American Economic Journal: Microeconomics 2020, 12(4): 99–113
https://doi.org/10.1257/mic.20180038

The Importance of Commitment Power in Games with Imperfect Evidence†

By Francisco Silva*

The literature initiated by Green and Laffont (1986) studies principal-agent models with hard evidence. Evidence is modeled by assuming that the message set of the agent is type dependent. In this setup, Glazer and Rubinstein (2004, 2006) and Sher (2011) show that when the agent's utility function is type independent there is no advantage for the principal in having commitment power. This paper shows that this way of modeling evidence implicitly assumes it to be perfectly accurate and that the result that commitment power has no value is not robust to making the evidence imperfect. (*JEL* C70, D82)

In standard principal-agent models, not only is the principal never harmed by having commitment power, he usually strictly benefits from being able to make commitments. However, a strand of the literature initiated by Green and Laffont (1986) has identified a setting where the principal can do just as well without commitment power; a setting where the agent's preferences are independent of his private type and where the agent can provide "hard" evidence about his type—see Glazer and Rubinstein (2004, 2006) and Sher (2011). Following Glazer and Rubinstein (2004), I refer to these as *persuasion* games.[1] Examples of applications that could be modeled by a persuasion game that have been mentioned in the literature include: a defendant who, regardless of his guilt, wants to minimize his punishment; an expert who, regardless of the state of the world, wants the same policy to be implemented; a project manager who, regardless of the quality of the project, wants to maximize the amount of investment it gets, etc.

Identifying such a general setting in which commitment power does not seem to be particularly valuable is relevant because it undermines the importance of the instruments that allow the principal to commit—rules, laws, tradition, etc. In a

* Department of Economics, Pontificia Universidad Catolica de Chile, Vicuna Mackenna 4860, Piso 3, Macul, Santiago, Chile (email: franciscosilva@uc.cl). Johannes Hörner was coeditor for this article. I would like to thank Francesc Dilme, Mustafa Dogan, Nicolas Figueroa, Ju Hu, Jeanne LaFortune, Samir Mamadehussene, Steven Matthews, and Joaquin Poblete, as well as the participants at the Midwest Economic Theory Conference (Philadelphia), Spanish Economic Association meeting (Madrid), Portuguese Economic Journal meeting (Lisbon), Universidad Alberto Hurtado seminar, Universidad de Santiago seminar, and Pontificia Universidad Catolica (Chile) seminar for their useful comments.

† Go to https://doi.org/10.1257/mic.20180038 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

[1] Different authors have given different names to the game that is played between the principal and the agent. For example, Bull and Watson (2007) call it a "disclosure" game, while Hart, Kremer, and Perry (2017) call it an "evidence" game.

recent paper, Hart, Kremer, and Perry (2017) discuss this very issue. They present a refinement of sequential equilibria for persuasion games—what they call "truth-leaning equilibria"—and show that truth-leaning equilibria are always the principal's preferred sequential equilibria and are just as good for the principal as being able to commit. The following is a quote from them regarding the importance of commitment power:

> The second application concerns the judicial system. The system (the "principal") commits itself through constitutions, laws, legal doctrines, precedents … Our result says that the power of these commitments does not, however, go beyond selecting among all equilibria the truth-leaning equilibria, which are most natural in this setup. (Hart, Kremer, and Perry 2017)

The suggestion that is made is that constitutions, laws, legal doctrines, and precedents are much less important for the criminal system than one would think; they serve only to select the appropriate sequential equilibria. The purpose of this paper is to push back on this notion that commitment power is more or less irrelevant by showing that the result that commitment power does not help the principal is a product of implicitly modeling the evidence as "perfect." I do this by generalizing the persuasion game as modeled by Bull and Watson (2007) to include imperfect evidence. I find that, under the same general conditions described in Sher (2011), commitment power does matter for the principal. Furthermore, I show that the value of commitment power for the principal need not converge to zero as the evidence becomes perfect.

At first glance, it might not be obvious that persuasion games implicitly assume that the evidence is perfect. The following example, based on example 1 of Hart, Kremer, and Perry (2017), is particularly useful in that regard. Imagine that the agent is a job market candidate (JMC) and the principal is a university who must choose a salary $x \in \mathbb{R}$ for the JMC. The JMC's private type $\theta \in \{\theta_L, \theta_H\}$, with $\theta_L < \theta_H$, and his utility function is just $u(x) = x$ (so, it is type independent). The university's utility function is $v(\theta, x) = -(\theta - x)^2$—the university wants to match the salary to the type of the JMC.[2]

In this context, the evidence that is typically provided by the JMC is his job market paper (JMP). In persuasion games, evidence is modeled by assuming that the agent's choice set is type dependent. Specifically, the game is as follows: the agent, privately informed of his type θ, chooses a message m from his type dependent message set $M(\theta)$; the principal observes m and chooses x. It is the fact that the message set $M(\theta)$ is type dependent that generates hard evidence: if some type θ of the agent sends some message that does not belong to the message set of some other type θ', type θ is providing evidence that he is not type θ'. Seeing as, presumably, high-type JMCs are better in some sense than low-type JMCs in writing JMPs, one would model this interaction by assuming that there is some message d,

[2] In example 1 of Hart, Kremer, and Perry (2017), the agent is a university professor rather than a JMC.

interpreted as being a "good JMP," that is only available to the high-type JMC—i.e., $d \in M(\theta_H)$ but $d \notin M(\theta_L)$.

It is then easy to see that, in such a model, commitment power does not matter for the university: the university can simply follow the policy that it pays a salary of θ_H if a good JMP is provided (i.e., if d is sent) and a salary of θ_L otherwise. Such a policy induces the high-type JMC to provide a good JMP, which leads to the first best outcome for the university without it needing commitment power. I say that persuasion games implicitly assume that the evidence is perfect because there is perfect correlation between the agent's type and his ability to produce a good JMP: the high type is certain to be able to produce a good JMP, while the low type is certain to be unable to produce a good JMP. But, I would argue, it would be more reasonable to think that high-type JMCs, while more likely to write good JMPs than low-type JMCs, are not guaranteed to do so; when they write their JMP, they are not certain that their JMP will be considered good.

The same type of argument could be made about most of the applications that have served as motivation to study persuasion games. In particular, for the case of the defendant and the criminal justice system that Hart, Kremer, and Perry (2017) discuss, it seems clear that there is doubt at the time of the verdict; in fact, the threshold that is used in American criminal law is that of "reasonable" doubt, an acknowledgment that doubt is likely to persist until even after the trial.

Assuming the evidence is imperfect makes commitment power matter because of uncertainty: there is only uncertainty if the evidence is imperfect. I show in the text that as long as the principal is somewhat risk averse, she will attempt to reduce the risk she faces in a way that can only be done if she can commit.

This paper is most closely related to the literature on principal-agent games with evidence (Green and Laffont 1986, Deneckere and Severinov 2008, Strausz 2016). Within this literature, special attention has been given to persuasion games: Glazer and Rubinstein (2004, 2006) show that commitment power has no value for the principal when the principal's choice set is binary; Sher (2011) shows that this result holds even if the principal's choice set is not binary provided the principal's utility function is a concavification of the agent's utility function; Hart, Kremer, and Perry (2017) propose a refinement of sequential equilibria that is always the principal's preferred equilibria. The paper also builds on the analysis of Silva (2019) and Siegel and Strulovici (2019), who explicitly consider an environment with a defendant and a benevolent principal acting as a representative of the criminal justice system.[3]

The paper is organized as follows. In Section I, I describe the model; in Section II, I present and discuss the main results; in Section III, I conclude. All proofs are in the online Appendix.

[3] I will discuss the connection with these two papers after presenting the results.

I. Model

A. *Fundamentals*

There are two players: the principal (she) and the agent (he). Players' preferences depend on the choice $x \in \mathbb{R}$ made by the principal. The agent's utility function is denoted by $u : \mathbb{R} \to \mathbb{R}$, so that it only depends on choice x. I assume that u is differentiable and strictly increasing: x is the agent's reward. As for the principal, her preferences are assumed to depend not only on x but also on the private type of the agent, which is denoted by $\theta \in \Theta \subset \mathbb{R}$. Set Θ is assumed to be finite and the prior probability of $\theta \in \Theta$ is denoted by $q(\theta)$. The principal's utility function is denoted by $v : \Theta \times \mathbb{R} \to \mathbb{R}$, so that $v(\theta,x)$ represents the utility of the principal when the agent's type is θ and the choice made is x.

Following Sher (2011), I assume that there is a differentiable function $g : \Theta \times \mathbb{R} \to \mathbb{R}$ such that $v(\theta,x) = g(\theta,u(x))$ for all $(\theta,x) \in \Theta \times \mathbb{R}$ and with the property that $g(\theta,\cdot)$ is concave for all $\theta \in \Theta$. This assumption implies that if the agent is indifferent between two lotteries of rewards, the principal will prefer the one with less variance. Regardless of how "realistic" this assumption may be, it is crucial to follow Sher (2011) in making it because this is the condition that is used to show the result that commitment power does not matter in persuasion games. In fact, as Sher (2011) shows, it is possible to find counterexamples to the result if the condition fails.

For each $\theta \in \Theta$, $v(\theta,\cdot)$ is assumed to have a unique maximizer denoted by $x^*(\theta)$, which without loss of generality is assumed to be nondecreasing. Finally, to ensure the existence of optimal mechanisms, I assume that for all $\theta \in \Theta$,

$$\lim_{x \to \pm\infty} v(\theta,x) = -\infty.$$

A simple example of a function v that has the assumed properties that the reader might want to keep in mind is the following:

$$v(\theta,x) = -(\theta - u(x))^2.$$

The idea behind these preferences is to have the principal wanting to give larger rewards to the larger types of the agent.

B. *Persuasion Games*

Let me start by describing in more detail how evidence has been modeled in the persuasion game literature. I follow Bull and Watson (2007) in my description. Hard evidence is modeled as a finite and type dependent set E^θ that is available to the agent. Bull and Watson mention this set as being a set of documents: the agent has access to some documents for some types and some other documents for some other types. The persuasion game (what they call "disclosure game") is as follows. First, the agent simultaneously chooses a document $d \in E^\theta$ and a (cheap-talk) message $m \in M$.

Following the tradition of cheap-talk games, M is some arbitrarily large and type independent set. After observing (d,m), the principal chooses reward $x \in \mathbb{R}$.

From this description, one can identify two properties of persuasion games that I want to preserve: (i) the agent has the power to choose what evidence to present to the principal, and (ii) the ability to present certain evidence depends on the agent's type. As will become clear, my model preserves these two properties while allowing for imperfect evidence.

C. *The Signaling Game*

The signaling game that the agent and the principal play in my model can be divided into three periods.

Period 1: The agent simultaneously chooses three things: a document $d \in D$, where $D \equiv \bigcup_{\theta\in\Theta} E^{\theta}$ denotes what would be the total set of documents in Bull and Watson (2007); a cheap-talk message $m \in M$, where M represents the set of cheap-talk messages; and a private action $a \in \{0,1\}$ (i.e., an action that will not be observed by the principal). I assume that M is arbitrarily large; in particular, I assume that $\Theta \subseteq M$.

Period 2: A public random variable $s \in \{0,1\}$ is realized. The distribution of s depends on the agent's type θ, on the document d he chose to present and on action a. In particular, I assume that for all $(\theta,d,a) \in \Theta \times D \times \{0,1\}$,

$$\Pr\{s = 1|\theta,d,a\} = ap(\theta,d)$$

for some $p(\theta,d) \in [0,1]$. Put differently, if $a = 0$, then $s = 0$ for sure, while if $a = 1$, the probability that $s = 1$ is given by $p(\theta,d)$. Function p is also called an "information structure."

Period 3: The principal observes document d, cheap-talk message m, and random variable s and then chooses an outcome $x = f(d,m,s)$.

There are a few substantial differences to Bull and Watson (2007). First, the agent is able to choose any document regardless of his type, because D is type independent. In the JMC example, if $d \in D$ is the JMP, this means that any type can write a JMP by choosing d. Even though any type can choose any document, not every type is as likely to generate $s = 1$, because $p(\theta,d)$ depends on θ. So signal s works as a measure of the quality of the document; if $s = 1$ the document has good quality, and if $s = 0$ it has bad quality. In the example, everyone can write a JMP but the likelihood that the JMP is good ($s = 1$) depends on the agent's type. Function p also depends on d, because the likelihood that the document is of high quality naturally depends on the document that is being presented. As for the principal, she gets to condition her choice of a reward not only on the document d that is presented by the agent but also on s, the signal about the document's quality (in addition to the cheap-talk message m). Finally, adding variable a and making it private allows the agent to generate a document of bad quality with certainty without letting the principal know that he did so on purpose. I postpone the discussion over the need to model decision variable a to Section IIA.[4]

D. *Definitions*

Following Sher (2011), mapping $f: D \times M \times \{0,1\} \rightarrow \mathbb{R}$ is called a persuasion rule. If the principal has commitment power, she commits to a persuasion rule f at the outset before the agent chooses $(d,m,a) \in D \times M \times \{0,1\}$. A strategy for the agent is a function $\sigma : \Theta \rightarrow \Delta(D \times M \times \{0,1\})$. Strategy σ is a "best response" to f if and only if, for all $\theta \in \Theta$ and $(d,m,a) \in D \times M \times \{0,1\}$,

$$\sigma(\theta)(d,m,a) > 0 \Rightarrow ap(\theta,d)u(f(d,m,1)) + (1 - ap(\theta,d))u(f(d,m,0))$$
$$\geq a'p(\theta,d')u(f(d',m',1)) + (1 - a'p(\theta,d'))u(f(d',m',0))$$

for all $(d',m',a') \in D \times M \times \{0,1\}$. A profile (f,σ) is *feasible* if σ is a best response to f.

The principal's expected utility under profile (f,σ) is denoted by $V(f,\sigma)$. A profile (f,σ) is *optimal* if and only if it is feasible and there is no other feasible profile (f',σ') for which $V(f,\sigma) < V(f',\sigma')$.[5] Finally, a profile (f,σ) is *credible* if and only if it is feasible and, for all $(d,m,s) \in D \times M \times \{0,1\}$,

$$f(d,m,s) \in \underset{x \in \mathbb{R}}{\operatorname{argmax}} E^{\sigma}(v(\theta,x) | d,m,s),$$

where $E^{\sigma}(\cdot)$ represents the expectation formed by the principal, knowing that the agent's strategy is given by σ. Put differently, (f,σ) is credible if and only if it is a (weak) perfect Bayesian equilibrium (PBE) of the signalling game.

E. *Modified Revelation Principle*

There is a preliminary result regarding optimal persuasion rules that is quite convenient for later—a sort of modified revelation principle.

LEMMA 1: *For every feasible profile (f,σ) there is another feasible profile (f',σ') such that $V(f,\sigma) \leq V(f',\sigma')$, where σ' is such that the agent does not randomize, reports truthfully, and chooses $a = 1$, i.e., for all $\theta \in \Theta$, the agent chooses $(d,m,a) = (\hat{d}(\theta),\theta,1)$ for some mapping $\hat{d} : \Theta \rightarrow D$.*

[4] After completing this article, I became aware of Ball and Kattwinkel (2019), who study a similar model of imperfect evidence. However, they use the model to study very different problems of revenue maximization and do not discuss issues of commitment power. Both papers were completed independently.

[5] In principle, it would be possible for the principal to commit to randomizing over the space of rewards for each (d,m,s) observed. However, because of the assumption that $g(\theta,\cdot)$ is concave for each $\theta \in \Theta$, one can show by the same argument as in Theorem 2 of Sher (2011) that it is optimal for the principal not to randomize.

Lemma 1 states that when looking for optimal profiles, one need only focus on those for which the agent does not randomize, truthfully announces his type, and attempts to produce good-quality evidence.[6]

II. Discussion

A. *Perfect Evidence*

I start the discussion by considering the special case where $p(\theta,d) \in \{0,1\}$ for all $(\theta,d) \in \Theta \times D$. Under these assumptions, my model and the standard persuasion game are fundamentally the same in that there is no uncertainty: when a certain type picks a certain document, he knows whether the following s will be 1 or 0. Therefore, it is no surprise that, also in my model, commitment power does not matter. I say that the evidence is *perfect* if and only if, for all $\theta \in \Theta$, (i) $p(\theta,d) \in \{0,1\}$ for all $d \in D$ and (ii) there is some d for which $p(\theta,d) = 1$.

PROPOSITION 1: *If the evidence is perfect, there is a credible profile* (f,σ) *that is optimal.*

The idea of the argument is to use the result from Sher (2011) about persuasion games to show that the same holds in my model. In order to use Sher (2011), one has to establish an equivalence between persuasion games and my model; for that purpose, modeling the decision variable a is essential, as the following example illustrates.

Example 1: Suppose the JMC and the university play a standard persuasion game. The JMC's utility function is $u(x) = x$ and the university's utility function is $v(\theta,x) = -(\theta - x)^2$. Assume that there are two documents that can be presented by the JMC: a good JMP labeled d_1 and a good essay on the Spanish Civil War labeled d_2. For argument's sake, assume that $E^{\theta_L} = \{d_1,d_2\}$, while $E^{\theta_H} = \{d_1\}$, so that the high type is not able to write a good essay on the Spanish Civil War, but the low type can. For completeness, suppose that $M = \{\theta_L,\theta_H\}$.

What is the best the university can do? The university would like to reward type θ_H more than type θ_L, but it cannot because anything that the high type can do, the low type can also do. So, regardless of commitment power, the best the university can do is to give each type a reward $x = E(\theta)$.

Now, let us consider how the same example would look like in my model. Set $D = \{d_1,d_2\}$, where d_1 is interpreted as a JMP (though not necessarily a good one) and d_2 as an essay on the Spanish Civil War (again, not necessarily a good one). The information structure is such that $p(\theta_H,d_1) = p(\theta_L,d_1) = p(\theta_L,d_2) = 1$, while $p(\theta_H,d_2) = 0$, so that the high type is able to write a good JMP but is not able to write a good essay, while the bad type is able to write both a good JMP and a good essay. Suppose that the decision variable a was not modeled (or equivalently, suppose that the agent had no other option than to choose $a = 1$).

[6]Notice that Lemma 1 does not say anything about optimal *credible* profiles, only about optimal ones. In general, there are no optimal credible profiles with the properties described in Lemma 1.

In that case, the university could actually do better than in the original persuasion game; it could implement the following policy:

$$f(d,m,s) = \begin{cases} \theta_H & \text{if } (d,m,s) = (d_2,\theta_H,0) \\ \theta_L & \text{otherwise,} \end{cases}$$

which would induce both types to choose $(d,m) = (d_2,\theta_H)$. In words, the university ignores the JMP (because it is not informative) and simply evaluates the essay that the agent presents. However, because the ability to write the essay is negatively correlated with the agent's type and the agent is not allowed to write a bad essay $(a = 1)$, it is best for the university to reward the agent if he writes a bad essay. In addition to this outcome being unintuitive, it does not match the outcome that arises from the original persuasion game; in this version of the model the university would achieve its first best.

A similar outcome could be achieved if decision variable a was modeled but assumed to be public. In that case, the university could follow a policy in which it only rewards the agent in θ_H if $a = 1$ and $(d,m,s) = (d_2,\theta_H,0)$, essentially forcing the agent to attempt to write a good essay and rewarding him only if he fails.

By contrast, if decision variable a is modeled as being private, the university is unable to reward the high type more than the low type just like in the original persuasion game. The only way to distinguish between the two types would be to make them write an essay but, in that event, the low type would simply write a bad essay on purpose without being caught.

The only nontrivial aspect of the argument of Proposition 1 has to do with the fact that even though persuasion games and this paper's game are fundamentally alike once one models decision variable a, they are not technically the same, because in my model, the agent has more choices available—remember that each type θ can choose to send any document he wants, including some document d for which $p(\theta,d) = 0$. However, I show in the online Appendix that one can restrict attention to persuasion rules where that does not happen—i.e., without loss of generality, one can focus on persuasion rules for which each type θ always reports some document d for which $p(\theta,d) = 1$. Given that, the extra freedom the agent has ends up not mattering so that the result from Sher (2011) holds. Assumption (ii), which states that for every type θ there is some document d for which $p(\theta,d) = 1$, simply guarantees that in the persuasion game counterpart of my game, every type's message set is nonempty.

B. *Imperfect Evidence*

The main purpose of the paper is to study whether the result that commitment power does not matter holds when there is imperfect evidence. I say that a profile (f,σ) is *constant* if there is $k \in \mathbb{R}$ such that $f(d,m,s) = k$ for all $(d,m,s) \in D \times M \times \{0,1\}$.

PROPOSITION 2: *If* $p(\theta,d) \in (0,1)$ *for all* $(\theta,d) \in \Theta \times D$, $g(\theta,\cdot)$ *is strictly concave for all* $\theta \in \Theta$ *and there is no optimal profile* (f,σ) *that is constant*, *then there is no optimal profile that is credible.*

Proposition 2 states that, under essentially the same conditions as in Sher (2011), commitment power matters as long as one assumes that the evidence is imperfect. The one difference to Sher (2011) is that $g(\theta,\cdot)$ is assumed to be strictly concave for all $\theta \in \Theta$ rather than just concave, so that the principal strictly benefits from reducing the variance of the agent's lottery of rewards. I would note, however, that assuming that $g(\theta,\cdot)$ is strictly concave for all $\theta \in \Theta$ is a sufficient but not necessary condition for commitment power to matter; i.e., one can build examples where $g(\theta,\cdot)$ is not strictly concave for all $\theta \in \Theta$ where commitment power matters. I would also note that the condition that there is no optimal profile that is constant is necessary—if there is a constant profile that is optimal, it is trivially credible through a "babbling equilibrium." Nevertheless, in general, constant profiles are not optimal.

While the complete proof of the result is provided in the online Appendix, I provide an outline of the proof below.

PROOF (OUTLINE):

Consider an optimal profile and let $\lambda \in \mathbb{R}$ and $\eta \in \mathbb{R}$ denote the lowest and second-lowest expected utilities attained by some types of the agent (so $\eta > \lambda$). Let sets Θ^λ and Θ^η denote the sets of types that have a utility of λ and η, respectively, given that optimal profile. There are three steps to the proof.

Step 1: It must be that the lottery of rewards each type $\theta \in \Theta^\lambda$ is given is constant. This follows because no type wants to mimic the type with the lowest expected utility, which allows the principal to reduce the risk of the types in set Θ^λ without it affecting the incentives of the agent. Naturally, this means that types outside of set Θ^λ receive a lottery $(y_1(\theta), y_0(\theta))$, where $y_1(\theta) > u^{-1}(\lambda) > y_0(\theta)$; the agent receives a reward of y_1 if and only if, following his choice of document, $s = 1$.

Step 2: If the optimal profile is credible, it must be that

$$u^{-1}(\lambda) = \underset{x\in\mathbb{R}}{\operatorname{argmax}}\, E\big(v(\theta,x) | \theta \in \Theta^\lambda\big).$$

The idea of the argument is that the only types that get a constant reward are those in set Θ^λ. Therefore, for the principal to allocate that constant reward to those types, she must be able to distinguish them from all other types. If she cannot commit, that constant reward must be the one that maximizes her expected utility given that the agent's type is in Θ^λ.

Step 3: To find a contradiction, one must perturb the optimal profile to find another feasible profile that is strictly preferred by the principal. The perturbation is as follows. Pick some type $\hat{\theta} \in \Theta^\eta$, who is getting some lottery $\big(y_1(\hat{\theta}), y_0(\hat{\theta})\big)$. Perturb that lottery as follows: if $s = 1$, then the agent's reward is $y_1(\hat{\theta}) - \varepsilon$, while if $s = 0$, then the reward is $y_0(\hat{\theta}) + \delta(\varepsilon)$, where for each small $\varepsilon > 0$, $\delta(\varepsilon) > 0$ is such that type $\hat{\theta}$ is indifferent between the two lotteries. If ε is small enough, the principal is better off by this change because she decreases type $\hat{\theta}$'s risk. Furthermore, if ε is small enough, no type outside of set Θ^λ will want to mimic type $\hat{\theta}$. As for types in set Θ^λ, notice that if there is no type in set Θ^λ that was indifferent to mimicking type $\hat{\theta}$ before the perturbation, then it follows that if ε is small enough, the perturbed profile will be feasible and strictly better for the principal—a contradiction. Instead, take the other possibility. In that case, one has to perturb the profile further to achieve a feasible profile. In particular, say that one increases the constant reward of all types in set Θ^λ by $\xi(\varepsilon)$, where $\xi(\varepsilon) > 0$ is the smallest value such that no type in set Θ^λ wants to mimic type $\hat{\theta}$. It follows that if ε is sufficiently small, the perturbed profile is feasible. What is less clear is whether it is preferred by the principal to the unperturbed profile; the principal gains on type $\hat{\theta}$ but loses on set Θ^λ, because their (constant) reward is larger than what the principal would prefer. The proof is completed by showing that when ε is sufficiently small, what the principal gains from type $\hat{\theta}$ more than compensates what she loses on set Θ^λ, which is a contradiction to the optimality of a credible profile. ∎

To give some additional intuition on the result, let us go back to the example of the JMC from before.

Example 2: Consider the same example from before, with the same two types and the same preferences but, for simplicity, assume that $D = \{d_1\}$, where d_1 represents the JMP. Assume also that $1 > p(\theta_H, d_1) > p(\theta_L, d_1) > 0$, so that the high type is more likely to write a good JMP than the low type. For simplicity, assume that $q(\theta_H) = q(\theta_L) = 1/2$ (each type is equally likely). One can show that the principal's preferred credible profile (f^c, σ^c) is as follows:

$$f^c(d_1, m, s) = \begin{cases} x^h & \text{if } (m,s) = (\theta_H, 1) \\ x^l & \text{if } (m,s) = (\theta_H, 0), \\ \theta_L & \text{if } m = \theta_L \end{cases}$$

where

$$x^h \equiv \frac{p(\theta_H, d_1)\,\theta_H + p(\theta_L, d_1)\,\theta_L}{p(\theta_H, d_1) + p(\theta_L, d_1)}$$

and

$$x^l \equiv \frac{\left(1 - p(\theta_H, d_1)\right)\theta_H + \left(1 - p(\theta_L, d_1)\right)\theta_L}{2 - p(\theta_H, d_1) - p(\theta_L, d_1)},$$

while

$$\sigma^c(\theta)(d, m, a) = \begin{cases} 1 & \text{if } (d, m, a) = (d_1, \theta_H, 1) \\ 0 & \text{otherwise,} \end{cases}$$

for any $\theta \in \{\theta_L, \theta_H\}$. The agent chooses $(d, m, a) = (d_1, \theta_H, 1)$ for any type $\theta \in \{\theta_L, \theta_H\}$, because $x^h > x^l > \theta_L$. The profile is credible because x^h is the sequentially optimal reward for the university after observing $(d, m, s) = (d_1, \theta_H, 1)$ and x^l is the sequentially optimal reward for the university after observing $(d, m, s) = (d_1, \theta_H, 0)$. Essentially, the university just asks for the JMP, assesses its quality, and then chooses the reward that it finds optimal after updating its beliefs about the JMC's type. It is also easy to see that when $p(\theta_H, d_1) = 1$ and $p(\theta_L, d_1) = 0$, the profile is actually first best for the university, so that it is optimal. However, if $1 > p(\theta_H, d_1) > p(\theta_L, d_1) > 0$, profile (f^c, σ^c) is not optimal, as I show below.

Consider the following alternative profile (f', σ'), where

$$f'(d_1, m, s) = \begin{cases} x^h & \text{if } (m, s) = (\theta_H, 1) \\ x^l & \text{if } (m, s) = (\theta_H, 0), \\ \hat{x} & \text{if } m = \theta_L \end{cases}$$

where

$$\hat{x} \equiv p(\theta_L, d_1)\, x^h + \big(1 - p(\theta_L, d_1)\big)\, x^l,$$

and

$$\sigma'(\theta)(d, m, a) = \begin{cases} 1 & \text{if } (d, m, a) = (d_1, \theta_H, 1) \text{ and } \theta = \theta_H \\ 1 & \text{if } (d, m, a) = (d_1, \theta_L, 0) \text{ and } \theta = \theta_L \\ 0 & \text{otherwise.} \end{cases}$$

In other words, the university's policy is the following: if the JMC claims to be the high type, the university will look at his JMP and will give him a reward of x^h if it is good, and of x^l if it is bad. If, however, the JMC admits to being the low type, the university will give him a reward of $\hat{x} \in (x^l, x^h)$ regardless of the quality of his JMP. By design, the low type is indifferent between the lottery of x^h and x^l and $\hat{x}$, while the high type prefers the former.

Notice that the lottery of rewards that the high type receives is the same under both profiles. However, under profile (f', σ'), the low type receives a constant reward that leaves him indifferent to the lottery of rewards that he receives under profile (f^c, σ^c). Seeing as the principal strictly prefers the agent to take less risk whenever he is indifferent, profile (f', σ') is preferred to profile (f^c, σ^c) by the principal. But profile (f', σ') is not credible: there is complete separation on the reporting stage by the two types, with each type truthfully reporting their type, which would induce the university to choose a reward of θ_H after observing $(d, m) = (d_1, \theta_H)$ and a reward of θ_L after observing $(d, m) = (d_1, \theta_L)$.[7]

[7] In the example (and in general), it is key that when the agent chooses what to report he does not know the quality of his JMP (i.e., he does not know whether $s = 1$ or $s = 0$). If he did, the game would be equivalent to a standard persuasion game. There would be four types (θ, s): $(\theta_H, 1)$, $(\theta_L, 1)$, $(\theta_H, 0)$, and $(\theta_L, 0)$, and only the first two would be able to generate a high-quality JMP. In that case, the principal would simply select rewards based on the quality of the JMP and would not require commitment power.

The previous example is also helpful in that it allows for a natural comparison to the judicial system application that is discussed by Hart, Kremer, and Perry (2017) mentioned in the introduction. Instead of a JMC who is trying to convince the university that his type is high, imagine that there is a defendant who is trying to convince the judicial system that he is innocent. In this context, the profile (f', σ') from the example looks a lot like "plea bargaining": the defendant is given the opportunity to forgo the risk of going to trial, which could have lead to an acquittal or a large punishment, in exchange for a reduced punishment following a confession of guilt. The institution of plea bargaining most definitively requires commitment power. The rights of confessing defendants must be protected; otherwise, the judicial system would be tempted to punish them more harshly after learning of their guilt.

C. *Almost-Perfect Evidence*

The last issue I address has to do with what happens when there is almost-perfect evidence—i.e., when $p(\theta, d)$ is close to either 0 or 1 for all $(\theta, d) \in \Theta \times D$. In particular, imagine that one takes a sequence of functions p that converges to some perfect evidence function $\hat{p}$. Must the value of commitment power converge to zero? In this section, I show that if the agent's utility function is unbounded, then the answer is "no."

For each information structure $p : \Theta \times D \rightarrow [0,1]$, let V_p^O denote the expected utility of the principal of an optimal profile under information structure p (O stands for "optimal") and let V_p^{OC} denote the expected utility of the principal of her preferred credible profile under information structure p (OC stands for "optimal credible"). Let also set P denote the set of information structures for which there is perfect evidence, i.e., for all $p \in P$, for all $\theta \in \Theta$, $p(\theta, d) \in \{0,1\}$ for all $d \in D$ and there is $d \in D$ such that $p(\theta, d) = 1$.

PROPOSITION 3: *If* $\lim_{x \rightarrow -\infty} u(x) = -\infty$ *and function* $g(\theta, \cdot)$ *is strictly concave for all* $\theta \in \Theta$, *then for any* D *and* $\hat{p} \in P$ *for which* $V_{\hat{p}}^O < E\big(v\big(\theta, x^*(\theta)\big)\big)$, *there are sequences* $\{p^t\} \rightarrow \hat{p}$ *such that* $\{V_{p^t}^O - V_{p^t}^{OC}\} \nrightarrow 0$.

Proposition 3 states sufficient conditions for there to be a sequence of information structures converging to perfect evidence for which the value of commitment power does not converge to zero. The condition that $V_{\hat{p}}^O < E\big(v\big(\theta, x^*(\theta)\big)\big)$ simply states that the "first best" for the principal is not attainable when $p = \hat{p}$, which is a feature of all nontrivial problems. The role of the first condition—that the agent's utility function is unbounded—is better understood by going back to the example.

Example 3: Consider the same example as before with $D = \{d_1\}$ and consider information structure p^τ, where

$$p^\tau(\theta, d_1) = \begin{cases} 1 & \text{if } \theta = \theta_H \\ \tau & \text{if } \theta = \theta_L \end{cases}$$

for some $\tau \in (0,1)$. Notice that $\lim_{\tau \to 1} p^{\tau} = \hat{p}$, where $\hat{p} \in P$ is such that $\hat{p}(\theta, d_1) = 1$ for $\theta \in \{\theta_L, \theta_H\}$, i.e., $\hat{p}$ is an information structure for which there is perfect evidence. By Proposition 1, $V_{\hat{p}}^{O} = V_{\hat{p}}^{OC}$, i.e., commitment power has no value if there is perfect evidence. In this example, it is easy to see that if $\tau = 1$, it is impossible for the principal to separate between the two types with or without commitment power, so that

$$V_{\hat{p}}^{O} = V_{\hat{p}}^{OC} = -\frac{\left(\theta_H - \frac{\theta_H + \theta_L}{2}\right)^2 + \left(\theta_L - \frac{\theta_H + \theta_L}{2}\right)^2}{2} = -\left(\frac{1}{2}\theta_H - \frac{1}{2}\theta_L\right)^2.$$

Let us now consider the case where $\tau < 1$. If the principal cannot commit, the best she can do is to use the JMP as a way to differentiate between the two types, just like in the previous section. In particular, for each $\tau < 1$, one can show that the optimal credible profile is (f^c, σ^c). For each τ, it follows that

$$x^h = \frac{\theta_H + \tau\theta_L}{1 + \tau} \quad \text{and} \quad x^l = \theta_L$$

so that

$$V_{p^{\tau}}^{OC} = -\frac{\left(\theta_H - \frac{\theta_H + \tau\theta_L}{1+\tau}\right)^2 + \tau\left(\theta_L - \frac{\theta_H + \tau\theta_L}{1+\tau}\right)^2}{2} \underset{\tau \to 1}{\longrightarrow} V_{\hat{p}}^{OC}.$$

If, however, the principal can commit, then she can actually achieve her first best for any $\tau \in (0,1)$ through the following feasible profile (f^{τ}, σ''), where

$$f^{\tau}(d_1, m, s) = \begin{cases} \theta_H & \text{if } (m,s) = (\theta_H, 1) \\ \underline{x}(\tau) & \text{if } (m,s) = (\theta_H, 0) \\ \theta_L & \text{if } m = \theta_L \end{cases},$$

where $\underline{x}(\tau)$ is small enough that

$$\tau\theta_H + (1-\tau)\underline{x}(\tau) < \theta_L,$$

and

$$\sigma''(\theta)(d_1, m, a) = \begin{cases} 1 & \text{if } (d_1, m, a) = (d_1, \theta_H, 1) \text{ and } \theta = \theta_H \\ 1 & \text{if } (d_1, m, a) = (d_1, \theta_L, 1) \text{ and } \theta = \theta_L \\ 0 & \text{otherwise} \end{cases}.$$

In other words, if the JMC claims to being the high type, he receives a reward of θ_H if his JMP is good and of $\underline{x}(\tau)$ if it is bad. Seeing as only the low-type JMC is likely to write a bad JMP, $\underline{x}(\tau)$ can be chosen to be arbitrarily small to prevent the low type from mimicking the high type—the assumption that the agent's utility function is unbounded guarantees that, for all τ, such an $\underline{x}(\tau)$ exists. Seeing as the low type has no incentives to mimic the high type, the principal can then give him a reward of θ_L in the event that he admits to being the low type. Therefore, it is clear that for any $\tau < 1$, $V^O_{p^\tau} = 0 \nrightarrow V^O_{\hat{p}}$, so that $V^O_{p^\tau} - V^{OC}_{p^\tau} \nrightarrow 0$, i.e., the value of commitment power does not converge to zero.[8]

III. Conclusion

In this paper, I show that the result that commitment power does not matter crucially depends on the evidence being perfect because it guarantees that there is no uncertainty for either player. I also show that the value of commitment power need not vanish as the evidence becomes perfect. This is an important result as it shows that traditions and rules of law have an importance beyond equilibrium selection: they allow the principal to commit, which makes her strictly better off.

Throughout, I use the example of a defendant to illustrate the logic behind these results as a way to contrast my work with that of Hart, Kremer, and Perry (2017), who use the same application. This paper is also related to Silva (2019) and Siegel and Strulovici (2019), who explicitly consider an environment with a defendant and a benevolent principal acting as a representative of the criminal justice system but assume (i) that the evidence is public and (ii) a simple information structure (only two types and the equivalent to one document, very much like in the example).[9] While in the example where there is only one document ($D = \{d_1\}$, where d_1 was the JMP), the fact that the evidence is public or private does not make that much of a difference, provided the high type is more likely to write a good JMP than the low type (because, in the principal's preferred equilibrium, both types attempt to write a good JMP, i.e., both types choose $a = 1$), the example with $D = \{d_1, d_2\}$ shows that is not always the case, so that the difference between private and public evidence is not always trivial.[10] Nevertheless, it can be trivially shown (by slightly altering the proof of Proposition 2) that the result that commitment power matters when the evidence is imperfect (Proposition 2) also holds when the evidence is public.

REFERENCES

Ball, Ian, and Deniz Kattwinkel. 2019. "Probabilistic Verification in Mechanism Design." Unpublished.

Bull, Jesse, and Joel Watson. 2007. "Hard Evidence and Mechanism Design." *Games and Economic Behavior* 58 (1): 75–93.

Crémer, Jacques, and Richard P. McLean. 1988. "Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions." *Econometrica* 56 (6): 1247–57.

Deneckere, Raymond, and Sergei Severinov. 2008. "Mechanism Design with Partial State Verifiability." *Games and Economic Behavior* 64 (2): 487–513.

[8] The argument is similar to Crémer and McLean (1988): the discontinuity arises because the small difference between perfect evidence and almost-perfect evidence gets magnified by the possibility of having arbitrarily small rewards, just like the small difference between independent types and almost-independent types got magnified by the possibility of having arbitrarily large transfers.

[9] The model they consider is as follows: the defendant starts by announcing whether they confess to being guilty; if they refuse to confess, the principal will exogenously gain access to a (one-dimensional) signal—this paper's signal s—that is correlated with the defendant's guilt; after observing the signal, the principal allocates a punishment.

[10] In the particular application of the defendant, it is unclear which of the two assumptions would be more realistic—the evidence being private or public. I suspect that a combination of both would be true—some evidence is privately held by the defendant, while some evidence can be exogenously obtained by the judicial system.

Glazer, Jacob, and Ariel Rubinstein. 2004. "On Optimal Rules of Persuasion." *Econometrica* 72 (6): 1715–36.

Glazer, Jacob, and Ariel Rubinstein. 2006. "A Study in the Pragmatics of Persuasion: A Game Theoretical Approach." *Theoretical Economics* 1 (4): 395–410.

Green, Jerry R., and Jean-Jacques Laffont. 1986. "Partially Verifiable Information and Mechanism Design." *Review of Economic Studies* 53 (3): 447–56.

Hart, Sergiu, Ilan Kremer, and Motty Perry. 2017. "Evidence Games: Truth and Commitment." *American Economic Review* 107 (3): 690–713.

Sher, Itai. 2011. "Credibility and Determinism in a Game of Persuasion." *Games and Economic Behavior* 71 (2): 409–19.

Siegel, Ron, and Bruno Strulovici. 2019. "Judicial Mechanism Design." Unpublished.

Silva, Francisco. 2019. "If We Confess Our Sins." *International Economic Review* 60 (3): 1389–1412.

Strausz, Roland. 2016. "Mechanism Design with Partially Verifiable Information." Cowles Foundation Discussion Paper 2040.