}$ or γ is strictly larger than every $\beta \in \mathcal{A}$.

Define a new mechanism P^* by setting $P^*(\cdot \mid s_i, e_i, t_{-i}, e_{-i}) = P(\cdot \mid s_i, e_i, t_{-i}, e_{-i})$ if $(s_i, e_i) \notin \cup_{\beta \in \mathcal{A}} R_i^\beta$ and otherwise,

$$P^*(a \mid s_i, e_i, t_{-i}, e_{-i}) = \mathrm{E}_{t_i}[P(a \mid t_i, M_i(t_i), t_{-i}, e_{-i}) \mid (t_i, M_i(t_i)) \in \cup_{\beta \in \mathcal{A}} R_i^\beta],$$

for all $a \in A$ and all $(t_{-i}, e_{-i}) \in R_{-i}$. We now show that P^* is incentive compatible and gives the principal the same payoff as P, establishing the claim.

To see that P^* is incentive compatible, note that the interim payoff to t_j for any feasible (s_j, e_j) for $j \neq i$ is unaffected by this change. Hence we have incentive compatibility for any $j \neq i$.

So fix any t_i and any $(s_i, e_i) \neq (t_i, M_i(t_i))$ with $e_i \in \mathcal{E}_i(t_i)$. If neither $(t_i, M_i(t_i))$ nor (s_i, e_i) is contained in $\cup_{\beta \in \mathcal{A}} R_i^\beta$, then the response to either report is unaffected, so incentive compatibility of P implies that t_i prefers reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i). If both are contained in $\cup_{\beta \in \mathcal{A}} R_i^\beta$, then the expected payoff under P^* is the same in response to either report, so this incentive compatibility constraint holds.

So suppose $(t_i, M_i(t_i)) \in \cup_{\beta \in \mathcal{A}} R_i^\beta$ and (s_i, e_i) is not. Then $(s_i, e_i) \in R_i^\gamma$ for some γ that is either below every $\beta \in \mathcal{A}$ or above every $\beta \in \mathcal{A}$. If γ is below every $\beta \in \mathcal{A}$, then $\hat{u}_i(t_i, M_i(t_i) \mid P^*) > \hat{u}_i(s, e_i \mid P^*)$ and $\hat{u}_i(t, M_i(t_i) \mid P) > \hat{u}_i(s_i, e_i \mid P)$. The latter inequality and the incentive compatibility of P implies $t_i \in T_i^+$, so that the former inequality implies t_i prefers reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i). Similarly, If γ is above every $\beta \in \mathcal{A}$, then both inequalities are strictly reversed, implying $t_i \in T_i^-$ and that t_i prefers reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i). A similar argument holds for the case where $(s_i, e_i) \in \cup_{\beta \in \mathcal{A}} R_i^\beta$ and $(t_i, M_i(t_i))$ is not. Hence P^* is incentive compatible.

To see that P^* yields the same expected payoff to the principal as P, recall that by the Revelation Principle, the specification of P^* on messages other than those of the form $(t, M(t))$ are irrelevant to the principal's payoffs. Fix any t_{-i} and let $\bar{v}_j = \bar{v}_j(R_j^{\alpha_j})$ for the α_j with $(t_j, M_j(t_j)) \in R_j^{\alpha_j}$. For any $(t_i, M_i(t_i)) \in R_i^\beta$, $\beta \in \mathcal{A}$, let $p^\beta = P(\cdot \mid t_i, e_i, t_{-i}, M_{-i}(t_{-i}))$. By Lemma 4, P is sequentially rational for the principal. Hence for every $\beta \in \mathcal{A}$, p^β maximizes

$$\sum_a p^\beta(a)[u_i(a)\bar{v}_i(\beta) + \sum_{j \neq i} u_j(a)\bar{v}_j].$$

Since $\bar{v}_i(\alpha) = \bar{v}_i(\beta)$ for all $\beta \in \mathcal{A}$, the only way we can have $p^\alpha \neq p^\beta$ is if the principal is

indifferent between p^α and p^β. Obviously, then, the fact that P^* differs from P in such situations has no payoff consequences. Hence P^* yields the principal the same payoff as P. ∎

In light of this result, we henceforth assume P is measurable with respect to $\bar{v}$ in the sense defined above.

Lemma 8. *P is robustly incentive compatible.*

Proof. Suppose not. Then either there exists $t_i \in T_i^+$, $e_i \in \mathcal{E}_i(t_i)$, $s_i \in T_i$, $\bar{t}_{-i} \in T_{-i}$, and $\bar{e}_{-i} \in \mathcal{E}_{-i}$ such that

$$\sum_a P(a \mid t_i, M_i(t_i), \bar{t}_{-i}, \bar{e}_{-i}) u_i(a) < \sum_a P(a \mid s_i, e_i, \bar{t}_{-i}, \bar{e}_{-i}) u_i(a) \tag{2}$$

or some $t_i \in T_i^-$, $e_i \in \mathcal{E}_i(t_i)$, $s_i \in T_i$, $\bar{t}_{-i} \in T_{-i}$, and $\bar{e}_{-i} \in \mathcal{E}_{-i}$ with the opposite strict inequality. Since these cases are entirely symmetric, we consider only the former, so fix t_i, s_i, e_i, $\bar{t}_{-i}$, and $\bar{e}_{-i}$ satisfying equation (2). Assume $(s_i, e_i) \in R_i^\alpha$ and $(t_i, M_i(t_i)) \in R_i^\beta$. By measurability with respect to the mechanism partition for i, we know that $\alpha \neq \beta$. By Lemma 7, $\bar{v}_i(\alpha) \neq \bar{v}_i(\beta)$. By incentive compatibility and the fact that $t_i \in T_i^+$, we have $\alpha < \beta$ and hence by Lemma 6, $\bar{v}_i(\alpha) < \bar{v}_i(\beta)$.

As in the proof of Lemma 6, for each j, including $j = i$, let

$$\bar{u}_j^\alpha = \sum_a P(a \mid t_i', M_i(t_i'), t_{-i}', M_{-i}(t_{-i}')) u_j(a)$$

for any $(t_i', M_i(t_i')) \in R_i^\alpha$ and any $(t_{-i}', M_{-i}(t_{-i}')) \in \prod_{j \neq i} R_j^{\alpha_j}$. (By Lemma 2, such t_i' and t_{-i}' must exist.) Similarly, define $\bar{u}_j^\beta$ using some t_i' with $(t_i', M_i(t_i')) \in R_i^\beta$. Finally, let $\bar{v}^\alpha$ denote the vector $(\bar{v}_i(\alpha), \bar{v}_{-i})$ and define $\bar{v}^\beta$ analogously. By Lemma 5, $\bar{v}_i(\beta) > \bar{v}_i(\alpha)$ implies $\bar{u}_i^\beta \geq \bar{u}_i^\alpha$. But $\bar{u}_i^\beta$ is the left–hand side of equation (2) and $\bar{u}_i^\alpha$ is the right–hand side, a contradiction. ∎

Lemma 9. *There exists an optimal mechanism P which is robustly incentive compatible and is deterministic in the sense that $P(a \mid t, M(t)) \in \{0, 1\}$ for all $a \in A$ and $t \in T$.*

Proof. Given an arbitrary mechanism $\tilde{P}$, let $\Pi_i(\tilde{P})$ denote the mechanism partition of T_i induced by $\tilde{P}$. Given $r_i \in R_i$, let $\pi_i(r_i \mid \tilde{P})$ denote the event of $\Pi_i(\tilde{P})$ containing r_i. By Lemmas 3 and 7, we know that there exists an optimal mechanism P which is strictly measurable with respect to $\Pi_i(P)$ for all i in the sense that if $\mathrm{E}_{t_i}[v_i(t_i) \mid (t_i, M_i(t_i)) \in \pi_i(r_i \mid P)] = \mathrm{E}_{t_i}[v_i(t_i) \mid (t_i, M_i(t_i)) \in \pi_i(r_i' \mid P)$, then

$$P(\cdot \mid r_i, r_{-i}) = P(\cdot \mid r_i', r_{-i}), \ \forall r_{-i}.$$

Let $\mathcal{P}$ denote the set of optimal mechanisms $\tilde{P}$ such that $\tilde{P}$ is strictly measurable with respect to each $\Pi_i(\tilde{P})$ in this sense. Finally, let P denote any mechanism in $\mathcal{P}$ which is minimal in the sense that there is no $P' \in \mathcal{P}$ which is strictly measurable with respect to each $\Pi_i(P')$ and for which $\Pi_i(P')$ has weakly fewer elements than $\Pi_i(P)$ for all i, strictly fewer for some i. By finiteness of T, such a P must exist.

If P is deterministic, we are done, so suppose it is not. In light of Lemma 4, this can only occur when the principal is indifferent ex post. In other words, if $P(a^* \mid r) > 0$ for some $a^* \in A$ and some $r \in R$, then

$$\sum_a P(a \mid r)\mathrm{E}_t[v(a,t) \mid (t, M(t)) \in \pi(r \mid P)] = \mathrm{E}_t[v(a^*,t) \mid (t, M(t)) \in \pi(r \mid P)].$$

Hence there must exist a deterministic mechanism, say P^*, which is strictly measurable with respect to each $\Pi_i(P)$ which yields the same expected payoff for the principal.

We now show that P^* is incentive compatible and strictly measurable with respect to each $\Pi_i(P^*)$. To show incentive compatibility, suppose P^* is not incentive compatible. Then either there exists $t_i \in T_i^+$, $s_i \in T_i$, and $e_i \in \mathcal{E}_i(t_i)$ such that

$$\hat{u}_i(t_i, M_i(t_i) \mid P^*) < \hat{u}_i(s_i, e_i \mid P^*)$$

or $t_i \in T_i^-$, $s_i \in T_i$, and $e_i \in \mathcal{E}_i(t_i)$ with the reverse strict inequality. Because P^* is measurable with respect to each $\Pi_i(P)$, this implies $\hat{u}_i(t_i, M_i(t_i) \mid P) \neq \hat{u}_i(s_i, e_i \mid P)$. Hence incentive compatibility of P implies

$$\hat{u}_i(t_i, M_i(t_i) \mid P) > \hat{u}_i(s_i, e_i \mid P)$$

if $t_i \in T_i^+$ and the reverse strict inequality if $t_i \in T_i^-$.

Consider the mechanism $P^\lambda \equiv \lambda P + (1-\lambda)P^*$. For every $\lambda \in [0,1]$, this mechanism has the same payoff for the principal as P. Clearly, for λ sufficiently close to 1, P^λ is incentive compatible. Let λ^* be the smallest λ such that P^λ is incentive compatible. It is easy to see that such a λ^* exists and that it satisfies

$$\lambda^*\hat{u}_i(t_i, M_i(t_i) \mid P) + (1-\lambda^*)\hat{u}_i(t_i, M_i(t_i) \mid P^*) = \lambda^*\hat{u}_i(s_i, e_i \mid P) + (1-\lambda^*)\hat{u}_i(s_i, e_i \mid P^*)$$

for some $t_i, s_i \in T_i$ and $e_i \in \mathcal{E}_i(t_i)$ such that $\pi_i(t_i, M_i(t_i) \mid P) \neq \pi_i(s_i, e_i \mid P)$. Hence for every i, $\Pi_i(P^{\lambda^*})$ either equals or coarsens $\Pi_i(P)$, coarsening for some i. By the same reasoning as Lemmas 3 and 7, there exists another mechanism P^{**} with $\Pi_i(P^{\lambda^*}) = \Pi_i(P^{**})$ for every i which is strictly measurable with respect to each $\Pi_i(P^{**})$ and yields the principal the same expected payoff as P^{λ^*}. This contradicts the minimality of P. Hence P^* is incentive compatible.

To see that P^* is strictly measurable with respect to each $\Pi_i(P^*)$, suppose it is not. By construction, this means that each $\Pi_i(P^*)$ is either equal to or a coarsening of $\Pi_i(P)$

and is a coarsening for some i. Again following the same reasoning as Lemmas 3 and 7, there exists another mechanism P^{**} with $\Pi_i(P^*) = \Pi_i(P^{**})$ for every i which is strictly measurable with respect to each $\Pi_i(P^*)$ and yields the principal the same expected payoff as P^*. This again contradicts the minimality of P.

Finally, the same reasoning as in the proof of Lemma 8 shows that P^* is robustly incentive compatible. ∎

We now construct an equilibrium for the game which yields the same payoff for the principal as P. In particular, the strategy for agent i in this game is the same as i's strategy in an equilibrium of the artificial game for i. Recall that the artificial game for i is a two–player game between i and the principal. i has a set of types T_i where the prior over T_i is the same as in the mechanism design problem. If i is type t_i, then her set of feasible actions is $Z_i(t_i) \equiv T_i \times \mathcal{E}_i(t_i)$. The principal's set of feasible actions is $X = [\min_j \min_{t_j \in T_j} v_j(t_j), \max_j \max_{t_j \in T_j} v_j(t_j)]$. The game is sequential. First, agent i learns her type $t_i \in T_i$. Then she chooses an action $z_i \in Z_i(t_i)$. Next, the principal observes this action and chooses $x \in X$. If i's type is t_i and the principal chooses action x, then the principal's payoff is $-(x - v_i(t_i))^2$, while i's payoff is

$$\begin{cases} x, & \text{if } t_i \in T_i^+; \\ -x, & \text{otherwise.} \end{cases}$$

Denote a strategy for i in this game by $\sigma_i(\cdot \mid t_i)$, a function from T_i to $\Delta(Z_i(t_i))$. Let the principal's belief be denoted $q_i(\cdot \mid s_i, e_i)$ where this is a function from $R_i = T_i \times \mathcal{E}_i$ to $\Delta(T_i)$. Finally, the principal's action in response to (s_i, e_i) is denoted $x_i : R_i \to X$.

We construct the relevant equilibrium of the artificial game for i by first considering what we will call the *restricted artificial game*. In the restricted game, type t_i cannot choose any action in R_i but can only choose actions in R_i^α where α is the unique α such that $t_i \in T_i^\alpha$.

Fix any i and any perfect Bayesian equilibrium $(\sigma_i^*, x_i^*, q_i^*)$ of the restricted artificial game for i.[4] Obviously, sequential rationality implies that $x_i^*(s_i, e_i)$ is the expectation of $v_i(t_i)$ given the belief q_i^* or

$$\sum_{t_i \in T_i} v_i(t_i) q_i^*(t_i \mid s_i, e_i).$$

[4]To see that such an equilibrium must exist, consider the game where i is restricted to putting probability $\varepsilon > 0$ on each of her pure strategies. By standard results, this game has a Nash equilibrium. As $\varepsilon \downarrow 0$ (taking subsequences as needed), these strategies converge to a Nash equilibrium of the restricted artificial game by upper hemicontinuity of the Nash equilibrium correspondence. These strategies and the limiting beliefs for the principal must also be a perfect Bayesian equilibrium since the principal's limiting strategy must be optimal given his limiting belief.

Let $X_i^*(t_i)$ denote the action chosen by the principal when i is type t_i. That is, $X_i^* : T_i \to X$ and is given by

$$X_i^*(t_i) = x_i^*(s_i, e_i), \text{ for some } (s_i, e_i) \in \text{supp}(\sigma_i^*(\cdot \mid t_i)).$$

Note that the principal's optimal action is always pure and that t_i is never indifferent between two distinct actions by the principal. Hence every message in the support of t_i's mixed strategy must lead to the same response by the principal. Thus the definition above is unambiguous. Clearly, for this to be an equilibrium, it must be true that if $t_i \in T_i^+$,

$$X_i^*(t_i) = \max_{(s_i,e_i) \in Z_i(t_i) \cap R_i^\alpha} x_i^*(s_i, e_i),$$

while for $t_i \in T_i^-$,

$$X_i^*(t_i) = \min_{(s_i,e_i) \in Z_i(t_i) \cap R_i^\alpha} x_i^*(s_i, e_i).$$

By construction, in any equilibrium of the restricted artificial game for i, the principal learns at least which event of the mechanism partition for i that t_i lies in. This is true because if $t_i \in T_i^\alpha$, then t_i can only send $(s_i, e_i) \in R_i^\alpha$. Hence observing (s_i, e_i) reveals the relevant value of α. Since the optimal mechanism is measurable with respect to the mechanism partition, this means that the principal must have enough information to carry out the optimal mechanism if this is the information the agents reveal to him. On the other hand, the principal may learn more than just that $t_i \in T_i^\alpha$ in the equilibrium. The following lemma shows that this extra information, if any, cannot be useful for the principal.

Lemma 10. *For each i, fix any equilibrium of the restricted artificial game for i and any α_i such that $T_i^{\alpha_i} \neq \emptyset$. Then for every $t \in \prod_i T_i^{\alpha_i}$,*

$$P(\cdot \mid t_i, M_i(t_i)) \in \arg\max_{p \in \Delta(A)} \sum_a p(a) \sum_i u_i(a) X_i^*(t_i).$$

In other words, given the belief formed by the principal in the equilibria at profile t, it is optimal for him to follow the optimal mechanism.

Proof. Suppose not. For any $\hat{v} = (\hat{v}_1, \ldots, \hat{v}_I)$, let $\tilde{p}(\cdot \mid \hat{v})$ denote any $p(\cdot) \in \Delta(A)$ which maximizes

$$\sum_a p(a) \sum_i u_i(a) \hat{v}_i.$$

Clearly, there exists $p(\cdot \mid t)$ with

$$\sum_a p(a \mid t) \sum_i u_i(a) X_i^*(t_i) > \sum_a P(a \mid t_i, M_i(t_i)) \sum_i u_i(a) X_i^*(t_i) \tag{3}$$

if and only if this holds for $p(\cdot \mid t) = \tilde{p}(\cdot \mid X^*(t))$ where $X^*(t) = (X_1^*(t_1), \ldots, X_I^*(t_I))$.

Given any $(s_i, e_i) \in R_i^{\alpha_i}$, let

$$\hat{v}_i(s_i, e_i) = \begin{cases} X_i^*(s_i), & \text{if } e_i = M_i(s_i); \\ x_i^*(s_i, e_i), & \text{otherwise.} \end{cases}$$

Given $(s, e) \in \prod_i R_i^{\alpha_i}$, let $\hat{v}(s, e) = (\hat{v}_1(s_1, e_1), \ldots, \hat{v}_I(s_I, e_I))$. Fix a small $\varepsilon > 0$ and define a new mechanism P^* by

$$P^*(\cdot \mid s, e) = \begin{cases} \varepsilon\tilde{p}(\cdot \mid \hat{v}(s, e)) + (1 - \varepsilon)P(\cdot \mid s, e), & \text{if } (s_i, e_i) \in R_i^{\alpha_i},\ \forall i; \\ P(\cdot \mid s, e), & \text{otherwise.} \end{cases}$$

In other words, for those types $t \in \prod_i T_i^{\alpha_i}$, we assign a convex combination of the $\tilde{p}$ that will be optimal for the principal given the belief they will induce in the restricted artificial games and the original mechanism, assuming they report honestly and provide maximal evidence. If they deviate from maximal evidence, we assign a convex combination of the $\tilde{p}$ optimal for the principal given the induced beliefs in the restricted artificial games given those deviations and the original mechanism. Finally, for all other type profiles, the mechanism is unchanged.

We now show that P^* is incentive compatible. So fix some $t_i \in T_i$ and (s_i, e_i) such that $e_i \in \mathcal{E}_i(t_i)$. If t_i strictly prefers reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i) under P, then for ε sufficiently small, t_i still has this strict preference. So suppose that t_i is indifferent between reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i) under P, so $(t_i, M_i(t_i))$ and (s_i, e_i) are in the same event of the mechanism partition for i. Clearly, if that event is not $R_i^{\alpha_i}$, then P^* still treats these reports the same way, so t_i is still indifferent.

So assume $(t_i, M_i(t_i)), (s_i, e_i) \in R_i^{\alpha}$. The only way t_i would not be indifferent is if $x_i^*(s_i, e_i) \neq X_i^*(t_i)$. If $t_i \in T_i^+$, we know that $X_i^*(t_i) \geq x_i^*(x_i, e_i)$. By Lemma 5, this implies

$$\begin{aligned} &\mathrm{E}_{t_{-i}}\left[\sum_a \tilde{p}(a \mid X_i^*(t_i), X_{-i}^*(t_{-i}))u_i(a) \mid t_{-i} \in \prod_{j \neq i} T_j^{\alpha_j}\right] \\ &\quad \geq \mathrm{E}_{t_{-i}}\left[\sum_a \tilde{p}(a \mid x_i^*(s_i, e_i), X_{-i}^*(t_{-i}))u_i(a) \mid t_{-i} \in \prod_{j \neq i} T_j^{\alpha_j}\right]. \end{aligned}$$

Since P is incentive compatible, this implies t_i prefers reporting maximal evidence to reporting (s_i, e_i) in P^*. A similar argument applies to $t_i \in T_i^-$. Hence P^* is incentive compatible.

But then we have a contradiction. By hypothesis, P is the optimal incentive compatible mechanism, so the fact that P^* is also incentive compatible implies that it cannot yield the principal a strictly higher payoff than P. ∎

Lemma 11. *Fix $\alpha > \beta$ such that $T_i^\alpha \neq \emptyset$ and $T_i^\beta \neq \emptyset$ and any equilibrium of the restricted artificial game for i. Then for every $t_i \in T_i^\alpha$ and $t_i' \in T_i^\beta$, we have $X_i^*(t_i) \geq X_i^*(t_i')$.*

Proof. Since $\alpha > \beta$, there exists $\hat{t}_{-i} \in T_{-i}$ such that

$$u_i^\alpha \equiv \sum_a P(a \mid t_i, M_i(t_i), \hat{t}_{-i}, M_{-i}(\hat{t}_{-i}))u_i(a) > \sum_a P(a \mid t_i', M_i(t_i'), \hat{t}_{-i}, M_{-i}(\hat{t}_{-i}))u_i(a) \equiv u_i^\beta.$$

For each $j \neq i$, let

$$u_j^\alpha = \sum_a P(a \mid t_i, M_i(t_i), \hat{t}_{-i}, M_{-i}(\hat{t}_{-i}))u_j(a),$$

and define u_j^β analogously. By Lemma 10, we know that $p^\alpha \equiv P(\cdot \mid t_i, M_i(t_i), \hat{t}_{-i}, M_{-i}(\hat{t}_{-i}))$ maximizes over $p(\cdot) \in \Delta(A)$

$$\sum_a p(a) \left[u_i(a)X_i^*(t_i) + \sum_{j \neq i} u_j(a)X_j^*(\hat{t}_j) \right]$$

and p^β defined analogously maximizes the analog for t_i'. Hence by Lemma 5, $u_i^\alpha > u_i^\beta$ implies $X_i^*(t_i) \geq X_i^*(t_i')$. ∎

We now show how we can modify an equilibrium of the restricted artificial game for i to construct an equilibrium of the unrestricted artificial game for i with the same equilibrium path. So fix any equilibrium of the restricted artificial game for i, say $(\sigma_i^*, x_i^*, q_i^*)$. We first show that in the unrestricted artificial game for i, agent i does not has a profitable deviation from these strategies to any (s_i, e_i) which has positive probability in equilibrium (i.e., with $\sigma_i^*(s_i, e_i \mid t_i) > 0$ for some t_i).

Fix any $t_i \in T_i^\alpha$. Since these strategies are an equilibrium of the restricted game, t_i does not have a profitable deviation to any $(s_i, e_i) \in R_i^\alpha$ with $e_i \in \mathcal{E}_i(t_i)$. So consider a deviation by t_i to some $(s_i, e_i) \in R_i^\beta$ for $\beta \neq \alpha$ such that (s_i, e_i) has positive probability under the equilibrium strategies. By Lemma 11, we know that a deviation to any $(s_i, e_i) \in R_i^\beta$, $\beta \neq \alpha$, which has positive probability in equilibrium must at least weakly increase the principal's belief if $\beta > \alpha$ and decrease it if $\beta < \alpha$. If $t_i \in T_i^+$, then Lemma 1 implies that for every $(s_i, e_i) \in R_i^\beta$ with $\beta > \alpha$, we have $e_i \notin \mathcal{E}_i(t_i)$. Hence t_i cannot deviate to an (s_i, e_i) which has positive probability in equilibrium and increases the principal's belief. Similarly, a negative type cannot feasibly deviate to any (s_i, e_i) which has positive probability in equilibrium and decreases the principal's belief.

So we only need to ensure that there is no profitable deviation to an (s_i, e_i) which has zero probability in the restricted game. Fix any such (s_i, e_i) and suppose $(s_i, e_i) \in R_i^\alpha$. Let $F_i = \{t_i \in T_i \mid e_i \in \mathcal{E}_i(t_i)\}$. Since we have an equilibrium of the restricted game, we know that

$$\min_{\bar{t}_i \in F_i \cap T_i^+ \cap T_i^\alpha} X_i^*(\bar{t}_i) \geq x_i^*(s_i, e_i) \geq \max_{\bar{t}_i \in F_i \cap T_i^- \cap T_i^\alpha} X_i^*(\bar{t}_i).$$

Let $\hat{T}_i^+$ denote the set of $t_i \in F_i \cap T_i^+$ with $t_i \in T_i^\beta$ for some $\beta \neq \alpha$. By Lemma 1, we must have $\beta > \alpha$ for each $t_i \in \hat{T}_i^+$. If $F_i \cap T_i^+ \cap T_i^\alpha \neq \emptyset$, then by Lemma 11, we have

$$X_i^*(t_i) \geq \min_{\bar{t}_i \in F_i \cap T_i^+ \cap T_i^\alpha} X_i^*(\bar{t}_i) \geq x_i^*(s_i, e_i), \quad \forall t_i \in \hat{T}_i^+,$$

so no positive type can gain by deviating to (s_i, e_i). So assume $F_i \cap T_i^+ \cap T_i^\alpha = \emptyset$.

Fix $\hat{t}_i \in \hat{T}_i^+$ such that $X_i^*(\hat{t}_i) \leq X_i^*(t_i)$ for all $t_i \in \hat{T}_i^+$. By the same reasoning as above, if $X_i^*(\hat{t}_i) \geq x_i^*(s_i, e_i)$, no positive type can gain by deviating to (s_i, e_i). So assume $x_i^*(s_i, e_i) > X_i^*(\hat{t}_i)$.

Suppose there is any $t_i' \in F$ such that $X_i^*(\hat{t}_i) \geq v_i(t_i')$. If so, redefine the belief in response to (s_i, e_i) by setting it equal to a convex combination of the equilibrium belief from the restricted game and a degenerate distribution on t_i' chosen to make the expectation of v_i equal to $X_i^*(\hat{t}_i)$. By construction, then, no positive type will be able to gain by deviating to (s_i, e_i). For any $t_i \in F_i \cap T_i^-$ with $t_i \in T_i^\gamma$, $\gamma \neq \alpha$, Lemma 1 implies $\gamma < \alpha$. Hence by Lemma 11, $X_i^*(t_i) \leq X_i^*(\hat{t}_i)$. Hence no negative type can gain by deviating to (s_i, e_i) either.

Hence we can assume that there is no $t_i' \in F_i$ with $X_i^*(\hat{t}_i) \geq v_i(t_i')$. That is, every type t_i' who can send e_i has $v_i(t_i') > X_i^*(\hat{t}_i)$. Clearly, if $\hat{t}_i$ sends $M_i(\hat{t}_i)$, an option which must be feasible in the restricted artificial game for i, she must prove at least as much as e_i. Hence she has a strategy available in the restricted game which must lead to a belief by the principal above $X_i^*(\hat{t}_i)$, a contradiction.

Summarizing, we see that either no positive type can gain by sending (s_i, e_i) given the belief from the restricted artificial game for i this leads to or we can change that belief in such a way that no positive or negative type can gain. The symmetric argument for negative types then shows that by only changing off equilibrium beliefs, we can turn an equilibrium for the restricted artificial game for i into an equilibrium of the unrestricted game.

To complete the proof, we now use the equilibrium strategies of the artificial games to construct equilibrium strategies for the real game. The strategy for agent i is the same as her strategy in the equilibrium of the artificial game for i. Similarly, the principal's belief about t_i when he observes (s_i, e_i) is given by his belief in the artificial game for i. Given the principal's beliefs, sequential rationality tells us what his action must be at any information set where he has a unique optimal choice given his beliefs. However, we need to specify his actions at information sets with multiple optimal choices. On the equilibrium path, we will specify his actions to follow the optimal mechanism, but information sets off the equilibrium path are more subtle.

To construct the principal's equilibrium strategy, we divide the possible (s, e) profiles

he may observe into three sets. First, if (s, e) has positive probabilty under the equilibrium strategies of the agents, the principal chooses what the mechanism prescribes given the types. To be more specific, if $(s_i, e_i) \in R_i^{\alpha_i}$ and $\sigma_i^*(s_i, e_i \mid t_i) > 0$ for some t_i for each i, then the principal chooses $P(\cdot \mid \hat{t}, M(\hat{t}))$ for any $\hat{t}$ such that $\hat{t}_i \in T_i^{\alpha_i}$ for all i. Second, if (s, e) has the property that (s_i, e_i) has zero probability under the equilibrium strategies of the agents for at least two i, then the principal chooses any optimal $p(\cdot)$ given his beliefs. Obviously, the specification of the principal's strategy on such histories does not affect equilibrium considerations for the agents.

Third, consider any (s, e) such that (s_i, e_i) has zero probability under the equilibrium strategies for exactly one i. If $x_i^*(s_i, e_i) \neq X_i^*(t_i)$ for all t_i, then we can take the principal's response to (s, e) to be any optimal $p(\cdot) \in \Delta(A)$ given his beliefs. If $(s_i, e_i) \in R_i^\alpha$ and there exists $t_i \in T_i^\alpha$ with $x_i^*(s_i, e_i) = X_i^*(t_i)$, then we can treat (s_i, e_i) the same way as any positive probability $(s_i', e_i') \in R_i^\alpha$ as specified above. Next, suppose $(s_i, e_i) \in R_i^\alpha$ but the only t_i's satisfying $x_i^*(s_i, e_i) = X_i^*(t_i)$ have $t_i \notin T_i^\alpha$. If all such t_i are in the same T_i^β, then we treat (s_i, e_i) the same way as any positive probability (s_i', e_i') in R_i^β.

Finally, suppose $(s_i, e_i) \in R_i^\alpha$, there is no $t_i \in T_i^\alpha$ with $x_i^*(s_i, e_i) = X_i^*(t_i)$, and there exists $\bar{t}_i^k \in T_i^{\beta_k}$ with $x_i^*(s_i, e_i) = X_i^*(\bar{t}_i^k)$, $k = 1, 2$, with $\beta_1 > \beta_2$. By Lemma 11, $\beta_1 > \beta_2$ implies that every expectation of v_i induced in equilibrium by a type in $T_i^{\beta_1}$ must weakly exceed every expectation induced by a type in $T_i^{\beta_2}$. Hence it must be true that $X_i^*(\bar{t}_i^1) = \min_{t_i \in T_i^{\beta_1}} X_i^*(t_1)$ and $X_i^*(\bar{t}_i^2) = \max_{t_i \in T_i^{\beta_1}} X_i^*(t_i)$. If $\beta_2 > \alpha$, then take the principal's response to (s_i, e_i) to be the same as his response to any $(s_i', e_i') \in R_i^{\beta_2}$ which has positive probability. If $\alpha > \beta_1$, then take the principal's response to (s_i, e_i) to be the same as his response to any $(s_i', e_i') \in R_i^{\beta_1}$ which has positive probability. Finally, if $\beta_1 > \alpha > \beta_2$, take the principal's response to be a 50–50 mixture between his response given any $(s_i', e_i') \in R_i^{\beta_1}$ on the equilibrium path and the response given any $(s_i'', e_i'') \in R_i^{\beta_2}$ on the equilibrium path. (This case can only arise if $X_i^*(\bar{t}_i^1) = X_i(\bar{t}_i^2) = X_i^*(t_i')$ for all $t_i' \in T_i^\alpha$.)

To see that these strategies form an equilibrium, first note that Lemma 10 implies that the principal is choosing a best reply given his beliefs in response to every (s, e) which has positive probability in equilibrium. The construction above ensures that the principal is also sequentially rational in response to any (s, e) which has zero probability in equilibrium. Turning to the agents, consider any $t_i \in T_i$ and consider a deviation by t_i to some (s_i, e_i) with $\sigma_i^*(s_i, e_i \mid t_i) = 0$. If $x_i^*(s_i, e_i) \neq X_i^*(t_i)$, then the fact that these strategies are an equilibrium of the artificial game for i implies that if $t_i \in T_i^+$, we have $X_i^*(t_i) > x_i^*(s_i, e_i)$ and the reverse strict inequality if $t_i \in T_i^-$. By Lemma 5, this implies that t_i is at least weakly worse off deviating to (s_i, e_i).

So suppose $x_i^*(s_i, e_i) = X_i^*(t_i)$. That is, suppose the principal has the same belief about v_i under the deviation as he would following equilibrium play by t_i. For concrete-

ness, assume $t_i \in T_i^+$ — an analogous argument covers the case where $t_i \in T_i^-$. Assume that $t_i \in T_i^\alpha$. Since $e_i \in \mathcal{E}_i(t_i)$ by hypothesis, Lemma 1 implies that $(s_i, e_i) \in R_i^\beta$ for $\beta \leq \alpha$. If there is some type $t_i' \neq t_i$ who sends (s_i, e_i) with positive probability in equilibrium, then the outcome is the same as in the optimal mechanism given any $t_i' \in T_i^\beta$ while the outcome if t_i follows the strategy from the equilibrium of the artificial game is the same in the optimal mechanism given t_i. By incentive compatibility, t_i does not gain by deviating to (s_i, e_i).

So assume (s_i, e_i) is not sent with positive probability by any type in equilibrium. If $\beta = \alpha$ or if there is no $\gamma \neq \alpha$ with $t_i' \in T_i^\gamma$ and $x_i^*(s_i, e_i) = X_i^*(t_i')$, then (s_i, e_i) is treated the same way as any $(s_i', e_i') \in R_i^\alpha$ which does have positive probability, so the outcome is the same as if t_i followed the equilibrium strategy from the artificial game. Hence, again, he does not gain by deviating.

Finally, suppose (s_i, e_i) has zero probability in equilibrium, $\alpha > \beta$, and there is some $\gamma \neq \alpha$ with $t_i' \in T_i^\gamma$ and $x_i^*(s_i, e_i) = X_i^*(t_i')$. If $\gamma > \alpha$, then, again, (s_i, e_i) is treated the same way as any $(s_i', e_i') \in R_i^\alpha$ which has positive probability in the equilibrium of the artificial game, so the outcome is again the same as if t_i followed the equilibrium strategy from the artificial game. Hence, again, she does not gain by deviating. If $\alpha > \beta > \gamma$, the response to (s_i, e_i) is a 50–50 randomization between the way the principal would respond to positive probability $(s_i', e_i') \in R_i^\alpha$ and the way he would respond to positive probability $(s_i', e_i') \in R_i^\gamma$. This is strictly worse for t_i than the response to t_i's equilibrium strategy. Finally, if $\alpha > \gamma > \beta$, the principal's response is the same as his response to any positive probability $(s_i', e_i') \in R_i^\gamma$, again worse for t_i than following the equilibrium strategy. ▌

B Proof of Theorem 2

We first show there exists v_i^* solving

$$v_i^* = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \text{ or } v_i(t_i) \leq v_i^*]. \tag{4}$$

If $T_i = T_i^0$, then it is easy to see that $v_i^* = \mathrm{E}(v_i(t_i))$ satisfies (4). On the other hand, if $T_i^0 = \emptyset$, then $v_i^* = \min_{t_i \in T_i} v_i(t_i)$ satisfies (4). In what follows, assume $T_i^0 \neq \emptyset$ and $T_i^0 \neq T_i$.

Write $T_i \setminus T_i^0$ as $\{t_i^1, \ldots, t_i^N\}$ where without loss of generality $v_i(t_i^n) < v_i(t_i^{n+1})$. (If we have $t_i, t_i' \in T_i \setminus T_i^0$ with $t_i \neq t_i'$ and $v_i(t_i) = v_i(t_i')$, we can treat these two types as if they were one type for the purposes of this calculation.) For $n = 1, \ldots, N$, let

$$g_i^n = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \text{ or } t_i = t_i^k, \text{ for } k \leq n]$$

and let $g_i^0 = \mathrm{E}(v_i(t_i) \mid t_i \in T_i^0)$.

Suppose that there is no solution to equation (4). If $g_i^0 \leq v_i(t_i^1)$, then g_i^0 satisfies (4). Hence $g_i^0 > v_i(t_i^1)$. But g_i^1 is a convex combination of $v_i(t_i^1)$ and g_i^0, so $v_i(t_i^1) < g_i^1 < g_i^0$. Again, if $g_i^1 \leq v_i(t_i^2)$, then g_i^1 satisfies (4), so we must have $g_i^1 > v_i(t_i^2)$, implying $v_i(t_i^2) < g_i^2 < g_i^1$. Similar reasoning gives $g_i^{n-1} > g_i^n > v_i(t_i^n)$ for $n = 1, \ldots, N$. In particular, $g_i^N > v_i(t_i^N)$. But $g_i^N = \mathrm{E}[v_i(t_i)]$, so this implies g_i^N solves equation (4), a contradiction. Hence a solution exists.

To see that the solution is unique, suppose to the contrary that v_i^1 and v_i^2 both solve (4) where $v_i^1 > v_i^2$. Let

$$T_i^k = T_i^0 \ \cup \ \{t_i \in T_i \setminus T_i^0 \mid v_i(t_i) \leq v_i^k\},$$

so $v_i^k = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^k]$. Since $v_i^1 > v_i^2$, we have $T_i^2 \subset T_i^1$ and

$$T_i^1 \setminus T_i^2 = \{t_i \in T_i \setminus T_i^0 \mid v_i^2 < v_i(t_i) \leq v_i^1\}.$$

Note that v_i^1 is a convex combination of v_i^2 and $\mathrm{E}[v_i(t_i) \mid t_i \in T_i^1 \setminus T_i^2]$. But every $t_i \in T_i^1 \setminus T_i^2$ has $v_i(t_i) \leq v_i^1$, so we must have

$$\mathrm{E}[v_i(t_i) \mid t_i \in T_i^1 \setminus T_i^2] \leq v_i^1 \leq v_i^2,$$

contradicting $v_i^1 > v_i^2$.

To construct equilibrium strategies, first note that we must have $x_i^*(s_i, \{t_i\}) = v_i(t_i)$. That is, if t_i proves her type, the principal must infer correctly. Thus we only need to determine the principal's beliefs in response to reports of the form (s_i, T_i) where the agent proves nothing.

It is easy to see that if $T_i^0 = \emptyset$, then $v_i^* = \min_{t_i \in T_i} v_i(t_i)$ and that the essentially unique equilibrium has every type proving her type. This is the usual unraveling argument. Any type t_i' with $v_i(t_i') = \max_{t_i \in T_i} v_i(t_i)$ must strictly prefer proving her type to pooling with lower types and so must prove her type. But then any type with the next highest possible value of $v_i(t_i)$ cannot pool with higher types and so must prove her type, etc. So for the rest of this proof, assume $T_i^0 \neq \emptyset$.

Clearly, we cannot have (s_i, T_i) and (s_i', T_i), both with positive probability in equilibrium with $x_i^*(s_i, T_i) \neq x_i^*(s_i', T_i)$. Since all types are positive, every type strictly prefers whichever of these reports yields the larger x in response. Hence we may as well fix some s_i^* and suppose that the only (s_i, T_i) sent with positive probability in equilibrium is (s_i^*, T_i) where $x_i^*(s_i^*, T_i) \geq x_i^*(s_i, T_i)$ for all $s_i \in T_i$.

Let $\tilde{v}_i = x_i^*(s_i^*, T_i)$. From the above, we know that types $t_i \in T_i^0$ send report (s_i^*, T_i). Any type $t_i \notin T_i^0$ can send either (s_i^*, T_i) and obtain response $\tilde{v}_i$ or can send some $(s_i, \{t_i\})$

and receive response $v_i(t_i)$. Hence t_i chooses the former only if $\tilde{v}_i > v_i(t_i)$. Ignoring indifference for a moment, we see that this implies that $\tilde{v}_i$ must be the v_i^* defined in equation (4). To address indifference, note that v^* is not changed if we add or remove from the set of types sending this message a type with $v_i(t_i) = v_i^*$. Hence we have the same outcome regardless. ▌

C Proof of Theorem 3

The existence and uniqueness of v_i^+ follows from Theorem 2 taking the set of types to be T_i^+. For v_i^-, note that Theorem 2 applied to the function $-v_i(t_i)$ and types T_i^- implies that there is a unique v_i^- satisfying

$$-v_i^- = \mathrm{E}[-v_i(t_i) \mid t_i \in T_i^0 \cap T_i^- \text{ or } (t_i \in T_i^- \setminus T_i^0 \text{ and } -v_i(t_i) \leq -v_i^-)]$$

which can be rewritten as the definition of v_i^-.

Next, we show that there exists v_i^* solving

$$\begin{aligned} v_i^* = &\mathrm{E}\left[v_i(t_i) \mid (t_i \in T_i^0) \text{ or } (t_i \in T_i^- \setminus T_i^0 \text{ and } v_i(t_i) \geq v_i^*)\right. \\ &\left. \text{or } (t_i \in T_i^+ \setminus T_i^0 \text{ and } v_i(t_i) \leq v_i^*)\right]. \end{aligned} \tag{5}$$

Let the function of v_i^* on the right–hand side be denoted $g_i(v_i^*)$. So we seek to prove that there exists v_i^* solving $v_i^* = g_i(v_i^*)$.

As with Theorem 2, the proof is by contradiction. So suppose there is no v_i^* solving this equation. Let $v_i^1, \ldots, v_i^N$ denote the values of $v_i(t_i)$ for $t_i \notin T_i^0$. Without loss of generality, assume $v_i^k < v_i^{k+1}$ for $k = 1, \ldots, N-1$.

First, note that for $v_i^* \leq v_i^1$, we have $g_i(v_i^*) = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^-]$. So the assumption that there is no v_i^* with $v_i^* = g_i(v_i^*)$ implies $\mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^-] > v_i^1$ as otherwise we can set $v_i^* = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^-]$ and obtain a solution to equation (5).

Clearly, the function $g_i(v_i^*)$ is constant in v_i^* for $v_i^* \in (v_i^k, v_i^{k+1})$. Hence if $g_i(v_i^k) \in (v_i^k, v_i^{k+1})$, we have a solution to equation (5) in this interval.

Also, if $g_i(v_i^k) > v_i^{k+1}$, then $g_i(v_i^{k+1}) > v_i^{k+1}$. To see this, first suppose that $v_i^{k+1} \in v_i(T_i^-)$. In this case, $g_i(v_i^k)$ is a convex combination of $g_i(v_i^k + 1)$ and v_i^{k+1}. Since $g_i(v_i^k) > v_i^{k+1}$ by assumption, we must have $g_i(v_i^{k+1}) \geq g_i(v_i^k) > v_i^{k+1}$, implying the claim. Alternatively, suppose $v_i^{k+1} \in v_i(T_i^+)$. In this case, $g_i(v_i^{k+1})$ is a convex combination of $g_i(v_i^k)$ and v_i^{k+1}. Since $g_i(v_i^k) > v_i^{k+1}$, this implies $g_i(v_i^{k+1}) > v_i^{k+1}$.

As shown above, we start with $g_i(v_i^1) > v_i^1$, so by induction, we have $g_i(v_i^N) > v_i^N$. But $g_i(v_i^*) = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^+]$ for all $v_i^* \geq v_i^N$. So there exists $v_i^* > v_i^N$ solving (5), a contradiction.

To show uniqueness, suppose to the contrary that v_i^1 and v_i^2 are both solutions to equation (5) where $v_i^1 > v_i^2$. Let

$$T_i^{k+} = \{t_i \in T_i^+ \setminus T_i^0 \mid v_i(t_i) \leq v_i^k\}, \quad k = 1, 2$$

and

$$T_i^{k-} = \{t_i \in T_i^- \setminus T_i^0 \mid v_i(t_i) \geq v_i^k\}, \quad k = 1, 2.$$

Clearly, since $v_i^1 > v_i^2$, we have $T_i^{2+} \subseteq T_i^{1+}$ and $T_i^{1-} \subseteq T_i^{2-}$. But

$$v_i^k = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^{k+} \cup T_i^{k-}].$$

Let

$$\tilde{v}_i = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^{2+} \cup T_i^{1-}].$$

Then v_i^1 is a convex combination of $\tilde{v}_i$ and $\mathrm{E}[v_i(t_i) \mid t_i \in T_i^{1+} \setminus T_i^{2+}]$, while v_i^2 is a convex combination of $\tilde{v}_i$ and $\mathrm{E}[v_i(t_i) \mid t_i \in T_i^{2-} \setminus T_i^{1-}]$. It is easy to see that

$$v_i^2 \leq \mathrm{E}[v_i(t_i) \mid t_i \in T_i^{1+} \setminus T_i^{2+}] \leq v_i^1$$

since $v_i^2 \leq v_i(t_i) \leq v_i^1$ for all $t_i \in T_i^{1+} \setminus T_i^{2+}$. Similarly,

$$v_i^2 \leq \mathrm{E}[v_i(t_i) \mid t_i \in T_i^{2-} \setminus T_i^{1-}] \leq v_i^1.$$

Since v_i^1 is a convex combination of $\tilde{v}_i$ and something smaller than v_i^1, we must have $\tilde{v}_i \geq v_i^1$. But since v_i^2 is a convex combination of $\tilde{v}_i$ and something larger than v_i^2, we must have $v_i^2 \geq \tilde{v}_i$. Hence

$$v_i^1 \leq \tilde{v}_i \leq v_i^2,$$

contradicting $v_i^1 > v_i^2$.

Turning to equilibrium strategies, first note that if $x_i^*(s_i, T_i) > x_i^*(s_i', T_i)$, then no positive type will send report (s_i', T_i) and no negative type will send (s_i, T_i). Hence there are, at most, two distinct values of $x_i^*(s_i, T_i)$ observed on the equilibrium path. Let $\tilde{v}_i^+ = \max_{s_i \in T_i} x_i^*(s_i, T_i)$ and $\tilde{v}_i^- = \min_{s_i \in T_i} x_i^*(s_i, T_i)$. For the moment, assume $\tilde{v}_i^+ > \tilde{v}_i^-$. Then it is easy to see that every positive type $t_i \in T_i^0$ sends a report generating $\tilde{v}_i^+$ as does every positive type $t_i \notin T_i^0$ with $v_i(t_i) < \tilde{v}_i^+$. Similarly, every negative type $t_i \in T_i^0$ or not in T_i^0 but with $v_i(t_i) > \tilde{v}_i^-$ sends some report generating $\tilde{v}_i^-$. All other types t_i send a report of the form $(s_i, \{t_i\})$. Given this, it is clear that $\tilde{v}_i^+$ must equal v_i^+ and $\tilde{v}_i^-$ must equal v_i^-. This is an equilibrium iff $v_i^+ \geq v_i^-$. Note that if $v_i^+ = v_i^-$, then the expectation of v_i given the set of types sending either report must also be the same value. Thus in this case, we have $v_i^- = v_i^+ = v_i^*$.

Regardless of the relationship between v_i^- and v_i^+, there is also an equilibrium where the principal's beliefs ignore the type report and condition only on the evidence. Letting $\tilde{v}_i$ denote the principal's expected value of v_i condition on the evidence report $e_i = T_i$, we see that positive types with $v_i(t_i) > \tilde{v}_i$ will prove their types as will negative types with $v_i(t_i) < \tilde{v}_i$. Hence $\tilde{v}_i$ must satisfy equation (5), so $\tilde{v}_i = v_i^*$. ∎

D Proof of Corollary 2

When $v_i^+ \leq v_i^-$, there is only one equilibrium in the artificial game for i, so the claim follows. When $v_i^+ > v_i^-$, however, there are (essentially) two equilibria. In one, type reports are used to separate positive types from negative types. All positive types with evidence and $v_i(t_i) > v_i^+$ prove their types, as do all negative types with evidence and $v_i(t_i) < v_i^-$. All other positive types send one type report and evidence $e_i = T_i$, while all other negative types send another type report and the same evidence. In what follows, we refer to this equilibrium as the *cheap talk equilibrium* as it uses the "cheap talk" of type reports to help separate. In the other equilibrium, the principal's beliefs depend only on the evidence presented, so type reports are irrelevant. All positive types with evidence and $v_i(t_i) > v_i^*$ prove their types as do all negative types with evidence and $v_i(t_i) < v_i^*$. All other types report some fixed type report and evidence $e_i = T_i$. We refer to this equilibrium as the *non–talk equilibrium.*

Since there are two equilibria in the artificial game for i in this case, we need to determine which strategies for i are used in the equilibrium of the game which has the same outcome as the optimal mechanism. Clearly, if the principal is better off under one set of strategies than the other, then these must be the strategies used since the equilibrium corresponding to the optimal mechanism must be the best possible equilibrium for the principal.

We now show that the principal's payoff is always at least weakly larger in the cheap talk equilibrium, completing the proof of Corollary 2.

First, we show that $v_i^+ > v_i^-$ implies $v_i^+ \geq v_i^* \geq v_i^-$ with at least one strict inequality. To see this, suppose to the contrary that $v_i^* > v_i^+ > v_i^-$. Define the following sets of types:

$$\hat{T}_i^- = \{t_i \in T_i^- \mid t_i \in T_i^0 \text{ or } v_i(t_i) > v_i^-\}$$
$$\hat{T}_i^+ = \{t_i \in T_i^+ \mid t_i \in T_i^0 \text{ or } v_i(t_i) < v_i^+\}$$
$$\hat{T}_i^{*-} = \{t_i \in T_i^- \mid t_i \in T_i^0 \text{ or } v_i(t_i) > v_i^*\}$$
$$\hat{T}_i^{*+} = \{t_i \in T_i^+ \mid t_i \in T_i^0 \text{ or } v_i(t_i) < v_i^*\}$$

In other words, the types in $\hat{T}_i^-$ are the negative types who "pool" together in the cheap talk equilibrium, while $\hat{T}_i^+$ is the set of positive types who pool together in this equilibrium. Similarly, $\hat{T}_i^{*-}$ and $\hat{T}_i^{*+}$ are, respectively, the set of negative and positive types who all pool together in the non–talk equilibrium. By definition,

$$v_i^- = \mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^-]$$

$$v_i^+ = \mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^+]$$

$$v_i^* = \mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^{*-} \cup \hat{T}_i^{*+}]$$

Hence v_i^* is a convex combination of $\mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^{*-}]$ and $\mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^{*+}]$.

Since $v_i^- < v_i^*$, we see that $\hat{T}_i^{*-} \subseteq \hat{T}_i^-$. Note that if $t_i \in \hat{T}_i^-$ but $t_i \notin \hat{T}_i^{*-}$, then $v_i^- \leq v_i(t_i) < v_i^*$. Hence $v_i^- = \mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^-]$ is a convex combination of $\mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^{*-}]$ and the expectation of $v_i(t_i)$ for a set of types all with $v_i(t_i) > v_i^-$. Hence

$$v_i^* > v_i^- = \mathrm{E}[v_i(t_i) \mid t_i \in T_i^-] > \mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^{*-}].$$

Similarly, $v_i^+ < v_i^*$ implies that $\hat{T}_i^+ \subseteq \hat{T}_i^{*+}$. Since the types in $\hat{T}_i^{*+} \setminus \hat{T}_i^+$ all satisfy $v_i^+ \leq v_i(t_i) < v_i^*$, we see that $\mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^{*+}]$ is a convex combination of $v_i^+ = \mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^+] < v_i^*$ and an expectation of $v_i(t_i)$ for a set of types with $v_i(t_i) < v_i^*$. Hence

$$\mathrm{E}[v_i(t_i) \mid t_i \in \hat{T}_i^{*+}] < v_i^*.$$

But then we have v_i^* is a convex combination of two terms which are strictly smaller than v_i^*, a contradiction. A similar argument rules out the possibility that $v_i^+ > v_i^- > v_i^*$.

Consider the game between the agents and the principal. We know that there is a robust PBE with the same outcome as in the optimal mechanism. We know i's strategy in this equilibrium must either be the one she uses in the cheap talk equilibrium or the one she uses in the non–talk equilibrium. Fix the strategies of all agents other than i. We know these strategies are defined from the artificial games for these agents, independently of which strategy i uses or the principal's response to i. Thus we can simply determine which strategy by i leads to a higher payoff for the principal.

Note that the principal's payoff for a fixed a is linear in his expectation of v_i. Hence his maximized payoff is convex in his expectation of v_i. We now show that the distribution of beliefs for the principal in the cheap talk equilibrium is a mean–preserving spread of the distribution in the non–talk equilibrium, completing the proof. To be precise, let (σ_i^1, x_i^1) denote the cheap talk equilibrium strategies and (σ_i^2, x_i^2) the non–talk equilibrium strategies from the artificial game for i. For $k = 1, 2$, define probability distributions B^k over $\mathbf{R}$ by

$$B^k(\hat{v}_i) = \rho_i\left(\left\{t_i \in T_i \mid X_i^k(t_i) = \hat{v}_i\right\}\right).$$

(Recall that $X_i^k(t_i) = x_i^k(s_i, e_i)$ for any (s_i, e_i) with $\sigma_i^k(s_i, e_i \mid t_i) > 0$ and that ρ_i is the prior over T_i.) The law of iterated expectations implies

$$\sum_{\hat{v}_i \in \text{supp}(B^k)} \hat{v}_i B^k(\hat{v}_i) = \text{E}[v_i(t_i)], \quad k = 1, 2.$$

Hence the two distributions have the same mean.

Consider any $\hat{v}_i < v_i^-$. Since $v_i^- \leq v_i^*$, for $k = 1$ or $k = 2$, we have $X_i^k(t_i) = \hat{v}_i$ if and only if there is a negative type with evidence who has $v_i(t_i) = \hat{v}_i$. Similarly, since $v_i^* \leq v_i^+$, for any $\hat{v}_i > v_i^+$, we have $X_i^k(t_i) = \hat{v}_i$ iff there is a positive type with evidence who has $v_i(t_i) = \hat{v}_i$. Hence $B^1(\hat{v}_i) = B^2(\hat{v}_i)$ for any $\hat{v}_i \notin [v_i^-, v_i^+]$.

Also, we have $B^1(\hat{v}_i) = 0$ for all $\hat{v}_i \in (v_i^-, v_i^+)$. Any type with v_i in this range either (1) is positive and chooses to induce belief v_i^+ or (2) is negative and chooses to induce belief v_i^-. Under B^2, however, many of the types generating beliefs concentrated at v_i^- or v_i^+ in the cheap talk equilibrium instead generate beliefs in (v_i^-, v_i^+). In particular, types without evidence or types with evidence they prefer not to show induce the belief v_i^*, a positive type with evidence who has $v_i(t_i) \in (v_i^*, v_i^+)$ generate the belief $v_i(t_i)$, and similarly for negative types. Hence B^1 is a mean–preserving spread of B^2. ▌

References

[1] Ben-Porath, E., E. Dekel, and B. Lipman, "Optimal Allocation with Costly Verification," *American Economic Review*, **104**, December 2014, 3779–3813.

[2] Ben-Porath, E., and B. Lipman, "Implementation and Partial Provability," *Journal of Economic Theory*, **147**, September 2012, 1689–1724.

[3] Bull, J., and J. Watson, "Hard Evidence and Mechanism Design," *Games and Economic Behavior*, **58**, January 2007, 75–93.

[4] Deneckere, R. and S. Severinov, "Mechanism Design with Partial State Verifiability," *Games and Economic Behavior*, **64**, November 2008, 487–513.

[5] Dye, R. A., "Disclosure of Nonproprietary Information," *Journal of Accounting Research*, **23**, 1985, 123–145.

[6] Erlanson, A., and A. Kleiner, "Costly Verification in Collective Decisions," working paper, November 2015.

[7] Erlanson, A., and A. Kleiner, "A Note on Optimal Allocation with Costly Verification," working paper, May 2016.

[8] Fudenberg, D., and J. Tirole, "Perfect Bayesian Equilibrium and Sequential Equilibrium," *Journal of Economic Theory*, **53**, April 1991, 236–260.

[9] Glazer, J., and A. Rubinstein, "On Optimal Rules of Persuasion," *Econometrica*, **72**, November 2004, 1715–1736.

[10] Glazer, J., and A. Rubinstein, "A Study in the Pragmatics of Persuasion: A Game Theoretical Approach," *Theoretical Economics*, **1**, December 2006, 395–410.

[11] Green, J., and J.-J. Laffont, "Partially Verifiable Information and Mechanism Design," *Review of Economic Studies*, **53**, July 1986, 447–456.

[12] Guttman, I., I. Kremer, and A. Skrzypacz, "Not Only What but also When: A Theory of Dynamic Voluntary Disclosure," *American Economic Review*, **104**, August 2014, 2400–2420.

[13] Hart, S., I. Kremer, and M. Perry, "Evidence Games: Truth and Commitment," working paper, July 2015.

[14] Jung, W., and Y. Kwon, "Disclosure When the Market is Unsure of Information Endowment of Managers," *Journal of Accounting Research*, **26**, 1988, 146–153.

[15] Kartik, N., and O. Tercieux, "Implementation with Evidence," *Theoretical Economics*, **7**, May 2012, 323–355.

[16] Lipman, B., and D. Seppi, "Robust Inference in Communication Games with Partial Provability," *Journal of Economic Theory*, **66**, August 1995, 370–405.

[17] Lipman, B., "An Elementary Proof of the Optimality of Threshold Mechanisms," working paper, July 2015.

[18] Sher, I., "Credibility and Determinism in a Game of Persuasion," *Games and Economic Behavior*, **71**, March 2011, 409–419.

[19] Sher, I., and R. Vohra, "Price Discrimination through Communication," *Theoretical Economics*, **10**, May 2015, 597–648.

[20] Shin, H. S., "The Burden of Proof in a Game of Persuasion," *Journal of Economic Theory*, **64**, October 1994, 253–264.

[21] Shin, H. S., "Disclosures and Asset Returns," *Econometrica*, **71**, January 2003, 105–133.