A Principal-Agent Relationship with No Advantage to Commitment

Rajiv Vohra ⓡ Francisco Espinosa ⓡ Debraj Ray[†]

September 2021

Published in *Pure and Applied Functional Analysis* **6**, 1043–1064 (2021); Special Issue on Mathematical Economics, Part II, Dedicated to Professor M. Ali Khan on the occasion of his 70th birthday.

Abstract. This paper explores conditions under which the ability to commit to a menu of contracts in a principal-agent relationship creates no additional benefit for the principal, over and above simultaneous interaction without commitment. A central assumption is that the principal's payoff depends only on the *payoff* to the agent and her type.

1. The Setting

Contracts are often written in anticipation of informational asymmetries that can make it impossible to achieve first-best efficiency. The principal-agent model applies to a class of such problems where one agent, the principal, designs a mechanism to influence an agent's actions. For instance, an owner of a firm may design a payment scheme for a worker whose effort is unobservable but correlated with output. Moreover, it is also possible that the agent has private information. In this setting, the principal aims to design *and commit to* a mechanism that serves her interest by providing the agent with the appropriate incentives. That ability to commit is often critical. In contrast, we identify a class of principal-agent problems in which the principal's commitment power is unnecessary in the sense that the outcome of an optimal mechanism with commitment can be achieved as an equilibrium of a simultaneous move game between the principal and the agent.[1]

The principal's action set X_P is taken to be a subset of a Banach space, while the agent's action set is X_A, also a subset of a Banach space. The principal's action can be thought of as a "contract" or a "reward function," typically an object that combines with the agent's action to produce an outcome. For instance, in the optimal taxation problem of Mirrlees (1971), the principal's action is a tax function, so X_P is the space of continuous real-valued functions on some interval of potential incomes, with the sup norm, while X_A might represent labor-leisure choices made by the agent.

[†]Vohra: Brown University, rajiv_vohra@brown.edu, Espinosa: University of Chicago, espinosa.fran@gmail.com; Ray: New York University and University of Warwick, debraj.ray@nyu.edu. Ray acknowledges funding under NSF grant SES-1851758. We thank Bart Lipman and an anonymous referee for constructive and helpful comments on an earlier draft. We dedicate this paper to Ali Khan on the occasion of his 70th birthday. He is a mentor and close friend – a *guru* – to the first and third authors (presented above in random order). His unbounded enthusiasm for mathematical economics, philosophy and poetry continues to inspire us. On that last area of expertise, Khan *sahib* would likely invoke the words of an immortal Bollywood song: "Maĩ shāyar to nahĩ." We respectfully disagree.

[1]For similar results in different settings, see Ben-Porath, Dekel and Lipman (2019, 2020), Deb, Pai and Said (2018), Espinosa ⓡ Ray (2018), Glazer and Rubinstein (2004, 2006), and Hart, Kremer and Perry (2017).

Following standard practice, we model the agent's private information through a finite set of types, $\{1, \ldots, T\}$. Given actions p and a of the principal and agent respectively, an agent of type t has payoff function $u_A(p, a, t)$, and the principal has payoff function $u_P(p, a, t)$. We assume:

[A] The agent's choice set X_A is compact, and for all t, $u_A(p, a, t)$ is continuous in (p, a) and Gâteaux differentiable with respect to p. Moreover, $D_p u_A(p, a, t)$, the Gâteaux partial derivative of $u_A(p, a, t)$, is continuous in (p, a).

Our central assumption is that the principal's payoff depends only on the payoff to the agent and her type. In particular, it does not depend *directly* on the actions of either the principal or the agent. Specifically, we assume that there is a function $f(u, t)$, differentiable in its first argument, such that

[U] $u_P(p, a, t) = f(u_A(p, a, t), t)$.

The principal's payoff could depend positively, negatively or in a non-monotonic way on the payoff of the agent. It's just not allowed to depend *directly* on the actions. This assumption is not general: there are many situations in which it fails. But there are other situations in which this restriction is salient; see Ray ⓡ Vohra for a discussion.[2]

2. Mechanism Design With Commitment

Both the mechanism design problem and an accompanying game without commitment to be studied in Section 3 rely on messages (about agent types) communicated by the agent to the principal. While finite, we allow this message space, R, to be rich enough so as to communicate as much or as little information as the agent wishes. While the revelation principle (Myerson 1982) tells us that R can be set equal to T, we allow R to contain 2^T, so that in the corresponding game without commitment, the agent could report, for instance, that her type lies in a particular subset of T.

In the design problem with commitment, the principal selects a *mechanism* — a function $\pi : R \to P$ — thereby committing to an action $\pi(r)$ for every agent report $r \in R$. The agent freely chooses her report $\rho(t)$ and her action $\alpha(t)$ as functions of her type $t \in T$. We refer to (ρ, α) as the agent's *participation strategy*. The principal's expected utility from mechanism π and the agent's participation strategy (ρ, α) is

$$U_P(\pi, \rho, \alpha) = \sum_t \psi_t u_P(\pi(\rho(t)), \alpha(t), t),$$

[2] A literature in game theory and welfare economics studies "nonpaternalistic externalities," where a person's payoff depends on others' payoffs, not directly their actions; see, for example, Pearce (1983), Ray (1987), Bergstrom (1999), Ray ⓡ Vohra (2020), and Vasquez and Weretka (2020). The present paper departs from that literature in several respects: (1) the principal-agent model involves asymmetric information, (2) the agent's payoff is allowed to depend on the actions of the principal, (3) our focus here is in examining how this restriction on the principal's payoff may make commitment unnecessary. In Ray ⓡ Vohra (2019) the focus is on deriving efficiency implications of non-paternalistic externalities.

where ψ_t is her strictly positive prior on agent type t. The constraint on mechanism design comes from the incentive compatibility of the agent's report and action. For every $t \in T$, it must be that:

$$u_A(\pi(\rho(t)), \alpha(t), t) \geqslant u_A(\pi(r), a, t) \tag{1}$$

for all $r \in R$ and $a \in X_A$.

An *optimal mechanism* maximizes $U_P(\pi, \rho, \alpha)$ subject to (1).

This (standard) notion of an optimal mechanism presumes that the principal can not only choose a mechanism but also induce the agent to choose any participation strategy that satisfies (1). In our context, the principal's power to choose among participation strategies satisfying (1) is not necessary, thanks to assumption (U). Given π, *all* (ρ, α) that are incentive-compatible for the agent in the sense of (1), are payoff-equivalent for the principal. Whenever the agent is indifferent between (ρ, α) and (ρ', α') so is the principal.

3. Equilibrium Without Commitment

Suppose now that the two parties play a simultaneous move game with no commitment. Prior to play, the agent makes a costless announcement regarding her type, using the space R. Following such an announcement, agent and principal simultaneously take actions $a \in X_A$ and $p \in X_P$.

Equivalently, and using the notation already introduced, we can view this game as the simultaneous play of some π by the principal, and (ρ, α) by the agent. By *equilibrium* we refer to a Nash equilibrium of this "augmented game."[3]

Certainly there are equilibria of the augmented game which involve ignoring the announcements entirely. But there are other situations in which "credible neologisms" are available so that types can usefully separate (Farrell 1993). For instance, suppose that there are two types of agent, and agents actually have no action to play. If the principal knows that the agent is of a particular type, she can take an action that yields both the principal and the agent of that type a payoff of 1 each, while yielding a payoff of 0 to the "wrong" type. If, on the other hand, the type is unknown (say with a uniform prior), then the principal chooses some third action which yields, say, 1/2 to each type of agent and to the principal. Then there is an equilibrium of the augmented game in which the principal will listen to the agent, who in turn will find it profitable to reveal her type.

4. Zero Value of Commitment

Suppose the principal chooses a mechanism π and an associated participation strategy (ρ, α) satisfying (1). This means that the agent of type t is subject to principal action $\pi(\rho(t))$. By pretending

[3]This is equivalent to a sequential equilibrium of the two-stage game with agent announcements.

to be of type t' the agent would elicit action $\pi(\rho(t'))$. Thus, (π, ρ, α) effectively allows the agent to "choose" the principal's action from the menu $(\pi(\rho(t)_{t\in T})$. We will refer to elements of this menu as contracts, and $\pi(\rho(t))$ as the contract assigned to t. Of course, incentive compatibility implies that no type can gain by choosing a contract different from the one assigned to her. But indifferences are possible. To compare agent payoffs across the contracts in a menu it will be useful to consider the agent's value function, $v_A(p,t) = \max_{a\in X_a} u_A(p,a,t)$. Our next condition asks that the menu corresponding to an optimal mechanism is reducible (if needed) to a smaller set or "sub-menu," in which each type obtains the same payoff as before, and *strictly* prefers her assigned contract to any other in this sub-menu.

[C] Suppose π is an optimal mechanism with associated, incentive compatible participation strategy (ρ, α). Then there exists a partition $\mathcal{T} = \{T_1, \ldots T_m\}$ of T, and a menu $\{q_1, \ldots, q_m\}$, where $q_i \in \{\pi(\rho(\tau))\}_{\tau\in T}$ for all $i = 1, \ldots, m$, such that

(i) Each $t \in T_i$ is indifferent between q_i and $\pi(\rho(t))$: $v_A(q_i, t) = v_A(\pi(\rho(t), t)$.

(ii) Each $t \in T_i$ *strictly* prefers q_i to any other contract in $\{q_1, \ldots q_m\}$: $v_A(q_i, t) > v_A(q_j, t)$ for $j \neq i$.

Remark 1. Clearly, if the optimal mechanism is strictly incentive compatible, so every agent strictly prefers her assigned contract to others in the menu, [C] holds trivially — no subset need be chosen or re-assignment contemplated.[4]

Remark 2. Many models of interest effectively involve two types of agents, e.g., Deb, Pai and Said (2018), Espinosa ⓡ Ray (2018), Glazer and Rubinstein (2004, 2006) and Hart, Kremer and Perry (2017).[5] In all such cases [C] is satisfied. As Espinosa ⓡ Ray (2018) point out, if type 1 is indifferent to the contract that type 2 accepts, offer both types the contract designed for type 2. The restricted subset is just that consisting of type 2's contract, and [C] holds trivially.

Remark 3. With more than two types, if an optimal mechanism is such that each type is indifferent to at most one other type's contract and there are no indifference cycles, [C] hold. More generally still, it can be shown that [C] holds if and only if every such cycle is of even length. A proof of this assertion is available from the authors upon request.

---

[4] An anonymous referee drew our attention to the Dutta and Sen (2012) notion of partial honesty as a simple way of making incentive compatibility constraints strict. The agent is said to be partially honest if whenever her optimal strategies include truth-telling she chooses truth-telling over lying. In this formulation, the preference for lying is lexicographic in the sense that it applies only in cases of indifference. However, this is not adequate for our purposes because in the proof of our main theorem Condition [C] is used along with continuity of the utility function to argue that an agent whose type is in T_i strictly prefers q_i to any q' arbitrarily close to q_j, $j \neq i$. Partial honesty, as a tie-breaking rule, will clearly not be sufficient for such an argument, unless one were to additionally presume that a sufficiently small gain will not dissuade the agent from telling the *exact* truth.

[5] Glazer-Rubinstein, as well as Hart-Kremer-Perry actually allow for more than two types, but these differ in terms of the evidence supplied to a principal. Once reduced to "equivalent" types after evidence provision is removed, this is a two-type model in our reduced-form setting.

Remark 4. It's possible to identify assumptions on the preferences of the types that imply [C]. Suppose that for any pair of types t and t', and principal actions p and p',

$$v_A(p,t) = v_A(p',t) \text{ implies } v_A(p,t') = v_A(p',t'). \tag{2}$$

So if π is an optimal mechanism and t is indifferent between her contract and that of t', then t' is also indifferent between her contract and t's. Then types can be partitioned into sets such that all types within each set are indifferent across contracts within their set and strictly prefer any such contract to one in a different set: [C] is satisfied. Condition (2) is related to the *simple type dependence* property used by Ben-Porath, Dekel and Lipman (2019) to show, in a somewhat different context, that commitment is not necessary.[6] They study a model in which the agent can present evidence about his type but doesn't have any other action. Simple type dependence is then the same as (2).[7]

In general, [C] is not to be had for free. In Example 2 below, [C] doesn't hold. There are three types, and types 1 and 2 prefer their assigned contracts, while type 3 is indifferent across all three contracts. Because types 1 and 2 have their unique optima under their original contracts, these cannot be dropped from the reduced set, so there is no way to give type 3 a unique optimum. As we shall see, our main result also fails in this example.

We make one more (technical) assumption. Fix an action p for the principal, a set of types $W \subseteq T$ and an action profile α for the agents. Say that the principal's payoff can be *improved* at (p, α, W) if there is $\hat{p} \in P$ such that

$$\sum_{t \in W} \psi_t u_P(\hat{p}, \alpha(t), t) > \sum_{t \in W} \psi_t u_P(p, \alpha(t), t).$$

The assumption that follows asks for the existence of a "local improvement" whenever an improvement is possible. By [A] and [U], $u_P(p,a,t)$ is Gâteaux differentiable in p, allowing us to measure local changes through $D_p u_P(p,a,t)$.[8] Let $S(p,t) = \arg\max_{a \in X_A} u_A(p,a,t)$ denote the set of best responses for an agent of type t given p. Say that the principal's payoff can be *locally improved* at (p, α, W) if, given $a(t) \in S(p,t)$ for every $t \in W$, there exists a direction $\hat{d} \in X_P$ such that

$$\sum_{t \in W} \psi_t D_p u_P(p, a(t), t)\hat{d} > 0.$$

[6] However, Condition [C] and their overall assumptions are not comparable: neither implies the other.

[7] It is always possible to write the agent's payoff function as an indirect utility function that depends only on the principal's action, by optimizing out the agent's action. Given (U), this leaves the principal's optimization problem unchanged. However, (2) can then acquire different meaning. For instance, type independence over the principal's "elementary actions" (e.g., retain, not retain) may not translate into type independence over the principal's actions in X_p. Consider the Espinosa-Ray setting with costless choice of noise. The principal's action, p, is a contract specifying the probability with which he retains the agent as a function of the signal emitted by the agent. And the distribution of the signal is function of the agent's action. In a menu consisting of two such contracts, the agents may well be opposed in their preferences over the two functions, even though they have the same preference for retention.

[8] Since $f : R \to R$ is differentiable and u_A is Gâteaux differentiable, the chain rule applies (see for example Proposition 2.47, Bonnans and Shapiro 2000) and $D_p u_P(p,a,t) = f' D_p u_A(p,a,t)$.

[I] If u_P can be *improved* at (p, α, W), then it can be locally improved at (p, α, W).

Condition [I] automatically holds if we think of the principal's actions p as randomizations over some finite set of pure actions, with $\hat{d} = (\hat{p} - p)$.

We can now state our result.

THEOREM **1.** *Assume* [A], [U] and [I]. *If an optimal mechanism satisfies* [C]*, then there exists an equilibrium of the augmented game which replicates precisely the same principal and type-specific agent payoffs.*

To obtain some intuition for this result, consider the setting in Espinosa ⓡ Ray (2018). There, an agent who privately knows his type (good or bad) seeks to be retained by a principal. The principal wishes to retain a good type, and to remove a bad type. The agent generates a noisy but informative scalar signal with full-support density centered on his type; specifically, he can amplify or reduce the precision of this process, though the mean of the signal is fixed by his type. The principal observes the signal realization (but not the signal structure, or at least not fully), and makes a retention decision. In this model, the principal's payoff can be expressed as a function of the payoffs of the agents. She wants to retain the good type, so these payoffs are perfectly and positively aligned. She wants to remove the bad type, so once again these payoffs are perfectly — though negatively — aligned. A commitment solution specifies a set of received signals for which the principal will retain the agents. In part because signal realizations are continuous, it can be shown that from any principal action that is not a best (no-commitment) response, the principal can always profitably move in the direction of her best response by changing the contract by a tiny amount. The consequent responses of the agents will have no first-order effect on the principal's payoff, by the envelope theorem. The assumption that principal payoffs can be written as a function of the agent's payoffs allows the envelope theorem to play this crucial role.

In the next Section we explore the role that our assumptions play in our main result. In particular, we provide examples to show that none of our substantive assumptions — [U], [I] or [C] — can be dropped from the statement of the Theorem.[9]

5. Two Examples

A principal's ability to commit to a mechanism can be valuable to her. Our first example illustrates this well-known fact, but essentially to explain why our result does not apply:

Example 1. (Failure of [U] or [I].) The principal can offer one of two jobs to an agent: an specialist position S, or a generalist position G, or she can decide not to hire the agent at all (action N). The

[9]We view [A] as a technical restriction, one that permits us to state the substantive condition [I]. It also permits us to use a general envelope theorem of Bonnans and Shapiro (2000). Of course, [A] cannot be dropped free of charge, but it is of little economic import and we do not explore it in the examples.

agent has two types: s, with probability q, or g with probability $1-q$. The specialist is ideally suited to position S, and his preferences are perfectly aligned with those of the principal. The principal would prefer not to hire the generalist, but conditional on doing so, he does less damage in position G, which is also the generalist's top choice. The agent takes no separate action a.

The following table describes principal and agent payoffs. The first entry pertains to the agent; the second to the principal. We assume that $0 < \epsilon < \frac{2q}{(1-q)}$.

	Principal's Action		
	S	G	N
Specialist s	(2, 2)	(1, 1)	(0, 0)
Generalist g	(1, $-\epsilon$)	(2, $-\epsilon/2$)	(0, 0)

TABLE 1. Payoffs to principal and agent types in Example 1.

Note that u_P can be written as a function f of u_A and agent type t: $f(u_A, e) = u_A$, while

$$f(u_A, g) = \begin{cases} -\epsilon & \text{if } u_A = 1 \\ -\epsilon/2 & \text{if } u_A = 2 \\ 0 & \text{if } u_A = 0 \end{cases}$$

Commitment Solution. The first-best for the principal is to have type s in position S and not hire type g, but that's not incentive compatible; type g will then pretend to be of type s. The principal must therefore bear an incentive cost. The optimal mechanism is one in which she commits position S to an announcement of s and G to an announcement of g, with expected payoff $2q - \frac{\epsilon}{2}(1-q)$.

Nash Equilibrium. If, in the augmented game, the types were to reveal themselves, then the principal wouldn't hire type g: she would offer N rather than G in response to a revealing g-announcement. In fact, the only Nash equilibrium is for the two types to not reveal themselves (say, "announce" $\{s, g\}$) and for the principal to offer S.[10] The principal's equilibrium payoff being $2q - \epsilon(1-q)$, less than the payoff from the optimal mechanism, and commitment matters.

Theorem 1 fails to apply because [I] fails. At the optimal solution, offering N to the generalist instead of G is an improvement for the principal, but no *local* improvement is possible. This issue is easily circumvented by allowing the principal to randomize over actions: assumption [I] is now satisfied. It can also be verified that the Nash equilibrium is unchanged: again, both types pool and are offered S. But now commitment has even more value. The optimal mechanism offers the g-announcement the positions G or N with equal probability, and S to the s-announcement. The generalist is still willing to report truthfully, because his expected payoff from either announcement equals 1. The principal earns still higher payoff than in the earlier commitment mechanism.

[10]This makes use of our assumption that $\epsilon < 2q/(1-q)$.

So Theorem 1 "fails" again, despite [I] being met. This time it does so because assumption [U] does not hold on the mixed space of principal actions. The generalist is indifferent between the uniform lottery on $\{G, N\}$, and the sure receipt of S. But the principal is *not* indifferent: she obtains $-\epsilon/4$ in the first situation, and $-\epsilon$ in the second. So the conditions of Theorem 1 do not apply, and indeed its conclusion doesn't hold.

That commitment has value even if mixed strategies are allowed is interesting in light of a recent result of Ben-Porath, Dekel and Lipman (2020). Allowing for mixed strategies, they show that the payoff corresponding to an optimal mechanism can be achieved through a Nash equilibrium of the non-commitment game. Their main assumption is that preferences are *semi-aligned*: there is a function $\nu(t)$ such that $u_P(p,t) = \nu(t)u_A(p,t)$. This is stronger than [U] in that it requires the principal's payoff to be *linear* in the agent's payoff, which is not the case in our example. Thus, our example shows that semi-alignedness cannot be dropped from the statement of their theorem.

Example 2. (Failure of [C]). The principal has three kinds of positions: a hardware specialist, H, a software specialist, S, and in addition to offering one of these positions, can also mix it with a general job G, using a time allocation of $\lambda \in [0,1]$ for the specialist position and $1-\lambda$ for G. (The no-job option N is not available.) Moreover, each of the specialist jobs H and S comes on a continuum of grades or *scales* $x \in [0,1]$, while job G can only be done at one level. The principal's action set can therefore be described by

$$X_P = \{\{(H,x,\lambda) \mid x \in [0,1], \lambda \in [0,1]\}, \{(S,x) \mid x \in [0,1], \lambda \in [0,1]\}\}$$

where an offer of $\lambda = 0$ is equivalent to offering G full-time. The agent has no action and is of three possible types: a hardware specialist h, a software specialist s (each with probability q) and a generalist g (with probability r). Assume $r > q > 0$.

	Principal's Action	
	(H,x,λ)	(S,x,λ)
Hardware specialist h	$(\lambda[x+1], \lambda[x+1])$	$(0,0)$
Software specialist s	$(0,0)$	$(\lambda[x+1], \lambda[x+1])$
Generalist g	$(\lambda x+1, 1-\lambda x)$	$(\lambda x+1, 1-\lambda x)$

TABLE 2. Payoffs to principal and agent types in Example 2.

Note that the principal's payoff is also fully aligned with the specialists, and also with the generalist (though with opposite sign). Therefore [U] holds: $u_P(u_A,t) = u_A$ for $t = h,s$ and $u_P(u_A,g) = 2-u_A$. It can also be checked that [I] is satisfied: every improvement comes from changing the scale x of the specialists or the time λ allocated to each, and then a suitable adjustment of x and λ will produce a local improvement. Nevertheless, we will show that commitment has value. Of course, given our main result, this must be a result of [C] being violated at an optimal mechanism.

Pure-Strategy Nash Equilibrium in the Augmented Game. There is (always) an equilibrium in which the principal disregards all type announcements, and offers a specialist position full time ($\lambda = 1$) at level $x = 0$. To see this, we only need to verify that the principal is playing a best response. If the principal offers any combined specialist-generalist position at scale x and time allocation λ, her payoff is $q\lambda(x + 1) + r(1 - \lambda x)$. Because $r > q$, she must set $x = 0$ and $\lambda = 1$, with resulting principal payoff $q + r$.

Moreover, there is no other equilibrium. If the agents separate fully, then the best response of the principal to each announcement of s or h is to offer that specialist job at scale $x = 1$ and time $\lambda = 1$, but then the generalist will deviate by announcing that he is s or h. If the agents separate partially, then (without loss) there are two cases: one is $\{hg, s\}$, and the other is $\{hs, g\}$. In the former case, the principal will offer S at $x = 1$ and $\lambda = 1$ in response to the announcement s. In response to hg, her unique best response is to offer H at $x = 0$ and $\lambda = 1$, obtaining a conditional expected payoff of 1 (H at any other scale and time allocation is dominated by H at $x = 0$ and $\lambda = 1$). This yields the generalist a payoff of 1, but then he will deviate and announce s to pick up the job S at scale $x = 1$ and time $\lambda = 1$, which yields him a payoff of 2.

In the latter case, the principal will offer either H or S to the announcement hs, with $x = 1$ and $\lambda = 1$ (she takes advantage of the 50-50 chance that the specialist is the right one). To g she will offer either of the specialist jobs but *only* at scale $x = 0$, but at any time allocation. In either case the generalist will deviate to announcing hs.

Commitment Solution. There is a feasible commitment solution in which the types all reveal themselves, and the principal offers full-time jobs ($\lambda = 1$) to match the specialist announcements, at scale $x = 0$. To the generalist she offers a full-time generalist job, with $\lambda = 0$. No type will want to deviate. This yields a payoff to the principal of 1, strictly higher than the Nash payoff.[11]

Since preferences are semi-aligned in this example, it follows from Ben-Porath, Dekel and Lipman (2020) that allowing the principal to mix must result in commitment value being 0.

6. Proof of Theorem 1

Suppose (π, ρ, α) is an optimal mechanism that satisfies [C]. Let $\mathcal{T} = \{T_1, \ldots, T_m\}$ and $\{q_{i=1}^m\}$ be as specified in [C]. Consider the simultaneous move game without commitment, and define strategies (π', ρ', α') as follows, where for each t, $r = \rho'(t)$ lies in R, and the domain of π' is R. Recall that $R \supseteq 2^T$.

(i) An agent of type $t \in T_i$ chooses $\rho'(t) = T_i$ and action $\alpha'(t) \in S(q_i, t)$.

(ii) The principal chooses strategy π' where $\pi'(T_i) = q_i$ for $i = 1, \ldots m$ and $\pi'(r) = q_1$ for $r \notin \mathcal{T}$.

[11] In fact, this is an optimal mechanism if $q \leqslant 0.25$. Otherwise it is optimal to offer each of the two specialist jobs at level 1 and $\lambda = 1$, which yields $4q$ to the principal.

Since $\pi'(\rho'(t)) = q_i$ for $t \in T_i$ and $\alpha'(t) \in S(q_i, t)$, choosing (ρ', α') under π' yields agent t

$$v_A(q_i, t) = v_A(\pi(\rho(t), t) = u_A(\pi(\rho(t), \alpha(t), t), \tag{3}$$

where the first equality is a consequence (from [C]) of the fact that type $t \in T_i$ is indifferent between $\pi(\rho(t))$ and q_i. Since the range of π' is $\{q_{i=1}^m\}$, (3) implies that given π', $(\rho'(t), \alpha'(t))$ satisfies the incentive compatibility constraints (1). This has two important implications:

(a) (ρ', α') is a best response to π',

(b) (π', ρ', α') is an optimal mechanism since (3) and [U] imply that $U_P(\pi, \rho, \alpha) = U_P(\pi', \rho', \alpha')$.

To complete the proof of the theorem we will show that (π', ρ', α') is an equilibrium of the game. Given (a), it suffices to show that π' is a best response of the principal. Suppose it is not. Then there is strategy $\hat{\pi}$ such that

$$U_P(\hat{\pi}, \rho', \alpha') > U_P(\pi', \rho', \alpha').$$

There must therefore exist $T_i \in \mathcal{T}$ with

$$\sum_{t \in T_i} \psi_t u_P(\hat{\pi}(\rho'(t)), \alpha'(t), t) > \sum_{t \in T_i} \psi_t u_P(q_i, \alpha'(t), t).$$

Letting $\hat{q}_i = \hat{\pi}(T_i)$ and using the fact that $\rho'(t) = T_i$ for all $t \in T_i$, this means that

$$\sum_{t \in T_i} \psi_t u_P(\hat{q}_i, \alpha'(t), t) > \sum_{t \in T_i} \psi_t u_P(q_i, \alpha'(t), t).$$

In other words, $\hat{q}_i$ is an improvement at (q_i, α', T_i).

By Condition [I], given $a(t) \in S(q_i, t)$ for all $t \in W$, there exists a direction $\hat{d}$ such that

$$\sum_{t \in W} \psi_t D_p u_P(q_i, a(t), t)\hat{d} = \sum_{t \in W} \psi_t f' D_p u_A(q_i, a(t), t)\hat{d} > 0. \tag{4}$$

The expression on the left hand side of (4) is based on partial derivatives with respect to p and does not take into account the indirect effect of a differential change in p on the optimal actions of the agent. Because T_j has a strict best response at q_j among the set $\{q_k\}$, one must conclude that this will continue to hold for a differential change in q_j, so no one changes their preferred "location". In other words, ρ' remains an optimal reporting strategy for the agent. Of course, the actions of the agents in T_i will generally change in response to a differential change in q_i. All such changes are included in the agent's value function, $v_A(p, t)$. Letting $v_P(p, t) = f(v_A(p, t), t)$, to complete the proof it suffices to obtain a version of (4) in terms of $D_p v_P(q_i, t)$ instead of $D_p u_A(q_i, a(t), t)$. Indeed, if the value functions were differentiable, a suitable envelope theorem may suggest a way of showing that the indirect effects are of order 0, allowing us to argue that (4) implies

$$\sum_{t \in W} \psi_t D_p v_P(q_i, t)\hat{d} > 0,$$

which would clearly yield the contradiction to (b) that we seek. While $v_A(p,t)$ and $v_P(p,t)$ are not in generally differentiable, we can use a general envelope theorem of Bonnans and Shapiro (2000) to show that they possess directional derivatives that are related to the partial derivatives of $u_A(p,a,t)$ and $u_P(p,a,t)$, showing that the indirect effects are, in some sense, negligible.[12]

Let $v'_A(p,t,d) \equiv \lim_{r\to 0^+}[v_A(p+rd,t) - v_A(p,t)]/r$ denote the directional derivative of $v_A(p,t)$ in the direction d. If this limit exists for every direction, $v_A(p,a,t)$ is said to be directionally differentiable. Given assumption [A], we can apply Theorem 4.13 in Bonnans and Shapiro (2000) to assert that $v_A(p,t,d)$ is directionally differentiable and for every t, and for every $d \in X_p$

$$v'_A(p,t,d) = \max_{a\in S(p,t)} D_p u_A(p,a,t)d$$

Thus, for every $d \in X_P$, and every t, there exist $a(t) \in S(p,t)$ such that[13]

$$v'_P(p,t,d) = f' v'_A(p,t,d) = f' D_p u_A(p,a(t),t)d.$$

Combining this with (4), there exists a direction $\hat{d}$ for which

$$\sum_{t\in W} \psi_t v'_P(q_i,t,\hat{d}) > 0$$

which contradicts (b) and completes the proof.

References

Ben-Porath, E., Dekel, E. and Lipman, B. L. (2019) "Mechanisms With Evidence: Commitment and Robustness", *Econometrica* **87**, 529–566.

Ben-Porath, E., Dekel, E. and Lipman, B. L. (2020) "Acquisition of Stochastic Evidence", mimeo.

Bonnans, J. F., and A. Shapiro (2000): *Perturbation Analysis of Optimization Problems*. New York: Springer-Verlag,

Bergstrom, T. (1999), "Systems of Benevolent Utility Functions," *Journal of Public Economic Theory* **1**, 71–100.

Deb, R., Pai, M., and M. Said (2018), "Evaluating Strategic Forecasters," *American Economic Review* **108**, 3057–3103.

Dutta, B. and A. Sen. (2012), "Nash Implementation with Partially Honest Individuals," *Games and Economic Behavior* **74**, 154-169.

Espinosa, F. ⓡ Ray, D. (2018) "Noisy Agents", NBER Working Paper No. 24627, May 2018.

[12] Note that the envelope theorems in Milgrom and Segal (2002) do not apply directly to our framework because they pertain to the case $X_P = [0,1]$.

[13] This makes use of the chain rule for directional derivatives, as in Proposition 2.47, Bonnans and Shapiro (2000).

Farrell, J. (1993) “Meaning and Credibility in Cheap-Talk Games”, *Games and Economic Behavior* **5**, 514-531.

Glazer, J. and A. Rubinstein (2004), “On Optimal Rules of Persuasion,” *Econometrica* **72**, 1715–1736.

Glazer, J. and A. Rubinstein (2006), “A Study in the Pragmatics of Persuasion: A Game Theoretical Approach,” *Theoretical Economics* **1**, 395-410.

Hart, S., I. Kremer and M. Perry (2017) “Evidence Games: Truth and Commitment” *American Economic Review*, **107**, 690-713.

Milgrom, P. and I. Segal (2002) “Envelope Theorems for Arbitrary Choice Sets” *Econometrica*, **70**, 583-601.

Mirrlees, J. A. (1971), “An Exploration in the Theory of Optimum Income Taxation”, *The Review of Economic Studies* **38**, 175-208.

Myerson, R. B. (1982), “Optimal Coordination Mechanisms in Generalized Principal-Agent Problems”, *Journal of Mathematical Economics* **10**, 67-81.

Pearce, D. (1983), “Nonpaternalistic Sympathy and the Inefficiency of Consistent Intertemporal Plans,” Ph.D. dissertation, Princeton University, reprinted in *Foundations in Microeconomic Theory*, edited by M. Jackson and A. McLennan, Berlin, Heidelberg: Springer, 2008.

Ray, D. (1987), “Nonpaternalistic Intergenerational Altruism,” *Journal of Economic Theory* **41**, 112–132.

Ray, D. ⓡ Vohra, R. (2020), “Games of Love and Hate,” *Journal of Political Economy* **128**, 1789-1825.

Vasquez, J. and M. Weretka (2020), “Affective Empathy in Non-cooperative Games,” *Games and Economic Behavior* **121**, 548-564.