ECONOMETRICA

JOURNAL OF THE ECONOMETRIC SOCIETY

An International Society for the Advancement of Economic Theory in its Relation to Statistics and Mathematics

http://www.econometricsociety.org/

Econometrica, Vol. 87, No. 2 (March, 2019), 529–566

MECHANISMS WITH EVIDENCE: COMMITMENT AND ROBUSTNESS

ELCHANAN BEN-PORATH
Department of Economics and Federmann Center for the Study of Rationality, Hebrew University

EDDIE DEKEL
Economics Department, Northwestern University and School of Economics, Tel Aviv University

BARTON L. LIPMAN
Department of Economics, Boston University

Econometrica, Vol. 87, No. 2 (March, 2019), 529–566

MECHANISMS WITH EVIDENCE: COMMITMENT AND ROBUSTNESS

ELCHANAN BEN-PORATH
Department of Economics and Federmann Center for the Study of Rationality, Hebrew University

EDDIE DEKEL
Economics Department, Northwestern University and School of Economics, Tel Aviv University

BARTON L. LIPMAN
Department of Economics, Boston University

We show that in a class of I-agent mechanism design problems with evidence, commitment is unnecessary, randomization has no value, and robust incentive compatibility has no cost. In particular, for each agent i, we construct a simple disclosure game between the principal and agent i where the equilibrium strategies of the agents in these disclosure games give their equilibrium strategies in the game corresponding to the mechanism but where the principal is not committed to his response. In this equilibrium, the principal obtains the same payoff as in the optimal mechanism with commitment. As an application, we show that certain costly verification models can be characterized using equilibrium analysis of an associated model of evidence.

KEYWORDS: Mechanism design, evidence, commitment, robustness.

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—a form of incentive compatibility analogous to but stronger than ex post incentive compatibility and dominant strategy incentive compatibility—has no cost. Also, commitment is unnecessary 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. This equilibrium can be derived from a collection of I auxiliary games, where the ith game is a simple disclosure game between agent i and the principal. As an application, we show that certain mechanism design problems with costly verification can be solved via an associated evidence model.[1]

To understand these results, consider the following example, the *simple allocation problem*. The principal has one unit of an indivisible good which he can allocate to one of I agents. Each agent i has private information in the form of her type t_i which determines $v_i(t_i)$, the value to the principal of allocating the good to agent i. 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 evidence which proves some facts about her type. For example, the principal may be a dean with one job slot to

Elchanan Ben-Porath: benporat@math.huji.ac.il
Eddie Dekel: dekel@northwestern.edu
Barton L. Lipman: blipman@bu.edu

We thank numerous seminar audiences and four anonymous referees for useful comments and suggestions. We also thank the National Science Foundation, Grant SES-0820333 (Dekel), and the US–Israel Binational Science Foundation for support for this research.

[1]In a model with costly verification, the agents do not have evidence to present, but the principal can learn the true type of an agent at a cost.

 https://doi.org/10.3982/ECTA14991

allocate to a department in the College. Each department wants the slot and has private information about the person the department would likely hire with the slot, information that is relevant to the value to the dean of assigning the slot to the department. Alternatively, the principal may be a state government which needs to choose a city in which to locate a public hospital. The state wants to place the hospital where it will be most efficiently utilized, but each city wants the hospital and has private information on local needs.

In a mechanism design formulation, the principal commits to how he will allocate the good as a function of cheap-talk reports and evidence presentation by the agents. A version of the Revelation Principle implies we can restrict attention to mechanisms where each agent reports her type truthfully and, in a sense to be defined later, presents all her evidence.

Alternatively, we could consider a game in which agents send evidence to the principal without any commitment by the principal, which we call the game without commitment. In this game, the principal forms beliefs about the types of the agents and allocates the good optimally given these beliefs. That is, the principal responds to the evidence and claims presented by forming a belief about $v_i(t_i)$ for each agent i and allocates the good to that agent for whom his expectation of $v_i(t_i)$ is largest. Since all agents want the good, in an equilibrium of this game, each agent i tries to persuade the principal that $v_i(t_i)$ is large.

This last observation implies that we could find certain equilibria of the game without commitment by means of what we call auxiliary games. For each agent i, consider the two-player game between i and the principal where type t_i has available the same cheap-talk messages and evidence she has in the game without commitment. The principal chooses an action $x \in \mathbf{R}$. The principal's payoff is $-(v_i(t_i) - x)^2$ and the agent's payoff is x. In other words, x is the principal's "estimate" of $v_i(t_i)$ and the agent's utility is increasing in the principal's estimate. Intuitively, the auxiliary game identifies the best strategy for agent i to use to try to convince the principal that $v_i(t_i)$ is large, just as she wants to do in the game without commitment. For each agent i, find an equilibrium of the auxiliary game for i. Then we can find an equilibrium for the game without commitment by having each agent play her strategy from the auxiliary game with the principal choosing a best response to the information this reveals to him. This equilibrium will be robust in the sense that no agent's beliefs about other agents plays any role in the equilibrium. We will show that the outcome in the best equilibrium so constructed is the same as the outcome in the optimal mechanism.[2] This implies that the optimal mechanism is robust in a similar sense. Consequently, many other games generate the same results—for example, if agents speak sequentially observing previous speakers, each will still wish to persuade the principal that her v_i is large and so her equilibrium strategy will not change.

One way to understand why the equilibrium has the same outcome as the optimal mechanism is to ask how the principal might use commitment to improve on the equilibrium. Conditional on any type of agent i whose type is perfectly revealed to the principal in equilibrium, it is clear that the principal cannot improve his payoff in a mechanism since he optimizes in equilibrium given exact knowledge of the type.

So consider a set of types of agent i that pool in equilibrium. Since they pool, the principal treats these types as if the value of giving the good to them were the average of $v_i(t_i)$ across the pool. Given a type in this pool whose value is above the average, the principal would like to be able to separate this type from the pool and give her the good more

[2]There could be several equilibria in the auxiliary game for agent i, in which case this construction gives multiple equilibria for the game without commitment.

often. This type would also like this response, so if she did not separate in the equilibrium, it must be because she does not have evidence that would enable her to do so. This lack of evidence also makes it impossible for the principal to separate her in an incentive compatible manner in a mechanism, so, again, the principal cannot improve.

Finally, consider a type in the pool whose value is below the average. The principal would like to separate this type from the pool to give her the good less often. It is possible that this type has evidence that would separate her from the pool but that she withholds this evidence in equilibrium to avoid revealing her low value. In a mechanism, the principal can promise to reward the agent for this revelation and so can use commitment to induce her to separate from the pool. However, he does not want to: rewarding this type means giving her the good *more* often, but he wants the information so that he can give it to her *less* often. Hence, again, the principal cannot use commitment to improve the outcome.

Our results apply to a broader class of allocation problems. For example, consider our example of a dean, but suppose the dean has several job slots to allocate where each department can have at most one and there are fewer slots than departments. A related problem is the allocation of a budget across divisions by the head of a firm. Suppose the organization has a fixed amount of money to allocate and that the value produced by a division is a function of its budget and its privately known productivity. Here each division wants to persuade top management that its productivity is high. Alternatively, consider a task allocation problem where the principal is a manager who must choose an employee to carry out a particular job. Suppose none of the employees wants to do the task and each has private information about how well she would do it. Here each employee wishes to convince the manager that her productivity is low.

A more complex example is a task that some employees would and some would not want to do, where both the employee's ability and desire to do the job are private information. In this case, certain types of employees wish to persuade the manager that they would perform poorly, while others have the opposite incentive and these incentives could be correlated with the value to the manager of assigning them the task. Our results cover this case as well.

A different class of examples is public goods problems. The principal chooses whether or not to provide a public good. If the principal provides the good, the cost is 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 to be 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 about the value of the public good to her. 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. The principal maximizes a weighted sum of the agents' utilities, possibly including a benefit or cost of his own for providing the public good. Here some types wish to persuade the principal that they highly value the public good, while others wish to persuade him of the opposite.

The conclusion that the principal does not require commitment is important for several reasons. First, it is not always obvious whether commitment is an appropriate assumption for a given setting. Our result says that we obtain the same outcome either way. Second, in some settings, whether the principal is committed is endogenous. In the settings we consider, we predict that the principal would not invest to achieve commitment power.

Another useful implication of our results is that we can compute optimal mechanisms by considering equilibria of the game without commitment. In particular, as discussed above, we can characterize the relevant equilibrium by means of a collection of I auxiliary games, one for each agent, where the game for agent i is a simple disclosure game between agent i and the principal. The auxiliary game does not depend on the principal's preferences in the original mechanism design problem or the structure of the set of allocations. In some cases, the use of auxiliary games makes determining the optimal mechanism straightforward. In particular, if each auxiliary game has either a unique equilibrium or a unique "most informative" equilibrium, we can use these equilibria to directly compute the information the principal uses in the optimal mechanism. Given this information, it is straightforward to compute the outcome under the optimal mechanism.

To illustrate, we consider optimal mechanisms with the evidence technology proposed by Dye (1985). In Dye's model, each agent has some probability of having evidence that would enable her to exactly prove her type and otherwise has no evidence. When we apply this approach to the simple allocation problem or to the public goods problem, we find optimal mechanisms reminiscent of optimal mechanisms in a different context, namely, under costly verification. We discuss this connection to Ben-Porath, Dekel, and Lipman (2014) (henceforth BDL) and to Erlanson and Kleiner (2017) in Section 3.2 where we show that a class of costly verification models can be solved using our results for evidence models. This connection does not imply that all of our results for mechanisms with evidence carry over to costly verification models, only that optimal mechanisms for costly verification can be computed via Dye-evidence models.

The paper is organized as follows. Section 2 presents the formal model. In Section 2.5, we state the main results sketched above. The proof of this theorem is sketched and the roles of the assumptions explained in Section 4. In Section 3, we specialize to Dye (1985) evidence 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 also show that under some conditions, optimal mechanisms for costly verification can be solved using the optimal mechanisms for Dye evidence. We discuss the related literature in Section 5. Proofs not contained in the text are in the Appendix or the Supplemental Material (Ben-Porath, Dekel, and Lipman (2019)).

2. MODEL AND RESULTS

The set of agents is $\mathcal{I} = \{1, \ldots, I\}$ where $I \geq 1$. The principal has a finite set of feasible actions A and can randomize over these. For example, in the simple allocation problem, we have $A = I$ where $a = i$ means that the good is allocated to i.[3] More generally, $a \in A$ can be interpreted as an allocation of money (where money is finitely divisible) as well as other goods, public or private. It is notationally complex but not difficult to extend our results to the case where A is infinite. Each agent i has private information in the form of a type t_i where types are distributed independently across agents. The finite set of types of i is denoted T_i and ρ_i is the (full support) prior.[4]

[3]This formulation assumes the principal *must* allocate the good to some agent. Alternatively, we can set $A = \{0, 1, \ldots, I\}$ where $a = 0$ is interpreted as the principal keeping the good. Our results hold for either specification.

[4]Finiteness of T_i is primarily for tractability. As we will point out, there is one step in our proofs which does not obviously generalize to infinite type spaces.

2.1. *Preferences*

We first state our assumptions on preferences and then explain the interpretation. Given action a by the principal and type profile $t = (t_1, \ldots, t_I)$, agent i's utility is $u_i(a, t_i)$ and the principal's utility is $v(a, t)$ where

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

and

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

where

$$v_i(t_i) = \begin{cases} \bar{v}_i(t_i) & \text{if } t_i \in T_i^+, \\ -\bar{v}_i(t_i) & \text{if } t_i \in T_i^-. \end{cases}$$

For brevity, let $v_0(t_0) = 1$ and write this as $\sum_i u_i(a)v_i(t_i)$ with the convention that the sum runs from $i = 0$ to I.

The principal's utility function has two natural interpretations. First, we can interpret v as a social welfare function where $\bar{v}_i(t_i)$ reflects how much the principal "cares" about agent i's utility. Second, we can think of $\bar{v}_i(t_i)$ as measuring the extent to which the principal's interests are aligned with agent i's. That is, a high value of $\bar{v}_i(t_i)$ does not mean that the principal likes agent i but that the principal likes what agent i likes.

We do not restrict the sign of $\bar{v}_i(t_i)$ or $v_i(t_i)$. Thus, the principal's interests can be in conflict with those of some or all agents in a way which depends on the agents' types.

Turning to the agents, our formulation relaxes the restriction to type-independent preferences commonly used in the literature (see Section 5 for a survey), but requires that the type dependence takes a particularly simple form. Hence, we call our assumption on the agents' utility functions *simple type dependence*. It says that all types of agent i have the same indifference sets over A since if $u_i(a) = u_i(a')$, then every type of i is indifferent between a and a'. Thus, the only difference between types is the direction in which utility is increasing. Specifically, the types in T_i^+ have utility increasing in $u_i(a)$, while those in T_i^- have utility decreasing in this direction. We call the types in T_i^+ the *positive types* and those in T_i^+ the *negative types*. While restrictive, this formulation allows a broad range of interesting forms of type dependence.

For one thing, simple type dependence accommodates all the examples discussed in the Introduction. We illustrate with two examples. First, consider the simple allocation problem. Let $A = \{1, \ldots, I\}$ where $a = i$ means the principal allocates the good to agent i. Since every type desires the good, assume $T_i = T_i^+$, so $T_i^- = \emptyset$ and let

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

Let $u_0(a) \equiv 0$. Then our assumption on v implies $v_i(t_i)$ is the value to the principal of allocating the good to agent i when his type is t_i.

As another example, consider the public goods problem. Let $A = \{0, 1\}$, where 1 is providing the good and 0 is not providing it. Let the utility function for agent i be $av_i(t_i)$ and

the utility function for the principal be the sum of the agent's utilities or $\sum_i a v_i(t_i)$. For simplicity, assume $v_i(t_i) \neq 0$ for every t_i and every i. Then we can renormalize the utility function for t_i by dividing through by $|v_i(t_i)|$. After renormalizing, the utility function of agent i is

$$\begin{cases} a & \text{if } v_i(t_i) > 0, \\ -a & \text{if } v_i(t_i) < 0. \end{cases}$$

Letting $u_i(a) = a$ and defining $T_i^+ = \{t_i \in T_i \mid v_i(t_i) > 0\}$, we obtain a case of simple type dependence. The principal's utility function equals $\sum_i u_i(a) v_i(t_i)$, as assumed.

An example not discussed in the Introduction but commonly used in the literature has $u_i(a) = a$ for every i and

$$v(a, t) = -\sum_{i=1}^{I} \alpha_i \big(a - \beta_i(t_i)\big)^2.$$

Here the principal wants to guess the agents' types ($\beta_i(t_i)$'s) and all agents want to be thought of as having a high β_i. This does not look like the principal's utility function we assumed, but we can rewrite this as

$$v(a, t) = -a^2 \sum_i \alpha_i + \sum_i a 2\beta_i(t_i) - \sum_i \alpha_i \big(\beta_i(t_i)\big)^2.$$

The last term is not relevant to the principal's choice of mechanism, so we can renormalize by dropping it. Letting $u_0(a) = -a^2 \sum_i \alpha_i$ and $v_i(t_i) = 2\beta_i(t_i)$ yields our model.

As noted above, a special case of simple type dependence is type-independent preferences, the case studied in much of the literature including all previous work on commitment in mechanisms with evidence. Simple type dependence also holds trivially when the agent has only two type-independent indifference curves over A. For example, if the principal has only two actions, as in Glazer and Rubinstein (2004, 2006), then there can only be two indifference curves (at most). 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 others prefer not to.[5] Here the principal has as many actions as there are agents, but each agent has only two indifference curves over A. Again, there are only two (nontrivial) preferences over $\Delta(A)$, so simple type dependence is without loss of generality.

2.2. *Evidence*

Each agent may have evidence which would prove some claims about herself. Formally, 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 $e_i \subset T_i$. For example, if agent i presents a house deed with her name on it, she proves that she is one of the types who owns a house. 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

[5]This formulation is natural if the "good" is a task assignment as discussed in the Introduction.

$e_i \in \mathcal{E}_i(s_i)$. 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. Otherwise, the evidence does rule out s_i. In short, for any $e_i \in \bigcup_{s_i \in T_i} \mathcal{E}_i(s_i)$, we have $t_i \in e_i$ if and only if $e_i \in \mathcal{E}_i(t_i)$.

We also assume *normality* (Bull and Watson (2007), Lipman and Seppi (1995)). This convenient simplification, used in much of the literature, says that t_i can prove an event which summarizes all the evidence she has. Intuitively, there are no time or other restrictions on the evidence an agent can present, so she can present everything she has. Formally, 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, if t_i can prove that her type is in e_i, e_i', etc., then she can prove that her type is in all of these sets and hence in their intersection. More precisely, this intersection is itself an event that t_i can prove. Henceforth, we denote this maximally informative event by

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

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

As usual, we assume that *all* of an agent's private information is summarized by her type. Thus, it is common knowledge what evidence each agent has as a function of her type. On the other hand, our robustness result implies that no agent needs to know *anything* about other agents—in particular, no agent needs to understand what evidence others might have.

2.3. *Mechanisms*

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 version of the Revelation Principle has been shown by, among others, Bull and Watson (2007) and Deneckere and Severinov (2008). As in the usual model, we might not need agents to reveal this much information, but it is without loss of generality to induce them to do so as the principal can commit to ignoring some of it. Formally, let $\mathcal{E}_i = \bigcup_{t_i \in T_i} \mathcal{E}_i(t_i)$ and $\mathcal{E} = \prod_i \mathcal{E}_i$. A *mechanism* is then a function $P : T \times \mathcal{E} \to \Delta(A)$.

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

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

In words, this is agent i's expected utility under mechanism P when her type is t_i but she reports type s_i, presents evidence e_i, and expects all other agents to claim types t_{-i} and report evidence e_{-i}.

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

$$\mathrm{E}_{t_{-i}} \mathcal{U}_i\big(t_i, M_i(t_i), t_{-i}, M_{-i}(t_{-i}) \mid t_i, P\big) \geq \mathrm{E}_{t_{-i}} \mathcal{U}_i\big(s_i, e_i, t_{-i}, M_{-i}(t_{-i}) \mid t_i, P\big),$$

for all $s_i, t_i \in T_i$ and all $e_i \in \mathcal{E}_i(t_i)$. In words, the agent prefers reporting her type truthfully and presenting maximal evidence to any other report and any other evidence she has available given that all other agents report truthfully and present maximal evidence.

For mechanisms with evidence, it is useful to define a mapping giving the outcome of the mechanism as a function of the type profile. In the literature on mechanism design without evidence, there is no need to do so since, given truth-telling, a direct mechanism *is* such a function. To be specific, given an incentive compatible mechanism P, we say that the *mechanism outcome* is the function $O_P : T \to \Delta(A)$ defined by $O_P(t)(a) = P(a \mid t, M(t))$. In other words, the mechanism outcome gives the probability distribution over A as a function of t which results when all agents report truthfully and provide maximal evidence. The principal's expected payoff from an incentive compatible mechanism P is

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

Before defining robust incentive compatibility, we recall more standard notions. A mechanism is *ex post incentive compatible* if for every agent i,

$$\mathcal{U}_i\big(t_i, M_i(t_i), t_{-i}, M_{-i}(t_{-i}) \mid t_i, P\big) \geq \mathcal{U}_i\big(s_i, e_i, t_{-i}, M_{-i}(t_{-i}) \mid t_i, P\big),$$

for all $s_i, t_i \in T_i$, all $t_{-i} \in T_{-i}$, and all $e_i \in \mathcal{E}_i(t_i)$. That is, 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 the other agents' types and that they are reporting truthfully and presenting maximal evidence.

Say that a reporting strategy $\sigma_i : T_i \to T_i \times \mathcal{E}_i$ is *feasible* if whenever $\sigma_i(t_i) = (s_i, e_i)$, we have $e_i \in \mathcal{E}_i(t_i)$. A mechanism is *dominant strategy incentive compatible* if for every agent i,

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

for all $s_i, t_i \in T_i$, all feasible $\sigma_{-i} : T_{-i} \to T_{-i} \times \mathcal{E}_{-i}$, and all $e_i \in \mathcal{E}_i(t_i)$. That is, a mechanism is dominant strategy incentive compatible if every type of every agent has a dominant strategy to report honestly and present maximal evidence.

Neither of these notions of incentive compatibility implies the other. In an ex post incentive compatible mechanism, an agent might want to deviate if she knew another agent were going to report (s_i, e_i) where $e_i \neq M_i(s_i)$. In a dominant strategy incentive compatible mechanism, an agent could prefer to deviate if she knew the types of her opponents. Our robustness notion combines the ex post and dominant strategy properties above.

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

$$\mathcal{U}_i\big(t_i, M_i(t_i), t_{-i}, e_{-i} \mid t_i, P\big) \geq \mathcal{U}_i(s_i, e_i, t_{-i}, e_{-i} \mid t_i, P),$$

for all $s_i, t_i \in T_i$, all $t_{-i} \in T_{-i}$, all $e_{-i} \in \mathcal{E}_{-i}$, and all $e_i \in \mathcal{E}_i(t_i)$. In other words, even if i knew the type and evidence reports of other agents, it would be optimal to report truthfully and provide maximal evidence regardless of what those reports are. Robust incentive compatibility implies ex post incentive compatibility and dominant strategy incentive compatibility, but is not implied by either. See Part SA of the Supplemental Material for details.

A robustly incentive compatible mechanism has the desirable property that it does not rely on the principal knowing the beliefs of the agents about each other's types or strategies. Furthermore, the outcome of the mechanism need not change if the agents report publicly and sequentially, rather than simultaneously, regardless of the order in which they report.

Since robust incentive compatibility implies incentive compatibility, the best robustly incentive compatible mechanism for the principal yields a weakly lower expected payoff than the best incentive compatible mechanism. Under our assumptions, there is no difference—there is an optimal incentive compatible mechanism for the principal which is robustly incentive compatible.

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, the principal chooses an $a \in A$ without randomizing. Of course, randomization is an important feature of optimal mechanisms in some settings. Under our assumptions, there is an optimal mechanism which is deterministic.

2.4. *Games*

Our result that commitment is not needed says that an equilibrium of a particular game between the principal and the agents has the same outcome as an optimal mechanism. The interest in this result depends on the game. The game we consider seems natural as it is just like the mechanism "game" in that the agents all make reports of types and send evidence to the principal, after which he chooses an outcome. The difference from the mechanism is that the principal is not committed to his response to these reports. We refer to this as the *game without commitment*. Our robustness property implies that the same result holds for a wide range of other games, such as games with sequential reports instead of simultaneous.

Our result is *not* that the agents and principal use the same strategies in the game as in the optimal mechanism. Fix the optimal (direct) mechanism and a profile of types t. In the mechanism, given this profile, the agents will report t truthfully and will present maximal evidence. The mechanism specifies a response to this, say $a^*(t)$. In the game, given this same profile of types t, the agents will send some reports, typically not truthful, and some evidence, typically not maximal. Furthermore, the principal's response to a given profile of reports and evidence will not generally be what he would commit to in the mechanism. For example, in the mechanism, he may commit to disregarding certain evidence, something he cannot do in the equilibrium of the game.[6]

In the game, the principal reacts to the reports and evidence by forming a belief based on the agents' equilibrium strategies and choosing a best action for himself conditional on these beliefs, say $\hat{a}(t)$. The surprising result is that the equilibrium and optimal mechanism we construct have the property that $a^*(t) = \hat{a}(t)$ for every profile t.

Formally, the game without commitment is as follows. The strategy set for agent i, Σ_i, is the set of functions $\sigma_i : T_i \to \Delta(T_i \times \mathcal{E}_i)$ such that $\sigma_i(s_i, e_i \mid t_i) > 0$ 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.[7] The principal's strategy set, Σ_P, is the set of functions $\sigma_P : T \times \mathcal{E} \to \Delta(A)$. A belief by the principal is a function $\mu : T \times \mathcal{E} \to \Delta(T)$ giving the principal's beliefs about t as a function of the profile of reports and evidence presentation.

[6]While the agents' strategies in the equilibrium differ from their strategies in the direct mechanism, there is an indirect mechanism with the same outcome and the same strategies. Specifically, take the indirect mechanism defined by having the principal commit to his equilibrium strategy of the game without commitment. Clearly, it is an equilibrium of this mechanism for the agents to play the same strategies as in the game without commitment, giving an indirect mechanism with the same outcome and the same strategies by the agents.

[7]We do not require t_i to report truthfully and do not require his claim of a type to be consistent with the evidence he presents. That is, we could have $\sigma_i(s_i, e_i \mid t_i) > 0$ even though $s_i \neq t_i$ and $e_i \notin \mathcal{E}_i(s_i)$.

We study perfect Bayesian equilibria of the game without commitment. 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. See Part SB of the Supplemental Material for details.

The equilibria of interest also satisfy a robustness property. We call a perfect Bayesian equilibrium (σ, μ) *robust* if, for every i and every $t_i \in T_i$, $\sigma_i(s_i, e_i \mid t_i) > 0$ implies

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

That is, $\sigma_i(t_i)$ is optimal for t_i given *any* actions by the other agents and the *equilibrium* strategy of the principal. A robust equilibrium generates an equilibrium in any of a wide range of other games—for example, where agents report sequentially with each agent observing the earlier reports. Similarly, there is no need for any agent to know the preferences or evidence of other agents, even as a function of their types.

Given a perfect Bayesian equilibrium (σ, μ), the *equilibrium outcome* is the function $O_{(\sigma,\mu)} : T \to \Delta(A)$ given by

$$O_{(\sigma,\mu)}(t)(a) = \sum_{(s,e) \in T \times \mathcal{E}} \prod_i \sigma_i(s_i, e_i \mid t_i) \sigma_P(a \mid s, e).$$

In other words, analogously to the mechanism outcome, the equilibrium outcome gives the probability distribution over A as a function of t generated by the equilibrium strategies. Given (σ, μ), the principal's expected utility is

$$\mathrm{E}_t \sum_a O_{(\sigma,\mu)}(t)(a) v(a, t).$$

We show that there is a robust perfect Bayesian equilibrium of this game (σ, μ) and an optimal mechanism P with the same outcome—that is, such that $O_P(\cdot) = O_{(\sigma,\mu)}(\cdot)$. In this sense, the principal does not need commitment.

The proof constructs an equilibrium from a set of I one-agent games which do not depend on A or preferences over A. Specifically, we define the *auxiliary game for agent i* as follows. There are 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 $[\min_j \min_{t_j \in T_j} v_j(t_j), \max_j \max_{t_j \in T_j} v_j(t_j)]$. The principal's utility given t_i and x is $-(x - v_i(t_i))^2$, while agent i's payoff is

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

In other words, the principal's action, x, is his "estimate" of $v_i(t_i)$. Positive types of the agent prefer larger estimates of $v_i(t_i)$ and negative types have the opposite preference. As in the game without commitment, 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 $X_i : T_i \times \mathcal{E}_i \to X$. By strict concavity of the principal's utility function in x, he has a unique optimal pure strategy given any belief. Hence he will never mix in equilibrium, so we only consider pure strategies for him.

To see the link between the auxiliary games and the game without commitment, recall that the principal's utility function is $\sum_i u_i(a)v_i(t_i)$. Hence, given some belief about each t_i, the principal maximizes the sum of the $u_i(a)$'s weighted by his expectation of $v_i(t_i)$. If the principal's belief about t_i goes up in the sense of generating a higher expected value of $v_i(t_i)$, then his action choice changes in the direction of increasing $u_i(a)$. A positive type is made better off by this, while a negative type is hurt. Hence positive types want to persuade the principal that v_i is large and negative types want him to believe it is small, incentives captured by the auxiliary game.

2.5. *Results: Commitment, Determinism, and Robust Incentive Compatibility*

Our main results are stated in the following theorem.

THEOREM 1: *If every u_i exhibits simple type dependence, then 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 of the game without commitment with the same outcome as in this optimal mechanism. In this equilibrium, agent i's strategy is also a perfect Bayesian equilibrium strategy in the auxiliary game for agent i.*

Theorem 1 is proved in the Appendix. See Section 4 for a proof sketch.

These results tell us that we can use equilibrium analysis to characterize the optimal mechanism. By Theorem 1, the outcome of the best perfect Bayesian equilibrium for the principal in the game without commitment is the same as the outcome of the best mechanism for the principal. This identifies how the mechanism must respond to any profile of type reports t when the evidence presented is the maximal evidence for t. To finish identifying the optimal mechanism, we only need to specify its response to profiles of type reports t some of which are accompanied by the "wrong" evidence.

Also, we can use the auxiliary games to identify the information revealed by the agents in the game without commitment. From this, we can compute the best reply of the principal, completing the specification of the equilibrium of the game without commitment, an approach we illustrate in the next section.

3. OPTIMAL MECHANISMS WITH DYE EVIDENCE

3.1. *Characterizing the Optimal Mechanism*

In this section, we show how one can use equilibria in the auxiliary games to characterize optimal mechanisms with Dye's (1985) evidence structure, a structure extensively studied in the economics and accounting literatures. We also show that this characterization can be used to characterize optimal mechanisms in a different setting. Specifically, we show that in certain models without evidence but where the principal can verify the type of an agent at a cost, the optimal mechanism can be computed from the optimal mechanism for an associated Dye-evidence model.

We say 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 has access to perfect evidence and can choose between proving nothing (proving T_i) and proving her type. Let T_i^0 denote the set of $t_i \in T_i$ with $\mathcal{E}_i(t_i) = \{T_i\}$. We sometimes refer to these types as having no evidence and types with $\mathcal{E}_i(t_i) = \{T_i, \{t_i\}\}$ as having evidence.

A complication is that the auxiliary games have multiple, essentially equivalent equilibria. Since type reports are cheap talk, any permutation of agent i's type reports and the principal's interpretation of them yields another equilibrium. Note, though, that this permutation does not affect the information the principal acquires about i's type or his choices given his information.

Given a perfect Bayesian equilibrium (σ_i^*, X_i^*) of the auxiliary game for agent i, define the equilibrium outcome to be the function $O^i_{(\sigma_i^*, X_i^*)} : T_i \to \Delta(\mathbf{R})$ given by[8]

$$O^i_{(\sigma_i^*, X_i^*)}(t_i)(x) = \sum_{(s_i, e_i) \in T_i \times \mathcal{E}_i | X_i^*(s_i, e_i) = x} \sigma_i^*(s_i, e_i \mid t_i).$$

Two equilibria of the auxiliary game are *essentially equivalent* if they generate the same equilibrium outcome. If there is an equilibrium such that every other equilibrium is essentially equivalent to it, the equilibrium is *essentially unique*.

First consider *type-independent utility* where $u_i(a, t_i)$ is independent of t_i for all i. That is, $T_i^- = \emptyset$, so $u_i(a, t_i) = u_i(a)$ for all t_i.

The following builds on well-known characterizations of equilibria with Dye evidence.

THEOREM 2: *Given Dye evidence, for every i, there exists a unique v_i^* such that*

$$v_i^* = \mathrm{E}_{t_i}\big[v_i(t_i) \mid t_i \in T_i^0 \text{ or } v_i(t_i) \le v_i^*\big].$$

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

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

To see this, note first that cheap talk is not credible since every type wants the principal to believe that v_i is large. Also, 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 $v_i(t_i)$ pool with the types without evidence, leading to an expectation of $v_i(t_i)$ equal to v_i^*.

In equilibrium, the principal's expectation of $v_i(t_i)$ is v_i^* given a type who presents no evidence and equals the true value otherwise. Let

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

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

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

That is, $\hat{p}(\cdot \mid \hat{v})$ is an optimal distribution over A for the principal when $\hat{v}$ is his profile of expectations of the v_i's. Then an equilibrium outcome of the game without commitment is the function $O(t)(a) = \hat{p}(a \mid \hat{v}(t))$.

[8] Since each T_i is finite, the set of type reports and the set of events that can be proven are also finite.

COROLLARY 1: *With type-independent utility and Dye evidence, there is an optimal mechanism P with mechanism outcome $O_P(t) = \hat{p}(\cdot \mid \hat{v}(t))$. Thus $\hat{p}(\cdot \mid \hat{v}(t))$ is both an optimal mechanism outcome and an equilibrium outcome.*

Corollary 1 yields characterizations of optimal mechanisms in many interesting cases.

EXAMPLE 1—The simple allocation problem with Dye evidence: Here $\hat{p}(i \mid t) > 0$ iff $\hat{v}_i(t_i) = \max_j \hat{v}_j(t_j)$, so the good is given to an agent with the highest $\hat{v}_j(t_j)$.

One way to turn this outcome function into a specification of a mechanism yields a *favored-agent mechanism*. 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).

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 $v_j(t_j) \leq v_j^*$, 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 we discuss below, this mechanism is reminiscent of the favored-agent mechanism discussed by Ben-Porath, Dekel, and Lipman (2014) (BDL) for the allocation problem with costly verification.

EXAMPLE 2—The multi-unit allocation problem with Dye evidence: Suppose the principal has $K < I$ identical units to allocate and that he must allocate all of them. Suppose each agent can have either 0 or 1 unit. Then the principal's action is selecting a set $\hat{\mathcal{I}} \subset \{1, \ldots, I\}$ of cardinality K specifying which agents get a unit. The principal's utility is $\sum_{i \in \hat{\mathcal{I}}} v_i(t_i)$. Again, agent i's utility is 0 if she does not get a unit and 1 if she does. So the principal allocates units to the K agents with the highest values of $\hat{v}_i(t_i)$ as computed above. One can interpret this as a recursive favored-agent mechanism.[9]

EXAMPLE 3—Allocating a "bad": Suppose the principal has to choose one agent to carry out an unpleasant task (e.g., serve as department chair). This problem is equivalent to having $I - 1$ goods to allocate since not receiving the assignment is receiving a good. One can apply the analysis of the previous example for $K = I - 1$ to characterize the optimal mechanism.

Turning to simple type dependence, consider the auxiliary game for i where some types wish to persuade the principal that $v_i(t_i)$ is large and others that $v_i(t_i)$ is small. Suppose that when the agent does not prove her type, she makes a cheap-talk claim either that 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

[9]Specifically, 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. After removing this agent and unit, we follow the same procedure for the second unit, and so on. 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, the agent with the second-highest v_i^* is the second-most favored, etc.

but says it is positive and let v_i^- be the analog for a negative declaration. If $v_i^+ > v_i^-$, then positive types prefer to truthfully report they are positive and similarly negative types prefer truthful reporting. If i is a positive type with evidence, she will prove her type only if $v_i(t_i) > v_i^+$, while a negative type with evidence will prove her type only if $v_i(t_i) < v_i^-$. For this to be an equilibrium, we must have

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

and

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

Suppose this gives a unique v_i^+ and v_i^-. If these values satisfy $v_i^+ < v_i^-$, we cannot have such an equilibrium as the positive types without evidence will imitate the negative and vice versa. Hence all types who do not present evidence must pool. (The pooling strategies are described further in Lemma 1 and Theorem 3.) If $v_i^+ \ge v_i^-$, then these strategies form an equilibrium. When $v_i^+ = v_i^-$, the cheap talk does not convey any extra information, so this is effectively the same as pooling. When $v_i^+ > v_i^-$, cheap talk is useful, but there is another equilibrium as well where cheap talk is treated as "babbling," as in all models with cheap talk.

The following lemma provides the background for the equilibrium characterization.

LEMMA 1: *With Dye evidence, for every i, there exist a unique v_i^+, v_i^-, and v_i^* such that*

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

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

and

$$v_i^* = \mathrm{E}_{t_i}\left[v_i(t_i) \mid \left(t_i \in T_i^0\right) \textit{ or } \left(t_i \in T_i^- \setminus T_i^0 \textit{ and } v_i(t_i) \ge v_i^*\right) \textit{ or } \left(t_i \in T_i^+ \setminus T_i^0 \textit{ and } v_i(t_i) \le v_i^*\right)\right].$$

THEOREM 3: *If $v_i^+ \le v_i^-$, then there is an essentially unique equilibrium in the auxiliary 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*

$$e_i^*(t_i) = \begin{cases} T_i & \textit{if } t_i \in T_i^0 \textit{ or } \left(t_i \in T_i^+ \textit{ and } v_i(t_i) \le v_i^*\right) \textit{ or } \left(t_i \in T_i^- \textit{ and } v_i(t_i) \ge v_i^*\right), \\ \{t_i\}, & \textit{otherwise}. \end{cases}$$

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. The first is the same strategy profile as above. In the second equilibrium, there are types $\hat{s}_i^+$ and $\hat{s}_i^-$ with $\hat{s}_i^+ \ne \hat{s}_i^-$ such that $t_i \in T_i^k$ sends $(\hat{s}_i^k, e_i^k(t_i))$, $k \in \{-, +\}$, where

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

and

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

When $v_i^+ > v_i^-$, we can always compare the two equilibria for the principal and we show that he prefers the one which separates the positive and negative types. Hence this equilibrium corresponds to the optimal mechanism. We characterize the principal's beliefs about v_i as a function of the true type t_i along the equilibrium path, $\hat{v}_i(t_i)$, as follows. If $v_i^+ > v_i^-$, we let

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

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

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

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

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

Then an equilibrium outcome of the game without commitment is the function $O(t)(a) = \hat{p}(a \mid \hat{v}(t))$.

COROLLARY 2: *In any model with simple type dependence and Dye evidence, there is an optimal mechanism P with mechanism outcome $O_P(t) = \hat{p}(\cdot \mid \hat{v}(t))$. 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 does not follow from Theorems 1 and 3 is the claim that when $v_i^+ > v_i^-$, the better equilibrium for the principal is the one that separates the positive and negative types. This is shown in Part SC of the Supplemental Material.

EXAMPLE 4—The public-goods problem: Consider the public goods model from Section 1. In equilibrium, given a profile of types t, the principal's expectation of v_i is $\hat{v}_i(t_i)$ defined in equation (1). The principal provides the public good iff $\sum_i \hat{v}_i(t_i) > 0$. Again, this describes the optimal outcome function; the rest of the optimal mechanism is straightforward.

While Example 1 above is reminiscent of Ben-Porath, Dekel, and Lipman's (2014) (BDL) analysis of allocation with costly verification, the optimal mechanism in Example 4 is reminiscent of the optimal mechanism under costly verification identified by Erlanson and Kleiner (2017) which leads us to discuss this connection more generally.

3.2. *Costly Verification*

BDL (2014) and Erlanson and Kleiner (2017) modeled costly verification by assuming the principal can pay a cost c_i to "check" or learn the realization of agent i's type, t_i. The agent cannot affect this process. By contrast, in the evidence model we consider here, the principal cannot acquire information about an agent without inducing the agent to reveal it.

Yet the optimal mechanisms in these papers look very similar to optimal mechanisms with Dye evidence. Compare BDL's optimal mechanism in the costly-verification version of the simple allocation problem to the mechanism in Example 1. In both cases, there is a favored agent and a threshold. If no non-favored agent "reports" above the threshold, the favored agent receives the object. Here, "reporting above the threshold" means to prove a value of $v_i(t_i)$ above the threshold. In BDL, it means to make a cheap-talk report of a type such that the type minus the checking cost is above the threshold. In both, if some non-favored agent "reports" above the threshold, the good goes to the agent with the highest such report. In the costly verification model, this is after checking this type.

Similarly, Erlanson and Kleiner considered the public goods model under costly verification. In their mechanism and in the optimal mechanism here when $v_i^+ > v_i^-$ for all i, we compute "adjusted reports" for each agent i given t_i. In both cases, the adjusted report for a positive type is $\max\{v_i^+, v_i(t_i)\}$, while the adjusted report for a negative type is $\min\{v_i^-, v_i(t_i)\}$ for certain cutoffs v_i^+ and v_i^-. Again, the difference between these scenarios is that the report is proven in the evidence model and is a cheap-talk claim adjusted by the verification cost in the costly-verification model. In both problems, these reports are summed to determine the principal's optimal action. Again, this includes some checking in the costly-verification model.

We generalize to show that certain costly-verification models can be rewritten as a Dye-evidence model, so that the optimal mechanism can be computed from our results about mechanisms with evidence. In the text, we explain this for the simple allocation problem. We give the general result and explain the connection to Erlanson and Kleiner in Appendix C. This connection does not imply that all properties of evidence models, such as the fact that the principal does not need commitment, carry over to costly-verification models.

So consider the simple allocation problem. For simplicity, assume $v_i(t_i) > 0$ for all t_i and all i and that no two types have the same value of $v_i(t_i)$. Now agents do not have evidence, but the principal can pay a cost $c_i > 0$ to learn the type of agent i, called *checking* i. BDL showed that an optimal mechanism specifies functions $p : T \to \Delta(\{1, \ldots, I\})$ and $q_i : T \to [0, 1]$ where $p(t)$ is the probability distribution over which agent the principal gives the good to and $q_i(t)$ gives the probability that the principal checks i given type reports t. The principal's objective function is

$$\mathrm{E}_t\left[\sum_i p_i(t)v_i(t_i) - q_i(t)c_i\right],$$

where $p(t) = (p_1(t), \ldots, p_I(t))$. The incentive compatibility constraints are

$$\hat{p}_i(t_i) \geq \hat{p}_i\big(t_i'\big) - \hat{q}_i\big(t_i'\big), \quad \forall t_i, t_i' \in T_i, \forall i,$$

where $\hat{p}_i(t_i) = \mathrm{E}_{t_{-i}} p_i(t)$ and $\hat{q}_i(t_i) = \mathrm{E}_{t_{-i}} q_i(t)$. To see this, note that if type t_i reports truthfully, he receives the good with expected probability $\hat{p}_i(t_i)$. If he misreports and claims to be type t_i', he is checked with expected probability $\hat{q}_i(t_i')$. In this case, the principal learns he has lied and does not give him the good. Thus, his probability of receiving the good is the same as t_i''s probability minus the probability of being checked.

For each i, let t_i^0 be the type with the smallest value of $v_i(t_i)$. It is not hard to show that the solution satisfies $\hat{p}_i(t_i) \geq \hat{p}_i(t_i')$ if $v_i(t_i) \geq v_i(t_i')$. Hence, if incentive compatibility holds for type t_i^0, then it holds for every other type of agent i. So we can rewrite incentive compatibility as

$$\hat{q}_i\big(t_i'\big) \geq \hat{p}_i\big(t_i'\big) - \hat{p}_i\big(t_i^0\big), \quad \forall t_i' \in T_i, \forall i.$$

The optimal solution sets $\hat{q}_i$ as small as possible since checking is costly, so $\hat{q}_i(t_i) = \hat{p}_i(t_i) - \hat{p}_i(t_i^0)$ for all t_i. Hence the objective function is

$$\sum_i \mathrm{E}_{t_i}\big[\hat{p}_i(t_i)v_i(t_i) - \hat{q}_i(t_i)c_i\big] = \sum_i \mathrm{E}_{t_i}\big[\hat{p}_i(t_i)\big(v_i(t_i) - c_i\big) + \hat{p}_i\big(t_i^0\big)c_i\big].$$

Thus, we can solve the principal's problem by choosing p to maximize the above subject to $\hat{p}_i(t_i) \geq \hat{p}_i(t_i^0)$ for all $t_i \in T_i$ and all i. We can write the objective function as

$$\sum_i \mathrm{E}_{t_i}\big[\hat{p}_i(t_i)\tilde{v}_i(t_i)\big] = \mathrm{E}_t\left[\sum_i p_i(t)\tilde{v}_i(t_i)\right],$$

where

$$\tilde{v}_i(t_i) = \begin{cases} v_i(t_i) - c_i & \text{if } t_i \neq t_i^0, \\ v_i\big(t_i^0\big) - c_i + \dfrac{c_i}{\rho_i\big(t_i^0\big)} & \text{if } t_i = t_i^0. \end{cases}$$

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

This is the same objective function as for the simple allocation problem with Dye evidence where the value to the principal of allocating the good to agent i is $\tilde{v}_i(t_i)$. Construct the evidence functions by assuming $\mathcal{E}_i(t_i^0) = \{T_i\}$ and $\mathcal{E}_i(t_i) = \{\{t_i\}, T_i\}$ for all $t_i \neq t_i^0$. In this case, the incentive compatibility constraint is $\hat{p}_i(t_i) \geq \hat{p}_i(t_i^0)$, just as in the costly-verification model. We can apply our characterization of optimal mechanisms with Dye evidence to obtain the solution to this problem. One can then "invert" the $\tilde{v}_i$'s, writing the solution as a function of the v_i's, to give the solution for the costly-verification model.

Specifically, for each i, define the cutoffs $\tilde{v}_i^*$ from the $\tilde{v}_i$ functions as before—that is, $\tilde{v}_i^*$ is the expectation of $\tilde{v}_i$ conditional on t_i not having evidence (type t_i^0) or having $\tilde{v}_i(t_i) \leq \tilde{v}_i^*$. As shown above, the optimal mechanism for this problem with evidence is to select a favored agent who has $\tilde{v}_i^* \geq \tilde{v}_j^*$ for all $j \neq i$, set threshold $\tilde{v}_i^*$, giving the good to i if $\tilde{v}_j(t_j) \leq \tilde{v}_i^*$ for all $j \neq i$ and to that agent j who maximizes $\tilde{v}_j(t_j)$ otherwise. It is easy to show that this is equivalent to the optimal mechanism in BDL.

This approach yields optimal mechanisms with costly verification for Examples 2 and 3 and the model of Erlanson and Kleiner, as discussed in Appendix C.

4. UNDERSTANDING THE RESULTS

We provide intuition for our results in two ways. In Section 4.1, we sketch the proof of Theorem 1 in the context of the simple allocation problem. In Section 4.2, we discuss the roles our assumptions play in the results.

4.1. *Proof Sketch for Simple Allocation Problem*

One simplification in type-independent settings like the simple allocation problem is that we can write a mechanism as a function only of type reports, where it is understood that if i claims type t_i, she also reports maximal evidence for t_i, $M_i(t_i)$. If i claims type t_i but does not show evidence $M_i(t_i)$, type independence implies that the principal knows the worst possible outcome for i—here, not giving her the good—and can use this to punish. This deters any "obvious" deviations, leaving only more subtle deviations of the form of reporting some $s_i \neq t_i$ and providing evidence $M_i(s_i)$. So for this proof sketch, a mechanism is a function $P : T \to \Delta(A)$.

Fix an optimal mechanism P. The probability that type t_i receives the good under P is

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

where action $a = i$ is the action of the principal to give the good to agent i. Partition each T_i according to equality under $\hat{p}_i$. That is, for each $\alpha \in [0, 1]$, let

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

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 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. We call $\mathcal{T}$ the *mechanism partition*.

Incentive compatibility is equivalent to the statement that $M_i(s_i) \in \mathcal{E}_i(t_i)$ implies $\hat{p}_i(t_i) \geq \hat{p}_i(s_i)$. That is, if t_i can imitate s_i 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. Hence, if $M_i(s_i) \in \mathcal{E}_i(t_i)$, $t_i \in T_i^\alpha$, and $s_i \in T_i^\beta$, we must have $\alpha \geq \beta$.

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

To see why this property holds, suppose it is violated. In other words, suppose we have a 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 $\{s_i, s_i'\}$. That is, there is $t_{-i} \in T_{-i}$ 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. More precisely, if q is the probability of type s_i conditional on $\{s_i, s_i'\}$, then for every $a \in A$ and $t_{-i} \in T_{-i}$, we set

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

By assumption, the payoffs to agents $j \neq i$ do not 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 by this change.

So consider agent i. Her payoff from reporting anything other than s_i or s_i' is unchanged. The expected payoff from reporting s_i was $\hat{p}_i(s_i)$ in the original mechanism, while the expected payoff from reporting s_i' was $\hat{p}_i(s_i')$. The new mechanism "averages" these two types together, so 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 $\hat{p}_i(s_i) = \hat{p}_i(s_i')$, the probability 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.[10]

Finally, consider the principal. Recall that his utility function is

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

[10]More generally, suppose all types have the same indifference curves. Then if s_i is indifferent between reporting s_i or claiming to be type s_i', s_i' would also be indifferent between these reports. Hence neither type's payoff changes if we replace the response to either report with the averaged response. This is a key implication of simple type dependence.

Under the original mechanism, the principal's expected payoff is

$$\mathrm{E}_t \sum_a P(a \mid t) \sum_j u_j(a) v_j(t_j) = \sum_j \mathrm{E}_{t_j} \left[\mathrm{E}_{t_{-j}} \sum_a P(a \mid t) u_j(a) \right] v_j(t_j).$$

But since $u_j(a)$ is 1 if $a = j$ and 0 otherwise,

$$\mathrm{E}_{t_{-j}} \sum_a P(a \mid t) u_j(a) = \hat{p}_j(t_j),$$

so the principal's expected payoff in the original mechanism is just $\sum_j \mathrm{E}_{t_j} \hat{p}_j(t_j) v_j(t_j)$. Since the probability t_j receives the good is unchanged in the new mechanism for every j and every type $t_j \in T_j$, the expected payoff of the principal is unchanged. Hence P^* is also an optimal mechanism. Repeating as needed, we construct an optimal mechanism which is measurable with respect to $\mathcal{T}$.

This property is critical because we can construct an equilibrium of the game without commitment where the principal obtains at least the information embodied in the mechanism partition. Since the optimal mechanism is measurable with respect to this partition, this means the principal receives enough information to carry out the optimal mechanism. We construct such an equilibrium and use it to complete the proof.

Specifically, we use the auxiliary games to construct the equilibrium strategies. This construction has four steps. First, we consider equilibria in the *restricted auxiliary game for i*. In this game, type t_i is restricted to sending evidence which is maximal for some s_i in the same event of the mechanism partition as t_i. That is, if $s_i, t_i \in T_i^\alpha$ for some α, then in the restricted auxiliary game for i, t_i can send evidence $M_i(s_i)$ if $M_i(s_i) \in \mathcal{E}_i(t_i)$. For $s_i' \notin T_i^\alpha$, t_i cannot send evidence $M_i(s_i')$ even if $M_i(s_i') \in \mathcal{E}_i(t_i)$. The principal's action in this game is the choice of a number x where his payoff is $-(x - v_i(t_i))^2$ and the agent i's utility is x, as in the unrestricted case described above. In the restricted game, the principal *must* learn at least that $t_i \in T_i^\alpha$ since, by construction, the only messages available to t_i reveal that $t_i \in T_i^\alpha$.

Second, we show that, given this information, the principal cannot do better than to implement the outcome of the mechanism. More specifically, for each i, fix an equilibrium of the restricted auxiliary game. For any $t_i \in T_i$, t_i's equilibrium strategy in the restricted game for i determines the principal's equilibrium expected value of v_i which we denote $\hat{v}_i(t_i)$.[11] For a profile of types t, let $\hat{v}(t) = (\hat{v}_1(t_1), \ldots, \hat{v}_I(t_I))$. Typically, the evidence presented will not reveal the type profile t, but must reveal at least the event of the mechanism partition containing t and hence what the optimal mechanism specifies given t. The second step is to show that for every type profile t, following the allocation prescribed by the optimal mechanism for this type profile is optimal for the principal when his expectation of v is $\hat{v}(t)$. In this sense, the equilibrium does not give him information he can use to improve on the mechanism.

To see this, suppose to the contrary that there is a strategy $p^* : \mathbf{R}^I \to \Delta(A)$ for the principal as a function of the expected values $\hat{v}$ which gives him a strictly higher expected payoff than the optimal mechanism. Consider the following alternative mechanism. Given reports $(t, M(t))$, the principal chooses the allocation $p^*(\hat{v}(t))$ with probability ε and the

[11] If agent i's equilibrium strategy in the restricted game is mixed, optimality for i requires that the principal has the same belief in response to every pure strategy in the support. Hence the principal's belief in response to the equilibrium strategy of any type is unambiguously defined.

original mechanism $P(\cdot \mid t, M(t))$ otherwise. If following the alternative strategy yields the principal a strictly higher expected payoff than following the mechanism, then this mechanism, if incentive compatible, yields a higher payoff than the optimal mechanism. Since this is a contradiction, the new mechanism must not be incentive compatible.

But the new mechanism is incentive compatible. To see this, fix any $s_i, t_i \in T_i$ with $M_i(s_i) \in \mathcal{E}_i(t_i)$. By incentive compatibility of P, we must have $\hat{p}_i(t_i) \geq \hat{p}_i(s_i)$. If this inequality is strict, then for ε sufficiently small, t_i prefers not to imitate s_i in the new mechanism. So suppose $\hat{p}_i(t_i) = \hat{p}_i(s)$, so that t_i and s_i are in the same event of the mechanism partition. It is easy to show that $M_i(s_i) \in \mathcal{E}_i(t_i)$ implies $\mathcal{E}_i(s_i) \subseteq \mathcal{E}_i(t_i)$. Hence in the restricted auxiliary game, t_i must get a weakly larger payoff than s_i. That is, we must have $\hat{v}_i(t_i) \geq \hat{v}_i(s_i)$. But then p^* must give the good to t_i at least as often as s_i. Therefore, t_i gets the good weakly more often than s_i in the new mechanism, so it is incentive compatible, a contradiction.

This result also has implications for the optimal mechanism. Since the alternative strategy p^* must give the good to one of the agents with the highest expected v_i, the optimal mechanism must be doing the same. Otherwise, it would give the principal a lower expected payoff. One implication of this is that if $t_i \in T_i^\alpha$ and $s_i \in T_i^\beta$ for $\alpha > \beta$, then we must have $\hat{v}_i(t_i) \geq \hat{v}_i(s_i)$. (Recall that T_i^α is the set of t_i who receive the good with probability α in the optimal mechanism.). If $\hat{v}_i(t_i) < \hat{v}_i(s_i)$, then given the information revealed by the restricted auxiliary game equilibria, the principal would want to give the good to s_i at least as often as to t_i. But the optimal mechanism gives t_i the good strictly more often and gets the same payoff as p^*, so this cannot hold.

The third step is to show that by appropriately specifying beliefs in response to evidence which has zero probability in the restricted auxiliary game, we obtain an equilibrium of the *unrestricted auxiliary game for i*, where the payoffs are the same as in the restricted game but where t_i can send any evidence she possesses. Specifically, in the unrestricted auxiliary game for i, if i presents evidence e_i which is off path in the sense that it is not presented by any type in equilibrium, then the principal responds by setting $x = \min_{t_i | e_i \in \mathcal{E}_i(t_i)} v_i(t_i)$.

To see that this gives an equilibrium of the unrestricted auxiliary game, consider a deviation by type t_i to a message that was not available to her in the restricted game. First, consider a deviation to evidence which is not chosen in the equilibrium of the restricted auxiliary game by any type. Since t_i could present $M_i(t_i)$ in the restricted game, her equilibrium payoff must be at least $\min_{s_i | M_i(t_i) \in \mathcal{E}_i(s_i)} v_i(s_i)$. Since $M_i(t_i)$ rules out the largest number of types t_i can rule out,

$$\min_{s_i | M_i(t_i) \in \mathcal{E}_i(s_i)} v_i(s_i) \geq \min_{s_i | e_i \in \mathcal{E}_i(s_i)} v_i(s_i),$$

for any $e_i \in \mathcal{E}_i(t_i)$. Hence sending $M_i(t_i)$ yields a weakly higher payoff than any off path evidence t_i can send, so t_i would not deviate to such evidence.

So consider a deviation to evidence t_i could not have used in the restricted game but which is used in equilibrium by some other type, s_i. That is, if t_i is in partition event T_i^α, then the deviation is to $M_i(s_i)$ which is sent in equilibrium by type t_i' where s_i need not equal t_i'. Since this is evidence t_i could not have used in the restricted game, we must have s_i and t_i' in partition event T_i^β for $\beta \neq \alpha$. Since t_i can send s_i's maximal evidence, incentive compatibility implies $\alpha > \beta$. As noted above, this implies $\hat{v}_i(t_i) \geq \hat{v}_i(s_i)$, so t_i does not gain from the deviation.

The final step in constructing an equilibrium of the game without commitment is to put the pieces together. Set the agents' strategies and the principal's beliefs to be those in the equilibria of the unrestricted auxiliary games. Similarly to the construction above, for

each $\hat{v} = (\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_I)$, let $a^*(\hat{v})$ select one of the agents with the highest $\hat{v}_i$ to give the good to. Given any reports that lead in equilibrium to expected values $\hat{v}$, the principal's strategy is $a^*(\hat{v})$. Clearly, this strategy is sequentially rational for the principal. To see that this gives a robust equilibrium, fix any reports by the agents other than i. Obviously, the report for i which maximizes the principal's expected value of v_i will maximize her probability of getting the good. But this means that i will follow her equilibrium strategy, regardless of the reports of the other agents, giving us a robust equilibrium.

Note that the outcome of this equilibrium is not necessarily the same as the outcome of the optimal mechanism that was our starting point. However, the fact that the principal receives at least the information he needs to follow the optimal mechanism implies that his payoff in this equilibrium must be at least that in the optimal mechanism. Since it cannot be strictly larger, we see that the principal's payoff in this equilibrium is the same as in the optimal mechanism.

Hence if the principal commits to the strategy he uses in this equilibrium, we obtain an *indirect* mechanism with the same payoff as the equilibrium. As in the standard mechanism design model, it is not difficult to turn this into a direct mechanism with the same outcome. Note that the principal's strategy is deterministic in the equilibrium (both on and off the equilibrium path) and hence the implied mechanism is deterministic. It is also not hard to see that the robustness of the equilibrium implies that the mechanism is robustly incentive compatible.

4.2. *Role of Assumptions*

In this subsection, we explain the roles of our assumptions in generating the results. Part SD of the Supplemental Material illustrates these points with examples showing which results fail when we drop various assumptions.

First, consider the robustness properties. These properties say that each agent's optimal strategy does not depend on the other agents' types or behavior. Clearly, the independence of types across agents and the private-values assumption that agent i's utility depends only on t_i and a play important roles. If types are correlated, it will be optimal for the principal to use reports by one agent to help enforce incentive compatibility for others, so each agent's optimal strategy will depend on her beliefs about the others. Similarly, if an agent's utility depends on the types of other agents, her optimal strategy will depend on her beliefs about their types.

The functional form of the principal's utility function is also important for robustness. Given any belief about the types of the agents, the principal will choose a to maximize a weighted sum of the $u_i(a)$'s with weights given by the expectations of the $v_i(t_i)$'s. This implies that if the principal's expectation of $v_i(t_i)$ increases, his optimal action changes in the direction of increasing $u_i(a)$. Thus, agent i's incentives to signal about her type depend only on her preferences regarding the principal's expectation of $v_i(t_i)$, independently of his beliefs about the types of the other agents. Without this, robustness is unlikely: if what agent i wants the principal to believe about t_i depends on the principal's beliefs about t_{-i}, then i's optimal strategy depends on her beliefs about the other agents.

To clarify, the change in i's utility from changing the principal's beliefs about t_i depends on the principal's beliefs about t_{-i}. For example, in the simple allocation problem, whether an increase in the principal's expectation of $v_i(t_i)$ changes i's utility depends on the beliefs about the other agents. However, while the *magnitude* of the change in utility depends on the principal's beliefs about t_{-i}, the *sign* does not. Thus, for a positive type of i, increasing the principal's expectation of v_i has a positive effect for all t_{-i}.

Except for the private-values assumption, our assumptions on the form of the agent's utility are not essential for robustness. For example, suppose the principal's set of feasible actions A is a product space $A_1 \times \cdots \times A_I$. Write a typical action $a \in A$ as $a = (a_1, \ldots, a_I)$ where agent i's utility depends only on her type and a_i and the principal's utility function is $\sum_i v_i(a_i, t_i)$. In this case, we effectively have I different principal–agent problems and robust incentive compatibility will not cost the principal anything, regardless of what else we assume about the agents' utility functions.

Simple type dependence and our assumptions on the principal's utility function are both important for our result that commitment is not necessary. Conceptually, we can separate this result into two pieces. First, along the equilibrium path of the game without commitment, the principal finds it optimal to implement the outcome of the optimal mechanism. Second, there are beliefs for the principal off the equilibrium path which make it sequentially rational for him to choose outcomes which deter such deviations by the agents. Simple type dependence plays an important role in both parts.

For the first part, consider the one-agent case and contrast our analysis with the classical indifference curve analysis of a mechanism design problem as in Mas-Colell, Whinston, and Green's (1995) treatment of principal–agent models with adverse selection. In the classical analysis, one uses differences in the indifference curves for high and low types to identify incentive compatible allocations and then the principal optimizes over these. With simple type dependence, *all* types have the same indifference curves, so this approach does not work. If incentive compatibility were driven by differences in indifference curves, we would not obtain our result in general. To see why, consider the two-type case. Suppose, as in the classical adverse selection problem, the optimal mechanism gives the two types different allocations and that the incentive compatibility constraint is binding only for one type. That is, type t (omitting subscripts as we have one agent) is indifferent between the allocation for t and for t' and the constraint binds in the sense that the allocation for t' is not first-best. Then commitment is necessary. Without commitment, there is no game where the principal learns which type he is facing (necessary to choose different allocations for the two types) and does not choose the first-best allocation for type t'.

When indifference curves are the same and incentive compatibility is achieved by evidence, this situation cannot arise. If two types are given different allocations in the optimal mechanism, it cannot be true that one is indifferent between these allocations and the other type is not. If both types are indifferent between the two allocations, as our proof shows, our assumptions on the principal's utility function imply that he obtains the same payoff from giving the "expected allocation" to both types and so does not need to separate them. If both types are not indifferent, there are two possibilities. First, suppose they have the same preferences—either both are positive or both are negative. In this case, the type with the better allocation must have evidence the other type lacks. This use of evidence to separate the types is equally available in a mechanism or a game. Second, suppose one type is positive and the other negative. Then they have the opposite preferences regarding these allocations. If each strictly prefers the allocation she gets, it is easy to separate them, either in a mechanism or in equilibrium. If each type strictly prefers the other's allocation, it must be that each has evidence unavailable to the other which can be used to separate them, evidence which again is equally available in a mechanism or a game.

Simple type dependence also allows us to address off path behavior. For intuition, first, consider the simpler case of type independence. When the utility functions of the agents do not depend on their types, the principal knows how to punish deviations. To prevent deviations while maintaining sequential rationality, choose the principal's beliefs off path to

be those beliefs consistent with the evidence presented for which his best reply is the worst possible for the agent who deviated. Since the principal's actions on path are optimal for him given some beliefs, this generates off path behavior which punishes deviations.

Simple type dependence is more complicated, since the inferences which hurt positive types help negative types, making it harder to select off path beliefs to deter deviations. But this restricts both mechanisms and games. To see the intuition, suppose there is only one agent with two types, t' and t'', where t' is positive and t'' is negative. Fix any report and evidence, say $(\bar{t}, \bar{e})$, which is a deviation from equilibrium and is feasible for both types, so that we need to select a belief for the principal. Let $a(t)$ denote the action played by the principal in the proposed equilibrium (and, because we are assuming the only issue at stake is off path behavior, in the mechanism) as a function of the agent's type t. Since the types separate in the mechanism, we have $a(t') \neq a(t'')$. Let a^* be the response to $(\bar{t}, \bar{e})$ in the mechanism. Since the mechanism is incentive compatible, it must be true that $u_i(a(t')) \geq u_i(a^*) \geq u_i(a(t''))$. But this means we can construct the equilibrium to have the principal infer from $(\bar{t}, \bar{e})$ that the agent is type t' and choose action $a(t')$. This ensures sequential rationality and deters the deviation by either type.

5. CONNECTION TO THE LITERATURE

In this section, we give details on how our results relate to the literature. Green and Laffont (1986) began the literature on mechanism design with evidence. We make use of results in Bull and Watson (2007) and Deneckere and Severinov (2008). Below, we discuss in more detail a particularly relevant part of this literature which identifies conditions under which the principal does not need commitment to obtain the same outcome as under the optimal mechanism, a result first shown by Glazer and Rubinstein (2004, 2006) and extended by Sher (2011) and Hart, Kremer, and Perry (2017).

The first papers on games with evidence are Grossman (1981) and Milgrom (1981). We make particular use of Dye (1985) and Jung and Kwon (1988).[12] More recent papers of interest on this topic include Hagenbach, Koessler, and Perez-Richet (2014) and Guttman, Kremer, and Skrzypacz (2014). The papers most closely related to our application to costly verification models are BDL (2014) and Erlanson and Kleiner (2017).

Our results on robust incentive compatibility are related to earlier results on dominant strategy incentive compatible mechanisms without evidence. Manelli and Vincent (2010) and Gershkov, Goeree, Kushnir, Moldovanu, and Shi (2013) showed that in certain settings with transfers and quasi-linear utility, every incentive compatible allocation is equivalent (yields the same interim utilities for all types of all agents) to a dominant strategy incentive compatible allocation. In Part SE of the Supplemental Material, we show how their approach can be adapted to our setting to provide an alternative, though more complex, proof of our result that robust incentive compatibility is costless for the principal.

We extend the earlier results that commitment is not necessary in the one-agent setting in several ways. First, we consider multiple agents. Second, because we have multiple agents, we can consider robustness with respect to agents' beliefs about other agents, an issue absent in the one-agent setting. Third, our characterization of the equilibrium strategies is novel.

Even when we restrict our analysis to the one-agent case, our results are not nested by the previous literature. Most significantly, all previous results assume the agent's preferences are independent of her type, while we allow simple type dependence. To clarify, for

[12] See also Farrell (1986) which appears to have developed essentially the same model independently.

the remainder of this discussion, we consider the one-agent case, so t is the type of the single agent, T her set of types, and u her utility function.

Glazer and Rubinstein (2004, 2006), the first to show a result of this form, used weaker assumptions on evidence as they did not assume normality. However, they assumed that the principal only had two actions available and the agent's preference was type-independent. By contrast, in the one-agent, two-action case, our assumption of simple type dependence is without loss of generality.

Sher (2011) generalized Glazer–Rubinstein by assuming type-independent utility for the agent and that the principal's utility can be written as a concave function of the agent's utility. 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 this depends on a directly, not just through $u(a)$, even the type-independent version of our model is not nested by (nor does it nest) Sher's assumptions. In particular, if the agent is indifferent between a and $\hat{a}$, Sher's assumptions require the principal to be indifferent given any t, a restriction we do not impose.

Hart, Kremer, and Perry (2017), like us, assumed normality. Unlike us, they assumed type-independent utility for the agent and assumed that the principal *cannot* randomize. In addition, they weakened Sher's concavity assumption to the property that for each $t \in T$, the principal's utility function over A can be written as $v(a,t) = \varphi_t(u(a))$ where $\sum_t \mu(t)\varphi_t$ is single-peaked (equivalently, strictly quasi-concave) for any $\mu \in \Delta(T)$. Because we allow the principal's utility to depend on a directly, our model violates this assumption for the same reasons our model violates Sher's assumption. Also, we prove that the principal does not need to randomize.

In the Appendix of an earlier version of their paper, Hart, Kremer, and Perry (2016) allowed the principal to randomize. Their main assumption states that if we fix any indifference curve for the agent, then there is a point on that indifference curve which is best for the principal *independently* of t. In the one-agent version of our model, we have $v(a,t) = u_0(a) + u(a)v(t)$. Hence, holding fixed the agent's utility, for any t, the best lottery over a is any p on the indifference curve which maximizes $\sum_a p(a)u_0(a)$. Thus, except for the type dependence we allow, in the one-agent case, our assumptions are nested in their model.

Hart, Kremer, and Perry also gave a refinement of equilibrium in the disclosure game that identifies the principal's best equilibrium. Our result that the principal's best equilibrium in the game without commitment can be found using I one-agent disclosure games is analogous in that it also provides a means to understand this equilibrium.

APPENDIX A: PROOF OF THEOREM 1

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

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

The Revelation Principle for this class of problems says we can restrict attention to equilibria where each t_i sends $r_i = (t_i, M_i(t_i))$. Hence $\hat{\mathcal{U}}_i(r_i; P)$ is the expected utility of t_i from report r_i if t_i is a positive type and minus the expected utility if t_i is a negative type.

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

$$R_i^\alpha = \big\{r_i \in R_i \mid \hat{\mathcal{U}}_i(r_i; P) = \alpha\big\}.$$

Finiteness of T_i implies that $\mathcal{E}_i$ is finite and hence R_i is finite. References to R_i^α below assume this set is nonempty unless stated otherwise. The nonempty R_i^α's form a partition of

R_i, called the *mechanism partition for i*, denoted $\{R_i^\alpha\}$. The product partition of R formed by the cells $\prod_i R_i^{\alpha_i}$ is the *mechanism partition*, denoted $\{\prod_i R_i^{\alpha_i}\}$. Let

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

LEMMA 2: *P is incentive compatible iff the following hold for every $(s_i, e_i) \in R_i^\alpha$ and $(t_i, M_i(t_i)) \in R_i^\beta$.* (i) *If $t_i \in T_i^+$ and $\alpha > \beta$, then $e_i \notin \mathcal{E}_i(t_i)$.* (ii) *If $t_i \in T_i^-$ and $\beta > \alpha$, then $e_i \notin \mathcal{E}_i(t_i)$.*

PROOF: Immediate. *Q.E.D.*

LEMMA 3: *Without loss of generality, P has the property that for all i, if $(s_i, e_i) \in R_i^\alpha$, then there exists $t_i \in T_i^\alpha$ with $e_i \in \mathcal{E}_i(t_i)$. Hence if $R_i^\alpha \neq \emptyset$, then $T_i^\alpha \neq \emptyset$.*

PROOF: Suppose $(s_i, e_i) \in R_i^\alpha$. By Lemma 2, for any $t_i \in T_i^\beta$ with $e_i \in \mathcal{E}_i(t_i)$, we have $\beta \geq \alpha$ if $t_i \in T_i^+$ and $\beta \leq \alpha$ if $t_i \in T_i^-$. Thus, if there is no $t_i \in T_i^\alpha$ with $e_i \in \mathcal{E}_i(t_i)$, we can move (s_i, e_i) to the smallest $\beta > \alpha$ with $t_i \in T_i^+$ and $e_i \in \mathcal{E}_i(t_i)$ or to the largest $\beta < \alpha$ with $t_i \in T_i^-$ and $e_i \in \mathcal{E}_i(t_i)$ and will preserve incentive compatibility and the principal's expected payoff. We carry out this move by changing the mechanism so that $P(\cdot \mid s_i, e_i, t_{-i}, e_{-i}) = P(\cdot \mid t_i, M_i(t_i), t_{-i}, e_{-i})$ for all $(t_{-i}, e_{-i}) \in R_{-i}$ for the chosen t_i. *Q.E.D.*

LEMMA 4: *Without loss of generality, P is measurable with respect to the mechanism partition for each i, $\{R_i^\alpha\}$, in the sense that if $(s_i, e_i), (s_i', e_i') \in R_i^\alpha$, then $P(\cdot \mid s_i, e_i, t_{-i}, e_{-i}) = P(\cdot \mid s_i', e_i', t_{-i}, e_{-i})$ for all $(t_{-i}, e_{-i}) \in R_{-i}$. Hence P is 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: Suppose P is not measurable with respect to the mechanism partition for some i. We construct an incentive compatible mechanism which is measurable and has the same payoff for the principal as P. Fix i and α such that $R_i^\alpha \neq \emptyset$. By Lemma 3, $T_i^\alpha \neq \emptyset$.

Define a mechanism P^* by

$$P^*(\cdot \mid s_i, e_i, t_{-i}, e_{-i}) = \begin{cases} P(a \mid s_i, e_i, t_{-i}, e_{-i}) & \text{if } (s_i, e_i) \notin R_i^\alpha, \\ \mathrm{E}_{t_i}\big(P(a \mid t_i, M_i(t_i), t_{-i}, e_{-i}) \mid (t_i, M_i(t_i)) \in R_i^\alpha\big), & \text{otherwise.} \end{cases}$$

The expected payoff to any type t_j of agent $j \neq i$ from any report is the same in P and P^*. So incentive compatibility of P implies incentive compatibility of P^* for $j \neq i$.

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

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

So every t_i receives the same expected payoff from every report in P and P^*, so incentive compatibility of P implies incentive compatibility of P^*. Also, P^* gives the principal the same expected payoff as P. Hence P^* is an optimal mechanism. Iterating gives an optimal mechanism measurable with respect to the mechanism partition for i; iterating over i gives an optimal mechanism measurable with respect to the mechanism partition. *Q.E.D.*

Note that the principal's expected payoff is linear in the expected values of the v_i's. The following lemma gives a standard but useful implication regarding optimal actions.

LEMMA 5: *Let*

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

Given any belief of the principal over each T_i, let $\hat{v}_i$ denote the expectation of $v_i(t_i)$ 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 the principal's expected utility, $\hat{v} \cdot u$. Suppose v and v' satisfy $v_i > v'_i$ and $v'_j = v_j$ for $j \neq i$. Then for any $u \in \mathcal{U}^*(v)$ and $u' \in \mathcal{U}^*(v')$, we have $u_i \geq u'_i$.

PROOF: Standard. *Q.E.D.*

We now construct an equilibrium for the game without commitment which yields the same payoff for the principal as P. The strategy for agent i in this equilibrium is the same as i's strategy in an equilibrium of the auxiliary game for i. The auxiliary 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 $z_i \in Z_i(t_i)$. 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 (behavioral) 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 $q_i : T_i \times \mathcal{E}_i \to \Delta(T_i)$. The principal's strategy for the game is denoted $X_i : R_i \to X$.

We construct an equilibrium of the auxiliary game for i via the *restricted auxiliary game*. In the restricted game, type t_i can only choose actions in R_i^α where α is the unique α such that $t_i \in T_i^\alpha$. That is, t_i's strategy set is $Z_i(t_i) \cap R_i^\alpha$. Note that every $(s_i, e_i) \in R_i$ is contained in at least one $Z_i(t_i) \cap R_i^\alpha$ by Lemma 3.

Fix i and a perfect Bayesian equilibrium $(\sigma_i^*, X_i^*, q_i^*)$ of the restricted auxiliary game for i.[13] Sequential rationality for the principal implies that $X_i^*(s_i, e_i) = \sum_{t_i \in T_i} v_i(t_i) q_i^*(t_i \mid s_i, e_i)$, the expectation of $v_i(t_i)$ given the belief q_i^*.

[13] 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 auxiliary 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 $\hat{X}_i^*(t_i)$ denote the action chosen by the principal in equilibrium when i is type t_i. That is, $\hat{X}_i^* : T_i \to X$ and is given by

$$\hat{X}_i^*(t_i) = X_i^*(s_i, e_i), \quad \text{for some } (s_i, e_i) \in \operatorname{supp}\big(\sigma_i^*(\cdot \mid t_i)\big).$$

Because the principal's payoff function is strictly concave in his action, he always uses a pure strategy. Since t_i's payoff is either strictly increasing or strictly decreasing in the principal's actions, 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. For this to be an equilibrium, we require

$$\hat{X}_i^*(t_i) = \max_{(s_i,e_i)\in Z_i(t_i)\cap R_i^\alpha} X_i^*(s_i, e_i), \quad \forall t_i \in T_i^+,$$

$$\hat{X}_i^*(t_i) = \min_{(s_i,e_i)\in Z_i(t_i)\cap R_i^\alpha} X_i^*(s_i, e_i), \quad \forall t_i \in T_i^-.$$

By construction, if $t_i \in T_i^\alpha$, then t_i can only send $(s_i, e_i) \in R_i^\alpha$ in the restricted auxiliary game. Hence, in any equilibrium of this game, the principal at least learns the event of the mechanism partition for i that t_i lies in. Since the optimal mechanism is measurable with respect to the mechanism partition, this means that the principal has enough information to carry out the optimal mechanism. 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 6: *For each i, fix any equilibrium of the restricted auxiliary game for i. Then for every $t \in T$,*

$$P\big(\cdot \mid t, M(t)\big) \in \arg\max_{p\in\Delta(A)} \sum_a p(a) \sum_{i=0}^{I} u_i(a)\hat{X}_i^*(t_i).$$

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: For each $(\alpha_1, \ldots, \alpha_I)$ such that each $T_i^{\alpha_i} \neq \emptyset$, $P(\cdot \mid t, M(t))$ is constant over $t \in \prod_i T_i^{\alpha_i}$. Given any $t \in \prod_i T_i^{\alpha_i}$, the equilibria from the auxiliary games give the principal at least as much information as the fact that $t \in \prod_i T_i^{\alpha_i}$, so we must have

$$\max_{p\in\Delta(A)} \sum_a p(a) \sum_i u_i(a)\hat{X}_i^*(t_i) \geq \sum_a P\big(a \mid t, M(t)\big) \sum_i u_i(a)\hat{X}_i^*(t_i), \quad \forall t \in T.$$

The claim is that this holds with equality for all t. Suppose, to the contrary, that the inequality is strict for some t.

For each $\hat{v} = (1, \hat{v}_1, \ldots, \hat{v}_I) \in \mathbf{R}^{I+1}$, let $\tilde{p}(\cdot \mid \hat{v})$ denote any $p(\cdot) \in \Delta(A)$ which maximizes $\sum_a p(a) \sum_{i=0}^{I} u_i(a)\hat{v}_i$. In other words, $\tilde{p}(\cdot \mid \hat{v})$ is an optimal p for the principal given any beliefs over T such that $\hat{v}_i$ is the expected value of $v_i(t_i)$. So we have

$$\sum_a \tilde{p}\big(a \mid \hat{X}_i^*(t_i)\big) \sum_i u_i(a)\hat{X}_i^*(t_i) \geq \sum_a P\big(a \mid t, M(t)\big) \sum_i u_i(a)\hat{X}_i^*(t_i)$$

for all $t \in T$, strictly so for some t. We complete the proof by using this to construct a mechanism superior to the optimal mechanism, a contradiction.

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

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

That is, $\hat{v}_i(s_i, e_i)$ is the equilibrium belief type s_i induces in the restricted auxiliary game if e_i is maximal evidence for s_i; otherwise, it is the equilibrium belief the principal has in the restricted auxiliary game in response to report and evidence (s_i, e_i). This construction is needed so that each type will be induced to report truthfully and provide maximal evidence in the mechanism in order to mimic the equilibrium. We need this to, in effect, turn an indirect mechanism into a direct mechanism.

Given $(s, e) \in \prod_i R_i$, 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

$$P^*(\cdot \mid s, e) = \varepsilon \tilde{p}\big(\cdot \mid \hat{v}(s, e)\big) + (1 - \varepsilon)P(\cdot \mid s, e).$$

To show that P^* is incentive compatible, fix $t_i \in T_i$ 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.[14]

So suppose that t_i is indifferent between reporting $(t_i, M_i(t_i))$ and 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. Since t_i is indifferent between these two reports under P, she prefers reporting $(t_i.M_i(t_i))$ under P^* iff she prefers reporting $(t_i, M_i(t_i))$ under $\tilde{p}(\cdot \mid \hat{v}(s, e))$. That is, if $t_i \in T_i^+$, t_i prefers reporting $(t_i, M_i(t_i))$ to reporting (s_i, e_i) under P^* iff

$$\mathrm{E}_{t_{-i}}\left[\sum_a \tilde{p}\big(a \mid \hat{X}_i^*(t_i), \hat{X}_{-i}^*(t_{-i})\big)u_i(a)\right] \geq \mathrm{E}_{t_{-i}}\left[\sum_a \tilde{p}\big(a \mid X_i^*(s_i, e_i), \hat{X}_{-i}^*(t_{-i})\big)u_i(a)\right]. \quad (2)$$

If $X_i^*(s_i, e_i) = \hat{X}_i^*(t_i)$, this holds with equality. So suppose $X_i^*(s_i, e_i) \neq \hat{X}_i^*(t_i)$. Since (s_i, e_i) and $(t_i, M_i(t_i))$ are in the same event of the mechanism partition, (s_i, e_i) is a feasible report for t_i in the restricted auxiliary game. Hence the fact that $t_i \in T_i^+$ implies $\hat{X}_i^*(t_i) > X_i^*(s_i, e_i)$. By Lemma 5, this implies

$$\sum_a \tilde{p}\big(a \mid \hat{X}_i^*(t_i), \hat{v}_{-i}\big)u_i(a) \geq \sum_a \tilde{p}\big(a \mid X_i^*(s_i, e_i), \hat{v}_{-i}\big)u_i(a),$$

for all $\hat{v}_{-i}$, implying that (2) holds.

A similar argument for $t_i \in T_i^-$ completes the proof that P^* is incentive compatible. But then P^* is an incentive compatible mechanism giving the principal a strictly higher payoff than P, a contradiction. *Q.E.D.*

LEMMA 7: *Fix $\alpha > \beta$ such that $T_i^\alpha \neq \emptyset$ and $T_i^\beta \neq \emptyset$ and any equilibrium of the restricted auxiliary game for i. Then for every $t_i \in T_i^\alpha$ and $t_i' \in T_i^\beta$, we have $\hat{X}_i^*(t_i) \geq \hat{X}_i^*(t_i')$.*

[14]This argument requires finiteness of each T_i. We conjecture that a more complex argument could substitute in the case where some T_i are infinite.

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

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

By Lemma 6, $p^\alpha \equiv P(\cdot \mid t_i, M_i(t_i), \hat{t}_{-i}, M_{-i}(\hat{t}_{-i}))$ maximizes over $p(\cdot) \in \Delta(A)$,

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

and p^β defined analogously maximizes the analog for t_i'. Hence by Lemma 5, $u_i^\alpha > u_i^\beta$ implies $\hat{X}_i^*(t_i) \geq \hat{X}_i^*(t_i')$. *Q.E.D.*

We complete the proof in two steps. First, we show how to modify an equilibrium of the restricted auxiliary game for i to construct an equilibrium of the unrestricted auxiliary game for i with the same equilibrium path. Second, we use these equilibria to construct a robust equilibrium of the game without commitment and an optimal mechanism which is deterministic, robustly incentive compatible, and has the same outcome as the equilibrium.

LEMMA 8: *For any i and any equilibrium of the restricted auxiliary game for i, there is an equilibrium of the unrestricted auxiliary game where i follows the same strategy.*

PROOF: Let $(\sigma_i^*, X_i^*, q_i^*)$ be an equilibrium of the restricted auxiliary game for i. Fix any $(\bar{s}_i, \bar{e}_i) \in R_i^\beta$. Let F denote the set of types for whom $(\bar{s}_i, \bar{e}_i)$ is feasible—that is, $F = \{t_i \mid \bar{e}_i \in \mathcal{E}_i(t_i)\}$. We show that for any $\bar{t}_i \in T_i^\alpha \cap F$ with $\sigma^*(\bar{s}_i, \bar{e}_i \mid \bar{t}_i) = 0$, $\bar{t}_i$ weakly prefers her equilibrium strategy to $(\bar{s}_i, \bar{e}_i)$. That is, $\hat{X}_i^*(\bar{t}_i) \geq X_i^*(\bar{s}_i, \bar{e}_i)$ if $t_i \in T_i^+$ and the reverse inequality for $t_i \in T_i^-$. Clearly, if $\alpha = \beta$, the fact that $\bar{t}_i$ did not have a profitable deviation in the restricted game implies that she does not wish to deviate to $(\bar{s}_i, \bar{e}_i)$ in the unrestricted game, so we only consider $\alpha \neq \beta$.

First, suppose $(\bar{s}_i, \bar{e}_i)$ has positive probability in equilibrium. That is, there is another t_i' with $\sigma_i^*(\bar{s}_i, \bar{e}_i \mid t_i') > 0$, so $X_i^*(\bar{s}_i, \bar{e}_i) = \hat{X}_i^*(t_i')$. For any $\bar{t}_i \in T_i^+ \cap F$, incentive compatibility implies $\beta < \alpha$. By Lemma 7, this implies $\hat{X}_i(\bar{t}_i) \geq \hat{X}_i^*(t_i') = X_i^*(\bar{s}_i, \bar{e}_i)$, so $\bar{t}_i$ has no incentive to deviate. A similar argument applies to all $\bar{t}_i \in T_i^-$.

So suppose $(\bar{s}_i, \bar{e}_i)$ has zero probability in equilibrium. Let $\hat{t}_i$ minimize $\hat{X}_i^*(\hat{t}_i)$ over $t_i \in T_i^+ \cap F$. Clearly, if $\hat{X}_i^*(\hat{t}_i) \geq X_i^*(\bar{s}_i, \bar{e}_i)$, then we do not need to consider positive types further. Suppose, then, that $\hat{X}_i^*(\bar{s}_i, \bar{e}_i) > \hat{X}_i^*(\hat{t}_i)$.

We claim that there must be some $t_i' \in F$ with $\hat{X}_i^*(\hat{t}_i) \geq v_i(t_i')$. If not, then $\hat{t}_i$ could send $M_i(\hat{t}_i)$ in the restricted auxiliary game, an option which must be feasible, and prove at least as much as $\bar{e}_i$. This would generate a belief over $t_i' \in F$ which would have an expected value strictly larger than $\hat{X}_i^*(\hat{t}_i)$, a contradiction.

So change the principal's beliefs in response to $(\bar{s}_i, \bar{e}_i)$ to $\lambda q_i^*(\bar{s}_i, \bar{e}_i) + (1 - \lambda)\delta_{t_i'}$ where $\delta_{t_i'}$ is the degenerate distribution putting probability 1 on t_i'. Choose λ so that the expected value of $v_i(t_i)$ under this belief is $\hat{X}_i^*(\hat{t}_i)$. Change the principal's strategy to reply to $(\bar{s}_i, \bar{e}_i)$ with $x = \hat{X}_i^*(\hat{t}_i)$. With this change, clearly, no $t_i \in T_i^+ \cap F$ gains by deviating to $(\bar{s}_i, \bar{e}_i)$. To see that no $t_i \in T_i^- \cap F$ has an incentive to deviate, let $\hat{t}_i' \in T_i^\gamma$ denote such a type and suppose $\hat{t}_i \in T_i^\alpha$. Recall that $(\bar{s}_i, \bar{e}_i) \in R_i^\beta$. Since $\hat{t}_i \in T_i^+ \cap F$ and $\hat{t}_i' \in T_i^- \cap F$, incentive

compatibility implies $\alpha \geq \beta \geq \gamma$. By Lemma 7, $\hat{X}_i^*(\hat{t}_i) \geq \hat{X}_i^*(\hat{t}_i')$, so $\hat{t}_i'$ has no incentive to deviate to $(\bar{s}_i, \bar{e}_i)$ after this change.

Negative types can be handled by a symmetric argument. *Q.E.D.*

We complete the proof by constructing a robust equilibrium of the game without commitment which gives the principal the same payoff as in the optimal mechanism. We then use this equilibrium to construct an optimal mechanism which is deterministic and robustly incentive compatible with the same outcome as the equilibrium.

To construct the equilibrium for the game without commitment, let the strategy for agent i be the same as her strategy in the equilibrium of the auxiliary 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 auxiliary game for i.

Similarly to the proof of Lemma 6, for each $\hat{v} = (1, \hat{v}_1, \ldots, \hat{v}_I) \in \mathbf{R}^{I+1}$, let $\hat{p}(\cdot \mid \hat{v})$ denote any $p(\cdot) \in \Delta(A)$ which (a) is a degenerate distribution and (b) maximizes

$$\sum_a p(a) \sum_{i=0}^{I} u_i(a)\hat{v}_i. \tag{3}$$

Let the principal's strategy given (s, e) be to choose $\hat{p}(\cdot \mid \hat{v}(s, e))$ where $\hat{v}_i(s, e) = X_i^*(s_i, e_i)$ for $i = 1, \ldots, I$ and $\hat{v}_0(s, e) = 1$. Clearly, this satisfies sequential rationality for the principal.

To see that this specification gives a robust equilibrium, consider any t_i and suppose the other agents report (t_{-i}, e_{-i}). If t_i deviates from her proposed equilibrium strategy to a different strategy inducing the same expected value of v_i, this does not change the principal's action by construction. Hence such a deviation is not profitable. If t_i deviates to a strategy which induces a different expected value, then, by the fact that we started from an equilibrium of the auxiliary game, this change must be against t_i. That is, the change must lower the expected value if t_i is a positive type and raise it if t_i is negative. By Lemma 5, such a deviation cannot be profitable. Hence we have a robust equilibrium.

Because the principal receives at least as much information from the equilibrium strategies as the mechanism partition, his expected payoff must be at least as large as in the optimal mechanism. Clearly, it cannot be strictly larger than the payoff to the optimal mechanism, so it must be equal.

Hence committing to this strategy is an optimal indirect mechanism. It is straightforward to rewrite this as an optimal direct mechanism. It is deterministic by construction. It is straightforward to show that the robustness of the equilibrium implies that this mechanism is robustly incentive compatible. qed

APPENDIX B: PROOF OF LEMMA 1 AND THEOREM 3

For Lemma 1, 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

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

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

Next, we show that there exists v_i^* solving

$$v_i^* = \mathrm{E}_{t_i}\big[v_i(t_i) \mid \big(t_i \in T_i^0\big) \text{ or } \big(t_i \in T_i^- \setminus T_i^0 \text{ and } v_i(t_i) \geq v_i^*\big) \text{ or } \big(t_i \in T_i^+ \setminus T_i^0 \text{ and } v_i(t_i) \leq v_i^*\big)\big]. \tag{4}$$

Let $g_i(v_i^*)$ be the function on the right–hand side. We show there is v_i^* solving $v_i^* = g_i(v_i^*)$.

Suppose not. Let $v_i^1 < v_i^2 < \cdots < v_i^N$ denote the values of $v_i(t_i)$ for $t_i \notin T_i^0$. First, note that for $v_i^* \leq v_i^1$, we have $g_i(v_i^*) = \mathrm{E}_{t_i}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^-]$. If $\mathrm{E}_{t_i}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^-] \leq v_i^1$, then $v_i^* = \mathrm{E}_{t_i}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^-]$ is a solution to equation (4). So our hypothesis that there is no solution implies $\mathrm{E}_{t_i}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^-] > v_i^1$.

The function $g_i(v_i^*)$ is constant in v_i^* for $v_i^* \in (v_i^k, v_i^{k+1})$ but may be discontinuous at each v_i^k. The important point is that if $g_i(v_i^k - \varepsilon) > v_i^k$ for all sufficiently small $\varepsilon > 0$, then $g_i(v_i^k + \varepsilon) > v_i^k$ as well. That is, the function can never jump from above the 45° line to below. To see this, first suppose $v_i^k \in v_i(T_i^-)$.[15] In this case, as v_i^* increases from just below to just above v_i^k, we remove v_i^k from the conditioning set. If $g_i(v_i^k) > v_i^k$, removing this point from the conditioning set implies that $g_i(v_i^k + \varepsilon) > g_i(v_i^k)$. If $v_i \in v_i(T_i^+)$, then as v_i^* increases from just below to just above v_i^k, we add v_i^k to the conditioning set. If $g_i(v_i^k - \varepsilon) > v_i^k$, adding this point to the conditioning set implies that $g_i(v_i^k - \varepsilon) > g_i(v_i^k) > v_i^k$. So, again, the function remains above the 45° line.

By hypothesis, we have no solution, so $g_i(v_i^1) > v_i^1$. Since g_i cannot jump below the 45° line, the lack of a solution implies $g_i(v_i^*) > v_i^*$ for all $v_i^* \geq v_i^1$. In particular, $g_i(v_i^N) > v_i^N$. But $g_i(v_i^*) = \mathrm{E}_{t_i}[v_i(t_i) \mid t_i \in T_i^0 \cup T_i^+]$ for all $v_i^* \geq v_i^N$. So there exists $v_i^* > v_i^N$ solving (4), a contradiction.

To show uniqueness, suppose v_i^1 and v_i^2 are solutions to (4) where $v_i^1 > v_i^2$. Let

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

and

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

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

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

Let

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

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

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

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,

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

[15] $v_i(T_i^-)$ is the set of v_i such that $v_i = v_i(t_i)$ for some $t_i \in T_i^-$ and $v_i(T_i^+)$ (see below) is defined analogously.

Since v_i^1 is a convex combination of $\tilde{v}_i$ and a term smaller than v_i^1, we have $\tilde{v}_i \geq v_i^1$. Since v_i^2 is a convex combination of $\tilde{v}_i$ and a term larger than v_i^2, we 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 Theorem 3, we construct equilibrium strategies. If $X_i^*(s_i, T_i) > X_i^*(s_i', T_i)$, no positive type sends report (s_i', T_i) and no negative type sends (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 $\tilde{v}_i^- = \min_{s_i \in T_i} x_i^*(s_i, T_i)$. First, assume $\tilde{v}_i^+ > \tilde{v}_i^-$. Then 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 $v_i(t_i) \leq \tilde{v}_i^+$. Similarly, every negative type $t_i \in T_i^0$ or not in T_i^0 but with $v_i(t_i) \geq \tilde{v}_i^-$ sends some report generating $\tilde{v}_i^-$. All other types t_i send a report of the form $(s_i, \{t_i\})$. Hence $\tilde{v}_i^+$ 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 $v_i^- = v_i^+ = v_i^*$.

There is also an equilibrium where the principal ignores the type report. Letting $\tilde{v}_i$ denote the principal's expected value of v_i given evidence report $e_i = T_i$, 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 (4), so $\tilde{v}_i = v_i^*$. *Q.E.D.*

APPENDIX C: COSTLY VERIFICATION

We show that for a class of costly-verification models with simple type dependence, the optimal mechanism can be computed using our results for optimal mechanisms with Dye evidence. Continue to let A denote the finite set of actions available to the principal, T_i the finite set of types of agent i with the same distributional assumptions as in the text, and continue to assume that agent i's utility function can be written as

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

and that the principal's utility function can be written as $v(a, t) = \sum_{i=0}^{I} u_i(a) v_i(t_i)$.

We add three assumptions on preferences. First, each agent has exactly two indifference curves in A.[16] That is, for each agent i, we can partition A into nonempty[17] sets A_i^0 and A_i^1 where

$$u_i(a) = \begin{cases} 0 & \text{if } a \in A_i^0, \\ 1 & \text{if } a \in A_i^1. \end{cases}$$

(Because u_i does not depend on t_i, the sets A_i^0 and A_i^1 are common knowledge.) For example, this assumption holds in the allocation example and most of the related problems discussed in Example 1 of Section 1 as well as the public goods problem discussed in Example 4. It also holds in the public goods problem discussed in Erlanson and Kleiner (2017) (after renormalizing).

Second, assume that for all i, either $T_i^- = \emptyset$ or $v_i(t_i) > v_i(t_i')$ for all $t_i \in T_i^+$ and $t_i' \in T_i^-$. That is, either i's preferences are type-independent or every positive type has a higher v_i than every negative type. Erlanson and Kleiner made the latter assumption.

[16] This also includes "agent 0"—that is, this also applies to the utility function $u_0(a)$.

[17] If either set is empty for $i \neq 0$, then the agent is indifferent over all choices by the principal and incentive compatibility is trivially satisfied. Hence we can disregard any such agent.

For the costly-verification model, agents do not have evidence to present. Instead, the principal can *check* agent i at a cost $c_i > 0$. "Checking" agent i means that the principal learns agent i's type t_i. We show that the optimal mechanism can be computed by an appropriate "translation" of a related mechanism design problem with Dye evidence instead of costly verification.

Note that our assumptions imply that if $v_i(t_i) = v_i(t_i')$, then either both are positive types or both are negative. Since agents do not have evidence, this means that t_i and t_i' are identical and there is no need to distinguish them. Hence we write the type set for i as $T_i = \{t_i^0, \ldots, t_i^{K_i}\}$ where $v_i(t_i^k) < v_i(t_i^{k+1})$ for $k = 0, \ldots, K_i - 1$.

One can show that it is without loss of generality to focus on mechanisms with the following structure. First, all agents simultaneously make cheap-talk reports of types to the principal. The mechanism specifies a probability distribution over which agents to check and what $a \in A$ to choose as a function of the reports. Each agent will have an incentive to report his type honestly, so when the principal checks an agent, he finds that the report was truthful. Off the equilibrium path, if the principal finds that an agent has lied, the principal chooses any action which is worst for that agent. (Since the agents all expect the other agents to report honestly, the specification of the mechanism for histories where multiple agents are found to have lied is irrelevant.)

Hence we can write a mechanism as a function $P : T \to \Delta(2^{\mathcal{I}} \times A)$ where $P(Q, a \mid t)$ is the probability that the principal checks the agents in the set $Q \subseteq \mathcal{I}$ and chooses action $a \in A$ when the type reports are t and the checking verifies the reports were honest. The expected payoff of the principal from such a mechanism is

$$\mathrm{E}_t\left[\sum_{(Q,a)\in 2^{\mathcal{I}}\times A} P(Q, a \mid t)\left(v(a, t) - \sum_{i\in Q} c_i\right)\right].$$

Let

$$p(a \mid t) = \sum_{Q\subseteq\mathcal{I}} P(Q, a \mid t),$$

$$q_i(t) = \sum_{a\in A}\sum_{Q\subseteq\mathcal{I}\mid i\in Q} P(Q, a \mid t).$$

Then we can rewrite the principal's expected payoff as

$$\mathrm{E}_t\left[\sum_{a\in A} p(a \mid t)v(a, t) - \sum_i q_i(t)c_i\right].$$

Using the fact that $v(a, t) = \sum_i u_i(a)v_i(t_i)$, we can rewrite this as

$$\mathrm{E}_t\left[\sum_i v_i(t_i)\sum_{a\in A} p(a \mid t)u_i(a) - \sum_i q_i(t)c_i\right].$$

Let $p_i(t) = \sum_{a\in A_i^1} p(a \mid t)$. That is, $p_i(t)$ is the probability that the principal selects an action a such that $u_i(a) = 1$ given type profile t. Then the principal's expected pay-

off is

$$\mathrm{E}_t\Big[\sum_i \big(p_i(t)v_i(t_i) - q_i(t)c_i\big)\Big] = \sum_i \mathrm{E}_{t_i}\big[\hat{p}_i(t_i)v_i(t_i) - \hat{q}_i(t_i)c_i\big],$$

where $\hat{p}_i(t_i) = \mathrm{E}_{t_{-i}} p_i(t)$ and $\hat{q}_i(t_i) = \mathrm{E}_{t_{-i}} q_i(t_i, t_{-i})$.

If agent i of type t_i reports truthfully, his expected utility in mechanism P is

$$\mathrm{E}_{t_{-i}} \sum_{(Q,a)\in 2^{\mathcal{T}}\times A} P(Q, a \mid t)u_i(a)$$

if $t_i \in T_i^+$ and this times -1 otherwise. So the expected payoff to a positive type from reporting truthfully is $\hat{p}_i(t_i)$, while the expected payoff to a negative type is $-\hat{p}_i(t_i)$.

If agent i is type t_i but reports $t_i' \neq t_i$, he may be caught lying. If so, as noted above, the principal chooses an action which minimizes his payoff. So if $t_i \in T_i^+$, his payoff will be 0 if he is caught lying, while if $t_i \in T_i^-$, it will be -1. Hence, for a positive type, the expected payoff to the deviation is

$$\begin{aligned}
&\mathrm{E}_{t_{-i}}\Big[\sum_{(Q,a)\in 2^{\mathcal{T}}\times A \mid i\notin Q} P\big(Q, a \mid t_i', t_{-i}\big)u_i(a)\Big] \\
&\quad = \mathrm{E}_{t_{-i}}\Big[\sum_{(Q,a)\in 2^{\mathcal{T}}\times A} P\big(Q, a \mid t_i', t_{-i}\big)u_i(a) - \sum_{(Q,a)\in 2^{\mathcal{T}}\times A \mid i\in Q} P\big(Q, a \mid t_i', t_{-i}\big)u_i(a)\Big] \\
&\quad = \hat{p}_i\big(t_i'\big) - \mathrm{E}_{t_{-i}}\Big[\sum_{(Q,a)\in 2^{\mathcal{T}}\times A \mid i\in Q, a\in A_i^1} P\big(Q, a \mid t_i', t_{-i}\big)\Big].
\end{aligned}$$

We will simplify this expression further below.

If a negative type is caught reporting falsely, the principal chooses an action setting $u_i(a) = 1$ so that the agent's payoff is -1. Hence the expected payoff to a negative type t_i from claiming to be $t_i' \neq t_i$ is

$$\begin{aligned}
&\mathrm{E}_{t_{-i}}\Big[\sum_{(Q,a)\in 2^{\mathcal{T}}\times A \mid i\notin Q} P\big(Q, a \mid t_i', t_{-i}\big)\big(-u_i(a)\big) - \sum_{(Q,a)\in 2^{\mathcal{T}}\times A \mid i\in Q} P\big(Q, a \mid t_i', t_{-i}\big)\Big] \\
&\quad = \mathrm{E}_{t_{-i}}\Big[-\sum_{(Q,a)\in 2^{\mathcal{T}}\times A} P\big(Q, a \mid t_i', t_{-i}\big)u_i(a) - \sum_{(Q,a)\in 2^{\mathcal{T}}\times A \mid i\in Q} P\big(Q, a \mid t_i', t_{-i}\big)\big(1 - u_i(a)\big)\Big] \\
&\quad = -\hat{p}_i\big(t_i'\big) - \mathrm{E}_{t_{-i}}\Big[\sum_{(Q,a)\in 2^{\mathcal{T}}\times A \mid i\in Q, a\in A_i^0} P\big(Q, a \mid t_i', t_{-i}\big)\Big].
\end{aligned}$$

Summarizing, the incentive compatibility constraint for agent i is that for all positive types $t_i \in T_i^+$, we have

$$\hat{p}_i(t_i) \geq \hat{p}_i\big(t_i'\big) - \mathrm{E}_{t_{-i}}\Big[\sum_{(Q,a)\in 2^{\mathcal{T}}\times A \mid i\in Q, a\in A_i^1} P\big(Q, a \mid t_i', t_{-i}\big)\Big], \quad \forall t_i' \neq t_i, \tag{5}$$

and for all negative types $t_i \in T_i^-$, we have

$$\hat{p}_i(t_i) \leq \hat{p}_i(t_i') + \mathrm{E}_{t_{-i}}\left[\sum_{(Q,a)\in 2^{\mathcal{T}}\times A | i\in Q, a\in A_i^0} P(Q,a \mid t_i', t_{-i})\right], \quad \forall t_i' \neq t_i. \tag{6}$$

Note that the right-hand side of each incentive compatibility constraint is independent of t_i. Hence (5) holds for all positive types t_i iff it holds for the positive type with the smallest $\hat{p}_i(t_i)$ and (6) holds for all negative types t_i iff it holds for that negative type with the largest $\hat{p}_i(t_i)$.

The optimal mechanism is monotonic in the sense that $\hat{p}_i(t_i^k) \leq \hat{p}_i(t_i^{k+1})$ for $k = 0, \ldots, K_i - 1$. To see this, recall that $v_i(t_i^k) < v_i(t_i^{k+1})$, so the principal is better off with higher values of p_i associated with higher values of t_i. Suppose we have an incentive compatible mechanism with $\hat{p}_i(t_i^k) > \hat{p}_i(t_i^{k+1})$ for some k and i. Consider the mechanism which reverses the roles of these types—that is, assigns the outcome (Q,a) to (t_i^k, t_{-i}) that it would have assigned to (t_i^{k+1}, t_{-i}) and vice versa.[18] This altered mechanism is also incentive compatible and yields the principal a higher expected payoff.[19]

By assumption, for every i, either $T_i^- = \emptyset$ or $v_i(t_i) > v_i(t_i')$ for all $t_i \in T_i^+$, $t_i' \in T_i^-$. Hence, if there are J_i negative types (where J_i can be zero), the negative types are $t_i^0, \ldots, t_i^{J_i-1}$ and the positive types are $t_i^{J_i}, \ldots, t_i^{K_i}$. Thus, the positive type with the lowest $\hat{p}_i(t_i)$ is $t_i^{J_i}$, while the negative type with the highest $\hat{p}_i(t_i)$ is $t_i^{J_i-1}$ and we have $\hat{p}_i(t_i^{J_1-1}) \leq \hat{p}_i(t_i^{J_i})$. So we can write the incentive compatibility constraints (5) and (6) as

$$\hat{p}_i(t_i^{J_i}) \geq \hat{p}_i(t_i') - \mathrm{E}_{t_{-i}}\left[\sum_{(Q,a)\in 2^{\mathcal{T}}\times A | i\in Q, a\in A_i^1} P(Q,a \mid t_i', t_{-i})\right], \quad \forall t_i' \neq t_i, \tag{7}$$

and

$$\hat{p}_i(t_i^{J_i-1}) \leq \hat{p}_i(t_i') + \mathrm{E}_{t_{-i}}\left[\sum_{(Q,a)\in 2^{\mathcal{T}}\times A | i\in Q, a\in A_i^0} P(Q,a \mid t_i', t_{-i})\right], \quad \forall t_i' \neq t_i. \tag{8}$$

The following lemma generalizes results in Ben-Porath, Dekel, and Lipman (2014) and Erlanson and Kleiner (2017).

LEMMA 9: *In any optimal mechanism, we have*

$$P(Q,a \mid t_i, t_{-i}) = 0, \quad \forall t_{-i} \text{ if } t_i \in T_i^+, i \in Q, \text{ and } a \in A_i^0, \tag{9}$$

$$P(Q,a \mid t_i, t_{-i}) = 0, \quad \forall t_{-i} \text{ if } t_i \in T_i^-, i \in Q, \text{ and } a \in A_i^1. \tag{10}$$

Consequently, we can rewrite the incentive compatibility constraints (7) *and* (8) *as*

$$\hat{p}_i(t_i^{J_i}) \geq \hat{p}_i(t_i) - \hat{q}_i(t_i), \quad \forall t_i \in T_i^+, \tag{11}$$

$$\hat{p}_i(t_i^{J_i-1}) \leq \hat{p}_i(t_i) + \hat{q}_i(t_i), \quad \forall t_i \in T_i^-. \tag{12}$$

[18] To be precise, this implicitly assumes the two types have the same prior probability. If not, we can reverse the role of one of the types and "part of" the other.

[19] If t_i^k and t_i^{k+1} are both positive or both negative, then it is easy to see from (5) or (6) that the altered mechanism is incentive compatible. Our assumptions imply that if one of these types is positive and one negative, then the negative type is t_i^k. It is easy to see that in this case, reversing the roles of the types makes incentive compatibility easier to satisfy.

PROOF: First, we show that we only require (7) for $t_i' \in T_i^+$ and (8) for $t_i' \in T_i^-$. Specifically, monotonicity of $\hat{p}_i$ implies that (7) holds for all $t_i' \in T_i^-$ and (8) holds for all $t_i' \in T_i^+$. To see this, fix any $t_i' \in T_i^-$. By assumption, $v_i(t_i') \leq v_i(t_i^{J_i})$, so monotonicity implies $\hat{p}_i(t_i^{J_i}) \geq \hat{p}_i(t_i')$. Since $\hat{p}_i(t_i')$ is weakly larger than the right-hand side of (7), this implies (7) holds. A similar argument gives (8) for $t_i' \in T_i^+$.

Next, suppose (9) fails, so we have an optimal mechanism P with $P(Q, a \mid t_i, t_{-i}) > 0$ for some $t_{-i} \in T_{-i}$, $t_i \in T_i^+$, $i \in Q$, and $a \in A_i^0$. In other words, there is a positive probability that the principal checks a positive type and chooses an action giving that agent a payoff of zero. Construct a new mechanism P^* as follows. For any $(Q', a') \neq (Q, a)$ or $t' \neq t$, let $P^*(Q', a' \mid t') = P(Q', a' \mid t')$. Let $P^*(Q, a \mid t) = 0$ and let $P^*(Q \setminus \{i\}, a \mid t) = P(Q, a \mid t) + P(Q \setminus \{i\}, a \mid t)$. In other words, if i is checked but gets a zero payoff at (Q, a), we shift this probability to $(Q \setminus \{i\}, a)$, where i does not get checked but still gets the same zero payoff. The incentive compatibility constraints for any agent $j \neq i$ are unaffected. Since t_i is a positive type, the only incentive compatibility constraint for i that is potentially affected is (7) at $t_i' = t_i$ or where $t_i = t_i^{J_i}$. But since we have only changed the checking probability and not the marginal probabilities over actions $a \in A$, $\hat{p}_i(t_i)$ is unaffected. Similarly, the second term on the right-hand side of (7) for $t_i' = t_i$ only involves actions in A_i^1, so this term also is unaffected. Hence P^* is incentive compatible. Finally, since the probability over A is unchanged but the principal checks less often, his payoff must be strictly larger, a contradiction. A symmetric argument establishes (10).

To conclude, consider equation (7) for t_i'. Since $P(Q, a \mid t_i', t_{-i}) = 0$ if $a \in A_i^0$, we see that

$$\sum_{(Q,a) \in 2^{\mathcal{I}} \times A \mid i \in Q, a \in A_i^1} P\big(Q, a \mid t_i', t_{-i}\big) = \sum_{(Q,a) \in 2^{\mathcal{I}} \times A \mid i \in Q} P\big(Q, a \mid t_i', t_{-i}\big) = q_i\big(t_i', t_{-i}\big).$$

Hence we can rewrite (7) as $\hat{p}_i(t_i^{J_i}) \geq \hat{p}_i(t_i') - \hat{q}_i(t_i')$ for all $t_i' \in T_i^+$. A similar argument applied to (8) completes the proof. *Q.E.D.*

We can compute $\hat{q}_i(t_i')$ for all t_i' using Lemma 9. Since $\hat{q}_i$ is costly for the principal, the inequalities in equations (11) and (12) must hold with equality, so

$$\hat{q}_i(t_i) = \begin{cases} \hat{p}_i(t_i) - \hat{p}_i\big(t_i^{J_i}\big) & \text{if } t_i \in T_i^+, \\ \hat{p}_i\big(t_i^{J_i - 1}\big) - \hat{p}_i(t_i) & \text{if } t_i \in T_i^-. \end{cases}$$

Substitute into the objective function for $\hat{q}_i$ and rewrite it as

$$\sum_i \mathrm{E}_{t_i}\big[\hat{p}_i(t_i) v_i(t_i) - \hat{q}_i(t_i) c_i\big] = \sum_i \Bigg[\sum_{k=0}^{J_i - 1} \rho_i\big(t_i^k\big)\big[\hat{p}_i\big(t_i^k\big)\big(v_i\big(t_i^k\big) + c_i\big) - \hat{p}_i\big(t_i^{J_i - 1}\big) c_i\big] + \sum_{k=J_i}^{K_i} \rho_i\big(t_i^k\big)\big[\hat{p}_i\big(t_i^k\big)\big(v_i\big(t_i^k\big) - c_i\big) + \hat{p}_i\big(t_i^{J_i}\big) c_i\big] \Bigg]. \tag{13}$$

The only remaining incentive constraints are that $\hat{p}_i(t_i) \leq \hat{p}_i(t_i^{J_i - 1}) \leq \hat{p}_i(t_i^{J_i})$ for all negative types t_i and $\hat{p}_i(t_i^{J_i - 1}) \leq \hat{p}_i(t_i^{J_i}) \leq \hat{p}_i(t_i)$ for all positive types t_i.

Now consider a different mechanism design problem, this one with evidence instead of costly verification. We have the same set of types as in the problem above and the same u_i

functions. As above, types $t_i^0, \dots, t_i^{J_i-1}$ are negative and types $t_i^{J_i}, \dots, t_i^{K_i}$ are positive. The principal's objective function is now $\sum_i u_i(a)\tilde{v}_i(t_i)$ where

$$\tilde{v}_i(t_i) = \begin{cases} v_i(t_i) - c_i & \text{if } t_i \in T_i^+ \text{ and } t_i \neq t_i^{J_i}, \\ v_i(t_i) + c_i & \text{if } t_i \in T_i^- \text{ and } t_i \neq t_i^{J_i-1}, \\ v_i\big(t_i^{J_i}\big) - c_i + \dfrac{c_i}{\rho_i\big(t_i^{J_i}\big)} \displaystyle\sum_{k=J_i}^{K_i} \rho_i\big(t_i^k\big) & \text{if } t_i = t_i^{J_i}, \\ v_i\big(t_i^{J_i-1}\big) + c_i - \dfrac{c_i}{\rho_i\big(t_i^{J_i-1}\big)} \displaystyle\sum_{k=0}^{J_i-1} \rho_i\big(t_i^k\big) & \text{if } t_i = t_i^{J_i-1}. \end{cases}$$

It is easy to see that this specification of $\tilde{v}_i$ makes the principal's objective function in this problem the same as the expression in equation (13).

We specify the evidence structure as follows. For any t_i other than $t_i^{J_i-1}$ or $t_i^{J_i}$, we have $\mathcal{E}_i(t_i) = \{\{t_i\}, T_i\}$. Also, $\mathcal{E}_i(t_i^{J_i-1}) = \mathcal{E}_i(t_i^{J_i}) = \{T_i\}$. The implied incentive compatibility constraints are the following. First, since types $t_i^{J_i-1}$ and $t_i^{J_i}$ can each claim to be the other and send the other's (trivial) maximal evidence, each must weakly prefer her own allocation. Since $t_i^{J_i-1}$ is a negative type and $t_i^{J_i}$ is positive, this implies $\hat{p}_i(t_i^{J_i-1}) \leq \hat{p}_i(t_i^{J_i})$. Hence any other negative type prefers imitating $t_i^{J_i-1}$ to imitating $t_i^{J_i}$, while any positive type has the opposite preference. So the only other incentive compatibility constraints are $\hat{p}_i(t_i) \leq \hat{p}_i(t_i^{J_i-1})$ for any negative type t_i and $\hat{p}_i(t_i) \geq \hat{p}_i(t_i^{J_i})$ for any positive type t_i, the same constraints as in the costly-verification model.

Hence we can apply our results on optimal mechanisms with Dye evidence to compute the optimal mechanism for the evidence model as a function of $\tilde{v}_i$. We can then substitute in terms of v_i to rewrite in terms of the original costly-verification model. It is straightforward to show that doing so for the case considered in BDL (2014) or for the case considered in Erlanson and Kleiner (2017) yields the optimal mechanism identified there.

Because the assumptions used here also cover these cases, we can use this approach and the characterization given in Examples 2 and 3 of Section 3.1 to characterize optimal mechanisms with costly verification for the case where the principal allocates multiple identical goods or the case where he allocates a "bad."

REFERENCES

BEN-PORATH, E., E. DEKEL, AND B. L. LIPMAN (2014): "Optimal Allocation With Costly Verification," *American Economic Review*, 104, 3779–3813. [532,541,543,563]

——— (2019): "Supplement to 'Mechanisms With Evidence: Commitment and Robustness'," *Econometrica Supplemental Material*, 87, https://doi.org/10.3982/ECTA14991. [532]

BULL, J., AND J. WATSON (2007): "Hard Evidence and Mechanism Design," *Games and Economic Behavior*, 58, 75–93. [535,551]

DENECKERE, R., AND S. SEVERINOV (2008): "Mechanism Design With Partial State Verifiability," *Games and Economic Behavior*, 64, 487–513. [535,551]

DYE, R. A. (1985): "Disclosure of Nonproprietary Information," *Journal of Accounting Research*, 23, 123–145. [532,539,551]

ERLANSON, A., AND A. KLEINER (2017): "Costly Verification in Collective Decisions," Working Paper, September 2017. [532,543,551,560,563,565]

FARRELL, J. (1986): "Voluntary Disclosure: Robustness of the Unraveling Result," in *Antitrust and Regulation*, ed. by R. Grieson. Lexington Books, 91–103. [551]

FUDENBERG, D., AND J. TIROLE (1991): "Perfect Bayesian Equilibrium and Sequential Equilibrium," *Journal of Economic Theory*, 53, 236–260. [538]

GERSHKOV, A., J. GOEREE, A. KUSHNIR, B. MOLDOVANU, AND X. SHI (2013): "On the Equivalence of Bayesian and Dominant Strategy Implementation," *Econometrica*, 81, 197–220. [551]

GLAZER, J., AND A. RUBINSTEIN (2004): "On Optimal Rules of Persuasion," *Econometrica*, 72, 1715–1736. [534,551,552]

——— (2006): "A Study in the Pragmatics of Persuasion: A Game Theoretical Approach," *Theoretical Economics*, 1, 395–410. [534,551,552]

GREEN, J., AND J.-J. LAFFONT (1986): "Partially Verifiable Information and Mechanism Design," *Review of Economic Studies*, 53, 447–456. [551]

GROSSMAN, S. J. (1981): "The Informational Role of Warranties and Private Disclosures About Product Quality," *Journal of Law and Economics*, 24, 461–483. [551]

GUTTMAN, I., I. KREMER, AND A. SKRZYPACZ (2014): "Not Only What but Also When: A Theory of Dynamic Voluntary Disclosure," *American Economic Review*, 104, 2400–2420. [551]

HAGENBACH, J., F. KOESSLER, AND E. PEREZ-RICHET (2014): "Certifiable Pre–Play Communication: Full Disclosure," *Econometrica*, 82, 1093–1131. [551]

HART, S., I. KREMER, AND M. PERRY (2016): "Evidence Games: Truth and Commitment," Working Paper. [551,552]

——— (2017): "Evidence Games: Truth and Commitment," *American Economic Review*, 107, 690–713. [551, 552]

JUNG, W., AND Y. KWON (1988): "Disclosure When the Market Is Unsure of Information Endowment of Managers," *Journal of Accounting Research*, 26, 146–153. [551]

LIPMAN, B., AND D. SEPPI (1995): "Robust Inference in Communication Games With Partial Provability," *Journal of Economic Theory*, 66, 370–405. [535]

MANELLI, A., AND D. VINCENT (2010): "Bayesian and Dominant–Strategy Implementation in the Independent Private Values Model," *Econometrica*, 78, 1905–1938. [551]

MAS-COLELL, A., M. WHINSTON, AND J. GREEN (1995): *Microeconomic Theory*. New York: Oxford University Press. [550]

MILGROM, P. (1981): "Good News and Bad News: Representation Theorems and Applications," *Bell Journal of Economics*, 12, 350–391. [551]

SHER, I. (2011): "Credibility and Determinism in a Game of Persuasion," *Games and Economic Behavior*, 71, 409–419. [551,552]

Co-editor Joel Sobel handled this manuscript.

Manuscript received 9 January, 2017; final version accepted 31 October, 2018; available online 2 November, 2018.