

Mechanism Design for Acquisition of/Stochastic Evidence¹

Elchanan Ben-Porath² Eddie Dekel³ Barton L. Lipman⁴

First Preliminary Draft
September 2019

Current Draft
December 2023

¹We 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.

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

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

⁴Department 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. We also give a novel result on conditions under which there is no value to commitment for the principal.

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 can then choose which realizations 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 a random signal 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. However, there are many situations of economic interest where an agent must take some action to generate evidence. In these situations, the evidence that is generated is typically random in the sense that the outcome of the agent’s action is uncertain.¹ For example, consider a department of an organization or an entrepreneur that wishes to obtain funding for a new product. The department/entrepreneur can run different tests on the current prototype or carry out market research to obtain evidence on the demand for the product to present to the central administration or venture capitalists. The evidence these tests will generate is random.

We study two important issues in the literature on evidence. First, we identify conditions under which we can restrict attention to a relatively simple class of mechanisms. Second, we identify conditions in which the outcome of the optimal mechanism can be obtained without commitment by the principal. The paper is organized as follows.

In Section 2, we present the “technology” of the two models, relate them to the literature, and introduce a running example. In Section 3, we briefly discuss game-theoretic versions of the models and show that in a natural game, the signal–choice model is a reduced form of the evidence–acquisition model.

In Section 4, we turn to mechanism design. First, we provide an analog of the Rev-

¹The usual model where the agent is endowed with evidence can be thought of as the special case of the signal–choice model where all signals are degenerate.

elation Principle for the evidence–acquisition model. The general class of mechanisms for these problems is quite complex, involving numerous steps of communication between the agent and the principal. We analyze conditions under which we can identify the optimal recommendation for the principal regarding what evidence he would like to see. Identifying this recommendation eliminates the need to optimize over it and also enables us to simplify the mechanism, reducing it to the signal–choice model. We also give conditions under which we can identify the principal’s recommendation regarding what signal to choose, leading to a further simplification.

In Section 5, we show that under certain conditions, the optimal mechanism does not require commitment by the principal. That is, the best mechanism for the principal yields the same outcome as the best equilibrium of the game where the principal is not committed. We show this by first giving a general result for Nash equilibrium which generalizes results where the agent is already endowed with evidence in Ben-Porath, Dekel, and Lipman (2019) and earlier results in Glazer and Rubinstein (2004, 2006), Sher (2011), and Hart, Kremer, and Perry (2017). Then we extend the result to perfect Bayesian equilibrium for more specific assumptions on preferences. We show that in the more general class of models we consider, this result requires stronger assumptions on the preferences of the principal than in our previous work. However, our assumptions are without loss of generality when there are only two outcomes. Hence our result applies to problems where the principal’s actions are to accept or reject, such as when the principal decides whether to hire a candidate or not, approve funding for a project or not, or provide a good or not.

Proofs not contained in the text are in the Appendix.

2 Models

In this section, we discuss the “primitives” of the model, reserving discussion of the specifics of the game or mechanism for later sections.

Running Example, Part 1. Throughout the paper, we will use the following example to illustrate ideas and results. We have an employer, also referred to as *the principal*, and an employee, also called *the agent*. The agent’s private information

is her productivity for the principal. We consider two variations. First, we consider what we will call the *wage-setting version* of this problem. Here, as in Spence (1973), the principal sets a wage for the agent and his payoff is maximized by setting the wage equal to the agent's true productivity. The agent's payoff is strictly increasing in the wage. Second, we consider the *hiring version* of the problem. Here there is a fixed wage, outside the control of the principal, and he can only decide whether or not to hire the agent. The principal prefers hiring to not hiring iff the agent's productivity is sufficiently high, while the agent strictly prefers being hired, regardless of her true type. Hence in both versions the agent wants the principal to think she has a high type and the principal wants to know the true type, but in the second, the decision is coarser. We consider various forms of evidence acquisition by the agent to try to persuade the principal she has high productivity. ■

As in the running example, the players in the model are an agent and a principal. 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 $x \in X$. 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 include cheap talk.

We consider two ways of modeling information transmission, one of which is a 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 The utility functions of the agent and principal are $u : T \times X \rightarrow \mathbf{R}$ and $v : T \times X \rightarrow \mathbf{R}$ respectively.⁴

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) allows observability of a 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.

Running Example, Part 2. In the example, we assume a very stylized evidence–acquisition technology. To see the idea, 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,

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

³As 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.

⁴For some purposes, it is natural to also let the agent’s and/or principal’s utility to depend on the realized evidence set and/or the evidence acquisition actions. Dependence on the action, of course, allows the possibility that evidence acquisition is costly to the agent. Dependence on the realized set (a) allows the possibility that the agent’s costs depend on the realized set and (b) reflects the idea that the realization itself may be informative. We avoid adding these to the utility functions as it would complicate the notation even further, but note that neither addition would affect Theorems 1, 2, or 4.

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 message set 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,

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

In our example, the agent wants to persuade 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). ■

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.

While we discuss the details of games or 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 3. For a signal–choice version of our running example, we “convert” the same technology as in the evidence–acquisition model described in Part 2 into a signal–choice model. Note that the agent in the evidence–acquisition model 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 the signal–choice version of our running example, we assume that S_t , the set of signal distributions for type t , is 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. It is easy to see that this process generates a signal distribution, that is, a distribution on \mathbf{R}_+ . ■

Related literature: The usual model of evidence considers games or mechanism design problems where the agent’s set of feasible messages depends on 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. For early contributions in game theory, see Grossman (1981), Milgrom (1981), and Dye (1985). For an early contribution in mechanism design theory, see Green and Laffont (1986). For more recent examples

of papers in game theory or mechanism design, see Shin (2003), Acharya, DeMarzo, and Kremer (2011), Ben-Porath and Lipman (2012), Kartik and Tercieux (2012), Guttman, Kremer, and Skrzypacz (2014), and Rappoport (2020). Finally, for closely related work related to both games and mechanisms, see Glazer and Rubinstein (2004, 2006), Sher (2011), Hart, Kremer, and Perry (2017), and Ben-Porath, Dekel, and Lipman (2019).

In these papers, the agent is endowed with evidence and only chooses which evidence to 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, as discussed above, 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.

A number of 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 (2020) 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. Ball and Kattwinkel (2023), by contrast, do consider a 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 4.

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 (2020). 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.⁵ 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) give a model which can be thought of as a signal–choice model. A forecaster has private information about the quality of the signals she receives about some random variable. She sees a sequence of signals, announcing a prediction about the random variable after each such observation. After this, the realization of the random variable is observed. The principal updates his beliefs about the quality of her information. To embed this in a signal–choice model, the forecaster’s “message” can be thought of as a tuple giving the sequence of forecasts together with the realization of the random variable. A choice of a strategy by the forecaster giving

⁵For work on Bayesian persuasion with privately informed agents, see Perez–Richet (2014), Hedlund (2017), Kosenko (2020), and Koessler and Skreta (2021).

her forecasts as a function of the signals she sees generates a probability distribution over such sequences and hence is a signal choice. Deb, Pai, and Said’s result that the optimal mechanism in this setting does not require commitment by the principal is a special case of our results in Section 5. Their proof restricts attention to deterministic mechanisms; our results show that no such restriction is needed.

Espinosa (R) Ray (2023), Silva (2020), and Perez-Richet and Skreta (2021) also develop models that can be thought of as signal-choice models. However, these papers, while broadly related, focus on issues very different from the ones we explore.

3 Games

There are many timing assumptions one could consider in modeling the interaction between the agent and principal. We focus on the following sequential game.

First, the agent learns her type. In the evidence-acquisition model, she then chooses $a \in A_t$ and $M \subseteq \mathcal{L}$ is realized. She then chooses $m \in M$. If we consider the signal-choice model instead, the agent simply chooses $s \in S_t$ and the realization m is determined. Either way, the principal observes m but not the agent’s type or other information. The principal then chooses $x \in X$.

It is straightforward to show that the signal-choice model is a reduced form of the evidence-acquisition model. 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 each distribution and messaging strategy is equivalent to a signal choice. This is exactly the conversion described in Part 3 of our running example. In light of this, we could analyze the game as an evidence-acquisition model or equivalently replace the set of actions and messaging strategies with the set of induced signal distributions and analyze the game as a signal-choice model.

Recall that the signal-choice model can also be thought of as an evidence-acquisition

model where every set of evidence is a singleton. Thus in the context of the game considered here, these two models are equivalent — from any game in one class, we can construct an equivalent game in the other.

Running Example, Part 4. We illustrate the game with our running example. Since the evidence–acquisition model reduces to the signal–choice model, we focus on the latter. Assume the agent has two equally likely types, h and ℓ where $h > \ell > 0$. For the wage–setting version, we assume $X = \mathbf{R}_+$, $u(t, x) = x$, and $v(t, x) = -(t-x)^2$. That is, the principal chooses a wage, the agent’s utility is equal to the wage and the principal wishes to set the wage equal to the agent’s true productivity. For the hiring version, we assume $X = \{0, 1\}$, $u(t, x) = x$, and $v(t, x) = x(t - \bar{w})$ where $h > \bar{w} > \ell$. In other words, the agent wants to be hired ($x = 1$), while the principal wants to hire the high type but not the low type.

For either version, the following strategies form a perfect Bayesian equilibrium. Type h chooses the signal distribution which puts probability 1 on h , while ℓ chooses a distribution with probability ℓ/h on h and $1 - (\ell/h)$ on 0. The principal’s belief puts probability 1 on ℓ unless the signal he sees is h . By Bayes’ rule, if the principal sees signal h , his belief puts probability $h/(\ell + h)$ on type h , so the expected productivity is $(h^2 + \ell^2)/(h + \ell)$. He then chooses his action accordingly. So in the wage–setting version, he chooses $x = \ell$ if he sees any message other than h and sets $x = (h^2 + \ell^2)/(h + \ell)$ otherwise. In the hiring version, he does not hire if he sees any $m \neq h$. If he sees $m = h$, then he hires if

$$\frac{h^2 + \ell^2}{h + \ell} > \bar{w},$$

doesn’t hire if the reverse strict inequality holds, and can choose any probability of hiring otherwise. It is easy to see that, given the principal’s strategy, both types want to maximize the probability on signal h and these signal choices do that. So these strategies form an equilibrium.

To introduce the next section on mechanism design, consider the case where the principal can commit to his reaction to the m he observes. In this case, he can achieve his best possible outcome in the wage–setting version. To be specific, suppose the principal commits to $x = m$ if m is either h or ℓ and to $x = 0$ otherwise. Given any s chosen by the agent, the agent’s expected payoff is less than or equal to the expectation of m since for any m , the principal chooses $x \leq m$. Since every $s \in S_t$ has

expectation weakly less than t , this implies that the agent’s payoff must be weakly less than t . Since the agent can obtain a payoff of exactly t by choosing the degenerate s which produces $m = t$ with probability 1, we see that this is an optimal reply for the agent. Clearly, this enables the principal to set $x = t$ always, achieving his highest possible payoff. It is not hard to show that no (perfect Bayesian) equilibrium of the game yields the principal this payoff, so the ability to commit strictly improves the principal’s payoff.⁶

On the other hand, commitment does not help the principal in the hiring version. This is demonstrated in Section 4.3 and generalized in Section 5. ■

4 Mechanism Design

While any order for communication is “allowed” when studying games, for mechanism design, it is more standard to assume the sequence of communication steps which allows the principal to obtain the highest possible payoff. 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.⁷

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

⁶To see this, suppose there is an equilibrium that gives the principal this payoff. Then he must choose h in response to any signal realization that comes from type h with positive probability and ℓ in response to any signal realization that comes from type ℓ with positive probability. Also, sequential rationality implies that even off path, the principal never chooses an action smaller than ℓ . Hence it must be true that every realization of every signal distribution available to type ℓ leads the principal to choose ℓ . But this is impossible: type ℓ can generate any signal realization possible for type h with strictly positive probability.

⁷Gerardi 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 (2020) show that such problems do not arise in our single–agent setting.

and Obedience Principle identifies the appropriate protocol.⁸ First, the agent reports a type. Then the principal recommends a signal distribution. Finally, the agent chooses some distribution, the principal observes 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 this section, we give 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.

We begin with the evidence–acquisition model. We give a verbal description of the protocol and state our main result for this section, 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 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.

⁸For similar results in the evidence literature, see Bull and Watson (2007) and Deneckere and Severinov (2008).

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.

This section’s main result gives a condition on the evidence–acquisition technology which implies that each possible evidence set M has a “best” message in the sense that, without loss of utility, 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 $b_T : T \rightarrow T$. 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 $b_A : T \times T \times A \rightarrow A$, 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 $b_M : T \times T \times A \times A \times \mathcal{M} \rightarrow \mathcal{M}$. Finally, for stage 6, the agent has an evidence presentation strategy, $b_L : T \times T \times A \times A \times \mathcal{M} \times \mathcal{M} \times \mathcal{L} \rightarrow \mathcal{L}$. Of course, we require that $b_L(t, r, a, a', M, \hat{M}, m) \in M$ — 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_M , and B_L denote the sets of these functions respectively.

Similarly, for stage 2, the principal chooses a recommendation strategy $g_A : T \rightarrow A$, giving his recommended action as a function of the reported type. For stage 5, he chooses a message request strategy $g_L : T \times A \times \mathcal{M} \rightarrow \mathcal{L}$. We require that $g_L(r, a, \hat{M}) \in \hat{M}$. 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 $g_X : T \times A \times \mathcal{M} \times \mathcal{L} \times \mathcal{L} \rightarrow X$. 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.⁹

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^*

⁹Note 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.

as the *mechanism outcome*.

4.1 Identifying the Recommended Message

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 two ways to simplify the protocol and conditions under which these simplifications are without loss of generality.

In both cases, the idea is to identify some choices by the principal in a way which depends on the evidence structure but uses little or 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.

The idea behind the first simplification is to 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.

One way to understand when we can identify the principal's response in this way is by comparison to the literature with exogenously given evidence. In such models, 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 technology 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 .

To understand this 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 . In this sense, the presentation of m is evidence directly about M' , the agent’s set of evidence, not about t . It provides evidence only indirectly about t since types differ in terms of which evidence sets they are able or likely to obtain. The definition says that 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). In this sense, showing m_M^* reveals exactly what showing every message in M would reveal.

To put it differently, note that if the true message set, say M , is contained in M' , then nothing the agent could show would ever refute the possibility that the agent’s message set is M' . However, if $M \not\subseteq M'$, then there is some message $m \in M \setminus M'$ which the agent could show and prove conclusively that M' is not the feasible set. Normality says that for every M , there is one message in M which could be used to simultaneously rule out *every* such M' , proving to the principal that the true set of messages is either M or something which contains M .

Running Example, Part 5. In our 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 preferences, as asserted earlier. ■

To see more concretely that normality is about the information content of messages regarding the set of available messages, 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 $1/3$, $\{m_2\}$ with probability $1/3$, and $\{m_1, m_2\}$ with probability $1/3$. 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 $\{m_1, m_2\} \not\subseteq \{m'\}$. ■

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.¹⁰

The following theorem shows that normality will enable us to identify the principal's message recommendations, a result we can then use to simplify the protocol. Recall that a mechanism for the principal is a probability distribution γ over G with associated behavior strategy representation $(\gamma_A, \gamma_{\mathcal{L}}, \gamma_X)$.

¹⁰Another 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.

Theorem 1. *In the evidence–acquisition model, fix any incentive compatible mechanism γ . If the evidence technology is normal, then there exists an incentive compatible mechanism $(\gamma_A^*, \gamma_L^*, \gamma_X^*)$ with the following properties. First, $\gamma_L^*(t, a, M)(m_M^*) = 1$. That is, the principal always recommends the maximal evidence message for any reported M . Second, for all t ,*

$$\mu^*(x \mid \gamma, t) = \mu^*(x \mid \gamma^*, t), \quad \forall x \in X,$$

so γ and γ^* have the same mechanism outcome.

This simplification is, in general, not possible when the evidence technology is not normal. For example, there are preferences for the non–normal evidence technology in Example 1 for which it is better for the principal to request m_1 and preferences where it is better for him to request m_2 , *even though m_1 perfectly reveals the agent’s type.*¹¹

Theorem 1 implies that we can simplify the 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.

Hence a corollary to Theorem 1 is that we can use a simpler protocol. We refer to the following as the *abbreviated protocol for evidence–acquisition models*:

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 .

¹¹Illustrative examples are available on request.

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 $b_T : T \rightarrow T$ and $b_A : T \times T \times A \rightarrow A$ as before. Also, $b_{\mathcal{L}} : T \times T \times A \times A \times \mathcal{M} \rightarrow \mathcal{L}$ 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 $g_A : T \rightarrow A$, with $g_A(t) \in A_t$ and $g_X : T \times A \times \mathcal{L} \rightarrow X$ 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 B:

Corollary 1. *Assume the evidence technology is normal. 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 mechanism outcome.*

4.2 Reduction to Signal Choice

The identification of the principal's recommended message under normality enables us to 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*. As before, the notation is summarized after we state the main result of this section.

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 $b_T : T \rightarrow T$. A pure strategy for the principal for Stage 2 is denoted $g_S : T \rightarrow S$. Let $b_S : T \times T \times S \rightarrow S$ with $b_S(t, r, s) \in S_t$ denote a typical pure strategy for the agent for Stage 3. Finally, let $g_X : T \times S \times \mathcal{L} \rightarrow X$ 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 outcomes, here defined as (s, x) pairs. I.e., we can write $\mu^*(s, x \mid \gamma, t)$ as the probability distribution over (s, x) induced by the strategies (\hat{b}^*, γ) given the agent's type is t .

As in our analysis of games in Section 3, we can think of the agent's strategy in the evidence–acquisition model as a choice of a distribution over evidence sets and a messaging strategy. Again, a distribution and messaging strategy generates a probability distribution over the message the agent shows the principal. 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, under normality, 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 $\sigma : \text{supp}(a) \rightarrow \mathcal{L}$ such that $\sigma(M) \in M$, we can define a signal $s \in \Delta(\mathcal{L})$ by

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

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

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

This is exactly the translation from evidence acquisition to signal choice discussed less formally in Section 3.

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

Theorem 2. *In the evidence–acquisition model, fix any incentive compatible mechanism γ . If the evidence technology is normal, then 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*

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

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

In short, given normality, 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 signal–choice model. This is analogous to our result on games in Section 3.

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 analog of Theorem 1, Corollary 1, and Theorem 2 for this model. More specifically, it is still true that under normality, one can restrict attention to mechanisms for which the principal always recommends the maximal evidence message for any evidence set, enabling us to use (an appropriately modified version of) the abbreviated protocol and reduce to a version of the signal–choice model.

4.3 Identifying the Recommended Signal

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 normality.

While normality 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 in $\mathbf{R}_+^{\#\mathcal{L}}$. 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 $s \in S_t$, $s\Lambda \in \text{conv}(S_t)$.¹²

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 “exper-

¹²A matrix is Markov if all entries are non–negative and every row sum is 1.

iment” \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 the sense in which 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 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} and that 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 2. 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

	s_1	s_2	s_3	s'_3
m_1	1	0	0	1/2
m_2	0	1	0	1/2
m_3	0	0	1	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 3. 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

	s_1	s_2	s'_2
m_1	1	0	1/2
m_2	0	1	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. ■

Example 4. As in Example 2, 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

	s_1	s_2	s'_2
m_1	1/2	1/4	2/3
m_2	1/2	3/4	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 ,*

$$\mu^*(x | \gamma, t) = \mu^*(x | \gamma^*, t), \quad \forall x \in X.$$

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

Remark 1. Theorems 1 and 3 are connected in the following sense. 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 particular 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 in the sense defined above. (See Section E 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 (2023) 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’s notion of more discerning tests is related to our notion of more informative signals but is not the same. 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 $\pi(\tau | t)$ 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

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

and

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

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 see the connection to our condition, we write the signal distribution corresponding to test τ , type t , and the agent taking effort as $\tau_+(t) = (1 - \pi(\tau | t), \pi(\tau | t))$,

so that this is a distribution over $\{0, 1\}$. We write the distribution given no effort as $(1, 0)$ for any test and type. Given k_0 and k_1 satisfying Ball and Kattwinkel’s definition, let

$$\Lambda = \begin{pmatrix} k_0 & 1 - k_0 \\ k_1 & 1 - k_1 \end{pmatrix}.$$

Then their definition says that $\hat{\tau}_+(t)\Lambda = \tau_+(t)$ and that for every $t' \neq t$, $\hat{\tau}_+(t')\Lambda \in \text{conv}(\{(1, 0), \tau_+(t')\})$. It is not hard to show that their requirement that $k_1 \geq k_0$ is equivalent to $(1, 0)\Lambda \in \text{conv}(\{(1, 0), \tau_+(t)\})$.

In other words, there are two differences between $\hat{\tau}$ being more t -discerning than τ in their sense and the signal distribution for t given by $\hat{\tau}_+(t)$ being more informative than $\tau_+(t)$ in our sense when $S_{t'} = \{\hat{\tau}_+(t'), \tau_+(t'), (1, 0)\}$ for all t' . First, their restriction on the garbled signals only applies to the distribution under test $\hat{\tau}$ (with and without effort), not also to test τ with effort. Second, the convex hulls the garbled signals must lie in is only the convex hull of τ with and without effort.

In our model, the agent can choose any distribution in her feasible set, while in Ball and Kattwinkel, the agent can only choose distributions that can be generated by a choice of an effort level through the test chosen by the principal. Equivalently, the principal can observe and punish any deviation by the agent to the “wrong” test. Consequently, Ball and Kattwinkel’s informativeness comparisons can ignore incentive constraints associated with signals that are not generated by the test specified by the principal. The principal only needs to compare what happens with test $\hat{\tau}$ to what would happen with τ , while we require the principal to consider what happens when both tests are available.

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 $b_T : T \rightarrow T$ and a signal distribution strategy by $b_S : T \rightarrow S$

with $b(t) \in S_t$.

Stage 2. The principal observes the report, the realized m , and chooses an outcome. Let $g_X : T \times \mathcal{L} \rightarrow X$ 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 6. We showed in Part 5 of the example that the 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 , where we also generalize the example, now letting T be any finite subset of \mathbf{R}_+ , not necessarily $\{\ell, h\}$. 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 .

Now that we have identified the signal choices for each type in the optimal mechanism, it is not difficult to compute the rest of the mechanism. We already showed that the principal can achieve his best possible outcome for each type when his utility function is $-(t-x)^2$, so consider the hiring version where the principal's choice is to hire the agent ($x=1$) or not ($x=0$) and his payoff is $x(t-\bar{w})$ where $\bar{w} \in (\ell, h)$. Recall that types are equally likely. The agent's payoff is x . Let $\gamma^*(t)$ denote the probability the principal chooses $x=1$ when the agent reports type t and the realized message m also equals t . Given that the mechanism will induce truthful reporting and will induce the agent to choose the degenerate distribution with $m=t$, the principal's expected payoff is

$$\frac{1}{2}\gamma^*(h)(h-w) + \frac{1}{2}\gamma^*(\ell)(\ell-w).$$

Clearly, we may as well assume the mechanism has $x=0$ if the message observed differs from the agent's type report. As we will see, type h never wishes to imitate ℓ , so we do not need to impose this incentive compatibility constraint. Hence the only

incentive compatibility constraint we require is

$$\gamma^*(\ell) \geq \gamma^*(h) \frac{\ell}{h},$$

since the maximum probability ℓ can put on $m = h$ is when she chooses the distribution with probability ℓ/h on h and the remaining probability on 0. Since the principal's utility is decreasing in $\gamma^*(\ell)$ and increasing in $\gamma^*(h)$, the constraint is binding. Hence the principal chooses $\gamma^*(h)$ to maximize

$$\gamma^*(h) \left[\frac{1}{2} (h - \bar{w}) + \frac{1}{2} \frac{\ell}{h} (\ell - \bar{w}) \right].$$

So if

$$\frac{h^2 + \ell^2}{h + \ell} > \bar{w},$$

the optimal mechanism has $\gamma^*(h) = 1$ and $\gamma_\ell^* = \ell/h$. If we have the opposite strict inequality, it has $\gamma^*(h) = \gamma^*(\ell) = 0$. In both cases, type h has no incentive to imitate type ℓ , as asserted.

Also, in both cases, the outcome is the same as in the equilibrium we computed for this example in Section 3. In this sense, there is no value to the principal from commitment: he obtains the same outcome when he is able to commit to his responses to the agent and in a particular equilibrium of the game where he cannot commit. We present a generalization of the result of this example in the following section. ■

5 Commitment

We just saw in Section 4 that there is no value to commitment in the hiring version of our running example. In this section, we generalize this result.

Our generalization works in two steps. First, we introduce an assumption on *endogenous* variables and show that when it holds, there is no value to commitment. More specifically, when this condition is satisfied, there is a Nash equilibrium in the game between the principal and the agent without commitment that gives the principal the same payoff as in the optimal mechanism. As we show, essentially all

of the results in the literature showing no value to commitment can be thought of as identifying various conditions on primitives which imply our assumption on the endogenous variables.

Second, we identify a condition on primitives for the stochastic evidence model which implies our assumption on endogenous variables and extend the first result to address the question of sequential rationality. As we explain further below, the main complication posed by sequential rationality is that the analysis requires much more detail about the structure of the protocol than our Nash equilibrium result uses.

The first of these results applies to *any* finite game between the principal and the agent. In particular, it applies to any protocol for the evidence–acquisition model, including the special case of the signal–choice model, but also to models that have nothing to do with evidence or signals at all. To state this result, fix any finite set of pure strategies for the agent, denoted B , and any finite set of pure strategies for the principal, denoted G . Let $U(\beta, \gamma)$ denote the agent’s expected payoff in the game given mixed strategy profile $(\beta, \gamma) \in \Delta(B) \times \Delta(G)$, where we take expectations over the randomization of the strategies as well as any randomness in the game itself, such as the realization of the agent’s type. Similarly, let $V(\beta, \gamma)$ denote the principal’s expected payoff given (β, γ) .

Given any $\gamma \in \Delta(G)$, let $BR(\gamma)$ denote the agent’s set of best replies — i.e.,

$$BR(\gamma) = \{\beta \in \Delta(B) \mid U(\beta, \gamma) \geq U(\beta', \gamma), \forall \beta' \in \Delta(B)\}.$$

Let

$$V^* = \max_{\gamma \in \Gamma} \max_{\beta \in BR(\gamma)} V(\beta, \gamma).$$

In other words, V^* is the principal’s maximal expected payoff when he can commit to any mixed strategy in the game *and* can choose the agent’s best reply to his strategy. If (β^*, γ^*) solves $V^* = V(\beta^*, \gamma^*)$ and $\beta^* \in BR(\gamma^*)$, we say γ^* is *optimal for the principal* and refer to β^* as the associated β .

The only assumption we make on the game is, as mentioned above, a condition on *endogenous* variables. We say that the game is *aligned* if there exists a γ^* which

is optimal for the principal with the property that

$$V^* = V(\beta, \gamma^*), \quad \forall \beta \in BR(\gamma^*).$$

That is, the game is aligned if changes in the agent's best response do not affect the principal's payoff at some optimal mechanism. We discuss this assumption in detail below.

The following is our result on value of commitment relative to Nash equilibrium.

Theorem 4. *Fix any aligned game. Then there exists γ^* which is optimal for the principal and $\hat{\beta} \in \Delta(B)$ such that $(\hat{\beta}, \gamma^*)$ is a Nash equilibrium of the game induced by the protocol and $V(\hat{\beta}, \gamma^*) = V^*$.*

Proof. Suppose not. Fix an optimal γ^* for the principal with the property that $V(\beta, \gamma^*) = V^*$ for all $\beta \in BR(\gamma^*)$. The assumption of alignment is precisely that such γ^* exists.

We construct a contradiction as follows. Consider the restricted game where the principal's set of pure strategies is G , but the agent's set of pure strategies is $B \cap BR(\gamma^*)$. By finiteness of B and G , the restricted game has a mixed equilibrium, say, $(\hat{\beta}, \hat{\gamma})$.

By construction, $\hat{\beta}$ can only put positive probability on b 's that are best replies to γ^* and hence $\hat{\beta}$ is a best reply to γ^* . Hence by alignment, $V(\hat{\beta}, \gamma^*) = V^*$. By hypothesis, there is no Nash equilibrium giving the principal a payoff as large as V^* , so $(\hat{\beta}, \gamma^*)$ must not be a Nash equilibrium. Since $\hat{\beta}$ is a best reply to γ^* , this means that γ^* must not be a best reply to $\hat{\beta}$. Hence $V(\hat{\beta}, \hat{\gamma}) > V(\hat{\beta}, \gamma^*) = V^*$.

Hence for every $\varepsilon \in (0, 1)$,

$$\varepsilon V(\hat{\beta}, \hat{\gamma}) + (1 - \varepsilon)V(\hat{\beta}, \gamma^*) > V(\hat{\beta}, \gamma^*) = V^*,$$

so

$$V(\hat{\beta}, \varepsilon \hat{\gamma} + (1 - \varepsilon)\gamma^*) > V^*.$$

We contradict this by showing that for all sufficiently small $\varepsilon > 0$, $\hat{\beta} \in BR(\varepsilon \hat{\gamma} + (1 - \varepsilon)\gamma^*)$. This is a contradiction because it implies that the principal would be strictly

better off committing to $\varepsilon\hat{\gamma} + (1 - \varepsilon)\gamma^*$ and choosing $\hat{\beta}$ for the agent's best reply.

We show that for ε sufficiently small, $\hat{\beta}$ is a better reply for the agent than any other pure strategy. To see this, fix any pure strategy b . Then

$$U(b, \varepsilon\hat{\gamma} + (1 - \varepsilon)\gamma^*) \leq U(\hat{\beta}, \varepsilon\hat{\gamma} + (1 - \varepsilon)\gamma^*)$$

if and only if

$$\varepsilon[U(b, \hat{\gamma}) - U(\hat{\beta}, \hat{\gamma}) + U(\hat{\beta}, \gamma^*) - U(b, \gamma^*)] \leq U(\hat{\beta}, \gamma^*) - U(b, \hat{\gamma}^*). \quad (1)$$

First, suppose $b \in BR(\gamma^*)$, so $U(b, \gamma^*) = U(\hat{\beta}, \gamma^*)$. In this case, equation (1) reduces to

$$\varepsilon[U(b, \hat{\gamma}) - U(\hat{\beta}, \hat{\gamma})] \leq 0.$$

By the definition of the reduced game, $\hat{\beta}$ is a better response to $\hat{\gamma}$ than any $b \in BR(\gamma^*)$, so equation (1) holds.

So suppose $b \notin BR(\gamma^*)$. In this case, $U(b, \gamma^*) < U(\hat{\beta}, \gamma^*)$, so the right-hand side of equation (1) is strictly positive. Hence if the term in brackets on the left-hand side is less than or equal to zero, this holds for all $\varepsilon > 0$. So assume b is such that the term in brackets is strictly positive. Let \hat{B} denote the set of such b and let

$$\Delta = \min_{b \in \hat{B}} \frac{U(\hat{\beta}, \gamma^*) - U(b, \hat{\gamma}^*)}{U(b, \hat{\gamma}) - U(\hat{\beta}, \hat{\gamma}) + U(\hat{\beta}, \gamma^*) - U(b, \gamma^*)}.$$

Since B and hence \hat{B} are finite, the minimum is well-defined and strictly positive. Hence for every $\varepsilon \in (0, \Delta)$, equation (1) holds.

Hence for every such ε , $\hat{\beta} \in BR(\varepsilon\hat{\gamma} + (1 - \varepsilon)\gamma^*)$, a contradiction. ■

As noted, the assumption that the game is aligned is a hypothesis about endogenous variables. As we now explain, the various results in the literature characterizing situations where commitment does not have value can be thought of as various ways to generate the property that the relevant game is aligned.

The earliest results showing no value to commitment were due to Glazer and Ru-

binstein (2004, 2006). They considered the case where the principal chooses between two outcomes, called accept ($x = 1$) and reject ($x = 0$). The agent’s utility is x . The principal’s utility is x if the agent’s type is in a certain set of types, $-x$ otherwise. They consider nonstochastic evidence — that is, each type has only a single degenerate distribution over evidence sets. As in our model and all the models discussed below, the agent knows her type from the outset of the game. Hence if we compare two best replies by the agent to a given mechanism, these strategies must be optimal for every type of the agent. In particular, in Glazer and Rubinstein, this implies that every type of the agent must get accepted with the same probability in each best response. But then the principal’s payoff is the same across agent best responses as well, so the game is aligned.

Sher (2011) and Hart, Kremer, and Perry (2017) generalize Glazer–Rubinstein to allow more than two possible actions but where the principal’s utility can be written as a function of the agent’s utility. More specifically, the agent’s utility function u depends only on the principal’s choice, x , while the principal’s utility v depends on both x and the agent’s type t . The sense in which the principal’s utility can be written as a function of the agent’s utility is that if $u(x) = u(x')$, then $v(x, t) = v(x', t)$ for all t .

These models also have deterministic evidence, so the only random element of the model is the prior over the agent’s type. Both models have other assumptions which imply the existence of an optimal *deterministic* mechanism. In a deterministic mechanism, we have a similar argument to the one above. If the agent has multiple best responses to the mechanism, then every type of the agent must get the same utility from each of these best responses. Because there is no randomness given the agent’s type, this constant utility across best responses implies that the principal’s utility given any type of the agent is also constant across the agent’s best responses. Hence the game is aligned.¹³

Ben-Porath, Dekel, and Lipman (2019) differs in part by considering the multi-agent case. Specializing to the single-agent case, we assumed that the principal’s

¹³Hart, Kremer, and Perry allow infinitely many actions, but this difference from our model is not relevant. They have finitely many types, so only finitely many actions would ever be chosen in a deterministic mechanism.

utility function could be written as $v(t, x) = \nu(t)u(t, x) + \psi(x) + \varphi(t)$.¹⁴ We also had deterministic evidence and, like the papers discussed above, proved the existence of an optimal deterministic mechanism. However, the simple argument above does not apply here. If the agent changes to a different best reply, this would necessarily leave $u(t, x)$ unchanged for each t but might change $\psi(x)$ for some t 's and hence could change the principal's utility. However, our characterization of the optimal mechanism in that paper shows that it is "measurable" with respect to the agent's payoff in the sense that if two types of the agent receive the same payoff in the optimal mechanism, then the outcome for them is the same as well. This turns out to imply that the optimal mechanism has the property that any alternative best reply for the agent does not change the outcome and hence leaves the principal indifferent. Consequently, the game is again aligned.

When we move from deterministic evidence models to allowing some stochastic components, these conditions are, in general, no longer sufficient to ensure that the game is aligned. The simplest way to see this is to consider the wage-setting version of our running example. In this example, $u(t, x) = x$ and $v(t, x) = -(x - t)^2 = 2tx - t^2 - x^2$. We can rewrite this as $v(t, x) = 2tu(t, x) - t^2 - x^2$. Letting $\nu(t) = 2t$, $\varphi(t) = -t^2$, and $\psi(x) = -x^2$, we obtain $v(t, x) = \nu(t)u(t, x) + \psi(x) + \varphi(t)$, showing that the condition in our 2019 paper is satisfied. However, in Section 3, we showed that commitment enables the principal to obtain a strictly higher payoff than in any equilibrium in this example. Thus Theorem 4 implies that this game is not aligned with stochastic evidence.

The reason for this is the unavoidable additional randomness when we move away from deterministic evidence. If the agent switches between best replies, the *expected* utility for any type does not change, but the distribution over her utility may. Unless the principal's utility is a *linear* function of the agent's utility, this will generally mean that changes in the agent's best reply *do* affect the principal's utility. In the wage-setting version of our running example, the principal's payoff is strictly concave in the agent's utility, so this condition is violated.¹⁵ In short, to accommodate such

¹⁴We did not include the $\varphi(t)$ term as it has no implications for behavior and hence can be included or omitted without changing any results. We include it here to simplify the discussion below.

¹⁵To see this point more concretely, note that in the equilibrium we computed in Part 4 of our running example, type ℓ receives wage h with probability ℓ/h and ℓ otherwise. If the principal could commit, she could offer to pay the expectation of this wage with probability 1 if the agent reports ℓ

randomness, we need stronger conditions on the utility functions to ensure that the game is aligned.

We now give an assumption on preferences in our evidence acquisition model which ensures that the protocol is aligned and allows us to extend Theorem 4 to perfect Bayesian equilibrium. The assumption is that there is some function $\nu : T \rightarrow \mathbf{R}$ such that $v(t, x) = \nu(t)u(t, x) + \psi(t)$ for all $(t, x) \in T \times X$. When this holds, we say the preferences are *semi-aligned*. In other words, we use the same assumption as our 2019 paper but without the additional $\psi(x)$ term allowed. It is easy to see that this implies that if $U(b, g, t) = U(b', g', t)$ for all $t \in T$, then $V(b, g) = V(b', g')$. That is, if all types of the agent are indifferent between *any* two outcomes, then the principal is as well. Clearly, this implies that the game is aligned.

While the assumption of semi-aligned preferences is nontrivial, it is without (further) loss of generality when the principal has two actions available — i.e., when $\#X = 2$. Thus it holds in the hiring version of our running example. Other natural settings where the principal has two feasible actions are cases where the principal has to decide whether to fund the agent’s project, to lend funds, to provide a resource, etc.

To see that preferences are semi-aligned when there are only two outcomes, denote the outcomes x_0 and x_1 . Then we can renormalize the agent’s payoffs so that $u(t, x_0) = 0$ for all t , $u(t, x_1) = 1$ for types t who prefer x_1 to x_0 , and $u(t, x) = -1$ for types who prefer x_0 to x_1 .¹⁶ We can renormalize the principal’s utility function so that $v(t, x_0) = 0$ for all t . Without loss of generality, assume $v(t, x_0) \neq v(t, x_1)$ for all t .¹⁷ Given these renormalizations, we can write $v(t, x) = \nu(t)u(t, x)$ where $\nu(t) = v(t, x_1)/u(t, x_1)$.

To extend Theorem 4 to perfect Bayesian equilibrium, we need to put more structure on the protocol. Otherwise, it is difficult to characterize what kind of choices the

and to follow the equilibrium otherwise. Clearly, the agent is indifferent, but the principal pays the expected value for sure rather than facing the gamble. Given the concavity of the principal’s utility function, this is an improvement.

¹⁶Types who are indifferent between x_0 and x_1 do not affect the arguments, so we can assume without loss of generality that there are no such types.

¹⁷If this is violated for some t , then the principal’s decisions are the same as those he would make if such t were impossible. Hence such types can be disregarded.

principal might have at certain information sets and therefore difficult to characterize sequential rationality at all information sets. We emphasize that the additional structure is to allow a relatively straightforward proof; we do not know of any counterexamples from protocols outside the class we consider. We see the particular protocol used here as a reasonably general but illustrative structure.

To avoid repetition, we state the definitions, result, and proof for the evidence-acquisition model, but it is not difficult to rewrite it for the signal-choice model instead. For simplicity, we assume the protocol is a multi-stage game with certain properties. To be specific, as in all our previous analysis, we assume that the agent learns her type first.

After this, we have some fixed finite number of stages. Each of these stages has one of two forms. The first possibility is that we have cheap talk messages, either one from the agent to the principal or one from the principal to the agent. The set of cheap-talk messages is fixed throughout, independently of the agent's type or any actions. At the end of such a stage, both players observe the message sent. The second possibility is that the agent chooses some unobserved action which may affect the set of evidence she'll end up with and she may privately observe some outcome of this action. At the end of such a stage, the principal does not observe either the agent's action or this outcome. For simplicity, we suppose that the order in which these various forms of stages occur is fixed exogenously, independently of the agent's type or actions.

After these stages, there's a last stage where the agent presents an evidence message to the principal and the principal responds by choosing an action from the set X . The set of evidence messages available to the agent depends stochastically on the agent's type and the sequence of actions and outcomes from the earlier stages.

We require, as above, that the principal and agent have finite sets of pure strategies. Hence we assume there is a finite K such that for all feasible histories of messages, no more than K stages are played. Similarly, we assume that the set of all possible cheap talk messages for either player is finite as is the set of actions and possible evidence messages for the agent.

We say that a protocol satisfying these properties is *allowable*.

Theorem 5. *Given any allowable protocol, under the assumptions of Theorem 4, there is a perfect Bayesian equilibrium $(\hat{\beta}, \hat{\gamma})$ with $V(\hat{\beta}, \hat{\gamma}) = V^*$.*

It is not straightforward to extend this result to multiple agents. For example, suppose we have two agents, $i = 1, 2$. Suppose the principal’s decision is which agent to give one unit of a good to. Let $X = \{0, 1, 2\}$ where $x = 0$ means the principal keeps the good and $x = i$ means the principal gives the good to agent i . Suppose agent i ’s utility function, $u_i(t_i, x)$ is 1 if i receives the good, 0 otherwise. Suppose the principal’s payoff is $v_i(t_i)$ if he gives the good to agent i . Then we can write the principal’s utility as $v(t, x) = \sum_i v_i(t_i)u_i(t_i, x)$, a natural generalization of our assumption of semi-aligned preferences for the multiple agent case. One can give examples (available on request) showing that the no-value-to-commitment result does not hold for this model even though the principal has only two actions. This is in contrast to results in Ben-Porath, Dekel, and Lipman (2019) for deterministic evidence.

Finally, there is one related result outside the evidence literature. Vohra (r) Espinosa (r) Ray (2021) consider a principal-agent model with no value to commitment. Because their model does not include evidence, all communication is “cheap talk,” so there are no issues related to ensuring “appropriate” behavior off the equilibrium path. Hence it is most natural to compare their result to our result for Nash equilibrium, Theorem 4. Their primary assumption is that we can write the principal’s utility as a function only of the agent’s utility and the agent’s type. Because the agent in their model has a hidden action, this is not itself sufficient to imply that the game is aligned in our sense. They add an additional assumption which is akin to the measurability property derived in Ben-Porath, Dekel, and Lipman (2019) for our model. These two assumptions together do imply that their game is aligned. However, this is not sufficient to imply their result as they do not allow randomization, an ingredient we use in our proof. Instead, they have a richness assumption which serves a similar purpose.

Appendix

A Proof of Theorem 1

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 .

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 $\hat{m} \neq m_M^*$. 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_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 $m \notin \{\hat{m}, m_M^*\}$. Finally, let

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

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

Let $\gamma_X^*(h, m, m')(x) = \gamma_X(h, m, m')(x)$ if $(h, m) \neq (\hat{h}, m_M^*)$. 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 $m \in \mathcal{L} \setminus \{m_M^*\}$, we set $\gamma_X^*(\hat{h}, m_M^*, m)(x)$ equal to

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

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 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 $z : \hat{M} \rightarrow M$ 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

$$\begin{aligned} & \sum_{(x,m) \in X \times (\mathcal{L} \setminus \{\hat{m}, m_M^*\})} \gamma_{\mathcal{L}}(\hat{h})(m) \gamma_X(\hat{h}, m, z(m))(x) u(t, x) \\ & + \gamma_{\mathcal{L}}^*(\hat{h})(m_M^*) \sum_{x \in X} \gamma_X^*(\hat{h}, m_M^*, z(m_M^*))(x) u(t, x). \end{aligned}$$

We have two cases. First, suppose $z(m_M^*) \neq m_M^*$. In this case, the last term is equal to

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

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_M^* by sending $z(m_M^*)$. 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 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, $\gamma_X^* : T \times A \times \mathcal{L} \rightarrow \Delta(X)$, while in the full protocol, $\gamma_X : T \times A \times \mathcal{M} \times \mathcal{L} \times \mathcal{L} \rightarrow \Delta(X)$ 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 .¹⁸ 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 mechanism to any deviation, so the mechanism is incentive compatible.

¹⁸It 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.

C 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 $\gamma_A : T \rightarrow \Delta(A)$ and $\gamma_X : T \times A \times \mathcal{L} \rightarrow \Delta(X)$. For the signal choice model, a mechanism is a pair of functions $\gamma_S^* : T \rightarrow \Delta(S)$ and $\gamma_X^* : T \times S \times \mathcal{L} \rightarrow \Delta(X)$.

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 directly 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,σ^*} , 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 $\hat{\sigma}$. Hence the mechanism is incentive compatible.

D 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 $\mathcal{L} = \{m_1, \dots, m_L\}$ 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 $\gamma_X^*(t_1, s_1, m_i)$ 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

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

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.

E Proof of Remark 1

Let $s = s_{(a,\sigma)}$ and $s^* = s_{(a,\sigma^*)}$ where $\sigma^*(M) = m_M^*$ for all $M \in \text{supp}(a)$. 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 the messages as m_1, \dots, 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 j th 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 j th element of $\hat{s} \Lambda$ is

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

or

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

In other words, $\hat{s}\Lambda$ is constructed as follows. We choose a message, say m_i , according to distribution \hat{s} . If $s^*(m_i) = 0$, then this message is sent. If $s^*(m_i) > 0$, 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_i , it must be able to send that message. So we need to show that the randomization is feasible — that is, that whenever m_i could be sent, every message in M_i is also feasible. But this follows from the fact that $m_i = m_{M_i}^*$. By definition, this means that if the feasible set is M and $m_i \in M$, then $M_i \subseteq M$. So if $\hat{s} \in S_t$, then $\hat{s}\Lambda \in S_t$, completing the proof.

F Proof of Theorem 5

F.1 Lemma

The following result will be useful. Let W be a finite set of states of the world and A a finite set of actions. Let $u : A \times W \rightarrow \mathbf{R}$ be a utility function. Say that $\sigma \in \Delta(A)$ is a *best reply* to $p \in \Delta(W)$ if

$$\sum_w p(w) \sum_a \sigma(a)u(a, w) \geq \sum_w p(w) \sum_a \sigma'(a)u(a, w), \quad \forall \sigma' \in \Delta(A).$$

Say that σ is a *best reply* if there exists $p \in \Delta(W)$ such that σ is a best reply to p . Say that σ is *strictly dominated* if there exists $\sigma' \in \Delta(A)$ such that

$$\sum_a \sigma'(a)u(a, w) > \sum_a \sigma(a)u(a, w), \quad \forall w \in W.$$

Standard results say that σ is a best reply if and only if it is not strictly dominated. (This is typically stated for pure strategies σ , but it applies to mixed as well.)

Lemma 1. *Suppose σ' is strictly dominated. Then there exists a mixed strategy $\hat{\sigma}$ which strictly dominates σ' and which is not itself strictly dominated.*

Proof. Suppose not. That is, suppose σ' is strictly dominated, but that there is no

undominated strategy which strictly dominates it. Let Σ^* denote the set of undominated strategies in $\Delta(A)$. Equivalently, Σ^* is the set of all $\sigma \in \Delta(A)$ that are best replies. By finiteness of A , this set is nonempty.

Let

$$\mathcal{U} = \text{conv} \left(\left\{ u \in \mathbf{R}^W \mid \exists \sigma \in \Sigma^* \cup \{\sigma'\} \text{ with } u_w = \sum_a \sigma(a)u(a, w), \forall w \right\} \right).$$

$$\mathcal{U}^D = \left\{ u \in \mathbf{R}^W \mid u_w \geq \sum_a \sigma'(a)u(a, w), \forall w \right\}.$$

By hypothesis, there is no mixed strategy in Σ^* which strictly dominates σ' . Hence $\mathcal{U} \cap \text{int}(\mathcal{U}^D) = \emptyset$, so the interiors of \mathcal{U} and \mathcal{U}^D are disjoint. Clearly, both sets are nonempty and convex. Hence there exists a separating hyperplane. That is, there is $p \in \mathbf{R}^W$ such that $p \neq 0$ and $p \cdot u \geq p \cdot \hat{u}$ for all $u \in \mathcal{U}^D$, $\hat{u} \in \mathcal{U}$.

Consider \hat{u} defined by $\hat{u}_w = \sum_a \sigma'(a)u(a, w)$. Obviously, this is an element of \mathcal{U} . Consider u defined by $u_w = \hat{u}_w$ for $w \neq w'$ and $u_{w'} = \hat{u}_{w'} + \varepsilon$ for some $\varepsilon > 0$ and some w' . Clearly, this is an element of \mathcal{U}^D . Hence the separating hyperplane satisfies $p_{w'}\varepsilon \geq 0$. Since w' is arbitrary, $p_w \geq 0$ for all w . Since $p \neq 0$, we can renormalize by replacing p with \hat{p} defined by $\hat{p}_w = p_w / \sum_{w'} p_{w'}$. Hence $\hat{p} \in \Delta(W)$.

Continuing with the same \hat{u} as above, we see that we have $\hat{p} \in \Delta(W)$ such that

$$\sum_w \hat{p}(w) \sum_a \sigma'(a)u(a, w) \geq \sum_w \hat{p}(w)u_w \quad \forall u \in \mathcal{U}.$$

In particular, for any $\sigma \in \Sigma^*$, we can let u be the vector defined by $u_w = \sum_a \sigma(a)u(a, w)$ to conclude that

$$\sum_w \hat{p}(w) \sum_a \sigma'(a)u(a, w) \geq \sum_w \hat{p}(w) \sum_a \sigma(a)u(a, w) \quad \forall \sigma \in \Sigma^*.$$

By hypothesis, σ' is strictly dominated by some mixed strategy, say $\hat{\sigma} \notin \Sigma^*$. Hence

$$\sum_w \hat{p}(w) \sum_a \hat{\sigma}(a)u(a, w) > \sum_w \hat{p}(w) \sum_a \sigma'(a)u(a, w) \geq \sum_w \hat{p}(w) \sum_a \sigma(a)u(a, w) \quad \forall \sigma \in \Sigma^*.$$

Hence no best reply to \hat{p} is contained in Σ^* , a contradiction. ■

F.2 Theorem 5

Let H_P denote the set of possible public histories — i.e., histories both the principal and the agent see. More specifically, H_P consists of the various possible sequences of cheap-talk messages as well as the possible complete public histories of all cheap-talk messages followed by the agent's evidence message. It will be convenient to write such a history in the form $h \cdot r \cdot h'$ where h is a history, r the next cheap talk message observed, and h' a continuation.

Let H_A denote the set of private histories for the agent and denote a typical element by h_A . Hence h_A lists what the agent observes that the principal does not — her type, her action choice at each evidence action stage, and the outcome of that action choice. The full history observed by the agent — the public plus the private — will be written as (h, h_A) and the set of these histories is denoted H_F .

As before, β is a behavior strategy for the agent and γ a behavior strategy for the principal. Thus β maps H_F to possible choices for the agent, while γ maps public histories H_P to actions for the principal. We let ρ denote a belief for the principal, where this is a function from H_P to beliefs over T .

Because the protocol is allowable, all information sets for the agent are singletons. Because there is no issue of beliefs for the agent, given any strategy γ for the principal, we can define the set of strategies for the agent which are sequentially rational best replies, denoted $BR^s(\gamma)$.

Fix (β^*, γ^*) with $V(\beta^*, \gamma^*) = V^*$ and $\beta^* \in BR(\gamma^*)$, so that (β^*, γ^*) is optimal for the principal. If we construct the restricted game used in the proof of Theorem 4 by restricting the agent to strategies in $BR^s(\gamma^*)$ instead of $BR(\gamma^*)$, nothing in the proof changes. So there exists $\hat{\beta} \in BR^s(\gamma^*)$ such that $(\hat{\beta}, \gamma^*)$ is a Nash equilibrium with $V(\hat{\beta}, \gamma^*) = V^*$. Since $\hat{\beta} \in BR^s(\gamma^*)$, the agent's strategy is sequentially rational at all information sets. Hence we can assume we have a Nash equilibrium (β^*, γ^*) satisfying $V(\beta^*, \gamma^*) = V^*$ such that the agent's strategy is sequentially rational at all information sets.

Without loss of generality, we can also assume that all possible cheap-talk messages have positive probability at every cheap-talk stage in equilibrium. In other words, we can assume without loss of generality that (β^*, γ^*) satisfy the following two properties. First, for every public history h leading to a stage where the principal sends cheap talk, for every feasible cheap talk message r at that stage, $\gamma^*(h)(r) > 0$. Second, for every public history h leading to a stage where the agent sends cheap talk, for every feasible cheap talk message r at that stage, there exists a private history for the agent h_A consistent with being at this stage¹⁹ such that $\beta^*(h, h_A)(r) > 0$.

To show this, first consider the principal. Fix any stage where the principal chooses a cheap-talk message and a public history h leading up to this stage. Suppose cheap talk message \bar{r} has zero probability — i.e., $\gamma^*(h)(\bar{r}) = 0$. Fix any \hat{r} with $\gamma^*(h)(\hat{r}) > 0$. Then we change β^* , γ^* , and ρ^* to $\hat{\beta}$, $\hat{\gamma}$, and $\hat{\rho}$ as follows. Let $\hat{\gamma}(h)(\bar{r}) = \hat{\gamma}(h)(\hat{r}) = \gamma^*(h)(\hat{r})/2$. For every other cheap talk message r that the principal could send at this stage, we let $\hat{\gamma}(h)(r) = \gamma^*(h)(r)$. In other words, we spread the probability the principal was putting on \hat{r} across \hat{r} and \bar{r} .

For any continuation public history h' , let $\hat{\gamma}(h \cdot \bar{r} \cdot h') = \hat{\gamma}(h \cdot \hat{r} \cdot h') = \gamma^*(h \cdot \hat{r} \cdot h')$. Note that h' includes the “empty continuation.” We set $\hat{\rho}(h \cdot \bar{r} \cdot h') = \hat{\rho}(h \cdot \hat{r} \cdot h') = \rho^*(h \cdot \hat{r} \cdot h')$. For any private history for the agent h_A such that full history (h, h_A) leads to this stage, we set $\hat{\beta}((h, h_A) \cdot \bar{r} \cdot (h', h'_A)) = \hat{\beta}((h, h_A) \cdot \hat{r} \cdot (h', h'_A)) = \beta^*((h, h_A) \cdot \hat{r} \cdot (h', h'_A))$ for all continuations (h', h'_A) . For any history that does not start with the public history h , we make no changes.

This construction simply changes the “interpretation” of cheap talk. The “meaning” of \bar{r} after public history h is not pinned down by equilibrium initially since it has zero probability, but the meaning of \hat{r} is identified in terms of its effects on equilibrium beliefs and continuation strategies. Essentially, this change has both players interpret \bar{r} after public history h the same way that they interpret \hat{r} after public history h .

It is easy to see that these changes do not change the equilibrium outcome. If the principal was sequentially rational on history $h \cdot \hat{r} \cdot h'$, he still is and is also sequentially rational on history $h \cdot \bar{r} \cdot h'$. The agent was originally sequentially rational on all histories and still is. We can iterate this construction to handle every stage at which

¹⁹To be clear, consistent simply means that the history is the right length.

the principal sends cheap talk.

Turning to the agent, fix any stage where the agent sends a cheap-talk message. Fix any public history h up to this stage. Let \hat{H}_A be the set of possible private histories of the agent up to this stage, that is, H_A minus histories that are the wrong length. Let \bar{r} be a particular cheap-talk message available to the agent at this stage and suppose that $\beta(h, h_A)(\bar{r}) = 0$ for all $h_A \in \hat{H}_A$. Fix any \hat{r} such that $\beta(h, h_A)(\hat{r}) > 0$ for some $h_A \in \hat{H}_A$. Change strategies as follows.

Let $\hat{\beta}(h, h_A)(\bar{r}) = \hat{\beta}(h, h_A)(\hat{r}) = \beta^*(h, h_A)(\hat{r})/2$ for all $h_A \in \hat{H}_A$. In other words, for private histories where the agent gives \hat{r} zero probability under β^* , we make no change. For private histories where the agent gives \hat{r} strictly positive probability, we divide this probability across \bar{r} and \hat{r} . Since \hat{r} has positive probability for some private history h_A , this ensures the desired property. For other cheap talk messages r , we have $\hat{\beta}(h, h_A)(r) = \beta^*(h, h_A)(r)$ for all consistent h_A .

As before, for any continuation public history h' , let $\hat{\gamma}(h \cdot \bar{r} \cdot h') = \hat{\gamma}(h \cdot \hat{r} \cdot h') = \gamma^*(h \cdot \hat{r} \cdot h')$. Again, we set $\hat{\rho}(h \cdot \bar{r} \cdot h') = \hat{\rho}(h \cdot \hat{r} \cdot h') = \rho^*(h \cdot \hat{r} \cdot h')$. We do the same for the agent's strategy, setting $\hat{\beta}((h, h_A) \cdot \bar{r} \cdot (h', h'_A)) = \hat{\beta}((h, h_A) \cdot \hat{r} \cdot (h', h'_A)) = \beta^*((h, h_A) \cdot \hat{r} \cdot (h', h'_A))$. For any history that doesn't start with the public history h , we make no changes.

As before, this does not change the equilibrium outcome and it leaves the agent sequentially rational at all information sets. In addition, it makes the principal sequentially rational at weakly more information sets than before. With abuse of notation, continue to let (β^*, γ^*) denote the Nash equilibrium strategies and ρ^* the principal's beliefs.

Summarizing to this point, we know that β^* satisfies sequential rationality for the agent at all information sets. By Nash, the principal is sequentially rational at all positive probability information sets. By the construction above, we've ensured that this covers all information sets where the principal has only observed cheap talk. Hence if there is any information set where some player is not sequentially rational, it must be that the principal's strategy γ^* is not sequentially rational at an information set where he has observed an evidence message and has to choose x .

So fix any such public history h . Let T^* be the set of types for whom h is feasible (that is, the types that can send the evidence message observed by the principal at h). Let $V(\beta^*, \gamma^* | t, h)$ denote the principal's expected utility at h when the strategies followed from h forward are (β^*, γ^*) and the agent's true type is t . (Note that t and h together determine the node of his information set that the principal is at.) Beliefs $\rho \in \Delta(T^*)$ make γ^* sequentially rational at this information set iff

$$\sum_{t \in T^*} \rho(t) \sum_{g \in G} \gamma^*(g) V(\beta^*, g | t, h) \geq \sum_{t \in T^*} \rho(t) \sum_{g \in G} \gamma(g) V(\beta^*, g | t, h)$$

for all $\gamma \in \Delta(G)$. If such a ρ exists, we can set the principal's beliefs at this information set to this ρ and we have sequential rationality at this information set.

So suppose no such ρ exists. By Lemma 1, γ^* is dominated with respect to T^* in the sense that there is some $\hat{\gamma} \in \Delta(G)$ such that

$$\sum_g \hat{\gamma}(g) V(\beta^*, g | t, h) > \sum_g \gamma^*(g) V(\beta^*, g | t, h), \quad \forall t \in T^* \quad (3)$$

and such that $\hat{\gamma}$ is not itself dominated in this sense. Since $\hat{\gamma}$ is not dominated in this sense, there exists $\hat{\rho} \in \Delta(T^*)$ such that $\hat{\gamma}$ maximizes the principal's expected utility. Set the principal's belief at this information set to equal $\hat{\rho}$ and change his strategy at this information set to $\hat{\gamma}(h)$. Call $(\beta^*, \hat{\gamma}^*, \hat{\mu}^*)$ the resulting assessment. Because we have only changed the principal's strategy at a last information set, one with zero probability, we know that $\hat{\gamma}^*$ is sequentially rational at every information set where γ^* was sequentially rational as well as at the information set h .

We now show that for any full history (h', h_A) with positive probability under (β^*, γ^*) (or, equivalently, under $(\beta^*, \hat{\gamma}^*)$), the agent is sequentially rational under $(\beta^*, \hat{\gamma}^*, \hat{\rho}^*)$. In other words, this change in the principal's strategy at history h does not lead the agent to wish to deviate from any on-path history. To see this, suppose not.

Let \hat{T} denote the (nonempty) set of t such that there is a full history of the form $(h', t \cdot h'_A)$ (i.e., the agent's type is t) such that β^* is not sequentially rational at $(h', t \cdot h'_A)$. It is easy to see that we must have $\hat{T} \subseteq T^*$ since no other type could play in such a way as to lead to information set h and hence no other type could be

affected by the change in the principal's strategy. Also, for all $t \in \hat{T}$,

$$\sum_g \hat{\gamma}(g)U(\beta^*, g | t, h) > \sum_g \gamma^*(g)U(\beta^*, g | t, h),$$

where $U(\beta^*, g | t, h)$ is the agent's expected utility from strategies (β^*, g) conditional on the agent's type being t and the history h . Equivalently, this is conditional on the node identified by (t, h) .

By equation (3) and the assumption that preferences are semi-aligned, we know that for all $t \in T^*$,

$$\sum_g \hat{\gamma}(g)\nu(t)U(\beta^*, g | t, h) > \sum_g \gamma^*(g)\nu(t)U(\beta^*, g | t, h),$$

so for any $t \in \hat{T}$, we must have $\nu(t) > 0$. Let $\hat{\beta}'$ denote the best reply of the agent which differs from $\hat{\beta}$ only in letting types $t \in \hat{T}$ deviate. By hypothesis, $\hat{T} \neq \emptyset$, so $\hat{\beta}' \neq \hat{\beta}$.

Note that

$$\begin{aligned} V(\hat{\beta}', \hat{\gamma}) &= \mathbb{E}_t[\nu(t)U(\hat{\beta}', \hat{\gamma}, t)] \\ &= \Pr[\nu(t) < 0]\mathbb{E}_t[\nu(t)U(\hat{\beta}, \gamma^*, t) | \nu(t) < 0] + \Pr[\nu(t) > 0]\mathbb{E}_t[\nu(t)U(\hat{\beta}', \hat{\gamma}, t) | \nu(t) > 0] \\ &> \Pr[\nu(t) < 0]\mathbb{E}_t[\nu(t)U(\hat{\beta}, \gamma^*, t) | \nu(t) < 0] + \Pr[\nu(t) > 0]\mathbb{E}_t[\nu(t)U(\hat{\beta}, \gamma^*, t) | \nu(t) > 0] \\ &= V(\hat{\beta}, \gamma^*) = V^*. \end{aligned}$$

The second equality uses the fact that only types in \hat{T} deviate and these all have $\nu(t) > 0$. The strict inequality comes from the fact that the types who deviate in response to $\hat{\gamma}$ are made strictly better off than they were at $(\hat{\beta}, \gamma^*)$.

But $\hat{\beta}'$ is a best reply to $\hat{\gamma}$, so this is not possible, by definition of V^* .

Summarizing, $(\beta^*, \hat{\gamma}^*, \hat{\rho}^*)$ has the property that β^* is sequentially rational for the agent at every full history with positive probability given $(\beta^*, \hat{\gamma}^*)$. We now show that this also holds at full histories with zero probability.

So suppose β^* is not sequentially rational at some full history (h', h_A) which has

zero probability under $(\beta^*, \hat{\gamma}^*)$. By construction, the public history h' must have positive probability, so it must be that h' is inconsistent with h_A . That is, it must be that some of the cheap talk messages in h' are not supposed to be sent given the private history h_A . Since this node in the tree (recall that the agent always knows everything) has zero probability, *every* node which is a successor to this one has zero probability as well. With this in mind, change the agent's strategy at this history to anything which is sequentially rational and call $\hat{\beta}^*$ the resulting behavior strategy for the agent. Because we are changing the agent's strategy only at a history which her own strategy prevents her from reaching, this does not affect the sequential rationality of the principal's strategy or the consistency of his beliefs. Hence proceeding this way, we can change the agent's strategy at such full histories as needed to ensure sequential rationality for the agent at all full histories without affecting sequential rationality for the principal or changing the equilibrium outcome.



Summarizing, we have shown that if there is any public history h where γ^* is not sequentially rational, we can adjust the strategies at this history and possibly others to ensure sequential rationality at h , at all histories for the agent, and at all positive probability histories for the principal without changing the equilibrium outcome. Hence we can construct a perfect Bayesian equilibrium with the same outcome as the Nash equilibrium (β^*, γ^*) .

References

- [1] Acharya, V., P. DeMarzo, and I. Kremer, “Endogenous Information Flows and the Clustering of Announcements,” *American Economic Review*, **101**, December 2011, 2955–2979.
- [2] Ball, I., and D. Kattwinkel, “Probabilistic Verification in Mechanism Design,” working paper, April 2023.
- [3] Ben-Porath, E., E. Dekel, and B. Lipman, “Mechanisms with Evidence: Commitment and Robustness,” *Econometrica*, **87**, March 2019, 529–566.
- [4] Ben-Porath, E., and B. Lipman, “Implementation and Partial Provability,” *Journal of Economic Theory*, **147**, September 2012, 1689–1724.
- [5] Blackwell, D., and M. Girshick, *Theory of Games and Statistical Decisions*, Wiley, 1954.
- [6] Bull, J., and J. Watson, “Hard Evidence and Mechanism Design,” *Games and Economic Behavior*, **58**, January 2007, 75–93.
- [7] Che, Y.-K., and N. Kartik, “Opinions as Incentives,” *Journal of Political Economy*, **117**, October 2009, 815–860.
- [8] Deb, R., M. Pai, and M. Said, “Evaluating Strategic Forecasters,” *American Economic Review*, **108**, October 2018, 3057–3103.
- [9] DeMarzo, P., I. Kremer, and A. Skrzypacz, “Test Design and Minimum Standards,” *American Economic Review*, **109**, June 2019, 2173–2207.
- [10] Deneckere, R. and S. Severinov, “Mechanism Design with Partial State Verifiability,” *Games and Economic Behavior*, **64**, November 2008, 487–513.
- [11] Dye, R. A., “Disclosure of Nonproprietary Information,” *Journal of Accounting Research*, **23**, 1985, 123–145.
- [12] Espinosa, F., (r) D. Ray, “Too Good To Be True? Retention Rules for Noisy Agents,” *American Economic Journal: Microeconomics*, **15**, May 2023, 493–535.

- [13] Felgenhauser, M., and E. Schulte, “Strategic Private Experimentation,” *American Economic Journal: Microeconomics*, **6**, November 2014, 74–105.
- [14] Gerardi, D., and R. Myerson, “Sequential Equilibria in Bayesian Games with Communication,” *Games and Economic Behavior*, **60**, July 2007, 104–134.
- [15] Glazer, J., and A. Rubinstein, “On Optimal Rules of Persuasion,” *Econometrica*, **72**, November 2004, 1715–1736.
- [16] Glazer, J., and A. Rubinstein, “A Study in the Pragmatics of Persuasion: A Game Theoretical Approach,” *Theoretical Economics*, **1**, December 2006, 395–410.
- [17] Green, J., and J.-J. Laffont, “Partially Verifiable Information and Mechanism Design,” *Review of Economic Studies*, **53**, July 1986, 447–456.
- [18] Grossman, S. J., “The Informational Role of Warranties and Private Disclosures about Product Quality,” *Journal of Law and Economics*, **24**, 1981, 461–483.
- [19] Guttman, I., I. Kremer, and A. Skrzypacz, “Not Only What but also When: A Theory of Dynamic Voluntary Disclosure,” *American Economic Review*, **104**, August 2014, 2400–2420.
- [20] Hart, S., I. Kremer, and M. Perry, “Evidence Games: Truth and Commitment,” *American Economic Review*, **107**, March 2017, 690–713.
- [21] Hedlund, J., “Bayesian Persuasion by a Privately Informed Sender,” *Journal of Economic Theory*, **167**, January 2017, 229–268.
- [22] Henry, E., and M. Ottaviani, “Research and the Approval Process: The Organization of Persuasion,” *American Economic Review*, **109**, March 2019, 911–955.
- [23] Kamenica, E., and M. Gentzkow, “Bayesian Persuasion,” *American Economic Review*, **101**, October 2011, 2590–2615.
- [24] Kartik, N., and O. Tercieux, “Implementation with Evidence,” *Theoretical Economics*, **7**, May 2012, 323–355.
- [25] Koessler, F., and V. Skreta, “Information Design by an Informed Designer,” working paper, January 2021.

- [26] Kosenko, A., “Noisy Bayesian Persuasion with Private Information,” working paper, July 2020.
- [27] Lipman, B., and D. Seppi, “Robust Inference in Communication Games with Partial Provability,” *Journal of Economic Theory*, **66**, August 1995, 370–405.
- [28] Matthews, S., and A. Postlewaite, “Quality Testing and Disclosure,” *RAND Journal of Economics*, **16**, Autumn 1985, 328–340.
- [29] McClellan, A., “Experimentation and Approval Mechanisms,” working paper, December 2020.
- [30] Milgrom, P., “Good News and Bad News: Representation Theorems and Applications,” *Bell Journal of Economics*, **12**, 1981, 350–391.
- [31] Perez-Richet, E., “Interim Bayesian Persuasion: First Steps,” *American Economic Review: Papers & Proceedings*, **104**, May 2014, 5.
- [32] Perez-Richet, E., and V. Skreta, “Test Design under Falsification,” working paper, current draft, January 2021.
- [33] Rappoport, D., “Evidence and Skepticism in Verifiable Disclosure Games,” Chicago Booth working paper, March 2020.
- [34] Sher, I., “Credibility and Determinism in a Game of Persuasion,” *Games and Economic Behavior*, **71**, March 2011, 409–419.
- [35] Shin, H. S., “Disclosures and Asset Returns,” *Econometrica*, **71**, January 2003, 105–133.
- [36] Shishkin, D., “Evidence Acquisition and Voluntary Disclosure,” working paper, December 2020.
- [37] Silva, F., “The Importance of Commitment Power in Games with Imperfect Evidence,” *American Economic Journal: Microeconomics*, **12**, November 2020, 99–1113.
- [38] Spence, M., “Job Market Signaling,” *Quarterly Journal of Economics*, **87**, August 1973, 355–374.

- [39] Sugaya, T., and A. Wolitzky, “The Revelation Principle in Multistage Games,” *Review of Economic Studies*, forthcoming, 2020.
- [40] Vohra, R.,  F. Espinosa  D. Ray, “A Principal–Agent Relationship with No Advantage to Commitment,” working paper, January 2021.