Mechanism Design for Acquisition of/Stochastic Evidence[1]

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

First Preliminary Draft
September 2019

Current Draft
July 2025

[1] We thank the National Science Foundation, grant SES–1919319 (Dekel and Lipman), the US–Israel Binational Science Foundation, and the Foerder Institute at Tel Aviv University for support for this research. We also thank Ricky Vohra and three anonymous referees for useful comments and suggestions.

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

[3] Economics Department, Northwestern University, and School of Economics, Tel Aviv University. Email: eddiedekel@gmail.com.

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

Abstract

We explore two interrelated models of "hard information." In the *evidence–acquisition model*, an agent with private information searches for evidence to show the principal about her type. In the *signal–choice model*, a privately informed agent chooses an action generating a random signal whose realization may be correlated with her type. The signal–choice model is a special case and, as we show, under certain conditions, a reduced form of the evidence–acquisition model. We develop tools for characterizing optimal mechanisms for these models by giving conditions under which some aspects of the principal's optimal choices can be identified only from the information structure, without regard to the utility functions or the principal's priors.

1 Introduction

We explore two models of “hard information.” In the first, the *evidence–acquisition model*, the agent chooses among actions that generate random signals depending on her type. The agent then chooses which realization to present to a principal who chooses an action affecting both of their utilities. The second model is a special case and, under some conditions, a reduced form of the evidence–acquisition model. In this *signal–choice model*, the agent chooses among random signals, the realization of which the principal observes.

Most of the literature on evidence analyzes a principal–agent model where the agent is endowed with evidence and the question is what evidence he will disclose. Our two models extend the usual model by considering decisions by the agent which generate evidence and where there is ex ante uncertainty regarding the evidence that will materialize. Both models are natural for applications. For an example of the evidence–acquisition model, consider a division within an organization which wants additional funding for a project it is developing, say, a new product. The division can develop and test a prototype or do other market research to obtain evidence regarding the profitability of the product. The evidence resulting from the research is random ex ante. The division may choose which parts of its results to share with the organization.

As an example of a signal–choice model in applications, consider a lawyer who has private information about the innocence or guilt of her client trying to persuade a judge. When the lawyer calls a witness to the stand, she may know more about what the witness will say than the judge does, but may not be able to perfectly predict the witness’ testimony. In this sense, the witness is a random signal, the realization of which depends stochastically on the lawyer’s private information. Similarly, when an agent gives the name of a recommender to the principal, she may not know exactly what the recommender will say. In both cases, the agent effectively chooses a random variable, the realization of which she and the principal will see together.

We develop conditions under which we can restrict attention to a relatively simple class of mechanisms in these two classes of problems. For the evidence–acquisition model, we identify a sufficient condition for the evidence structure to be *simplifiable* in the sense that the evidence the principal requests from the agent is independent of the utility functions and priors. Identifying this request eliminates the need to optimize over it, making the analysis much less complex. We develop this result in Section 3.2 and show in Section 3.3 that when the evidence structure is simplifiable, we can reduce the evidence–acquisition model to the signal–choice model.

In Section 4, we give conditions under which we can similarly identify the signal choice the principal requests from the agent, leading to a further simplification. Proofs are in the Appendix.

2 Models

In this section, we discuss the "primitives" of the model, reserving discussion of the specifics of the mechanism for later sections.

We consider a principal and an agent. The agent has a finite set of types T where the realization $t \in T$ is the agent's private information. The principal's prior over T is denoted τ and is assumed to have full support. The principal has a finite set of actions X. An element of X specifies all aspects of the principal's action, including allocation of goods, monetary transfers, provision of resources, or other activities. After possibly several rounds of information exchange between the agent and the principal, the principal chooses some $x \in X$. The utility functions of the agent and principal are $u : T \times X \to \mathbf{R}$ and $v : T \times X \to \mathbf{R}$ respectively.[1] In what follows, we refer to (X, τ, u, v) as the *payoff structure*.

There is a set $\mathcal{L}$ of all possible evidence messages which could potentially be shown by the agent. For simplicity, we assume $\mathcal{L}$ is finite, but this is not needed for the results. Information exchange includes the transmission of an evidence message and possibly also includes cheap talk reports by the agent.

We consider two ways of modeling information transmission, one of which is a

[1] For some purposes, it is natural to also let the agent's and/or principal's utility depend on aspects of the information transmission (discussed next). We avoid adding these to the utility functions as it would complicate the notation even further, but note that adding costs of evidence acquisition (for the principal or the agent) would not affect any of our results for the evidence–acquisition model.

special case and, under certain conditions, a reduced form of the other. First, we consider the *evidence–acquisition model*, a model where the agent searches to find evidence. The agent has a variety of ways to try to obtain evidence. This search process could be sequential or one–shot. Rather than model this process, we focus on its outcomes by treating the agent as choosing a probability distribution over the evidence set she ultimately obtains. Formally, let A_t denote the set of evidence–gathering actions available to type t, with typical element $a \in A_t$, where we identify the action a with the probability distribution over evidence sets it generates. That is, $a \in \Delta(2^{\mathcal{L}} \setminus \{\emptyset\})$.[2] We denote a typical set of evidence as $M \subseteq \mathcal{L}$. Let $\mathcal{M}$ be the set of possible message sets M that can be produced. That is, $\mathcal{M}$ is the collection of M such that there exists t and $a \in A_t$ with $M \in \text{supp}(a)$. The assumption that $\emptyset \notin \mathcal{M}$ means that the agent can always say *something*, even if it is not informative — e.g., "I have no evidence to present." If M is the realized set of messages, then the agent can present any one $m \in M$ to the principal.[3]

While we assume that the principal observes only the m sent by the agent and not the chosen evidence acquisition action a, the model (implicitly) includes the possibility that a is observable as well. To see this, suppose every set of messages that could be realized by the agent's choice of action a is disjoint from any set that could be realized from a'. Then observing message m reveals the evidence acquisition action to the principal. Similarly, we can assume that only some distribution choices are observable or that only some messages reveal a in this sense, so whether the distribution is observed is itself random and/or in the control of the agent.

The model incorporates the important specific case where there is a set of tests, say Q, where each $q \in Q$ and $t \in T$ define a probability distribution over sets of evidence messages (test results). In some settings (e.g., college admissions tests), it is natural to assume that the principal observes the test q selected by the agent. Again, our model allows but does not require such observability.

When we discuss the evidence–acquisition model, we refer to $\{A_t\}_{t \in T}$ as the *evidence structure.*

[2] For any set B, $\Delta(B)$ is the set of probability distributions over B.

[3] As in the usual deterministic evidence model, the assumption that the agent can present only one message is without loss of generality. For example, if the agent could present two messages, we would simply replace $\mathcal{L}$ with the set of pairs of messages.

A special case of the evidence–acquisition model is where the agent has no choice of what message to send at the last step. Formally, this special case is when for every $t \in T$ and every $a \in A_t$, every $M \in \text{supp}(a)$ is a singleton. For convenience, we write this special case, the *signal–choice model*, differently. Instead of referring to agent's choices as evidence acquisition actions, we write the set of options available to type $t \in T$ as a nonempty set $S_t \subseteq \Delta(\mathcal{L})$ and refer to an $s \in \Delta(\mathcal{L})$ as a *signal distribution*. The interpretation is that if the agent chooses $s \in \Delta(\mathcal{L})$, then the principal sees message $m \in \mathcal{L}$ with probability $s(m)$. Equivalently, we can think of this as the singleton message in the realized evidence set.

Similarly to our comments above about the observability of a, the model allows the possibility that the realized m reveals the agent's choice of s always, reveals it with some probability, or reveals it for some s choices but not others. We refer to $\{S_t\}_{t \in T}$ as the *signal structure*.

While we discuss the details of mechanisms below, we use the following timing structure throughout. In both models, we assume the agent knows her type at the outset. There may be cheap talk between the principal and the agent before the agent chooses an evidence action or a signal distribution. After this, the agent sees the realization of her action. In the evidence–acquisition case, this is a set of evidence messages and (perhaps after further cheap talk) she can then send one evidence message to the principal. In the signal–choice model, the principal also sees the realization, perhaps followed by more cheap talk. After this, the principal chooses $x \in X$.

Running Example, Part 1. We use the following example to illustrate ideas and results. The principal is an employer and the agent an employee. The agent's private information t is her productivity for the principal. Hence the agent wants the principal to think she has a high type and the principal wants to know the true type.

For an example of an evidence–acquisition technology in this context, suppose the agent of type t can choose a variety of ways to potentially demonstrate her ability. Each of these options gives a probability distribution over an "outcome" she generates, where this outcome is, on average, equal to her true type. However, she can also withhold part of this "outcome" and show a lower realization than what she actually generates. More formally, $a \in A_t$ if and only if the following two statements are true. First, every $M \in \text{supp}(a)$ takes the form $[0, m]$ for some $m \in \mathbf{R}_+$. (Note that this means the set $\mathcal{L}$ in this example is infinite, unlike in the general model. Nothing changes in the example if we take $\mathcal{L}$ to be a finite but "dense" subset of an appropriate interval of real numbers.) Note that any $a \in A_t$ corresponds to a probability distribution over $\mathbf{R}_+$ where if the realization of this random variable is m, this means the set of available evidence messages is $[0, m]$. The second property is that for any $a \in A_t$, the expectation of this associated random variable is t. That is, in the case where a has a finite support,

$$\sum_{[0,m] \in \text{supp}(a)} a([0, m])m = t.$$

The agent wants to persude the principal that her type is large, so it is natural to conjecture that the option of showing a lower outcome will never be used by the agent and hence is irrelevant. In fact, one of our results will be that only the upper bound of a given evidence set will be shown by the agent in an optimal mechanism. However, this result is independent of the preferences of the agent — the same is true even in a different problem where the agent wants to persuade the principal that her type is small (e.g., if the agent's type determines the level of effort the principal wants her to exert).

For an example of signal choice, we "convert" this evidence structure into a signal structure. Note that, in the acquisition model, the agent can pick a distribution over evidence sets and decide what message she will use from each set. That is, she can choose a particular distribution over sets of the form $[0, m]$ and decide for each upper bound m what message $m' \in [0, m]$ she will send to the principal. Recall that the agent of type t can only generate a distribution over sets of the form $[0, m]$ with the property that the expectation of the upper bound m is t. Hence when we convert to signals, this generates the set of signal distributions with expected value less than or equal to t. In other words, for a signal–choice version of this example, we let S_t, the set of signal distributions for type t, be the set of all probability distributions on $\mathbf{R}_+$ with expected value less than or equal to t. Thus signal distributions are either unbiased or biased "against" the agent. One can think of this as a stylized model where the agent can give the principal one name of a reference for the principal to contact. References cannot be systematically biased in the agent's favor, but the agent generally cannot predict exactly what a given reference will say. ∎

Related literature: The usual model of evidence considers games or mechanism design problems where the agent's set of feasible messages is a deterministic function of her type. Thus by presenting a message which is only feasible for a certain set of types, the agent proves her type is in this set. The usual model is a special case of our evidence–acquisition model where each type has a single evidence action that generates a single evidence set with probability 1 and a special case of our signal–choice model where every signal is degenerate. For early contributions in game theory, see Grossman (1981), Milgrom (1981), and Dye (1985). For early contributions in mechanism design theory, see Green and Laffont (1986), Glazer and Rubinstein (2004, 2006), Forges and Koessler (2005), Bull and Watson (2007), and Deneckere and Severinov (2008). More specific connections to some of these papers will be discussed below.

Several earlier papers consider models of evidence acquisition, but, with few exceptions, all assume the agent does not know her type and do not consider optimal mechanisms. Matthews and Postlewaite (1985), Che and Kartik (2009), Felgenhauser and Schulte (2014), DeMarzo, Kremer, and Skrzypacz (2019), and Shishkin (2024) consider models in which an uninformed agent chooses a test or experiment which may reveal information about her type. These papers vary in the specifics, but in all cases, the agent's action produces a probability distribution over a set of options for the agent to reveal, as in our model. While not a model of evidence acquisition, some similar issues arise in Banerjee and Chen (2025), which considers full implementation in a model with multiple agents who have exogenously determined probability distributions over the evidence they have available. The paper closest to ours, Ball and Kattwinkel (forthcoming), considers one privately informed agent and optimal mechanisms. It will be more convenient to discuss their model and its relationship to ours at the end of Section 3.

Our signal–choice model is related to several different literatures. There are a number of papers related to the testing/experimentation papers discussed above but where the principal directly observes the outcome of any experiments conducted by the agent — see, for example, Henry and Ottaviani (2019) or McClellan (2022). To the best of our knowledge, all of these papers consider uninformed agents, unlike our model.

Similarly, the signal–choice model can be thought of as an "informed agent" version of the Bayesian persuasion model of Kamenica–Gentzkow (2011). As in the Bayesian persuasion model, the agent chooses an "experiment" which reveals information to the principal. Our model differs from Kamenica–Gentzkow in four ways. First, we do not assume that every possible signal is feasible. Second, we assume the agent knows her type, though she may not know the outcome of the experiment.[4] Third, while Kamenica and Gentzkow assume the principal observes the full experiment, we do not assume this. Specifically, while we can allow the principal to observe the signal choice of the agent as discussed above, he cannot observe the signals that would have been chosen by other types. Finally, Kamenica and Gentzkow characterize the optimal structure for the agent, while our mechanism design results focus on the best choice for the principal.

Deb, Pai, and Said (2018), Silva (2020), Perez-Richet and Skreta (2022), and Espinosa ⓡ Ray (2023) also develop models that can be thought of a signal–choice models. However, these papers, while broadly related, focus on issues very different from the ones we explore.

3 Simplification in Evidence Acquisition

3.1 General Mechanisms

Using standard Revelation Principle type arguments, one can show that we can restrict attention to a certain class of direct truth–telling mechanisms. However, these mechanisms are rather complex for the signal–choice model and quite involved for the evidence–acquisition model. Henceforth we use the term *protocol* to refer to the sequence of stages of communication in a mechanism.[5]

[4] For work on Bayesian persuasion with privately informed agents, see Perez–Richet (2014), Hedlund (2017), Kosenko (2023), and Koessler and Skreta (2023).

[5] Gerardi and Myerson (2007) have shown that the Revelation Principle may not hold for sequential equilibrium in dynamic environments, raising questions about our multi–stage mechanisms. However, Sugaya and Wolitzky (2021) show that such problems do not arise in our single–agent setting.

For the signal–choice model, we have, in effect, an adverse selection problem (the agent's private knowledge regarding her type), followed by moral hazard (the agent's unobserved choice of a signal distribution). Thus a variation on Myerson's Revelation and Obedience Principle identifies the appropriate protocol.[6] First, the agent reports a type. Then the principal recommends a signal distribution. Finally, the agent chooses some distribution, the principal observes the realization m, and the principal chooses $x \in X$.

In the evidence–acquisition model, the problem is much more complex. We start with adverse selection (the agent's type), then have moral hazard (the agent's choice of a distribution over evidence sets), followed by more adverse selection (the realized set of evidence messages). Hence we start as in the signal choice case where the agent reports her type, the principal recommends an action, and the agent chooses an action. But after this, the agent makes a report of the realized evidence set, the principal recommends a message choice from this set, and the agent sends a message. Only then does the principal choose $x \in X$. One can show by examples (omitted for brevity) that, in general, each of these steps may be necessary for the principal to obtain the highest possible payoff.

In general, this protocol can be difficult to analyze. Not only are there numerous objects to choose with constraints that can be quite complex, but in addition (as is well–known — see, e.g., Glazer–Rubinstein (2004)), the optimum may require randomization by the principal over what recommendation to make. Our main results establish conditions under which we can identify the principal's recommendations in an optimal mechanism based only on the evidence/signal structure. Under these conditions, we can eliminate some of the above steps, greatly simplifying the class of mechanisms we need to consider and thus greatly simplifying the analysis.

In this section, we consider the evidence–acquisition model, developing our simplification of the signal–choice model in Section 4. We give a verbal description of the protocol and state our main result for this model, then develop the relevant notation.

The protocol for evidence–acquisition models has seven stages. We refer to this as the *full protocol for evidence–acquisition models.* Recall that $\mathcal{M}$ is the collection

[6] For similar results in the evidence literature, see Bull and Watson (2007) and Deneckere and Severinov (2008).

of M such that there exists t and $a \in A_t$ with $M \in \text{supp}(a)$.

Stage 1. The agent makes a report of a type $r \in T$.

Stage 2. Given the report, the principal requests a distribution a over evidence sets.

Stage 3. The agent chooses some feasible action a' and the evidence set M is realized.

Stage 4. The agent makes a report $\hat{M} \in \mathcal{M}$ of her realized message set.

Stage 5. The principal proposes a message $m \in \hat{M}$ for the agent to send.

Stage 6. The agent sends a message $\hat{m}$ from the set of messages she has available.

Stage 7. The principal chooses an action x as a function of the history he has observed.

Our main result for evidence acquisition gives a condition on the evidence structure which implies that each possible evidence set M has a "best" message in the sense that, without changing the mechanism's outcome, the principal can *always* ask for this message from M if the agent reports M. This allows us to drop Stages 4 and 5, going from the realization of the message set to the agent's choice of an evidence message in Stage 6. This simplification enables us to reduce the evidence–acquisition model to a signal–choice model.

The reader may prefer to skip the following notation (which continues to the end of this subsection) on first reading. To state the mechanism protocol formally, we use b's to denote the agent's pure strategies at various stages and g's to denote the principal's pure strategies. The agent chooses three objects. For stage 1, the agent chooses a reporting strategy $b_T : T \to T$. For stage 3, the agent chooses an action strategy giving her action as a function of her true type, her report, and the principal's recommendation, so $b_A : T \times T \times A \to A$, where we require the agent's choice to be feasible for her in the sense that $b_A(t, \cdot, \cdot) \in A_t$ for all t. For stage 5, the agent has a second reporting strategy, again a function of all she has seen and done, so $b_{\mathcal{M}} : T \times T \times A \times A \times \mathcal{M} \to \mathcal{M}$. Finally, for stage 6, the agent has an evidence presentation strategy, $b_{\mathcal{L}} : T \times T \times A \times A \times \mathcal{M} \times \mathcal{M} \times \mathcal{L} \to \mathcal{L}$. Of course, we require that $b_{\mathcal{L}}(t, r, a, a', M, \hat{M}, m) \in M$ — that is, if the agent's type is t, her report r, the recommended action a, her chosen action a', the realized message set M, the reported message set $\hat{M}$, and the requested message m, the evidence message the agent sends must be in M, the true message set. We let B_T, B_A, $B_{\mathcal{M}}$, and $B_{\mathcal{L}}$ denote the sets of these functions respectively.

Similarly, for stage 2, the principal chooses a recommendation strategy $g_A : T \to A$, giving his recommended action as a function of the reported type. For stage 5, he chooses a message request strategy $g_{\mathcal{L}} : T \times A \times \mathcal{M} \to \mathcal{L}$. We require that $g_{\mathcal{L}}(r, a, \hat{M}) \in \hat{M}$. That is, if the agent reported r, the principal requested action a, and the agent reported evidence set $\hat{M}$, the message the principal requests must be feasible for the agent given her reported evidence set. For stage 7, he chooses an action strategy $g_X : T \times A \times \mathcal{M} \times \mathcal{L} \times \mathcal{L} \to X$. Let G_A, $G_{\mathcal{L}}$, and G_X denote the sets of these functions.

Let the principal's set of pure mechanisms or pure strategies be denoted $G = G_A \times G_{\mathcal{L}} \times G_X$. Let $\Gamma = \Delta(G)$ with typical element γ. We let $(\gamma_A, \gamma_{\mathcal{L}}, \gamma_X)$ denote the equivalent behavior strategy to γ. Let $B = B_T \times B_A \times B_{\mathcal{M}} \times B_{\mathcal{L}}$ denote the agent's set of pure strategies. Let $\beta \in \Delta(B)$ denote a typical mixed strategy for the agent.

A version of the standard Revelation Principle for this class of models says that without loss of generality, we can restrict attention to mechanisms where it is optimal for the agent to report truthfully and to obey the principal's recommendations at every stage along the equilibrium path.

To define incentive compatibility more precisely, note that any (β, γ, t) induces a probability distribution over the principal's action x. We denote this distribution by $\mu(x \mid \beta, \gamma, t)$. Let $U(\beta, \gamma, t)$ denote the agent's expected utility in the mechanism γ given strategy β when her type is t or

$$U(\beta, \gamma, t) = \sum_{x \in X} u(t, x)\mu(x \mid \beta, \gamma, t).$$

We say that a pure strategy $\hat{b} = (\hat{b}_T, \hat{b}_A, \hat{b}_{\mathcal{M}}, \hat{b}_{\mathcal{L}})$ is *truthful and obedient* if for all t, a, M, and m, we have $\hat{b}_T(t) = t$, $\hat{b}_A(t, t, a) = a$, $\hat{b}_{\mathcal{M}}(t, t, a, a, M) = M$, and $\hat{b}_{\mathcal{L}}(t, t, a, a, M, M, m) = m$. That is, the agent reports truthfully and obeys the principal at all stages. Throughout, we use $\hat{b}^*$ to denote any such honest and obedient strategy.[7] Note that the outcome of the mechanism is the same for any choice of an honest and obedient strategy.

A mechanism γ for the evidence–acquisition model is *incentive compatible* if for all t,

$$U(\hat{b}^*, \gamma, t) \geq U(b, \gamma, t), \quad \forall b \in B$$

for any truthful and obedient strategy $\hat{b}^*$. (Clearly, this condition also implies that $\hat{b}^*$ is a better strategy for the agent than any mixed strategy $\beta \in \Delta(B)$.)

Given any incentive compatible γ, let $\mu^*(x \mid \gamma, t) = \mu(x \mid \hat{b}^*, \gamma, t)$. We refer to μ^* as the *mechanism outcome*.

3.2 Simplifiability

Clearly, this is a complex protocol, giving us a complex set of mechanisms and incentive compatibility constraints. In the rest of this section, we introduce a notion of simplifiability and identify conditions under which this holds.

The idea is to identify some choices by the principal in a way which depends on the evidence structure but uses no information about the preferences of the principal or the agent. The ability to identify such choices allows us to greatly reduce the complexity of the protocol and the mechanism design problem.

More specifically, we identify the principal's response at Stage 5. If for every possible $\hat{M}$, there is a specific $m \in \hat{M}$ that the principal will always ask for, regardless of the preferences or other details of the model, then we can take as given that the principal requests this message and delete Stage 5. This enables us to eliminate Stage 4 since the agent's report of a message set is needed only to give the principal the opportunity to make such a recommendation. Hence we can combine Stages 3 and 6, skipping Stages 4 and 5.

Thus we say that an evidence structure is *simplifiable* if for every set of messages

[7] Note that there are many such strategies since we do not specify how the agent behaves on histories inconsistent with her strategy. Truth–telling and obedience are without loss of generality on path, but not necessarily off path.

M that may be some type's evidence set, there is a message $m^*_M \in M$ that the principal can always ask for, regardless of the payoff structure. More precisely, evidence structure $\{A_t\}_{t \in T}$ is simplifiable if for every $M \in \mathcal{M}$, there exists $m^*_M \in M$ such that for every payoff structure (X, τ, u, v) and every incentive compatible γ given that payoff structure, there is an incentive compatible γ^* for that payoff structure with the following two properties. First, $\gamma^*_{\mathcal{L}}(t, a, M)(m^*_M) = 1$. That is, the principal always recommends message m^*_M when the reported message set is M. Second, $\mu^*(x \mid \gamma^*, t) = \mu^*(x \mid \gamma, t)$ for all $x \in X$ and $t \in T$, so that the two mechanisms have the same outcome for every t.

As the name of this property is intended to emphasize, when an evidence structure is simplifiable, the analysis required is indeed much simpler. By identifying the message the principal can *always* request, we eliminate the need to determine the best way to use evidence to incentivize truthful reporting. So we have the answer to the optimization at Stage 5 of the protocol. In addition, we eliminate the need for random requests by the principal, a complication that is necessary in general otherwise. This means the agent knows what message the principal will request as a function her report of her evidence set, so we no longer require that report, eliminating Stage 4 and its associated incentive constraints.

We show that a natural generalization of the notion of normal evidence in the literature gives a sufficient condition for an evidence structure to be simplifiable.

In the literature with exogenously given evidence, it is well–known that one may need the principal to randomize over which message to request in response to the agent's type report. The idea is to prevent the agent from knowing how the principal will check various possible lies, thus deterring misreporting. See Glazer and Rubinstein (2004) for illustrative examples. As shown by Bull and Watson (2007), though, under a condition they call normality which Lipman and Seppi (1995) had previously called the full reports condition, this request by the principal is not needed. Normality or full reports says that the agent has available a message which reveals as much information as all the messages the agent has available, a message equivalent to showing the entire set of available messages. Thus asking for this message is the "best" way to deter lies.

We generalize this property to evidence–acquisition models as follows. We say that the evidence structure satisfies *normality* if for every $M \in \mathcal{M}$, there exists $m_M^* \in M$ such that for every $M' \in \mathcal{M}$, we have

$$m_M^* \in M' \iff M \subseteq M'.$$

We refer to the message m_M^* as the *maximal evidence* for M. As the notation suggests, this will be the message used in simplifiability.

To understand the normality condition, note that $M \subseteq M'$ trivially implies $m_M^* \in M'$ since $m_M^* \in M$. However, we write the condition as an "if and only if," including this trivial direction, to emphasize the following idea. Intuitively, the only thing that presenting a particular message m proves to the principal is that the agent is able to present this message — that is, that the set of messages the agent has available includes m. With this idea in mind, think of M' as the principal's "guess" about the agent's evidence set and M as the true set. Then when the agent presents m_M^*, the principal learns that M' contains this message. The statement of normality says that this is equivalent to the principal learning that M' contains all of M. In this sense, showing m_M^* reveals exactly what showing every message in M would reveal. Put differently, learning that m_M^* is feasible (i.e., that the true evidence set contains it) reveals exactly the same information about the agent's set of messages as learning that every message in M is feasible (i.e., is contained in the true evidence set).

Running Example, Part 2. In our evidence–acquisition example, $\mathcal{M}$ contains every interval of the form $[0, m]$ for $m \in \mathbf{R}_+$ since each such interval can be generated with positive probability by some (actually, by any) type. Hence it is easy to see that the most informative message, m_M^*, for the interval $[0, m]$ is the upper bound, m. That is, $m_{[0,m]}^* = m$ or, equivalently, $M = [0, m_M^*]$. This is true as for any $m' \in \mathbf{R}_+$, we have $m_M^* \in [0, m']$ if and only if $[0, m_M^*] \subseteq [0, m']$. Hence our running example satisfies normality. As Theorem 1 below will indicate, this means that there is an optimal mechanism using only the upper bounds of the intervals, regardless of the payoff structure, as asserted earlier. ▌

Theorem 1. *If the evidence structure is normal, then it is simplifiable. In particular, the principal can restrict attention to mechanisms which for every reported evidence set M requests the maximal evidence from M.*

Remark 1. The proof of Theorem 1 is trivially adapted to show a stronger result. To be specific, say that evidence message m dominates evidence message m' if for every $M \in \mathcal{M}$ such that $m \in M$, we have $m' \in M$. In other words, m' is included in more evidence sets (in the sense of set inclusion). We prove Theorem 1 by showing that for any incentive compatible mechanism that asks for some message that is not maximal evidence when M is reported, we can find another incentive compatible mechanism with the same outcome that requests the maximal evidence message for M instead. The argument only uses the fact that the maximal evidence message for M dominates in this sense any other message in the set. Hence replacing the initial message with any m' and the maximal evidence message with any m dominating it gives a proof that we never need to use any message that is dominated by another.

It is worth emphasizing that the presentation of a message is evidence directly about the agent's set of evidence, not about the agent's type t. It provides evidence only indirectly about t since types differ in terms of which evidence sets they are likely to obtain. To see more concretely how normality reflects this fact, consider the following example.

Example 1. The agent has two types, t_1 and t_2. Each type has only one distribution over evidence sets. Type t_1 obtains evidence set $\{m_1\}$ with probability 1/2 and $\{m_1, m_2\}$ with probability 1/2. Type t_2 receives evidence set $\{m_2\}$ with probability 1. This evidence technology violates normality. First, note that any singleton evidence set trivially has a maximal evidence message since if $M = \{m\}$, then it is obviously true that for any M', $m \in M'$ iff $M \subseteq M'$. So if normality fails, it is because $\{m_1, m_2\}$ has no maximal evidence message. It is easy to see that this is the case. For either message $m' \in \{m_1, m_2\}$, the singleton $\{m'\}$ is also an element of $\mathcal{M}$. Clearly, then, m' cannot be maximal since $m' \in \{m'\}$ but $\{m_1, m_2\} \not\subseteq \{m'\}$. In Appendix B, we show that this evidence structure is not simplifiable.

To see why this is surprising, note that if the agent presents m_1 to the principal, she proves that her type is t_1 as type t_2 never has this message available. Yet m_1 is not maximal evidence from $\{m_1, m_2\}$. Intuitively, presentation of m_1 proves the agent's type but presenting both m_1 and m_2 would prove more about the agent's available messages than m_1 proves. ▌

One way to understand this is to observe that in standard deterministic evidence models, the agent's type identifies exactly her set of available messages. In a sense, in the current model, the agent's *full* type is the pair (t, M) where M is the set of messages the agent has. So in this example, unlike in deterministic evidence models, proving that the "type" is t does not prove the agent's full type.[8]

The following example shows that normality is *not* necessary for simplifiability.

Example 2. As in Example 1, the agent has two types, t_1 and t_2, each of which has only one distribution over evidence sets. Now type t_1 has evidence set $\{m_1, m_2\}$ with probability 1, while t_2 obtains evidence set $\{m_1\}$ with probability $p < 1/2$ and $\{m_2\}$ otherwise. This evidence technology violates normality for the same reasons as in Example 1. However, as we show in Appendix C, this evidence structure is simplifiable. More specifically, any outcome achievable by an incentive compatible mechanism with this evidence structure can be achieved from a mechanism which requests m_1 when the agent reports type t_1 and evidence set $\{m_1, m_2\}$. Intuitively, this is the message t_2 is least likely to be able to imitate and this means it is best for deterring deviations. Note that when an evidence structure is normal, the maximal evidence message also has the property that it is the message in the evidence set that other types are least likely to have available. ▌

Necessary and sufficient conditions for simplifiability are not straightforward. Necessary conditions, in particular, are difficult to obtain as they hinge on many details regarding the variety of distributions over evidence sets that are possible. For example, the properties of some evidence set M only matter if M has positive probability under the evidence–acquisition actions chosen in some optimal mechanism. Hence we cannot separate necessary properties on the evidence structure from a characterization of which actions are "needed," a difficult undertaking in its own right.

One way to understand normality is to observe that it is necessary and sufficient for a stronger version of simplifiability. Note that normality only depends on $\mathcal{M}$, the collection of possible message sets, not which type might have which set. Examples 1 and 2 give evidence structures with the same $\mathcal{M}$ but one is simplifiable and one is not. In fact, normality is the unique condition with this property.

[8] Another way to see this point is to redefine the type space to be the set of possible (t, M) and the set of feasible messages for "type" (t, M) to be M. Applying the standard definition of normality to this model yields our definition.

To state this more precisely, define an evidence structure to be *robustly simplifiable* if it is simplifiable and every type space and evidence structure with the same $\mathcal{M}$ is also simplifiable. Because normality only depends on $\mathcal{M}$ and implies simplifiability, it is clearly sufficient for robust simplifiability. In Appendix D, we show that normality is also necessary for robust simplifiability.

Theorem 1 implies that we can use a simpler protocol under normality. Since the principal can always recommend the maximal evidence message for any reported message set, we do not need the stage where he makes this recommendation. Hence we do not need the agent to report the message set since the mechanism does not depend on it.

We refer to the following as the *abbreviated protocol for evidence acquisition*:

Stage 1. The agent reports a $t \in T$.

Stage 2. Given the report, the principal recommends a distribution over evidence sets for the agent.

Stage 3. The agent chooses a distribution and the evidence set M is realized.

Stage 4. The agent sends a message m from the set of available messages M.

Stage 5. The principal chooses an action as a function of the history he has observed, namely the agent's report, the recommended distribution, and the message m.

Again, the reader may wish to skip the following definitions and proceed directly to Corollary 1 below. We abuse notation by using the same notation to denote strategies for this protocol. Hence a pure strategy for the agent is now $b = (b_T, b_A, b_{\mathcal{L}})$ where $b_T : T \to T$ and $b_A : T \times T \times A \to A$ as before. Also, $b_{\mathcal{L}} : T \times T \times A \times A \times \mathcal{M} \to \mathcal{L}$ where $b_{\mathcal{L}}(t, r, a, a', M) \in M$ gives the message the agent sends as a function of her true type t, her reported type r, the principal's recommended distribution a, the distribution she actually chose a, and the realized set M. A pure strategy for the principal is $g = (g_A, g_X)$ where $g_A : T \to A$, with $g_A(t) \in A_t$ and $g_X : T \times A \times \mathcal{L} \to X$ gives the principal's choice of x as a function of the agent's report, the recommended distribution, and the observed message. Again, we denote the agent's pure strategies by $B = B_T \times B_A \times B_{\mathcal{L}}$ and the principal's pure strategies by $G = G_A \times G_X$.

The definition of incentive compatibility for this class of mechanisms is similar to the preceding. Briefly, incentive compatibility requires that an optimal strategy for the agent is to report t truthfully (so $b_T(t) = t$), to follow the principal's recommendation (so $b_A(t, t, a) = a$), and to use maximal evidence (so $b_{\mathcal{L}}(t, t, a, a, M) = m^*_M$).

We have the following corollary, proved in Appendix E:

Corollary 1. *Assume the evidence structure is normal or, more generally, simplifiable. Then for any incentive compatible mechanism in the full protocol for evidence–acquisition models, there is an incentive compatible mechanism for the abbreviated protocol with the same outcome.*

3.3 Reduction to Signal Choice

When the evidence structure is simplifiable, we can reduce the mechanism design problem for the evidence–acquisition model to the mechanism design problem for the signal–choice model. To show this, we first describe the latter. It is easy to see that we can assume the following *protocol for signal–choice.*

Stage 1. The agent reports a $t \in T$.

Stage 2. Given the report, the principal requests a signal distribution.

Stage 3. The agent chooses a signal distribution s as a function of her type, her report, and the recommendation of the principal, with the resulting message seen by the principal.

Stage 4. The principal chooses an outcome as a function of what has been said.

Formally, let a reporting strategy for Stage 1 be denoted $b_T : T \to T$. A pure strategy for the principal for Stage 2 is denoted $g_S : T \to S$. Let $b_S : T \times T \times S \to S$ with $b_S(t, r, s) \in S_t$ denote a typical pure strategy for the agent for Stage 3. Finally, let $g_X : T \times S \times \mathcal{L} \to X$ denote a typical pure strategy for the principal for the last stage. Abusing notation, again let $B = B_T \times B_S$ denote the set of pure strategies for the agent and $G = G_S \times G_X$ the set of pure strategies for the principal in this protocol. By the Revelation Principle, we can focus on mechanisms $\gamma \in \Gamma$ with the property that any strategy $\hat{b}^* = (\hat{b}^*_T, \hat{b}^*_S)$ for the agent satisfying $\hat{b}^*_T(t) = t$ and $\hat{b}^*_S(t,t,s) = s$ is a best reply for the agent to γ. Again, we refer to any such $\hat{b}^*$ as truthful and obedient. Given an incentive compatible mechanism γ, we can define the mechanism outcome as the function mapping t to probability distributions over X, analogously to the above. I.e., we can write $\hat{\mu}^*(x \mid \gamma, t)$ as the probability distribution over x induced by the strategies $(\hat{b}^*, \gamma)$ given the agent's type is t.

In the evidence–acquisition model, we can think of the agent choosing a and simultaneously choosing her *messaging strategy* — that is, her strategy for which message m to send as a function of the realization of the message set M. As we vary the agent's choice of distribution and messaging strategy, we trace out a set of probability distributions over messages m that the principal will observe. Thus we can replace the selection of a distribution/messaging strategy with the selection of a signal distribution. In general, this change reduces the principal's ability to influence the agent's decisions and will lead to a less effective mechanism. However, when the evidence structure is simplifiable, the ability to reduce to the abbreviated protocol implies that this change does not harm the principal.

Formally, fix an evidence–acquisition model. We construct a signal–choice model from it as follows. For any $a \in A$ and any function $\sigma : \text{supp}(a) \to \mathcal{L}$ such that $\sigma(M) \in M$, we can define a signal $s \in \Delta(\mathcal{L})$ by

$$s(m) = a\left(\{M \mid \sigma(M) = m\}\right).$$

Let $\Sigma(a)$ denote the set of such σ functions given a and let $s_{(a,\sigma)}$ denote the distribution on $\mathcal{L}$ induced by (a, σ). Let

$$S_t = \{s_{(a,\sigma)} \mid a \in A_t,\ \sigma \in \Sigma(a)\}.$$

The following result explains the sense in which the signal–choice model so constructed is equivalent to the evidence–acquisition model under simplifiability.

Theorem 2. *In the evidence–acquisition model, fix any incentive compatible mechanism γ. If the evidence structure is simplifiable, there exists an incentive compatible mechanism γ^* in the signal–choice model constructed from it that is equivalent to γ in the following sense. For any truthful and obedient strategy $\hat{b}^*$ for the agent in the* *signal–choice model given γ^*, we have*

$$\mu^*(x \mid \gamma, t) = \hat{\mu}^*(x \mid \gamma^*, t), \ \forall x \in X, \ t \in T,$$

so γ and γ^* have the same mechanism outcomes for every $t \in T$.

In short, given normality or, more generally, simplifiability, any outcome that can be induced by a mechanism for the evidence–acquisition model can be induced by a mechanism in the protocol for the associated signal–choice model.

One can consider mechanisms with different timing. For example, perhaps the agent only comes to the principal *after* having generated evidence. Recognizing this, the optimal mechanism takes into account the way the rules of the mechanism affect these incentives. For example, this seems like a natural way to think about courts. The lawyers know the rules of the court in advance and work to obtain evidence before bringing the case to court. It is easy to show the analogs of Theorem 1, Corollary 1, and Theorem 2 for this model. More specifically, it is still true that normality implies an appropriate generalization of simplifiability, enabling us to use (an appropriately modified version of) the abbreviated protocol and reduce to a version of the signal–choice model.

4 Simplification in Signal Choice

In this section, we focus on the signal–choice model, where, as just shown, this can be interpreted as a reduced form of the evidence–acquisition model under simplifiability.

While simplifiability (as the name indicates) greatly simplifies the mechanism design problem, the problem is still complex. We next turn to conditions under which we can identify the signal choice the principal requests as a function of the type.

Recall that $\mathcal{L}$ is finite. In this section, we write a signal distribution $s \in S$ as a row vector of length $\#\mathcal{L}$. Fix t^* and $s^*, \hat{s}^* \in S_{t^*}$. We say that s^* is *more informative than* $\hat{s}^*$ if there exists an $\#\mathcal{L} \times \#\mathcal{L}$ Markov matrix Λ such that $s^*\Lambda = \hat{s}^*$ and for every t and every $s \in S_t$, $s\Lambda \in \text{conv}(S_t)$.[9]

In the case where each S_t is finite, we can give an equivalent statement which will aid in clarifying the intuition of this condition. Let $\mathcal{S}$ denote the matrix formed by "stacking" the signal distributions. In other words, this is a matrix with $\#\mathcal{L}$ columns and a number of rows equal to $\sum_t \#S_t$. The first $\#S_{t_1}$ rows are the signal distributions available to t_1, the next $\#S_{t_2}$ rows those available to t_2, etc. Note that if $s \in S_t \cap S_{t'}$ for $t \neq t'$, then s appears (at least) twice in the matrix. Then s^* is more informative than $\hat{s}^*$ if there exists a Markov matrix Λ such that $\mathcal{S}\Lambda = \hat{\mathcal{S}}$ where the matrix $\hat{\mathcal{S}}$ has $\hat{s}^*$ in the row corresponding to s^* in $\mathcal{S}$ and for any row s of $\hat{\mathcal{S}}$ corresponding to one of type t's signal distributions, we have $s \in \text{conv}(S_t)$.

To see the intuition, recall Blackwell–Girshick's (1954) (BG) comparison of experiments. In their model, there are n states of the world. An experiment gives a probability distribution over a finite set of observations as a function of the state of the world. If there are N possible observations, we can write this as an $n \times N$ matrix E where e_{ij} is the probability of observation j in state i. Suppose we have two experiments, E and F. BG say experiment E is more informative than experiment F if there exists a Markov matrix Λ such that $E\Lambda = F$. The matrix Λ defines a garbling of the results of experiment E, so this says that F can be obtained from E by adding random noise.

Thus we can interpret our informativeness comparison as saying that the "experiment" $\mathcal{S}$ is more informative than "experiment" $\hat{\mathcal{S}}$ in the sense that we can obtain the latter by adding noise to the former. To understand the sense in which $\mathcal{S}$ and $\hat{\mathcal{S}}$ can be thought of as experiments, note that the rows in an experiment correspond to states of the world, while a row in $\mathcal{S}$ corresponds to a (type, signal distribution) pair. Intuitively, just as we can think of (t, M) as the (partly endogenous) "full type" in the evidence–acquisition model, it is natural to think of (t, s) as the (partly endogenous) "full type" in the signal–choice model.

To see why the existence of Λ implies s is more informative than s', suppose we have a mechanism in which the principal recommends s' if the agent reports that her type is t. Suppose the principal changes the mechanism to recommend s

[9] A matrix is Markov if all entries are non–negative and every row sum is 1.

in this situation instead and changes no other recommendations. Suppose that the principal's response to messages he subsequently receives from the agent after this recommendation is to "garble" them according to the Markov matrix Λ and then to respond the way the original mechanism specified. If the agent uses signal s, then the resulting distribution over the garbled message will be $s\Lambda$. By hypothesis, this is s'. Thus the distribution over the principal's choice of x will be the same as in the original mechanism. Suppose that the agent's true type is $\hat{t}$, she reports t, and that after receiving a signal recommendation from the principal, she uses some signal $\hat{s} \in S_{\hat{t}}$. Then the induced distribution over garbled messages will be $\hat{s}\Lambda$. By hypothesis, this is an element of $\text{conv}(S_{\hat{t}})$. In other words, in the original mechanism, type $\hat{t}$ could have generated this distribution over messages by a particular randomization over her available signals. Thus the expected outcome this type would generate is something she could have generated in the original mechanism. If the original mechanism was incentive compatible, then this deviation is not profitable. Thus the new mechanism is incentive compatible and generates the same outcome as the original one.

To understand this condition better, consider the following examples.

Example 3. Suppose there are three types, so $T = \{t_1, t_2, t_3\}$, and three messages, so $\mathcal{L} = \{m_1, m_2, m_3\}$. The first two types have only one signal distribution each, so $S_{t_1} = \{s_1\}$ and $S_{t_2} = \{s_2\}$, but t_3 has two signal distributions so $S_{t_3} = \{s_3, s_3'\}$. The distributions are given by

		m_1	m_2	m_3
S_{t_1}	s_1	1	0	0
S_{t_2}	s_2	0	1	0
S_{t_3}	s_3	0	0	1
	s_3'	1/2	1/2	0

It seems very intuitive that if the agent claims to be t_3, the principal should insist on signal s_3. It is easy to see that there is a Markov matrix Λ establishing that s_3 is more informative than s_3'. In particular, if we let

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1/2 & 1/2 & 0 \end{pmatrix},$$

we get that $s_1\Lambda = s_1$, $s_2\Lambda = s_2$, and $s_3\Lambda = s_3'\Lambda = s_3'$, so the conditions are met. ▮

Example 4. Suppose $T = \{t_1, t_2\}$, $\mathcal{L} = \{m_1, m_2\}$, $S_{t_1} = \{s_1\}$, and $S_{t_2} = \{s_2, s_2'\}$ where

		m_1	m_2
S_{t_1}	s_1	1	0
S_{t_2}	s_2	0	1
	s_2'	1/2	1/2

Again, it seems intuitive that if the agent claims to be t_2, the principal should ask for signal s_2. However, s_2 is not more informative than s_2' according to our definition. To have s_2 more informative than s_2', we require the Markov matrix Λ to satisfy, among other properties, $s_1\Lambda = s_1$ and $s_2\Lambda = s_2'$. It's easy to show that the only Markov matrix satisfying these two properties is

$$\Lambda = \begin{pmatrix} 1 & 0 \\ 1/2 & 1/2 \end{pmatrix}.$$

But then $s_2'\Lambda = (3/4, 1/4)$ which is not in the convex hull of $(0, 1)$ and $(1/2, 1/2)$. Intuitively, our construction has the principal changing from a mechanism where t_2 sends s_2' to one where she sends s_2 by treating a message of m_2 as if it were a 50–50 randomization over m_1 and m_2 and treating m_1 as m_1. But then by playing s_2', t_2 can effectively put more probability on the principal interpreting her message as m_1 in this mechanism than in the original, potentially creating profitable deviations. In Appendix G, we give an example of an outcome that can only be achieved by requesting s_2' from t_2 to illustrate. ▮

Example 5. As in Example 4, suppose $T = \{t_1, t_2\}$, $\mathcal{L} = \{m_1, m_2\}$, $S_{t_1} = \{s_1\}$, and $S_{t_2} = \{s_2, s_2'\}$, but now we have

		m_1	m_2
S_{t_1}	s_1	1/2	1/2
S_{t_2}	s_2	1/4	3/4
	s_2'	2/3	1/3

Here it is not obvious what signal the principal should ask type t_2 to use since s_1 is "between" s_2 and s_2'. However, the fact that s_2' is "closer" to s_1 than is s_2 implies s_2

is more informative than s_2'. More specifically, letting

$$\Lambda = \begin{pmatrix} 1/6 & 5/6 \\ 5/6 & 1/6 \end{pmatrix},$$

we get $s_1\Lambda = s_1$, $s_2\Lambda = s_2'$, and $s_2'\Lambda = (7/18, 11/18) \in \text{conv}\{(1/4, 3/4), (2/3, 1/3)\}$. ▌

Theorem 3. *In the signal–choice model, fix any incentive compatible mechanism γ with marginal γ_S on G_S. If there exists t^* and $s^*, \hat{s}^* \in S_{t^*}$ such that s^* is more informative than $\hat{s}^*$, then there exists an incentive compatible mechanism (γ_S^*, γ_X^*) satisfying the following two properties. First,*

$$\gamma_S^*(t)(s) = \begin{cases} \gamma_S(t)(s), & \text{if } t \neq t^* \text{ or } s \notin \{s^*, \hat{s}^*\}; \\ \gamma_S(t^*)(s^*) + \gamma_S(t^*)(\hat{s}^*), & \text{if } t = t^* \text{ and } s = s^*; \\ 0, & \text{if } t = t^* \text{ and } s = \hat{s}^*. \end{cases}$$

That is, γ^* moves any probability on recommending $\hat{s}^*$ for t^* to recommending s^* instead. Second, for all t,

$$\hat{\mu}^*(x \mid \gamma, t) = \hat{\mu}^*(x \mid \gamma^*, t), \ \forall x \in X.$$

That is, γ and γ^* generate the same probability distribution over actions by the principal for every $t \in T$.

Remark 2. Theorems 1 and 3 are connected in the following way. Suppose we begin with an evidence–acquisition model satisfying normality. By Theorem 2, we can reduce this to a signal–choice model where each signal distribution corresponds to a choice of a distribution over evidence sets and a messaging strategy for which message to send as a function of the realized set. Fix a particular distribution over evidence sets and let s be a signal distribution generated from this choice and any messaging strategy which does *not* always select the maximal evidence message. Let s^* be the signal distribution generated from the same distribution over evidence sets and the message strategy which does always select the maximal evidence message. Then s^* is more informative than s. (See Section I in the Appendix for proof.) Thus the result in Theorem 1 that we can restrict attention to mechanisms where the principal always induces use of maximal evidence can be thought of as an implication of the result in Theorem 3 that we can restrict to mechanisms where the principal always induces more informative signals. We present these results separately since the reduction of the evidence–acquisition model to the signal–choice model requires showing Theorem 1, so we cannot present only Theorem 3. ▌

Ball and Kattwinkel (forthcoming) study a model where the agent reports her type and then the principal selects a probabilistic pass–fail test out of a given set of such tests. Ball and Kattwinkel propose a comparison of tests that is related to our notion of more informative signals. In their model, a given test τ together with a type t and an effort choice by the agent determines a probability distribution over results where the set of results is $\{0, 1\}$. If the agent takes effort, the agent passes the test (gets an outcome of 1) with probability $\pi(\tau \mid t)$ and fails otherwise. If the agent does not take effort, she fails with probability 1.

Ball and Kattwinkel say that a test $\hat{\tau}$ is more t–discerning than a test τ if there are probabilities k_1 and k_0 with $k_1 \geq k_0$ such that

$$k_1 \pi(\hat{\tau} \mid t) + k_0[1 - \pi(\hat{\tau} \mid t)] = \pi(\tau \mid t) \tag{1}$$

and

$$k_1 \pi(\hat{\tau} \mid t') + k_0[1 - \pi(\hat{\tau} \mid t')] \leq \pi(\tau \mid t'), \quad \forall t' \neq t.$$

Intuitively, this says that a certain kind of garbling of $\hat{\tau}$ (namely, one which puts more weight on the success probability than the failure) gives the same success probabilities as τ for type t and lower success probabilities for all other types.

To relate this to our more informative signals, note that they assume the test is observable by the principal (in fact, is chosen by the principal) but the agent's effort is not. To fit this into our framework, we think of the message observed by the principal as success or failure on a specific test. More formally, we let 1_τ denote the observation by the principal of the agent passing test τ and 0_τ the observation of the agent failing test τ. The signal distribution generated if type t takes test τ and exerts effort, then, puts probability $\pi(\tau \mid t)$ on 1_τ, $1 - \pi(\tau \mid t)$ on 0_τ, and 0 on all other messages. We denote this signal distribution by $s_\tau^+(t)$. If type t takes test τ but doesn't take effort, the signal distribution is the degenerate distribution on 0_τ, which we denote $s_\tau^0(t)$.

In Appendix J, we show that signal $s^+_{\hat{\tau}}(t)$ is more informative than $s^+_{\tau}(t)$ in our sense if and only if test $\hat{\tau}$ is t–more discerning than τ. In this sense, in Ball and Kattwinkel's setting, our comparison is equivalent to theirs.

Theorem 3 implies that if type t has some signal distribution $s^* \in S_t$ which is more informative than any other $s \in S_t$, then the principal may as well always recommend s^* to t. If every t has such a most informative signal distribution, then Stage 2 of the mechanism protocol is not needed as we can restrict attention to mechanisms where every type of the agent is induced to choose her most informative signal distribution. In such a case, we can focus on the following *succinct protocol*:

Stage 1. The agent reports a $t \in T$ and chooses a signal distribution s. Denote a reporting strategy by $b_T : T \to T$ and a signal distribution strategy by $b_S : T \to S$ with $b(t) \in S_t$.

Stage 2. The principal observes the report, the realized m, and chooses an outcome. Let $g_X : T \times \mathcal{L} \to X$ denote a typical pure strategy for the principal.

Abusing notation yet again, let $B = B_T \times B_S$ denote the set of pure strategies for the agent and G the set of pure strategies for the principal in this protocol. When each type t has a most informative signal distribution s^*_t, we can focus on mechanisms $\gamma \in \Gamma$ with the property that the strategy $\hat{b}_T(t) = t$ and $\hat{b}_S(t) = s^*_t$ is a best reply for the agent to γ.

Running Example, Part 3. We showed in Part 2 of the example that our evidence–acquisition technology is normal. In particular, given any realized message set of the form $[0, m]$, the upper bound m is the most informative message for the set. Hence Theorem 2 implies that we can focus on the signal–choice model where for each t, S_t is the set of all distributions on $\mathbf{R}_+$ with expectation less than or equal to t. Since $\mathbf{R}_+$ is not finite, we need to adjust the example to apply our condition. So let $\mathcal{L}$ be any finite subset of $\mathbf{R}_+$ containing at least T. Assume S_t is the set of all probability distributions on $\mathcal{L}$ with expectation less than or equal to t.

We now show that the most informative signal distribution for type t is the degenerate distribution on t. Fix any type t^*. Let $s^* \in S_{t^*}$ denote the degenerate distribution putting probability 1 on $m = t^*$ and fix any other $s \in S_{t^*}$. Let the Λ

matrix be an identity matrix but with the row corresponding to $m = t^*$ replaced by s. That is, we let

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ s(m_1) & s(m_2) & s(m_3) & \dots & s(m_{\#\mathcal{L}-1}) & s(m_{\#\mathcal{L}}) \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

Then $s^*\Lambda = s$. Fix any other type t and any $\hat{s} \in S_t$. Let $\tilde{s} = \hat{s}\Lambda$. For $m \neq t^*$, we have $\tilde{s}(m) = \hat{s}(m) + \hat{s}(t^*)s(m)$. For $m = t^*$, we have $\tilde{s}(t^*) = \hat{s}(t^*)s(t^*)$. So

$$\begin{aligned} \sum_m \tilde{s}(m)m &= \sum_{m \neq t^*} [\hat{s}(m) + \hat{s}(t^*)s(m)]m + \hat{s}(t^*)s(t^*)t^* \\ &= \sum_{m \neq t^*} \hat{s}(m)m + \sum_{m \neq t^*} \hat{s}(t^*)s(m)m + \hat{s}(t^*)s(t^*)t^* \\ &= \sum_{m \neq t^*} \hat{s}(m)m + \hat{s}(t^*) \sum_m s(m)m \\ &\leq \sum_{m \neq t^*} \hat{s}(m)m + \hat{s}(t^*)t^* \\ &= \sum_m \hat{s}(m)m \leq t. \end{aligned}$$

The next–to–last line follows from $s \in S_{t^*}$ and therefore $\sum_m s(m)m \leq t^*$. The last inequality on the last line follows from $\hat{s} \in S_t$ and therefore $\sum_m \hat{s}(m)m \leq t$. So for every $\hat{s} \in S_t$, $\hat{s}\Lambda$ is a probability distribution over $\mathcal{L}$ with expectation weakly less than t and hence is an element of S_t and therefore of $\text{conv}(S_t)$. Hence s^* is more informative than s.

Given this, the incentive compatibility constraints are that each type reports truthfully and chooses the signal distribution with probability 1 on her true type. ▌

Appendix

A Proof of Theorem 1

Assume the evidence structure is normal and fix any incentive compatible mechanism $(\gamma_A, \gamma_{\mathcal{L}}, \gamma_X)$. We show how to construct an incentive compatible mechanism with the same mechanism outcome with the property that the principal always recommends m^*_M when the agent reports message set M, establishing that the evidence structure is simplifiable.

Fix any profile $(\hat{t}, \hat{a}, \hat{M}, \hat{m})$ consisting of a type report $\hat{t} \in T$, a recommended distribution over evidence sets $\hat{a} \in \text{supp}(\gamma_A(\hat{t}))$, a reported message set $\hat{M} \in \mathcal{M}$, and a requested message $\hat{m} \in \text{supp}(\gamma_{\mathcal{L}}(\hat{t}, \hat{a}, \hat{M}))$ such that $\hat{m} \neq m^*_{\hat{M}}$. If there is no such tuple, then the principal always recommends maximal evidence, so there is nothing to prove. We construct an alternative mechanism which replaces the recommendation $\hat{m}$ with a recommendation of $m^*_{\hat{M}}$ in this situation and will show that this mechanism is incentive compatible and implements the same outcome as the original mechanism. For brevity, let $\hat{h} = (\hat{t}, \hat{a}, \hat{M})$, the history on which we are changing the recommendations. We use h to denote a typical element of $T \times A \times \mathcal{M}$.

Define the new mechanism, $(\gamma^*_A, \gamma^*_{\mathcal{L}}, \gamma^*_X)$, as follows. First, $\gamma^*_A = \gamma_A$. Let $\gamma^*_{\mathcal{L}}$ satisfy $\gamma^*_{\mathcal{L}}(h)(m) = \gamma_{\mathcal{L}}(h)(m)$ if $h \neq \hat{h}$. Similarly, let $\gamma^*_{\mathcal{L}}(\hat{h})(m) = \gamma_{\mathcal{L}}(\hat{h})(m)$ for $m \notin \{\hat{m}, m^*_{\hat{M}}\}$. Finally, let

$$\gamma^*_{\mathcal{L}}(\hat{h})(m) = \begin{cases} \gamma_{\mathcal{L}}(\hat{h})(m^*_{\hat{M}}) + \gamma_{\mathcal{L}}(\hat{h})(\hat{m}), & \text{if } m = m^*_{\hat{M}}; \\ 0, & \text{if } m = \hat{m}. \end{cases}$$

In other words, the probability that was on recommendation $\hat{m}$ is moved to $m^*_{\hat{M}}$.

Let $\gamma^*_X(h, m, m')(x) = \gamma_X(h, m, m')(x)$ if $(h, m) \neq (\hat{h}, m^*_{\hat{M}})$. In other words, on histories other than $\hat{h}$ and on $\hat{h}$ if the principal did not request maximal evidence, we do not change the mechanism's outcome. Also, for all $m \in \mathcal{L} \setminus \{m^*_{\hat{M}}\}$, we set

$\gamma^*_X(\hat{h}, m^*_{\hat{M}}, m)(x)$ equal to

$$\frac{\gamma_{\mathcal{L}}(\hat{h})(\hat{m})\gamma_X(\hat{h}, \hat{m}, m)(x) + \gamma_{\mathcal{L}}(\hat{h})(m^*_{\hat{M}})\gamma_X(\hat{h}, m^*_{\hat{M}}, m)(x)}{\gamma_{\mathcal{L}}(\hat{h})(\hat{m}) + \gamma_{\mathcal{L}}(\hat{h})(m^*_{\hat{M}})}.$$

Finally, we set $\gamma^*_X(\hat{h}, m^*_{\hat{M}}, m^*_{\hat{M}})(x)$ equal to

$$\frac{\gamma_{\mathcal{L}}(\hat{h})(\hat{m})\gamma_X(\hat{h}, \hat{m}, \hat{m})(x) + \gamma_{\mathcal{L}}(\hat{h})(m^*_{\hat{M}})\gamma_X(\hat{h}, m^*_{\hat{M}}, m^*_{\hat{M}})(x)}{\gamma_{\mathcal{L}}(\hat{h})(\hat{m}) + \gamma_{\mathcal{L}}(\hat{h})(m^*_{\hat{M}})}.$$

In other words, if $m^*_{\hat{M}}$ is requested and anything else is reported, then the response is the "average response" to this form of disobedience, averaging over the cases where $\hat{m}$ or $m^*_{\hat{M}}$ was requested in the original mechanism. On the other hand, if $m^*_{\hat{M}}$ is requested and reported, then the response is the average response to obedience in response to a request for either $\hat{m}$ or $m^*_{\hat{M}}$ in the original mechanism.

We first show that this change in the mechanism does not change the outcome if the agent is truthful and obedient. The only situation in which a truthful and obedient agent is affected by the change is when her type is $\hat{t}$, the principal recommends (and she chooses) action $\hat{a}$, and the resulting message set is $\hat{M}$. Conditional on history $\hat{h}$ and obeying the principal's recommendations, the probability of x in the new mechanism is

$$\begin{aligned}
&\sum_{m\in\mathcal{L}} \gamma^*_{\mathcal{L}}(\hat{h})(m)\gamma^*_X(\hat{h}, m, m)(x) \\
&= \sum_{m\in\mathcal{L}\setminus\{\hat{m}, m^*_{\hat{M}}\}} \gamma_{\mathcal{L}}(\hat{h})(m)\gamma_X(\hat{h}, m, m)(x) \\
&\quad + 0 + \gamma^*_{\mathcal{L}}(\hat{h})(m^*_{\hat{M}})\gamma^*_X(\hat{h}, m^*_{\hat{M}}, m^*_{\hat{M}})(x) \\
&= \sum_{m\in\mathcal{L}\setminus\{\hat{m}, m^*_{\hat{M}}\}} \gamma_{\mathcal{L}}(\hat{h})(m)\gamma_X(\hat{h}, m, m)(x) \\
&\quad + [\gamma_{\mathcal{L}}(\hat{h})(\hat{m}) + \gamma_{\mathcal{L}}(\hat{h})(m^*_{\hat{M}})]\gamma^*_X(\hat{h}, m^*_{\hat{M}}, m^*_{\hat{M}})(x) \\
&= \sum_{m\in\mathcal{L}} \gamma_{\mathcal{L}}(\hat{h})(m)\gamma_X(\hat{h}, m, m)(x).
\end{aligned}$$

Hence, as asserted, the outcome under truth–telling is the same in the new mechanism as in the original mechanism. Therefore, the agent's expected payoff from truth–telling and obedience is the same in the two mechanisms.

We now show that for any type t and any deviation feasible for t in the new mechanism, there is a deviation that is feasible for type t in the original mechanism which yields the same expected payoff. Since truth–telling is superior to any feasible deviation in the original mechanism, then, truth–telling is superior to any feasible deviation in the new mechanism.

To see this, fix any type t (which may equal $\hat{t}$) and consider any feasible deviation. Obviously, if the deviation involves reporting a type other than $\hat{t}$, this deviation is also available in the original mechanism and yields the same payoff in the new mechanism as in the original one since the way the mechanism responds to such a report has not changed. Hence we can restrict attention to deviations which involve reporting type $\hat{t}$. So fix any such deviation. Clearly, we may as well condition on the event that the principal requests the distribution $\hat{a}$, the agent chooses a (which may equal $\hat{a}$), the agent obtains message set M, and reports message set $\hat{M}$ (which may equal M). Let $z : \hat{M} \to M$ give the message the agent sends as a function of the message the principal requests from her. Then the agent's expected payoff conditional on this event is

$$\sum_{(x,m)\in X\times\mathcal{L}} \gamma^*_{\mathcal{L}}(\hat{h})(m)\gamma^*_X(\hat{h},m,z(m))(x)u(t,x).$$

We can write this as

$$\sum_{(x,m)\in X\times(\mathcal{L}\setminus\{\hat{m},m^*_{\hat{M}}\})} \gamma_{\mathcal{L}}(\hat{h})(m)\gamma_X(\hat{h},m,z(m))(x)u(t,x)$$
$$+\gamma^*_{\mathcal{L}}(\hat{h})(m^*_{\hat{M}})\sum_{x\in X}\gamma^*_X(\hat{h},m^*_{\hat{M}},z(m^*_{\hat{M}}))(x)u(t,x).$$

We have two cases. First, suppose $z(m^*_{\hat{M}}) \neq m^*_{\hat{M}}$. In this case, the last term is equal to

$$\sum_{(x,m)\in X\times\{\hat{m},m^*_{\hat{M}}\}} \gamma_M(\hat{h})(m)\gamma_X(\hat{h},m,z(m^*_{\hat{M}}))(x)u(t,x).$$

Thus the conditional payoff to the deviation in the new mechanism is the same as the conditional payoff in the original mechanism where the agent responds to a request for *either* $\hat{m}$ or $m^*_{\hat{M}}$ by sending $z(m^*_{\hat{M}})$. So in this case, the payoff to the deviation in the new mechanism is the same as the payoff to a certain deviation which was also feasible in the original mechanism.

Second, suppose $z(m^*_{\hat{M}}) = m^*_{\hat{M}}$. In this case, the last term is equal to

$$\sum_{(x,m)\in X\times\{\hat{m},m^*_{\hat{M}}\}} \gamma_M(\hat{h})(m)\gamma_X(\hat{h},m,m)(x)u(t,x).$$

In other words, the payoff in the new mechanism is the same as the payoff in the old mechanism where the agent responds to a request for $\hat{m}$ with $\hat{m}$ and a request for $m^*_{\hat{M}}$ with $m^*_{\hat{M}}$. Note that we are assuming that the deviation in the new mechanism is feasible for the agent, so $m^*_{\hat{M}} \in M$. By the definition of normality, this implies $\hat{m} \in M$. Hence this deviation has the same payoff as a feasible deviation in the original mechanism.

In either case, then, the best deviation payoff in the new mechanism cannot exceed the best deviation payoff in the original mechanism, so the new mechanism is incentive compatible.

Clearly, we can repeat this argument as needed to obtain an incentive compatible mechanism which has the same mechanism outcome as γ and which has the property that $\gamma_{\mathcal{L}}(t,a,M)(m^*_M) = 1$ for all $(t,a,M) \in T \times A \times \mathcal{M}$.

B Proof of Comment in Example 1

To see that this evidence structure is not simplifiable, suppose that it is. Consider the preference structure where $X = \{a,b\}$ and $u(t,a) = 1$ and $u(t,b) = 0$ for all t. Consider the incentive compatible mechanism which requests m_1 from $\{m_1,m_2\}$, always responds to m_1 with a regardless of the reports, and always responds to m_2 with b regardless of the reports. It is easy to see that the agent has no incentive to misreport her type or evidence set. Since m_1 yields the better outcome from $\{m_1,m_2\}$, the agent has no incentive to disobey, so the mechanism is incentive compatible. The outcome of the mechanism is that t_1 gets a and t_2 gets b. It is not hard to see that there is no incentive compatible mechanism achieving this outcome which asks for m_2 from $\{m_1, m_2\}$. Since this mechanism would need to give t_1 outcome a, it would have to respond to a type report of t_1, evidence set report of $\{m_1, m_2\}$, and evidence presentation of m_2 with a. But then t_2 would deviate to making this claim and presenting m_2, a contradiction. So if this evidence structure is simplifiable, it must be that the principal can always ask for m_1 from $\{m_1, m_2\}$.

But then consider the following preference structure. As above, $X = \{a, b\}$ and t_1 strictly prefers a to b. Now, though, suppose t_2 strictly prefers b to a. Consider the incentive compatible mechanism which requests m_2 from $\{m_1, m_2\}$, always responds to m_1 with b, and always responds to m_2 with a. Clearly, t_1 obeys when she has $\{m_1, m_2\}$ so this mechanism is incentive compatible. It gives a 50–50 lottery between a and b for t_1 and a with probability 1 for t_2. It is impossible to obtain this outcome from an incentive compatible mechanism that requests m_1 from $\{m_1, m_2\}$. To see this, note that t_2 gets her worst possible outcome from this mechanism. So any incentive compatible mechanism with this outcome must give a for any reports followed by evidence m_2. Otherwise, t_2 could send these reports and evidence and improve her payoff. Since this is t_1's preferred outcome, the only way to induce t_1 to send m_1 when she has $\{m_1, m_2\}$ is to also respond to m_1 with outcome a. But then t_1 will claim to have evidence $\{m_1, m_2\}$ when she actually has $\{m_1\}$ unless this report also leads to outcome a. Then t_1 is getting outcome a, not the 50–50 lottery, a contradiction.

C Proof of Comment in Example 2

Suppose we have some outcome that we implement asking t_1 for m_1 with probability $\alpha < 1$ and for m_2 with probability $1 - \alpha$. We can write the mechanism as

report	request	evidence	outcome
$t_1, \{m_1, m_2\}$	m_1	m_1	a
$t_1, \{m_1, m_2\}$	m_1	m_2	w_1
$t_1, \{m_1, m_2\}$	m_2	m_1	w_2
$t_1, \{m_1, m_2\}$	m_2	m_2	b
$t_2, \{m_1\}$		m_1	c
$t_2, \{m_1\}$		m_2	w_3
$t_2, \{m_2\}$		m_1	w_4
$t_2, \{m_2\}$		m_2	d

Incentive compatibility requires that t_1 prefers a to w_1 and b to w_2. Also, we can assume without loss of generality that t_2 prefers a to w_1 and b to w_2. To see this, suppose, for example, that t_2 strictly prefers w_1 to a. Then we could change w_1 to a, satisfy t_1's incentive constraint, and reduce t_2's incentive to claim to be t_1.

Note that incentive compatibility requires t_2 to prefer $pc + (1 - p)d$ (where this denotes the lottery giving c with probability p and d otherwise) to

$$\alpha[pa + (1 - p)w_1] + (1 - \alpha)[pw_2 + (1 - p)b].$$

Since t_2 prefers b to w_2, $p < 1/2$ implies that the second term above is better for her than $(1 - p)w_2 + pb$, so incentive compatibility implies that she prefers $pc + (1 - p)d$ to

$$\alpha[pa+(1-p)w_1]+(1-\alpha)[pb+(1-p)w_2] = p[\alpha a+(1-\alpha)b]+(1-p)[\alpha w_1+(1-\alpha)w_2].$$

So change the mechanism to always ask t_1 for m_1 and to give

report	request	evidence	outcome
$t_1, \{m_1, m_2\}$	m_1	m_1	$\alpha a + (1-\alpha)b$
$t_1, \{m_1, m_2\}$	m_1	m_2	$\alpha w_1 + (1-\alpha)w_2$
$t_2, \{m_1\}$		m_1	c
$t_2, \{m_1\}$		m_2	w_3
$t_2, \{m_2\}$		m_1	w_4
$t_2, \{m_2\}$		m_2	d

t_1 prefers $\alpha a+(1-\alpha)b$ to $\alpha w_1+(1-\alpha)w_2$, so this is incentive compatible for her if the original mechanism was. From the argument above, t_2 prefers reporting truthfully and getting $pc + (1-p)d$ to claiming to be t_1 and getting

$$p[\alpha a + (1-\alpha)b] + (1-p)[\alpha w_1 + (1-\alpha)w_2].$$

We haven't changed anything about what happens when she reports t_2, so if the original mechanism was incentive compatible, this one is as well.

Hence the evidence structure is simplifiable.

D Robust Simplifiability

Suppose, contrary to the claim, that there is an evidence structure which is not normal but which satisfies robust simplifiability. Let $\mathcal{M}$ denote the collection of possible evidence sets under this evidence structure. Since normality is violated, there is a set $M_0 \in \mathcal{M}$ such that for every $m \in M_0$, there is $M' \in \mathcal{M}$ such that $m \in M'$ but $M_0 \not\subseteq M'$. Let K be the number of messages in M_0 and write this set as $\{m_1, \ldots, m_K\}$. For each $m_k \in M_0$, let M'_k denote any set with $m_k \in M'_k$ and $M_0 \not\subseteq M'_k$.

Because the evidence structure is robustly simplifiable, we must also get simplifiability in the following model. Let $T = \{t_0, t_1, \ldots, t_K, t_{K+1}\}$. Types t_k for $k \leq K$, each have only a single evidence acquisition action yielding evidence set M_k with probability 1. Type t_{K+1} has a single evidence action whose support is all the other M's in $\mathcal{M}$. To show that this evidence structure is not simplifiable, consider the following payoff structure. Let $X = \{a, b, c\}$. Types t_k for $k \leq K$ get utility 1 from a, 0 from b, and $-K$ from c. t_{K+1} is indifferent over all outcomes.

Consider a mechanism which responds to type report t_0 by randomizing uniformly over which of the K messages in M_0 to request. If the evidence message provided matches the request, the outcome is a. If not, the outcome is c. For any other type report and evidence provision, the outcome is b. This mechanism is incentive compatible and gives outcome a for t_0 and b for every other type. Since type t_{K+1} is indifferent over all outcomes, truth telling is certainly optimal for her. Any type $t_k \neq t_0$, $k \leq K$, has at most $K-1$ of the messages in M_0. Hence for any such type, reporting t_0 gives, at best, an expected payoff of

$$\left(\frac{K-1}{K}\right)(1) + \left(\frac{1}{K}\right)(-K) < 0,$$

so reporting honestly is better.

For any message $m_k \in M_0$, no incentive compatible mechanism which requests m_k from t_0 can achieve the same outcome. Such a mechanism would have to respond to a report of t_0 and the presentation of evidence m_k with outcome a to give t_0 the same outcome. Since type t_k has m_k in her evidence set, t_k would deviate to this type report and evidence presentation rather than get outcome b. Hence this evidence structure is not simplifiable and hence the original evidence structure is not robustly simplifiable.

E Proof of Corollary 1

Fix an incentive compatible mechanism $\gamma = (\gamma_A, \gamma_{\mathcal{L}}, \gamma_X)$. By Theorem 1, we can assume without loss of generality that $\gamma_{\mathcal{L}}(t, a, M)(m^*_M) = 1$ for all $(t, a, M) \in T \times A \times \mathcal{M}$. We construct a mechanism (γ^*_A, γ^*_X) for the abbreviated protocol which is incentive compatible and has the same outcome as γ. To do so, first let $\gamma^*_A = \gamma_A$.

To construct γ^*_X, note that in the abbreviated protocol, $\gamma^*_X : T \times A \times \mathcal{L} \to \Delta(X)$, while in the full protocol, $\gamma_X : T \times A \times \mathcal{M} \times \mathcal{L} \times \mathcal{L} \to \Delta(X)$ since the choice of x can depend on the agent's report of an evidence set and the message the principal requests, in addition to the type report, distribution recommendation, and received message as in the abbreviated protocol.

Given any $m \in \mathcal{L}$, define $M^*(m)$ as follows. First, if there is any M such that $m = m^*_M$, then let $M^*(m)$ equal this message set M.[10] Otherwise, let $M^*(m)$ denote any $M \in \mathcal{M}$ such that $m \in M$. Given this, let

$$\gamma^*_X(t, a, m) = \gamma_X(t, a, M^*(m), m^*_{M^*(m)}, m).$$

In other words, if the agent reports t, the principal recommends a, and the agent shows message m, then the outcome is the same as in the original mechanism when the agent reports t, the principal recommends a, the agent reports evidence set $M^*(m)$, the principal requests the maximal evidence message for this set, and the agent provides message m.

If the agent truthfully reports her type, follows the principal's recommended distribution a, and provides the maximal evidence message from any evidence set she obtains, this construction implies that the resulting distribution over X in the new mechanism will be the same as in the original mechanism. Hence if this mechanism is incentive compatible, it yields the same outcome as the original mechanism.

So consider an agent of type t who reports type $\hat{t}$ (which may or may not equal t), has a recommended to her by the principal, chooses $\hat{a}$, obtains evidence set M, and sends message m from it. In this situation, the outcome under the new mechanism is $\gamma_X(\hat{t}, a, M^*(m), m^*_{M^*(m)}, m)$, exactly the same outcome the agent could have obtained by reporting $\hat{t}$, choosing $\hat{a}$, reporting $M^*(m)$ as her evidence set, and then sending m. That is, any outcome the agent can generate in the new mechanism using a strategy which deviates from truth–telling, obedience, and sending maximal evidence is an outcome she could have generated in the original mechanism using a certain strategy which deviated from truth–telling and obedience. Since the original mechanism was incentive compatible, truth–telling and obedience were superior to this deviation. Hence the agent prefers truth–telling, obedience, and maximal evidence in the new

[10] It is straightforward to show that if $m^*_M = m^*_{\hat{M}}$, then $M = \hat{M}$. That is, $M^*(m)$ is unambiguously defined in this case.

mechanism to any deviation, so the mechanism is incentive compatible.

F Proof of Theorem 2

Fix an incentive compatible mechanism for the evidence–acquisition model under normality. By Corollary 1, we can take this mechanism to be based on the abbreviated protocol. Hence it consists of a pair of functions $\gamma_A : T \to \Delta(A)$ and $\gamma_X : T \times A \times \mathcal{L} \to \Delta(X)$. For the signal choice model, a mechanism is a pair of functions $\gamma_S^* : T \to \Delta(S)$ and $\gamma_X^* : T \times S \times \mathcal{L} \to \Delta(X)$.

Given the incentive compatible mechanism for the abbreviated protocol, we construct an equivalent incentive compatible mechanism for the associated signal–choice model as follows. Let

$$\gamma_S^*(t)(s_{(a,\sigma^*)}) = \gamma_A(t)(a).$$

That is, given a report of t, the principal recommends the signal distribution generated by evidence distribution a followed by showing maximal evidence with the same probability he recommended a in the original mechanism. Let

$$\gamma_X^*(t, s_{(a,\sigma^*)}, m) = \gamma_X(t, a, m).$$

That is, if the agent report type t and the signal distribution the principal recommends is the one corresponding to a and maximal evidence, then the principal replies to message m in the new mechanism the same way he replied in the original mechanism given type report t and recommendation a.

It is easy to see that if the agent reports her type truthfully and follows the principal's recommended signal distribution, then the outcome is equivalent to that of the original mechanism as defined in the statement of the theorem. If the agent deviates, this corresponds to a particular deviation strategy in the original mechanism and hence cannot be profitable for her. In particular, if type t reports $\hat{t}$, receives the recommendation $s_{(a,\sigma^*)}$, and uses signal distribution $s_{(\hat{a},\hat{\sigma})}$ instead, she generates exactly the outcome she would have generated in the original mechanism if she reported $\hat{t}$, received the recommendation a, chose the distribution $\hat{a}$ instead, and selected a message to send using the function $\hat{\sigma}$. So the mechanism is incentive compatible.

G Proof of Comment in Example 4

Consider the following payoff structure. The set of outomes is $\{a, b, c\}$. The utilities of the two types are given by the following table:

	t_1	t_2
a	0	1
b	1	0
c	3	-3

Suppose the mechanism recommends s_2' to t_2 and gives outcome a in response to m_1 and b in response to m_2. Clearly, t_2 will choose s_2' as this gives a 50–50 lottery over a and b, while choosing s_2 would give b for sure. So we implement the outcome giving a for t_1 and the 50–50 lottery between a and b for t_2.

To see that no mechanism can implement this outcome with t_2 choosing s_2 instead, note that t_1 receives her worst possible outcome. So any report followed by evidence m_1 must lead to an outcome of a or else t_1 would deviate.

Clearly, if some incentive compatible mechanism does generate this outcome with t_2 choosing s_2, it must respond to a type report of t_2 and evidence presentation of m_2 with the 50–50 lottery between a and b. But t_2 could deviate to s_2' and generate a 50–50 lottery between this outcome and a, which she strictly prefers.

H Proof of Theorem 3

Fix an incentive compatible mechanism (γ_S, γ_X) where $\gamma_S(t_1)(\hat{s}_1) = \hat{\alpha} > 0$. Let $\alpha = \gamma_S(t_1)(s_1)$ (where this can be 0). We construct an incentive compatible mechanism (γ_S^*, γ_X^*) with the same outcome where the principal recommends s_1 to t_1 with probability $\alpha + \hat{\alpha}$ and never recommends $\hat{s}_1$ to t_1.

For any $t \neq t_1$, $\gamma_S^*(t) = \gamma_S(t)$ and $\gamma_X^*(t, s, m) = \gamma_X(t, s, m)$ for all (s, m). For $s \neq s_1, \hat{s}_1$, we have $\gamma_S^*(t_1)(s) = \gamma_S(t_1)(s)$ and $\gamma_X^*(t_1, s, m) = \gamma_X(t_1, s, m)$. That is, if the agent reports a type other than t_1, the new mechanism is the same as the original one and if the agent reports t_1, the principal recommends signals other than s_1 or $\hat{s}_1$ with the same probability and treats them the same way as in the original mechanism.

Let $\gamma_S^*(t_1)(\hat{s}_1) = 0$ and $\hat{\gamma}_S^*(t_1)(s_1) = \alpha + \hat{\alpha}$. Since the principal never recommends $\hat{s}_1$ in response to a report of t_1 in this mechanism, we only need to specify $\gamma_X^*(t, s, m)$ for $(t, s) = (t_1, s_1)$. For notational convenience, we enumerate the messages as $\mathcal{L} = \{m_1, \ldots, m_L\}$ and for the Markov matrix Λ, we write the entry corresponding to (m_i, m_j) as λ_{ij} rather than λ_{m_i, m_j}.

Let

$$\gamma_X^*(t_1, s_1, m_i) = \frac{\alpha}{\alpha + \hat{\alpha}} \gamma_X(t_1, s_1, m_i) + \frac{\hat{\alpha}}{\alpha + \hat{\alpha}} \sum_j \lambda_{ij} \gamma_X(t_1, \hat{s}_1, m_j).$$

Because all the λ_{ij}'s are non–negative and because $\sum_j \lambda_{ij} = 1$ for every i, we see that $\gamma^*(t_1, s_1, m_i)$ is a convex combination of probability distributions over X and hence is a probability distribution over X.

Given this specification, suppose all types report honestly and obey the principal's recommendations. Obviously, if the true type $t \neq t_1$, we have the same outcome as before. So suppose $t = t_1$. Then the expected outcome is

$$(\alpha + \hat{\alpha}) \sum_i s_1(m_i) \gamma_X^*(t_1, s_1, m_i) + \sum_{s \in S_{t_1} \setminus \{s_1, \hat{s}_1\}} \gamma_S^*(t_1)(s) \sum_M s(m) \gamma_X^*(t_1, s, m). \qquad (2)$$

Substituting for γ_X^*, the first term in equation (2) is

$$\begin{aligned} \alpha \sum_i s_1(m_i) \gamma_X(t_1, s_1, m_i) + \hat{\alpha} \sum_i s_1(m_i) \sum_j \lambda_{ij} \gamma_X(t_1, \hat{s}_1, m_j) \\ = \alpha \sum_i s_1(m_i) \gamma_X(t_1, s_1, m_i) + \hat{\alpha} \sum_j \gamma_X(t_1, \hat{s}_1, m_j) \sum_i s_1(m_i) \lambda_{ij}. \end{aligned}$$

But $s_1\Lambda = \hat{s}_1$, so that for every j, $\sum_i s_1(m_i)\lambda_{ij} = \hat{s}_1(m_j)$. Hence this is

$$= \alpha \sum_i s_1(m_i)\gamma_X(t_1, s_1, m_i) + \hat{\alpha} \sum_i \hat{s}_1(m_i)\gamma_X(t_1, \hat{s}_1, m_j).$$

Substituting this for the first term in equation (2) and substituting for γ_S^* and γ_X^* in the second term, we see that the expected outcome under truth–telling and obedience is the same as under the original mechanism.

To show that the new mechanism is incentive compatible, we show that any deviation from truth–telling and obedience by any type generates a distribution over outcomes that the same type could have generated in the original mechanism. Since the original mechanism was incentive compatible, this deviation is not profitable, so the new mechanism is incentive compatible.

To see that this holds, fix any type t and any signal $s' \in S_t$. If t makes any type report other than t_1, the mechanism has not changed, so the claim obviously holds. So suppose type t reports type t_1. If the mechanism makes any signal recommendation other than s_1, then, again, the mechanism is the same as before, so the claim holds. So suppose the mechanism recommends signal s_1 and the agent uses s'. The expected outcome times the probability of this event is

$$(\alpha+\hat{\alpha}) \sum_i s'(m_i)\gamma_X^*(t_1, s_1, m_i) = \alpha \sum_i s'(m_i)\gamma_X(t_1, s_1, m_i) + \hat{\alpha} \sum_i s'(m_i) \sum_j \lambda_{ij}\gamma_X(t_1, \hat{s}_1, m_j).$$

By assumption, $s'\Lambda \in \text{conv}(S_t)$. Hence we can write $s'\Lambda = \sum_k a_k s^k$ where $a_k \geq 0$ for all k, $\sum_k a_k = 1$, and $s^k \in S_t$ for all k. In particular, for every j,

$$\sum_i s'(m_i)\lambda_{ij} = \sum_k a_k s^k(m_j).$$

Hence we can rewrite the above as

$$\alpha \sum_i s'(i)\gamma_X(t_1, s_1, m_i) + \hat{\alpha} \sum_k a_k s^k(i)\gamma_X(t_1, \hat{s}_1, m_i).$$

This is exactly what t would generate in the original mechanism if she responded to a recommendation of s_1 with s' and a recommendation of $\hat{s}_1$ by randomizing with probability a_k on s^k. Thus, as asserted, any expected outcome t can generate in the new mechanism is identical to some outcome she could have generated in the original mechanism. Hence the new mechanism is incentive compatible.

I Proof of Remark 2

Fix any a and message strategy σ and let $s = s_{(a,\sigma)}$. For the same a, let $\sigma^*(M) = m^*_M$ for all $M \in \text{supp}(a)$ and let $s^* = s_{(a,\sigma^*)}$. Abusing notation, write $\sigma(M)$ not as the message s sends from M but as the probability distribution over M when M is realized. So write $\sigma(M)(m)$ as the probability that message m is sent from set M. Enumerate $\mathcal{L}$, the set of all evidence messages, as $m_1, \ldots, m_K$. If $m_i = m^*_M$, we write $M = M_i$. Since no message can be maximal evidence for more than one evidence set, we have $s^*(m_i) = a(M_i)$.

Define a Markov matrix Λ as follows. If $s^*(m_i) = 0$, then $\lambda_{ii} = 1$ and $\lambda_{ij} = 0$ for $j \neq i$. If $s^*(m_i) > 0$, then $\lambda_{ij} = \sigma(M_i)(m_j)$. In other words, if s^* sends m_i with positive probability, then λ_{ij} is the probability that m_j is the message s sends given message set M_i.

Note that the jth element of $s^*\Lambda$ is

$$\sum_i s^*(m_i)\lambda_{ji} = \sum_{M \in \mathcal{M}} a(M)\sigma(M)(m_j) = s(m_j).$$

Hence $s^*\Lambda = s$, as required. For any other $\hat{s}$, the jth element of $\hat{s}\Lambda$ is

$$\sum_{i|s^*(m_i)>0} \hat{s}(m_i)\sigma(M_i)(m_j) + \sum_{i|s^*(m_i)=0} \hat{s}(m_i)\lambda_{ji}$$

or

$$\begin{cases} \sum_{i|s^*(m_i)>0} \hat{s}(m_i)\sigma(M_i)(m_j), & \text{if } s^*(m_j) > 0; \\ \sum_{i|s^*(m_i)>0} \hat{s}(m_i)\sigma(M_i)(m_j) + \hat{s}(m_j), & \text{otherwise.} \end{cases}$$

In other words, $\hat{s}\Lambda$ is constructed as follows. We choose a message, say m_j, according to distribution $\hat{s}$. If $s^*(m_j) = 0$, then this message is sent. If $s^*(m_j) > 0$, then instead we randomize the message to send according to the distribution $\sigma(M_i)$.

We now show that this must be feasible for any type for whom $\hat{s}$ is feasible. Clearly, if $\hat{s}$ generates a message m_j, it must be able to send that message. So we need to show that the randomization is feasible — that is, that whenever m_j could be sent, every message in M_j is also feasible. But this follows from the fact that $m_j = m^*_{M_j}$. By definition, this means that if the feasible set is M and $m_j \in M$, then $M_j \subseteq M$. So if $\hat{s} \in S_t$, then $\hat{s}\Lambda \in S_t$, completing the proof.

J Proof of Relationship to Ball and Kattwinkel

We first show that if $s^+_{\hat{\tau}}(t)$ is more informative than $s^+_\tau(t)$ in our sense, then $\hat{\tau}$ is more t–discerning than τ. Since $s^+_{\hat{\tau}}(t)$ is more informative than $s^+_\tau(t)$, there exists a Markov matrix Λ such that (among other properties) $s^+_{\hat{\tau}}(t)\Lambda = s^+_\tau(t)$ and $s^0_{\hat{\tau}}(t'), s^+_{\hat{\tau}}(t') \in \text{conv}(S_{t'})$ for all t'.

Because $s^+_{\hat{\tau}}(t')$ puts positive probability only on $1_{\hat{\tau}}$ and $0_{\hat{\tau}}$ for any t', the only elements of Λ that are relevant to this calculation are the ones in the rows corresponding to these messages. For intuition, think of Λ as a Markov transition matrix where element in the kth row and nth column is the probability that message k "transitions" to message n. So we focus only on the transitions from $0_{\hat{\tau}}$ and $1_{\hat{\tau}}$. Let k_0 denote the transition probability from $0_{\hat{\tau}}$ to 1_τ and k_1 the transition probability from $1_{\hat{\tau}}$ to 1_τ.

Letting $\lambda_{m,m'}$ denote the transition probability from m to m', we can write the entry of the vector $s^+_{\hat{\tau}}(t)\Lambda$ corresponding to some message m' as

$$\sum_m s^+_{\hat{\tau}}(t)(m)\lambda_{m,m'} = s^+_{\hat{\tau}}(t)(0_{\hat{\tau}})\lambda_{0_{\hat{\tau}},m'} + s^+_{\hat{\tau}}(t)(1_{\hat{\tau}})\lambda_{1_{\hat{\tau}},m'} = [1-\pi(\hat{\tau} \mid t)]\lambda_{0_{\hat{\tau}},m'} + \pi(\hat{\tau} \mid t)\lambda_{1_{\hat{\tau}},m'}.$$

For $m' = 1_\tau$, $s^+_{\hat{\tau}}(t)\Lambda = s^+_\tau(t)$ implies

$$[1-\pi(\hat{\tau} \mid t)]k_0 + \pi(\hat{\tau} \mid t)k_1 = \pi(\tau \mid t),$$

Ball and Kattwinkel's first condition, equation (1).

Similarly, the component of the vector $s^+_{\hat{\tau}}(t')\Lambda$ giving the probability on 1_τ is $[1-\pi(\hat{\tau} \mid t')]k_0 + \pi(\hat{\tau} \mid t')k_1$. Since this vector is in the convex hull of $S_{t'}$, the probability on 1_τ must be weakly less than the maximum probability any $s \in S_{t'}$ puts on this message. But the only signal distribution in $S_{t'}$ with nonzero probability on 1_τ is $s^+_\tau(t')$. Hence we must have

$$[1 - \pi(\hat{\tau} \mid t')]k_0 + \pi(\hat{\tau} \mid t')k_1 \leq \pi(\tau \mid t'),$$

Ball and Kattwinkel's second condition.

Finally, we show that $k_1 \geq k_0$. We have that $s^0_{\hat{\tau}}(t)\Lambda$ is in the convex hull of S_t. Hence the component giving the probability on 1_τ must be less than or equal to $\pi(\tau \mid t)$. It is not hard to see that this probability is k_0 so $\pi(\tau \mid t) \geq k_0$. But equation (1) (which we already showed must hold) is equivalent to

$$(k_1 - k_0)\pi(\hat{\tau} \mid t) = \pi(\tau \mid t) - k_0.$$

So $\pi(\tau \mid t) \geq k_0$ implies $k_1 \geq k_0$.

For the converse, suppose $\hat{\tau}$ is t–more discerning than τ. Fix the k_0 and k_1 in the definition. Define the matrix Λ as follows. Set the transition probability from $0_{\hat{\tau}}$ to 1_τ equal to k_0, the transition probability from $1_{\hat{\tau}}$ to 1_τ equal to k_1. Similarly, let the transition probability from $0_{\hat{\tau}}$ to 0_τ be $1 - k_0$ and from $1_{\hat{\tau}}$ to 0_τ be $1 - k_1$. Finally, for any m other than $0_{\hat{\tau}}$ and $1_{\hat{\tau}}$, let the transition probability from m to itself be 1 (so the transition to any different message is 0). It is easy to see that for any signal s generated by a test other than $\hat{\tau}$, we have $s\Lambda = s$, so obviously $s\Lambda$ is in the appropriate convex hull. For any t' (including t), we have

$$s^+_{\hat{\tau}}(t')\Lambda = k_1\pi(\hat{\tau} \mid t') + k_0[1 - \pi(\hat{\tau} \mid t')].$$

For $t' = t$, equation (1) implies $s^+_{\hat{\tau}}(t)\Lambda = s^+_\tau(t)$. For other t', Ball and Kattwinkel's second equation implies that $s^+_{\hat{\tau}}(t')$ is a convex combination of $s^+_\tau(t')$ and $s^0_\tau(t')$ and so is in the convex hull of $S_{t'}$. Finally, just as above, $s^0_{\hat{\tau}}(t') = k_0$ and $k_1 \geq k_0$ implies this is below $\pi(\tau \mid t)$. Hence $s^0_{\hat{\tau}}(t')$ is also in the convex hull of $S_{t'}$. So $s^+_{\hat{\tau}}(t)$ is more informative than $s^+_\tau(t)$.

References

[1] Ball, I., and D. Kattwinkel, "Probabilistic Verification in Mechanism Design," *Theoretical Economics*, forthcoming.

[2] Banerjee, S., and Y.-C. Chen, "Implementation with Uncertain Evidence," working paper March 2025.

[3] Blackwell, D., and M. Girshick, *Theory of Games and Statistical Decisions*, Wiley, 1954.

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

[5] Che, Y.-K., and N. Kartik, "Opinions as Incentives," *Journal of Political Economy*, **117**, October 2009, 815–860.

[6] Deb, R., M. Pai, and M. Said, "Evaluating Strategic Forecasters," *American Economic Review*, **108**, October 2018, 3057–3103.

[7] DeMarzo, P., I. Kremer, and A. Skrzypacz, "Test Design and Minimum Standards," *American Economic Review*, **109**, June 2019, 2173–2207.

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

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

[10] Espinosa, F., ⓡ D. Ray, "Too Good To Be True? Retention Rules for Noisy Agents," *American Economic Journal: Microeconomics*, **15**, May 2023, 493–535.

[11] Felgenhauser, M., and E. Schulte, "Strategic Private Experimentation," *American Economic Journal: Microeconomics*, **6**, November 2014, 74–105.

[12] Forges, F., and F. Koessler, "Communication Equilibria with Partially Verifiable Types," *Journal of Mathematical Economics*, **41**, 2005, 793–811.

[13] Gerardi, D., and R. Myerson, "Sequential Equilibria in Bayesian Games with Communication," *Games and Economic Behavior*, **60**, July 2007, 104–134.

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

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

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

[17] Grossman, S. J., “The Informational Role of Warranties and Private Disclosures about Product Quality,” *Journal of Law and Economics*, **24**, 1981, 461–483.

[18] Hedlund, J., “Bayesian Persuasion by a Privately Informed Sender,” *Journal of Economic Theory*, **167**, January 2017, 229–268.

[19] Henry, E., and M. Ottaviani, “Research and the Approval Process: The Organization of Persuasion,” *American Economic Review*, **109**, March 2019, 911–955.

[20] Kamenica, E., and M. Gentzkow, “Bayesian Persuasion,” *American Economic Review*, **101**, October 2011, 2590–2615.

[21] Koessler, F., and V. Skreta, “Informed Information Design,” *Journal of Political Economy*, **131**, November 2023, 3186–3232.

[22] Kosenko, A., “Constrained Persuasion with Private Information,” *B.E. Journal of Theoretical Economics*, **23**, 2023(1), 345–370.

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

[24] Matthews, S., and A. Postlewaite, “Quality Testing and Disclosure,” *RAND Journal of Economics*, **16**, Autumn 1985, 328–340.

[25] McClellan, A., “Experimentation and Approval Mechanisms,” *Econometrica*, **90**, September 2022, 2215–2247.

[26] Milgrom, P., “Good News and Bad News: Representation Theorems and Applications,” *Bell Journal of Economics*, **12**, 1981, 350–391.

[27] Perez-Richet, E., "Interim Bayesian Persuasion: First Steps," *American Economic Review: Papers & Proceedings*, **104**, May 2014, 5.

[28] Perez-Richet, E., and V. Skreta, "Test Design under Falsification," *Econometrica*, **90**, May 2022, 1109–1142.

[29] Shishkin, D., "Evidence Acquisition and Voluntary Disclosure," working paper, June 2024.

[30] Silva, F., "The Importance of Commitment Power in Games with Imperfect Evidence," *American Economic Journal: Microeconomics*, **12**, November 2020, 99–1113.

[31] Sugaya, T., and A. Wolitzky, "The Revelation Principle in Multistage Games," *Review of Economic Studies*, **88**, May 2021, 1503–1540.