NBER WORKING PAPER SERIES

SUBGAME PERFECT IMPLEMENTATION WITH ALMOST PERFECT INFORMATION AND THE HOLD-UP PROBLEM

Philippe Aghion
Drew Fudenberg
Richard T. Holden

Working Paper 15167
http://www.nber.org/papers/w15167

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2009

We owe special thanks to Oliver Hart for numerous discussions, and to Johannes Horner and Andy Skrzypacz for pointing out an error in a previous version. We are also grateful to Mathias Dewatripont, Bob Gibbons, Philippe Jehiel, John Moore, Roger Myerson, Andrew Postlewaite, Olivier Tercieux, Jean Tirole, Ivan Werning and Muhamet Yildiz for helpful discussions and comments. Michael Powell provided excellent research assistance. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Subgame Perfect Implementation with Almost Perfect Information and the Hold-Up Problem
Philippe Aghion, Drew Fudenberg, and Richard T. Holden
NBER Working Paper No. 15167
July 2009
JEL No. C72,C73,D23,L22

ABSTRACT

The foundations of incomplete contracts have been questioned using or extending the subgame perfect implementation approach of Moore and Repullo (1988). We consider the robustness of subgame perfect implementation to the introduction of small amounts of asymmetric information. We show that Moore-Repullo mechanisms may not yield (even approximately) truthful revelation in pure or totally mixed strategies as the amount of asymmetric information goes to zero. Moreover, we argue that a wide class of extensive-form mechanisms are subject to this fragility.

Philippe Aghion
Department of Economics
Harvard University
1805 Cambridge St
Cambridge, MA 02138
and NBER
paghion@fas.harvard.edu

Drew Fudenberg
Department of Economics
Harvard Unviersity
1805 Cambridge St
Cambridge, MA 02138
dfudenberg@harvard.edu

Richard T. Holden
University of Chicago
Booth School of Business
Room 528
5807 S Woodlawn Ave
Chicago,IL 60637
and NBER
richard.holden@chicagobooth.edu

1 Introduction

Grossman and Hart (1986) argued that in contracting situations where states of nature are *observable* but not *verifiable,* asset ownership (or vertical integration) could help limit the extent to which one party can be held up by the other party, which in turn should encourage ex ante investment by the former. However, vertical integration as a solution to the hold-up problem has been questioned in various papers[1] which all use or extend the subgame perfect implementation approach of Moore and Repullo (1988). In particular, Maskin and Tirole (1999a)[2] argue that although parties may have difficulty foreseeing future physical contingencies they can write contracts which specify ex ante the possible payoff contingencies. Once the state of the world is realized, the parties can "fill in" the physical details. The latter step is subject to incentive-compatibility considerations. That is, each agent must be prepared to specify the details truthfully. Maskin and Tirole achieve this through a 3-stage subgame perfect implementation mechanism which induces truth-telling by all parties as the unique equilibrium outcome[3].

In this paper, we consider the robustness of the Moore-Repullo (MR) mechanism to the introduction of small amounts of asymmetric information. We find that the MR mechanism may not yield even approximately truthful revelation in pure or totally mixed strategies as the amount of informational asymmetry goes to zero. Moreover, we show that this non-robustness result does not require informational asymmetries on agents' types (call it the introduction of crazy types): small deviations from symmetric information on the buyer's willingness to pay for the good or on the seller's production cost, suffice to generate this non-robustness result. This, in turn, has important implications for the debate on the foundations of incomplete contracts: namely, while asymmetric information about agents' types can eliminate the hold-up problem at the same time as it introduces undesirable equilibria in sequential mechanisms, asymmetric information about valuation and costs is

[1] For example, see Aghion-Dewatripont-Rey (1999) and recently Maskin-Tirole (1999a, 1999b).

[2] See also Maskin and Tirole (1999b).

[3] Whereas Nash implementation (see Maskin 1977, 1999) does not guarantee uniqueness.

shown to preserve the hold-up problem even though it perturbs the MR mechanism.

We proceed in several steps. In Section 2 we introduce a simple example of ex-post bargaining and exchange drawn from Hart and Moore (2003)[4] to illustrate our first point on the robustness of the MR mechanism to the introduction of small amounts of asymmetric information. More precisely, we modify the signal structure of the game by assuming that each player receives private signals about the true value of the good, instead of knowing it perfectly; thus the value is "almost common knowledge" in the sense of being common p-belief (Monderer and Samet (1989)) for p near 1. Our main finding here is that the simple subgame-perfect implementation mechanism à la MR for this example, does not yield approximately truthful revelation in either pure or totally mixed strategies as the correlation between the private signals and the true value of the good becomes increasingly perfect, although truthful revelation can be approximated by a partially mixed equilibrium.

The basic idea behind this result is that even a small amount of uncertainty at the interim stage, when players have observed their signals but not yet played the game, can loom large ex post once a player has sent a message. This is closely related to the observation that backwards induction and related equilibrium refinements are not in general robust to perturbations of the information structure (see Fudenberg, Kreps and Levine (1988), Dekel and Fudenberg (1990) and Borgers (1994)) so that the predictions generated under common knowledge need not obtain under almost common knowledge. However, in this example we restrict attention to informational asymmetries about the value of the good as opposed to the more general perturbations considered in the robustness literature. More specifically, in our modification of the Hart-Moore-Repullo example, the Seller produces a good whose valuation is stochastic, and may be high or low. Each contracting party gets a private and yet almost perfect signal about the good's valuation; the players have a common prior on the joint distribution of values and signals. The Moore-Repullo mechanism requests that one party, say the Buyer, make an announcement about the value of the good, and then the

[4] This example itself illustrates the mechanism in Moore and Repullo (1988, Section 5).

Seller may either challenge or not challenge the Buyer's announcement. Obviously, under perfect information, the Buyer's announcement contains no information which the Seller did not have ex ante. However, when each player receives a private signal about the value of the good, the Buyer's announcement *does* contain information about her own signal of the good's valuation. The Seller will then update his belief about the value of the good, on the basis of both, his own signal and the announcement made by the Buyer. And the resulting Bayesian updating is what causes the subgame implementation logic to break down mechanism to break down. For example, if the two parties commit to play the MR mechanism and yet the Buyer announces a low value for the good, then the Seller will update his beliefs towards the notion that the true value of the good is indeed low. Consequently, the Seller will not challenge the Buyer by fear of being fined under the MR mechanism for having unfairly challenged the Buyer.

In Section 3 we extend our analysis to a general setting with n states of nature and transferable utility, and we show that in this setting there exist natural social choice functions which cannot be implemented in any totally mixed equilibrium of perturbed MR mechanisms with arbitrarily small amount of private information about the value of the good. In Section 4 we move beyond Moore-Repullo mechanisms and ask whether the same logic can apply to *any* extensive-form mechanism. Here we make a more general but weaker claim: for any game induced by an extensive form-mechanism, there exists a nearby game with almost perfect information in which at least one equilibrium is "undesirable", i.e. does not induce truth-telling by the contracting parties. While this conclusion can be easily established if we allow for private signals about payoffs, based on Fudenberg, Kreps and Levine (1988), henceforth FKL[5], we also prove non-robustness to the common p-belief perturbation considered in Section 2 when restricting attention to pure strategies and three-stage sequential mechanisms.[6].

[5] FKL show that any Nash equilibrium is the limit of strict (and thus sequential) equilibria when a small change is made to prior beliefs. So if agents have any doubts about the *payoff structure*, extensive-form mechanisms cannot robustly improve on Nash implementation.

[6] Parallel work by Kunimoto and Tercieux (2009), who show that only Maskin-monotonic social choice

Thus, Section 4 suggests that if we start from any subgame implementation mechanism under perfect information, one could show the existence of at least one undesirable equilibrium with arbitrarily small perturbations of the information structure, whereas Section 3 shows that for MR mechanisms, this perturbation leads to uniqueness of an undersirable equilibrium. Together these findings highlight the difficulties in moving beyond Nash implementation which, in contrast to subgame perfect implementation, is robust to these deviations from perfect information. However, most settings in contract theory involve non-Maskin monotonic social choice functions (as in Section 2 and 3), and hence cannot be Nash implemented.

In Section 5, we link our analysis of the non-robustness of subgame perfect implementation to the hold-up problem. In particular we argue that while the hold-up problem also may be “solved” by introducing even small amounts of private information about the agents’ types, in contrast the hold-up problem remains when introducing small amounts of private information about valuation and/or costs. In other words, it is the non-robustness of MR and other extensive-form mechanisms to introducing private information about valuation or costs which restores the Grossman-Hart logic once we allow for message games, not their non-robustness to the introduction of “crazy types”.

In addition to the contracting and mechanism design literatures mentioned above, our paper also relates to previous work by Cremer and McLean (1988), Johnson, Pratt and Zeckhauser (1990), and Fudenberg, Levine and Maskin (1991). These papers show how one can take advantage of the correlation between agents’ signals in designing incentives to approximate the Nash equilibrium under perfect information. These papers consider static implementation games with commitment, and look at fairly general information structures, as opposed to our focus on the robustness of subgame-perfect implementation to small deviations from complete information. Chung and Ely (2003) show that only Maskin-monotonic social choice functions are implementable in undominated Nash equilibrium.

functions can be implemented in the closure of the sequential equilibrium correspondence, suggests that this result may extend to mixed strategies.

The remainder of this paper is organized as follows. Section 2 illustrates our basic idea using the simple example of Hart and Moore. We first present the implementation result under perfect information; then we introduce (small) informational asymmetries and illustrate our non-convergence result in that context; then we discuss the robustness of the example. Section 3 establishes a more general non-convergence result for 3-stage Moore-Repullo (MR) mechanisms with transferable utility, and then it discusses the non-robustness of subgame perfect implementation through extensive-form mechanisms other than MR. Section 4 extends the discussion to other sequential mechanisms. Finally, Section 5 analyzes how the non-robustness of MR and other subgame perfect implementation mechanisms, affects the debate on the foundations of Grossman-Hart (1986).

2 An example

2.1 Setup

Consider the following simple example from Hart and Moore (2003). This example captures, in the simplest possible setting, the logic of subgame perfect implementation mechanisms.

There are two parties, a B(uyer) and a S(eller) of a single unit of an indivisible good. If trade occurs then B's payoff is

$$V_B = v - p,$$

where p is the price. S's payoff is

$$V_S = p - \psi,$$

where ψ is the cost of producing the good, which we normalize to zero.

The good can be of either high or low quality. If it is high quality then B values it at $v = \bar{v} = 14$, and if it is low quality then $v = \underline{v} = 10$.

2.2 Perfect information

Suppose first that the quality v is observable by both parties, but not verifiable by a court. Thus, no initial contract between the two parties can be made credibly contingent upon v.

Yet, as shown by Hart and Moore (2003), truthful revelation of v by the buyer can be achieved through the following contract/mechanism, which includes a third party T.

1. B announces either "high" or "low". If "high" then B pays S a price equal to 14 and the game then stops.

2. If B announces "low" then: (a) If S does not "challenge" then B pays a price equal to 10 and the game stops.

3. If S challenges then:

    (a) B pays a fine F to T

    (b) B is offered the good for 6

    (c) If B accepts the good then S receives F from T (and also the 6 from B) and we stop.

    (d) If B rejects at 3b then S pays F to T

    (e) B and S Nash bargain over the good and we stop.

When the true value of the good is common knowledge between B and S this mechanism yields truth-telling as the unique equilibrium. To see this, let the true valuation $v = \bar{v} = 14$, and let $F = 9$. If B announces "high" then B pays 14 and we stop. If, however, B announces "low" then S will challenge because at stage 3a B pays 9 to T and, this being sunk, she will still accept the good for 6 at stage 3b (since it is worth 14). S then receives $9+6 = 15$, which is greater than the 10 that she would receive if she didn't challenge. Thus, if B lies, she gets $14 - 9 - 6 = -1$, whereas she gets $14 - 14 = 0$ if she tells the truth. It is straightforward to verify that truthtelling is also the unique equilibrium if $v = \underline{v} = 10$. Any fine greater than 8 will yield the same result.

2.3 Less than perfect information

2.3.1 Setup

Now let us introduce a small amount of noise into the setting above. Suppose that the players have a common prior that $\Pr(v = 14) = \Pr(v = 10) = 1/2$. Each player receives an independent draw from a signal structure with two possible signals: θ' or θ''. Let the signal structure be as follows:

	$\theta'_B\theta'_S$	$\theta'_B\theta''_S$	$\theta''_B\theta'_S$	$\theta''_B\theta''_S$
$\Pr(v = 14)$	$\frac{1}{2}(1-\varepsilon)^2$	$\frac{1}{2}(1-\varepsilon)\varepsilon$	$\frac{1}{2}\varepsilon(1-\varepsilon)$	$\frac{1}{2}\varepsilon^2$
$\Pr(v = 10)$	$\frac{1}{2}\varepsilon^2$	$\frac{1}{2}(1-\varepsilon)\varepsilon$	$\frac{1}{2}\varepsilon(1-\varepsilon)$	$\frac{1}{2}(1-\varepsilon)^2$

For simplicity we will keep the payments under the mechanism the same as above and assume that B must participate in the mechanism. We could easily adjust the payments accordingly and assume voluntary participation.

2.3.2 Pure strategy equilibria

We first claim that there is no equilibrium in pure strategies in which the buyer always reports truthfully. By way of contradiction, suppose there is such an equilibrium, and suppose that B gets signal θ'_B. Then she believes that, regardless of what signal player S gets, the value of the good is greater than 10 in expectation. So she would like to announce "low" if she expects that subsequently to such an announcement, S will not challenge. Now, suppose B announces low. In a fully revealing equilibrium, S will infer that B must have seen signal θ''_B if she announces low. S now believes that there is approximately 1/2 probability that $v = 10$ and therefore she will not challenge. But if S will not challenge then B would prefer to announce "low" when she received signal θ'_B. Therefore there does not exist a truthfully revealing equilibrium in pure strategies.

2.3.3 Mixed strategies and Bayesian updating

One might wonder if the truthful revelation outcome can be approximated by a mixed equilibrium, in the way that the pure-strategy Stackelberg equilibrium can be approximated by a mixed equilibrium of a "noisy commitment game"(van Damme and Hurkens (1997)). We show below that this is not the case. Comparing their result with ours suggests that the assumption of common knowledge of payoffs is less robust to small changes than is the assumption of perfectly observed actions.

Specifically, denote by σ'_B, $\sigma'_B \in [0,1]$. the probability that B announces "low" after seeing signals θ'_B, and let σ''_B be the probability B announces "high" after seeing θ'_B , as in the following table:

	High	Low
θ'_B	$1-\sigma'_B$	σ'_B
θ''_B	σ''_B	$1-\sigma''_B$

The corresponding mixing probabilities for player S are

	Challenge	Don't Challenge
θ'_S	$1-\sigma'_S$	σ'_S
θ''_S	σ''_S	$1-\sigma''_S$

2.3.4 The result

Using the above payoff expressions, we will now show that the pure information equilibrium whereby the buyer announces the valuation truthfully, does not obtain as a limit of any equilibrium E_ε of the above imperfect information game as $\varepsilon \to 0$. More specifically:

Proposition 1 *For any fine F there is no sequence of totally mixed equilibrium strategies σ_B, σ_S such that $\sigma'_B, \sigma''_B, \sigma'_S$ and σ''_S all converge to 0 as $\varepsilon \to 0$.*

Proof. For the sake of presentation, here we prove the Proposition under the restriction that the challenging fine F is fixed (equal to 9 as in the above perfect information example), however this restriction is immaterial.

Suppose, contrary to the theorem, that as $\varepsilon \to 0$, there is a sequence of equilibria along which $\sigma'_B, \sigma''_B, \sigma'_S$ and σ''_S all converge to 0 as $\varepsilon \to 0$. Note that in this case

Consider the seller's decision whether or not to challenge the buyer when $\theta_S = \theta'_S$ and the buyer announces "low". Computations in the Appendix show that

$$\begin{aligned} V_S\left(C|\theta_S = \theta'_S, L\right) &= \delta(\varepsilon)[\alpha(\varepsilon)(-4) + (1-\alpha(\varepsilon))15] \\ &\quad +(1-\delta(\varepsilon))[\frac{1}{2}(-4) + \frac{1}{2}15], \end{aligned}$$

where

$$\begin{aligned} \delta(\varepsilon) &= \Pr\left(\theta_B = \theta'_B|\theta_S = \theta'_S, L\right) \\ &= \frac{\left(\frac{1}{2}(1-\varepsilon)^2 + \frac{1}{2}\varepsilon^2\right)\left(\sigma'_B(\varepsilon)\right)}{\left(\frac{1}{2}(1-\varepsilon)^2 + \frac{1}{2}\varepsilon^2\right)\left(\sigma'_B(\varepsilon)\right) + \varepsilon(1-\varepsilon)\left(1-\sigma''_B(\varepsilon)\right)} \end{aligned}$$

and

$$\begin{aligned} \alpha(\varepsilon) &= \Pr\left(v = 10|\theta'_B, \theta'_S\right) \\ &= 1 - \frac{\frac{1}{2}(1-\varepsilon)^2}{\frac{1}{2}(1-\varepsilon)^2 + \frac{1}{2}\varepsilon^2}. \end{aligned}$$

Note that $\alpha(\varepsilon) \to 0$ as $\varepsilon \to 0$.There are two cases to consider.

Case A: $\delta(\varepsilon) \to 0$.

In this case as $\varepsilon \to 0$ we have

$$V_S\left(C|\theta_S = \theta'_S, L\right) \to \frac{1}{2}(-4) + \frac{1}{2}15 < V_S\left(DC|\theta_S = \theta'_S, L\right) = 10.$$

Thus S does not challenge if the buyer announces "low" and $\theta_S = \theta'_S$.

Now consider the case where $\theta_S = \theta''_S$. We have

$$\begin{aligned} V_S\left(C|\theta_S = \theta''_S, L\right) &= m(\varepsilon)[\frac{1}{2}(-4) + \frac{1}{2}15] \\ &\quad +(1 - m(\varepsilon))[n(\varepsilon)(-4) + (1 - n(\varepsilon))15], \end{aligned}$$

where (see Appendix)

$$\begin{aligned} m(\varepsilon) &= \Pr\left(\theta_B = \theta'_B|\theta_S = \theta''_S, L\right) \\ &= \frac{\varepsilon(1-\varepsilon)\left(\sigma'_B(\varepsilon)\right)}{\varepsilon(1-\varepsilon)\left(\sigma'_B(\varepsilon)\right) + \left(\frac{1}{2}\varepsilon^2 + \frac{1}{2}(1-\varepsilon)^2\right)\left(1 - \sigma''_B(\varepsilon)\right)} \end{aligned}$$

and

$$\begin{aligned} n(\varepsilon) &= \Pr\left(v = 10|\theta'_B, \theta''_S\right) \\ &= 1 - \frac{\frac{1}{2}\varepsilon^2}{\frac{1}{2}\varepsilon^2 + \frac{1}{2}(1-\varepsilon)^2}. \end{aligned}$$

Thus $m(\varepsilon) \to 0$ and $n(\varepsilon) \to 1$ when $\varepsilon \to 0$. Thus in the limit, challenging yields -4 which is strictly less than the payoff (10) of not challenging. It follows that if $(\sigma'_B, \sigma''_B) \to (0, 0)$, then necessarily $(\sigma'_S, \sigma''_S) \to (1, 0)$ in equilibrium when $\varepsilon \to 0$, which contradicts the assumption that $\sigma'_S \to 0$.

Case B: $\delta(\varepsilon) \not\to 0$ and $\sigma'_S(\varepsilon) > 0$ for all ε.

It is possible that σ'_B can be of order ε, so that $\delta\left(\varepsilon\right)$ can be non-zero. Note that in such a case that, as $\varepsilon \to 0$, $V_S\left(C|\theta_S = \theta'_S, L\right) \to 5\frac{1}{2} + 9\frac{1}{2}\ \delta(\varepsilon)$.

Since $V_S\left(DC|\theta_S = \theta'_S, L\right) = 10$, if $\sigma'_S > 0$, it must be that

$$5\frac{1}{2} + 9\frac{1}{2}\delta(\varepsilon) = 10.$$

Thus along a sequence of equilibria $\sigma'_B(\varepsilon) \to 0, \sigma''_B(\varepsilon) \to 0, \sigma'_S(\varepsilon) \to 0$, $\sigma''_S(\varepsilon) \to 0$, and $\sigma^{'}_S(\varepsilon) > 0$, we know that $\lim_{\varepsilon \to 0} \delta(\varepsilon)$ must exist and equal 9/19. Because $\sigma''_B \to 0$, we have

$$\delta(\varepsilon) = \frac{\left(\frac{1}{2}(1-\varepsilon)^2 + \frac{1}{2}\varepsilon^2\right)(\sigma'_B(\varepsilon))}{\left(\frac{1}{2}(1-\varepsilon)^2 + \frac{1}{2}\varepsilon^2\right)(\sigma'_B(\varepsilon)) + \varepsilon(1-\varepsilon)},$$

so

$$\lim_{\varepsilon \to 0} \delta(\varepsilon) = \lim_{\varepsilon \to 0} \frac{\sigma'_B(\varepsilon)}{\sigma'_B(\varepsilon) + 2\varepsilon}$$

and thus $\lim_{\varepsilon \to 0} \delta(\varepsilon) = 9/19$ requires that $\lim_{\varepsilon \to 0} \sigma'_B(\varepsilon)/\varepsilon = 9/5$, . so in particular $\sigma^{'}_B(\varepsilon) > 0$. However, $\lim_{\varepsilon \to 0} V_B (H | \theta_B = \theta'_B) = 0$, while $\sigma'_S(\varepsilon) \to 0$ implies that $\lim_{\varepsilon \to 0} V_B (L | \theta_B = \theta'_B) = -1$, so for small ε, the buyer strictly prefers to announce H, a contradiction.

To keep B to be indifferent between announcing high and low given that she saw signal $\theta_B = \theta'_B$ requires

$$4\sigma'_S = (1 - \sigma'_S),$$

as $\varepsilon \to 0$, i.e. $\sigma'_S \to 1/5$. This contradicts the supposition that $\sigma'_S \to 0$. ■

2.4 Discussion of the example

It is easy to show that he uniform prior of $p = 1/2$ is essential for Proposition 1 when the mechanism designer can choose any value of F (i.e. potentially greater than $F = 9$ as in the example). If $p > 1/2$ (i.e. the good being high value has greater prior probability) then in this example F can be chosen sufficiently large so as to induce the seller to challenge when she observes the high signal but B announces "low".

Similarly, even if $p = 1/2$, one could amend the example to include a different fine[7] at stage 3d than the one at stage 3a (i.e. B and S pay different fines depending on whether B accepts the good at stage 3b). If the fine B pays is sufficient large relative to F then the conclusions of Proposition 1 do not hold (e.g. if B pays $F = 30$ if challenged and S pays

[7] We thank Ivan Werning for suggesting this possibility.

$F = 15$ if B subsequently accepts).

We return to both of these issues when discussing the general mechanism in the next section. As it turns out, neither asymmetric fines nor large fines will lead to approximately truthful revelation with almost perfect information in the general Moore-Repullo mechanism.

As we mentioned in the introduction, this Hart-Moore-Repullo example is sufficiently simple that a 2-stage mechanism in which each player acts only once can achieve approximate efficiency.

A notable feature of the example, or more precisely the statement of Proposition 1, is that the sequence of equilibrium strategies we rule out involves $\sigma'_S \to 0$ and $\sigma''_S \to 0$. In the context of the example this may, at first sight, appear rather odd: for if B announces truthfully then S does not make a move and hence σ'_S and σ''_S are moot on the equilibrium path. In the more general environment considered below, however, such considerations are crucial.

This also relates to the fact that while there are no approximately truthful totally mixed equilibria, it is possible to construct such an equilibrium where the low type of buyer and seller both play pure strategies, but the high types both mix[8]. Again, in the general environment this is not possible.

3 More general Moore-Repullo mechanisms

Moore and Repullo (1988) offer a class of mechanisms which, with complete information, work well in very general environments. They also outline a substantially simpler mechanism which yields truth telling in environments where there is transferable utility. Since this is the most hospitable environment for subgame perfect implementation, and because most incomplete contracting settings are in economies with money, we shall focus on it.

[8] We are grateful to Johannes Horner and Andy Skrzypacz for pointing this out.

3.1 Setup

Let Ω be the (finite) set of possible states of nature[9]. Let there be two agents: 1 and 2, whose preferences over a social decision $d \in D$ are given by $\omega_i \in \Omega_i$ for $i = 1, 2$. Let $\Omega_i = \{\omega_i^1, ..., \omega_i^n\}$. The agents have utility functions as follows:

$$u_1(d, \omega_1) - t_1,$$
$$u_2(d, \omega_2) + t_2$$

where d is a collective decision, t_1 and t_2 are monetary transfers. The agent's ωs are common knowledge among each other (but not "publicly" known in the sense that the third party introduced below does not know the agents ωs).

Let $f = (D, T_1, T_2)$ be a social choice function where for each $(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2$ the social decision is $d = D(\omega_1, \omega_2)$ and the transfers are $(t_1, t_2) = (T_1(\omega_1, \omega_2), T_2(\omega_1, \omega_2))$.

Moore and Repullo (1988) propose the following class of mechanism, which we shall refer to as "the MR mechanism." There is one phase for each agent and each phase consists of three stages. The game begins with phase 1, in which agent 1 announces a value ω_1 as we now outline.

1. Agent 1 announces a preference ω_1, and we proceed to stage 2.
2. If agent 2 agrees then the phase ends and we proceed to phase 2. If agent 2 does not agree and "challenges" by announcing some $\phi_1 \neq \omega_1$, then we proceed to stage 3.
3. Agent 1 chooses between

$$\{d; t_1\}(\omega_1) = \{x; t_x + \Delta\}$$

and

$$\{d; t_1\}(\phi_1) = \{y; t_y + \Delta\},$$

[9] Moore and Repullo (1988) allow for an infinite space but impose a condition bounding the utilitiy functions which is automatically satisfied in the finite case.

where these functions are specified by the mechanism such that

$$u_1(x,\omega_i) - t_x > u_1(y,\omega_1) - t_y$$

and

$$u_1(x,\phi_1) - t_x < u_1(y,\phi_1) - t_y.$$

Also, if agent 1 chooses $\{x; t_x + \Delta\}$, then agent 2 receives $t_2 = t_x - \Delta$ (and a third party receives 2Δ). If, however, agent 1 chooses $\{y; t_y + \Delta\}$ then agent 2 receives $t_2 = t_y + \Delta$.

Phase 2 is the same as phase 1 with the roles of players 1 and 2 reversed, i.e. agent 2 announces a ω_2. We use the notation stage 1.2, for example, to refer to phase 1, stage 2.

The Moore-Repullo logic is as follows. If agent 1 lied at stage 1.1 then agent 2 could challenge with the truth and then at stage 1.3 agent 1 will find it optimal to choose $\{y; t_y + \Delta\}$. If Δ is sufficiently large then this will be worse for agent 1 than telling the truth and having the choice function f implemented. Moreover, agent 2 will be happy with receiving $t_y + \Delta$. If agent 1 tells the truth at stage 1.1 then agent 2 will not challenge because she knows that agent 1 will choose $\{x; t_x + \Delta\}$ at stage 1.3 which will cause agent 2 to pay the fine of Δ.

3.2 Perturbing the information structure

We now show that this result does not hold for a small perturbation of the information structure. Consider the following information structure. For each agent's preferences there is a separate signal structure with n signals. For agent 1's preferences recall that the states are $\omega_1^1, ..., \omega_1^n$. The n signals are $\theta_1^1, ..., \theta_1^n$. The conditional probability of signal θ_1^j given state ω_1^j given is $1-\varepsilon$, and the probability of each signal θ_1^j conditional on state $k \neq j$ is $\varepsilon/(n-1)$. Similarly, for agent 2's preferences the states are $\omega_2^1, ..., \omega_2^n$. The n signals are $\eta_2^1, ..., \eta_2^n$. The conditional probability of state ω_2^j given signal η_2^j is $1-\varepsilon$, and the probability of each state $k \neq j$ conditional on signal η_2^j is $\varepsilon / (n-1)$. The following table illustrates this.

[TABLE 1 HERE]

The timing is as follows. Nature chooses a payoff parameter for each player from a uniform distribution. Then each player simultaneously and privately observes a conditionally independent signal from the above signal structure about player 1's preferences. They then play phase 1 of the MR mechanism to elicit player 1's preferences. They then simultaneously and privately observe a conditionally independent signal from the above signal structure about player 2's preferences. Then they play phase 2 of the MR mechanism to elicit player 2's preferences[10].Denote the probability that agent 1 announces θ_1^j conditional on seeing signal θ_1^k as σ_k^j. Similarly let the probability the agent 2 announces ϕ_j (at stage 2) conditional on observing signal θ_1^k be μ_k^j. In the second phase of the mechanism (designed to elicit agent 2's preferences) the corresponding mixing probabilities are as follows. The probability that agent 2 announces θ_2^j conditional on seeing signal θ_2^k is ρ_k^j and the probability the agent 1 announces ϕ_j (at stage 2) conditional on observing signal θ_2^k is τ_k^j.

Theorem 1 *Suppose that the agents' beliefs are formed according to the above signal structure. Then there exists a social choice function f such that there is no profile of totally mixed equilibrium strategies $\left\{\sigma_k^j, \mu_k^j, \rho_k^j, \tau_k^j\right\}$ such that $\sigma_j^j \rightarrow 1, \rho_j^j \rightarrow 1$ and $\sigma_k^j \rightarrow 0, \rho_k^j \rightarrow 0$ for all $k \neq j$.*

Proof. See appendix. ■

Remark 1 *If the strategies are not totally mixed then there is no guarantee that any particular $\sigma_\ell^k > 0$, and hence the above expression for $\delta(\varepsilon)$ may not be well defined. In other*

[10] One could also imagine the players receiving both signals and then playing the two phases of the mechanism. This would complicate the analysis because it would expand the number of payoff parameters for each player.

words, Bayes Rule offers no guide as to beliefs in this case. Consider, however, two sets of beliefs in such circumstances: (i) that if no type of player 1 *announces* $\hat{\theta}_1 = \theta_1^k$ *then such an announcement is considered to be truthful; or (ii) that beliefs about* $\hat{\theta}_1$ *are uniformly distributed. In the first case* $\Pr\left(\theta_1 = \theta_1^j \middle| \theta_2 = \theta_2^j, \hat{\theta}_1 = \theta_1^k\right) = 0 = \delta(\varepsilon)$. *In the second* $\sigma_j^k = 1/n$ *for all* k, *and therefore* $\lim_{\varepsilon->0} \delta = 0$, *which is the conclusion we obtain when Bayes Rule is applicable.*

The difficulty which arises under almost perfect information is that player 1 can announce a state which is not the one "suggested" by her signal and have player 2 not challenge. After seeing the likely signal and a different announcement from player 1, player 2 believes that there is now only a 50:50 chance that the actual state is consistent with her signal. She then believes that if she challenges half the time she will receive the fine of Δ, but half the time she will pay it. This eliminates the role of the fine which was crucial to the mechanism under perfect information. This in turn allows player 1 to announce whichever signal will lead to the best social choice function for her. If her preferences are aligned with player 2's then she will announce truthfully, but if not she will not. Thus, in general, not all social choice functions can be implemented under almost perfect information.

The Hart-Moore-Repullo buyer-seller example is a simple setting in which preferences are clearly not aligned. There are always gains from trade, so the social decision is that there be trade. But regardless of the quality of the good, the buyer would prefer to pay 10 for it, not 14. The seller obviously prefers to receive 14, no matter what the quality. We suggest that such conflict is common in the settings where Property Rights Theory has proved useful, and therefore that 3-stage mechanisms may not lead to private information being revealed.

Given the fact that the role of the fine is eliminated because Δ is received by player 2 (say) with probability 1/2 upon challenging, but also paid with probability 1/2, one naturally wonders why an asymmetric fine (whereby player 2 pays or receives different amount depending on the choice of player 1) works. In the example of section 2 this worked because if B announced "high" then S had no right to challenge. In the general MR mechanism, however, it is (necessarily) the case that player 2 can challenge any announcement that player 1 makes. Consider modifying the MR mechanism so that the final part of stage 3 reads as follows: " if agent 1 chooses $\{x; t_x + \Delta_1\}$, then agent 2 receives $t_2 = t_x - \Delta_1$. If, however, agent 1 chooses $\{y; t_y + \Delta_2\}$ then agent 2 receives $t_2 = t_y + \Delta_2$." Following the same reasoning as in the proof of Theorem 2, when player 1 announces something other than θ_1^j the payoff as $\varepsilon \to 0$ to player 2 from challenging is now

$$\left(\begin{array}{c} \frac{1}{2}\left(\frac{1}{n}\sum_{i=i}^{n} u_2\left(y, \omega_2^i\right) + t_y + \Delta_2\right) \\ +\frac{1}{2}\left(\frac{1}{n}\sum_{i=i}^{n}\left(u_2\left(x, \omega_2^i\right)\right) + t_x - \Delta_1\right) \end{array} \right).$$

By making Δ_2 large relative to Δ_1 a challenge can be encouraged. Unfortunately this may also make player 2 challenge player 1 when she announces truthfully, as we illustrate by example below.

3.3 An example

We now provide an example which illustrates two points: first, that asymmetric fines do not help matters, and second that there are very natural social choice functions in simple settings which cannot be implemented by totally mixed equilibria in the setting with imperfect information[11]. As an illustration of this suppose that $D = \{N, Y\}$, with the interpretation that $d = Y$ is the decision to provide a public good and $d = N$ is not to provide it. Let $u_1 = \beta_1 d + t_1$ and $u_2 = \beta_2 d + t_2$ with $\beta_i \in \{\beta^L, \beta^H\}$ for $i = 1, 2$ with $0 = \beta^L < \beta^H$. The betas have the interpretation of being the utility derived from the public good net of its production cost. The signal structure for *each* player is as follows

	$\theta_1'\theta_2'$	$\theta_1'\theta_2''$	$\theta_1''\theta_2'$	$\theta_1''\theta_2''$
β_i^H	$\frac{1}{2}\left(1-\varepsilon\right)^2$	$\frac{1}{2}\left(1-\varepsilon\right)\varepsilon$	$\frac{1}{2}\varepsilon\left(1-\varepsilon\right)$	$\frac{1}{2}\varepsilon^2$
β_i^L	$\frac{1}{2}\varepsilon^2$	$\frac{1}{2}\left(1-\varepsilon\right)\varepsilon$	$\frac{1}{2}\varepsilon\left(1-\varepsilon\right)$	$\frac{1}{2}(1-\varepsilon)^2$

[11] This is adapted from Bolton and Dewatripont (2005), pp.558-559.

The social choice function we would like to implement involves $d = 1$ if and only if $\beta_1 + \beta_2 > 0$, with associated transfers such that $\beta_1 + t_1 = \beta_2 + t_2$. That is, provide the good if and only if it has aggregate benefit and equate payoffs.

The first phase of the mechanism involves eliciting player 1's preferences, β_1. Let the probability that agent 1 announces β^L conditional on seeing signal θ_1' as σ_1' and the probability that she announces β^H conditional on seeing signal θ_1'' as σ_1''. Let the probability that agent 2 challenges be q. An equilibrium in which agent 1 truthful reveals and is not challenged involves a sequence of strategies such that $\sigma_1' \to 0, \sigma_1'' \to 0$ as $\varepsilon \to 0$. We will again consider totally mixed strategies.

The MR mechanism for this phase involves agent 1 announcing β_1 and then agent 2 challenging or not by announcing $\hat{\beta}_1 \neq \beta_1$. If agent 2 does not challenge then agent 1's preference is deemed to be β_1. If agent 2 challenges then agent 1 pays Δ_1 to the third party and then agent 1 chooses between the social choice functions

$$\left(d = N, t_N - \Delta_1, -t_N - \Delta_1\right),$$

and

$$\left(d = Y, t_Y - \Delta_1, -t_Y + \Delta_2\right),$$

such that

$$t_N > \beta_1 + t_Y,$$

and

$$t_N < \hat{\beta}_1 + t_Y.$$

Again we assume that if a challenge occurs agent 1 subsequently learns her true preference. Suppose by way of contradiction that $\left(\sigma_1', \sigma_1''\right) \to (0, 0)$. The payoff to agent 2 from challenging given that she observed signal θ_2' is

$$V_2\left(C|\theta_2=\theta_2',\beta_1^L\right)=\Pr\left(\theta_1=\theta_1'|\theta_2=\theta_2',\beta_1^L\right)[K]$$
$$+\left(1-\Pr\left(\theta_1=\theta_1'|\theta_2=\theta_2',\beta_1^L\right)\right)\left[\frac{1}{2}\left(-t_N-\Delta_1\right)+\frac{1}{2}\left(-t_Y+\Delta_2\right)\right]$$

The calculation of $\Pr\left(\theta_1=\theta_1'|\theta_2=\theta_2',\beta_1^L\right)$ is identical to the case considered in Proposition 1 (see section 6.1 of the appendix for these calculations) and hence $\lim_{\varepsilon\to 0}\Pr\left(\theta_1=\theta_1'|\theta_2=\theta_2',\beta_1^L\right)=0$, given the supposition that $(\sigma_1',\sigma_1'')\to(0,0)$. This means that the value of K is immaterial. Thus

$$\begin{aligned}V_2\left(C|\theta_2=\theta_2',\beta_1^L\right) &= \frac{1}{2}\left(-t_N-\Delta_1\right)+\frac{1}{2}u_2\left(d=1,-t_Y+\Delta_2\right).\\ &= \frac{1}{2}\left(-t_N-\Delta_1\right)+\frac{1}{2}\left(\frac{1}{2}\beta^H-t_Y+\Delta_2\right),\end{aligned}$$

where the last line comes from the fact that player 2 has a 50:50 chance of being type β^H.

The value to agent 2 of not challenging is

$$\begin{aligned}V_2\left(DC|\theta_2=\theta_2',\beta_1^L\right) &= \frac{1}{2}\left(\beta^H-\frac{\beta^H}{2}\right)\\ &= \frac{1}{4}\beta^H.\end{aligned}$$

since the social choice function specifies that the project be built if player 2's preference is $\beta_2=\beta^H$ given that $\beta_1=\beta^L$, agent 2 pays $t_2=\beta^H/2$. This in turn happens with probability 1/2 in a truthful equilibrium in phase 2. Thus to ensure a challenge requires

$$\frac{1}{2}\left(-t_N-\Delta_1\right)+\frac{1}{2}\left(\frac{1}{2}\beta^H-t_Y+\Delta_2\right)>\frac{1}{4}\beta^H, \qquad (1)$$

When $\theta_2=\theta_2''$ agent 2 will not challenge an announcement of β_1^L (the calculations are identical to those for proposition 1 in the appendix). Thus in order to have $(\sigma_2'\,\sigma_2'')\to(0,0)$ we require inequality (1) to hold.

Now suppose $\theta_2 = \theta_2'$ and agent 1 announces β_1^H. The payoff to agent 2 from not challenging is

$$\begin{aligned} V_2\left(DC|\theta_2 = \theta_2', \beta_1^H\right) &= \frac{1}{2}\left(\beta^H - \frac{\beta_H - \beta_H}{2}\right) - \frac{1}{2}\frac{\beta_H}{2} \\ &= \frac{1}{4}\beta^H. \end{aligned}$$

The payoff from challenging is

$$\begin{aligned} V_2\left(C|\theta_2 = \theta_2', \beta_1^H\right) &= \Pr\left(\theta_1 = \theta_1'|\theta_2 = \theta_2', \beta_1^H\right)\left[\begin{array}{c} \Pr\left(\theta_2 = \theta_2'|\theta_1 = \theta_1', \beta_1^H, C\right)\left(-t_N - \Delta_1\right) \\ +\Pr\left(\theta_2 = \theta_2''|\theta_1 = \theta_1', \beta_1^H, C\right) \\ \cdot u_2\left(d = 1, -t_Y + \Delta_2\right) \end{array}\right] \\ &\quad + \left(1 - \Pr\left(\theta_1 = \theta_1'|\theta_2 = \theta_2', \beta_1^H\right)\right)\left[K'\right], \end{aligned}$$

where $\Pr\left(\theta_2 = \theta_2'|\theta_1 = \theta_1', \beta_1^H, C\right)$ is the posterior probability that agent 1 assigns to agent 2 having observed the high signal given that she (agent 1) saw the high signal and announced truthfully but was challenged. The calculation of $\Pr\left(\theta_1 = \theta_1'|\theta_2 = \theta_2', \beta_1^H\right)$ is identical to the case considered in Proposition 1 (see section 5.1 of the appendix for these calculations) and hence $\lim_{\varepsilon \to 0} \Pr\left(\theta_1 = \theta_1'|\theta_2 = \theta_2', \beta_1^H\right) = 1$, given the supposition that $(\sigma_1', \sigma_1'') \to (0, 0)$. This means that the value of K' is immaterial. Note that the calculation of agent 1's posterior is identical to that in the proof of Theorem 2 and hence

$$\lim_{\varepsilon \to 0} \Pr\left(\theta_2 = \theta_2'|\theta_1 = \theta_1', \beta_1^H, C\right) = \frac{1}{2}.$$

Thus with probability 1/2 agent 1 will choose $(d = N, t_N - \Delta_1, -t_N - \Delta_1)$ and with probability 1/2 will choose $(d = Y, t_Y - \Delta_1, -t_Y + \Delta_2)$[12]. Thus

$$\begin{aligned} V_2\left(C|\theta_2 = \theta_2', \beta_1^H\right) &= \frac{1}{2}\left(-t_N - \Delta_1\right) + \frac{1}{2}u_2\left(d = Y, -t_Y + \Delta_2\right) \\ &= \frac{1}{2}\left(-t_N - \Delta_1\right) + \frac{1}{2}\left(\frac{1}{2}\beta^H - t_Y + \Delta_2\right). \end{aligned}$$

So to deter a false challenge requires

$$\frac{1}{2}\left(-t_N - \Delta_1\right) + \frac{1}{2}\left(\frac{1}{2}\beta^H - t_Y + \Delta_2\right) < \frac{1}{4}\beta^H,$$

which contradicts (1).

4 More general extensive-form mechanisms

4.1 Common p-belief

We have so far restricted attention to Moore-Repullo mechanisms, and one may naturally wonder whether a different extensive-form mechanism could lead to truthful revelation as the unique equilibrium outcome when allowing for (small) p-belief perturbations. As a first step towards answering this question, in this section we consider a simple two-player-two-state example but allow for fairly general extensive-form mechanisms.

Thus, suppose there are two players: 1 and 2, and two states of the world A and B. The players have a common prior that each state is equally likely. Each player receives a conditionally independent signal from the symmetric signal structure we have considered earlier in the paper so that when ε is small the true state of the world is common p-belief for p near 1. Consider any mechanism in the following class: there are three stages, and in

[12] Here we assume, as in the first example, that in the event of a false challenge agent 1 learns the true state at stage 3. Again, we could give here a 50:50 chance of making a take-it-or-leave-it offer with her information at the time without altering the conclusion.

each stage one player acts. Without loss of generality assume that each player's action set is identical to her signal. That is, she may announce A or B. Again, without loss of generality, assume that player 1 acts at time 1, then player 2 acts at time 2, then player 1 acts again at time 3.

Observe that truthful revelation of the state of the world ω requires one of the following: (a) player 1 announces ω at time 1 and player 2 announces ω at time 2, or (b) player 2 announces ω at time 2 and player 1 announces ω at time 3.

Denote the messages sent by the players at times 1,2 and 3 as: m_1, m_2 and m_3, and let the payoff to player i be $v_i(m_1, m_2, m_3)$. For simplicity we will restrict attention to pure strategies.

With two states of nature, equal prior $pr(A) = pr(B) = 1/2$, and our assumed signal structure, then when player 2 observes the true state, anticipates player 1 to announce truthfully but player 1 lies at time 1, then player 2's posterior belief is simply back her prior belief 1/2.

Finally, to fix ideas, suppose the true state of the world is A. In the complete information game, incentive compatibility for player 2 requires $v_2(B, A, A) > v_2(B, B, \cdot)$, (where $v_2(B, B, \cdot)$ indicates that once the two players agree the game optimally ends) and incentive compatibility for player 1 requires $v_1(B, A, A) > v_1(B, A, B)$. We will work with limiting payoffs as the amount of informational asymmetry goes to zero in the imperfect information game. If both players report truthfully, then player 2's expected payoff (after receiving the signal that the state of the world is A) from announcing A after player 1 lied is

$$\frac{1}{2}v_2(B, A, A) + \frac{1}{2}v_2(B, A, B) \tag{2}$$

whereas her expected payoff from announcing B is

$$\frac{1}{2}v_2(B, B, A) + \frac{1}{2}v_2(B, B, \cdot). \tag{3}$$

To see this, note that when player 2 receives the signal that the state is A, but believes that player 1 saw a signal that the state is B, her posterior is 1/2.

Incentive compatibility thus requires

$$\frac{1}{2}v_2(B,A,A) + \frac{1}{2}v_2(B,A,B) > \frac{1}{2}v_2(B,B,A) + \frac{1}{2}v_2(B,B,\cdot) \tag{4}$$

Since we have $v_2(B,A,A) > v_2(B,B,\cdot)$ (from above), a necessary condition for (4) to hold is

$$v_2(B,A,B) > v_2(B,A,A) \tag{5}$$

But then if the true state of the world is B and player 1 announces truthfully, player 2 will announce A–since incentive compatibility for player 1 requires $v_1(B,A,B) > v_1(B,A,A)$– which in turn contradicts truthful revelation. This establishes the following:

Proposition 2 *Consider a three stage mechanism which implements truth-telling under perfect information where two players act sequentially and there are two states of nature. Then, no common p-belief perturbation of that mechanism, with* $p < 1$, *can induce truthful revelation in pure strategies in both states.*

In parallel work, Kunimoto and Tercieux (2009) show that only Maskin-monotonic social choice functions can be implemented in the closure of the sequential equilibrium correspondence. This in turn suggests that our analysis in this section could be extended to an arbitrary number of states of nature and to allow for mixed strategies.

4.2 Crazy types

Following Fudenberg, Kreps and Levine (1988) (FKL), we now consider the possibility that there might not be common knowledge of the payoffs at terminal nodes of the game induced by some extensive-form mechanism- this is more general in allowing small probabilities of e.g. the players having a preference for truthtelling or for making one false report instead of another. FKL essentially show that any Nash equilibrium can be “justified” when players entertain doubts about other players’ payoffs. Under such payoff uncertainty, therefore, no extensive-form mechanism can do better than implement Nash equilibria. The key advantage of extensive-form mechanisms is to implement social choice functions which are not Nash implementable, and to do so as a unique equilibrium. But, precisely because such mechanisms rely on refinements of Nash equilibrium, they are not robust to introducing payoff uncertainty.

We illustrate this point in the context of the example of Section 2. Figure 1 depicts the extensive form game induced by the Hart-Moore-Repullo example discussed in Section 2.

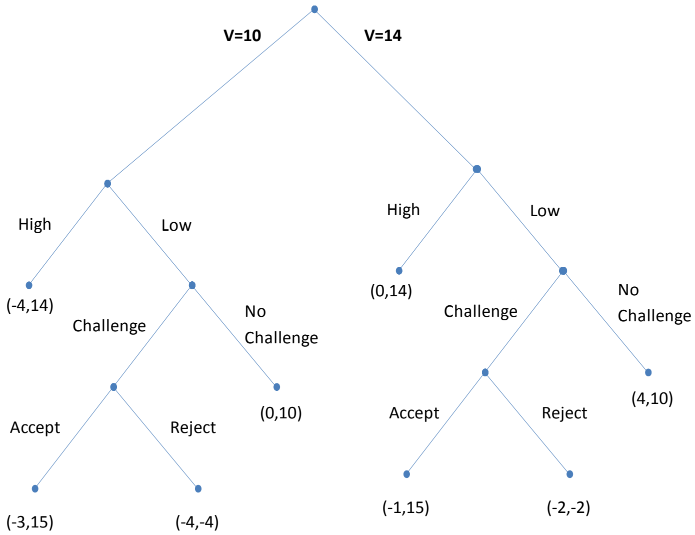

Figure 2: Hart-Moore-Repullo example

It is easy to see that when $v = 14$ the unique subgame perfect equilibrium is for the Buyer to announce “High”, achieving a payoff of 0, rather than announcing “Low”, being challenged by the Seller, then playing “Accept” and receiving a payoff of -1. However, by writing the game in normal form it is also easy to see that the outcome (Low, No Challenge), while not subgame perfect, is a Nash equilibrium (the bottom right entry in the left-hand payoff matrix below. Similarly, when $v = 10$, B announcing 14 is a Nash equilibrium, but not a subgame perfect equilibrium.

v=14	C	NC
14	$(0,14)^*$	$(0,14)$
A	$(-1,15)$	$(4,10)$
R	$(-2,-2)$	$(4,10)^*$

v=10	C	NC
14	$(-4,14)^*$	$(-4,14)$
A	$(-3,15)$	$(0,10)^*$
R	$(-4,-4)$	$(0,10)$

Figure 3: HMR example in normal form

But then one can appeal to Proposition 3 in Fudenberg, Kreps and Levine (1988), to conclude that there exists a "nearby game" (a *general elaboration* of this game) in which (Low, No Challenge) is a (subgame perfect) equilibrium.

5 Subgame perfect implementation and the hold-up problem

5.1 Hold-up without subgame perfect implementation but with common p-belief

In this section, we abstract from subgame implementation mechanisms and analyze whether the hold-up problem itself may or may not be affected by introducing small amounts of private information. To introduce the discussion, consider first what would happen if we allow from deviations in the sense of common p-belief, by introducing (small) amounts of private information. Without common knowledge of payoffs, we know from Fudenberg, Kreps and Levine (1988)) and our above discussion that truth-telling cannot be robustly implemented through extensive-form mechanisms. Yet, this does not provide foundations to Grossman and Hart (1986)'s analysis of the hold-up problem, simply because the hold-up problem also disappears once we allow for such deviations. The argument is straightforward: the hold-up problem involves a sequential game whose solution concept is a refinement of Nash equilibrium. FKL show that such refinements are not robust to deviations from common knowledge of payoffs in the following sense: all (pure strategy) Nash equilibria are the limit of strict (and hence sequential) equilibrium for an arbitrarily small change to prior beliefs about payoffs. But the hold-up problem arises precisely because, at the investment stage, players anticipate behavior at the (subsequent) bargaining stage. This sequential rationality is not robust to particular deviations from common knowledge, as FKL show.

This last result is somewhat disquieting as it implies that the non-robustness of subgame perfect implementation to deviations from common p-belief, is not so important for the analysis of the hold-up problem: the problem itself disappears when moving from common p-belief, for example when introducing crazy types. For instance, consider the standard hold-up where the are two players (1 and 2) who each make an investment, and then they bargaining sequentially over the split of the surplus. Now suppose that player 1 is, with very small probability, a crazy type in the sense that he will never accept any offer unless the other player invested at the first-best level. When making an offer player 1, even if not the crazy type, could reject an offer and player 2 will update and believe with non-negligible probability that he is the crazy type. Anticipating this, that player 2 is better off investing at the first-best level, rather than shading. Hence there is no hold-up problem.

However, we now argue that the hold-up problem does not disappear when introducing small amounts of private information about the cost or value of the good. More generally, with only common p-belief perturbations of the information structure, the hold-up problem persists. To see this, consider the following hold-up example. Again, there is a (B)uyer and a (S)eller of a good. The value of the good to B can be high or low (values H and L respectively). S is endowed with one unit of the good and values it at zero. S can make an investment, i at cost $c(i)$, which affects the prior probability that the good will be high value. Assume that c is increasing and convex, that $c(0) = 0, c'(0) = 0$ and $c(1) = \infty$. After this investment is made, suppose that B and S each receive a conditionally independent signal about v, according to the signal structure used before, i.e:

	$\theta'_B\theta'_S$	$\theta'_B\theta''_S$	$\theta''_B\theta'_S$	$\theta''_B\theta''_S$
$\Pr(v=H)$	$i\left(1-\varepsilon\right)^2$	$i\left(1-\varepsilon\right)\varepsilon$	$i\varepsilon\left(1-\varepsilon\right)$	$i\varepsilon^2$
$\Pr(v=L)$	$\left(1-i\right)\varepsilon^2$	$\left(1-i\right)\left(1-\varepsilon\right)\varepsilon$	$\left(1-i\right)\varepsilon\left(1-\varepsilon\right)$	$\left(1-i\right)\left(1-\varepsilon\right)^2$

The timing is as follows: S chooses i, B and S simultaneously observe their signals about v, then B and S bargaining over the transfer of the good according a bargaining protocol where each player has a 50:50 chance of being able to make a take-it-or-leave-it offer.

The first-best involves S solving

$$\max_i \left\{iH + \left(1-i\right)L - c\left(i\right)\right\}$$

The first-order condition is

$$\left(H-L\right) = c'\left(i^{FB}\right).$$

Now consider the second-best. If B gets to make the offer then he clearly offers zero and gets the good. If S gets to make the offer and saw signal θ', then her posterior belief about v is

$$\begin{aligned}\Pr\left(v=H|\theta_B=\theta'\right) &= \frac{i\left(1-\varepsilon\right)^2 + i\left(1-\varepsilon\right)\varepsilon}{i\left(1-\varepsilon\right)^2 + i\left(1-\varepsilon\right)\varepsilon + \left(1-i\right)\varepsilon^2 + \left(1-i\right)\left(1-\varepsilon\right)\varepsilon} \\ &= \frac{i-i\varepsilon}{\left(1-2i\right)\varepsilon + i}.\end{aligned}$$

And if she saw signal θ'' her posterior is

$$\begin{aligned}\Pr\left(v=H|\theta_B=\theta''\right) &= \frac{i\varepsilon\left(1-\varepsilon\right) + i\varepsilon^2}{i\varepsilon\left(1-\varepsilon\right) + i\varepsilon^2 + \left(1-i\right)\varepsilon\left(1-\varepsilon\right) + \left(1-i\right)\left(1-\varepsilon\right)^2} \\ &= \frac{i\varepsilon}{1-\left(1-2i\right)\varepsilon - i}.\end{aligned}$$

Now consider the following equilibrium for ε small but positive. If S gets the high signal she offers a price of

$$\frac{i-i\varepsilon}{\left(1-2i\right)\varepsilon + i}H, \tag{6}$$

and if she gets the low signal she offers a price of

$$\frac{i\varepsilon}{1-(1-2i)\,\varepsilon-i}L, \tag{7}$$

and B accepts price (6) if he got the high signal and price (7) if he got the low signal. S will not want to deviate from price (7) if she gets the low signal because if she raises it to price (6) B infers that S's signal was high, but her signal was almost surely low. So for ε small B's posterior belief is that v is close to $(H+L)/2$ and will reject the offer. Obviously, is S got the high signal she does not want to deviate to a lower price. Thus, we have an equilibrium.

Now consider the investment stage. S's expected payoff is

$$\frac{1}{2}0+\frac{1}{2}\left(iH+(1-i)\,L\right)-c\left(i\right).$$

The first-order condition for her maximization problem is therefore

$$\frac{1}{2}\left(H-L\right)=c'\left(i^{SB}\right).$$

By the convexity of c it follows that $i^{SB}<i^{FB}$, and hence the hold-up problem remains when introducing small amounts of private information about the value of the good.

5.2 Hold-up and the HM example

Finally, it is straightforward to show that the hold-up problem may no longer be solved by the MR mechanism once we introduce small amounts of private information about the good's valuation. To see this, let us simply introduce a stage prior to the mechanism considered in Section 2, where the Seller has the opportunity to make an investment which increases the probability that the good will be of high quality (i.e. that $v=14$). This is in the spirit of Che and Hausch (1999). Let S chooses investment i at cost $c(i)$, and let the $\Pr(v=14)=\beta i$.

The first-best benchmark involves maximizing total surplus from this investment. That is

$$\max_i \{\beta i 14 + (1-\beta i)10 - c(i)\}.$$

The first-order condition is

$$4\beta = c'(i).$$

Under the mechanism considered above the Seller solves the following problem for ε small

$$\max_i \left\{ \begin{array}{c} [\beta i(1-\Pr(L|v=\bar{v})) + (1-\beta i)\Pr(H|v=\underline{v})]\,14 \\ +\,[(1-\beta i)(1-\Pr(H|v=\underline{v})) + \beta i \Pr(L|v=\bar{v})]\,10 - c(i) \end{array} \right\},$$

where $\Pr(L|v=\bar{v})$ is the asymptotic probability that the buyer announces low when getting signal θ'_B and $\Pr(H|v=\underline{v})$ is the asymptotic probability that she announces high when getting signal θ''_B as $\varepsilon \to 0$.

Proposition 1 implies that at least one of these two probabilities remains bounded away from zero as $\varepsilon \to 0$. This in turn implies that the equilibrium investment under the above revelation mechanism, defined by the first-order condition

$$4\beta(1-\Pr(L|v=\bar{v}) - \Pr(H|v=\underline{v})) = c'(i),$$

remains bounded away from the first-best level of investment as $\varepsilon \to 0$.

Therefore, the Seller will not invest at the first-best level under non-integration of the Buyer and Seller. This is precisely in accordance with the conclusion of Grossman and Hart (1986).

6 Conclusion

Overall, our analysis provides some support for the Grossman-Hart-Moore approach to the hold-up problem. Namely, we started from a situation with perfect information where there is hold-up in the absence of a mechanism but the MR mechanism solves the problem; then we pointed to arbitrarily small deviations from perfect information about the good's valuation, for which the MR mechanism fails to induce truth-telling in pure or totally mixed strategies. We then argued that for a wide class of extensive-form mechanisms, there exist (arbitrarily small) deviations of this mechanism which involve common p-belief and therefore preserve the hold-up problem, and yet do not implement truth-telling as unique equilibrium in pure strategies.

If one allows for crazy types, in the sense of FKL, then there will generally be an equilibrium in which the hold-up problem disappears, and one where it is still present. In such settings one might still think of a potential role for asset ownership, namely as an equilibrium selection device and not as a device for providing incentives for specific investments.

References

[1] Aghion, P. M. Dewatripont and P. Rey. (1994), "Renegotiation Design with Unverifiable Information," *Econometrica* 62, 257-282.

[2] Borgers, T. (1994), "Weak Dominance and Almost Common Knowledge," *Journal of Economic Theory* 64, 265-276.

[3] Che, Y. and D. Hausch (1999), "Cooperative Investments and the Value of Contracting," *American Economic Review* 89, 125-147.

[4] Chung, K.S and J. Ely (2003), "Implementation with Near-Complete Information," *Econometrica* 71, 857-871.

[5] Cremer, J. and R.P. McLean (1988), "Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions", *Econometrica* 56, 1247-1257.

[6] van Damme, E. and S. Hurkens (1997), "Games with Imperfectly Observable Commitment," *Games and Economic Behavior* 21, 282-308.

[7] Dekel, E., D. Fudenberg and S. Morris, "Interim Correlated Rationalizability," *Theoretical Economics* 2, 15-40.

[8] Dekel, E. and D. Fudenberg (1990), "Rational Play Under Payoff Uncertainty," *Journal of Economic Theory* 52, 243-267.

[9] Farrell, J. and R. Gibbons (1989), "Cheap Talk Can Matter in Bargaining," *Journal of Economic Theory* 48, 221-237.

[10] Fudenberg, D., D. Kreps, and D.K. Levine (1988), "On the Robustness of Equilibrium Refinements," *Journal of Economic Theory* 44, 354-380.

[11] Fudenberg, D., D.K. Levine, and E. Maskin (1991), "Balanced-Budget Mechanisms for Adverse Selection Problems," unpublished working paper.

[12] Grossman, S, and O. Hart (1986), "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy* 94, 691-719.

[13] Hart, O. and J. Moore (1990), "Property Rights and the Nature of the Firm," *Journal of Political Economy* 98, 1119-1158.

[14] Hart, O. and J. Moore (2003), "Some (Crude) Foundations for Incomplete Contracts," unpublished working paper.

[15] Johnson, S, J.W. Johnson and R.J. Zeckhauser (1990), "Efficiency Despite Mutually Payoff-Relevant Private Information: The Finite Case," *Econometrica* 58, 873-900.

[16] Kunimoto, T. and O. Tercieux (2009), “Implementation with Near-complete Information: The Case of Subgame Perfection,” unpublished working paper.

[17] Maskin, E. “Nash Equilibrium and Welfare Optimality,” *Review of Economic Studies* 66, 23-38.

[18] Maskin, E. and J. Tirole (1999a), “Unforeseen Contingencies and Incomplete Contracts,” *Review of Economic Studies* 66, 83-114

[19] Maskin, E. and J. Tirole (1999b), “Two Remarks on the Property-Rights Literature,” *Review of Economic Studies* 66, 139-149.

[20] Monderer, D. and D. Samet (1988), “Approximating Common Knowledge with Common Beliefs,” *Games and Economic Behavior* 1, 170-190.

[21] Moore, J. (1992), “Implementation, contracts, and renegotiation in environments with complete information,” in J.J. Laffont (ed.), *Advances in Economic Theory: Sixth World Congress Vol 1,* 182-281.

[22] Moore, J. and R. Repullo (1988), “Subgame Perfect Implementation,” *Econometrica* 56, 1191-1220.

[23] Myerson, R.B. (1984). “Two-Person Bargaining Problems with Incomplete Information,” *Econometrica* 52, 461-487.

[24] Weinstein, J. and M. Yildiz. (2007). “A Structure Theorem for Rationalizability with Application to Robust Predictions of Refinements,” *Econometrica* 75, 365-400.

7 Appendix: Bayesian updating and ex post payoffs

Note: The following calculations are for the general case of prior probability of the good being high value of p*, as opposed to* $1/2$.

7.1 Preliminaries

In the derivation of posterior beliefs and ex post payoffs, we shall make use of the fact that B updates her beliefs about S's signal according to:

$$\begin{aligned}
\Pr\left(\theta_S=\theta'_S|\theta_B=\theta'_B\right) &= \frac{p(1-\varepsilon)^2+(1-p)\varepsilon^2}{p(1-\varepsilon)^2+(1-p)\varepsilon^2+\varepsilon(1-\varepsilon)}, \\
\Pr\left(\theta_S=\theta''_S|\theta_B=\theta'_B\right) &= \frac{\varepsilon(1-\varepsilon)}{p(1-\varepsilon)^2+(1-p)\varepsilon^2+\varepsilon(1-\varepsilon)}, \\
\Pr\left(\theta_S=\theta''_S|\theta_B=\theta''_B\right) &= \frac{p\varepsilon^2+(1-p)(1-\varepsilon)^2}{p\varepsilon^2+(1-p)(1-\varepsilon)^2+\varepsilon(1-\varepsilon)}, \\
\Pr\left(\theta_S=\theta'_S|\theta_B=\theta''_B\right) &= \frac{\varepsilon\left(1-\varepsilon\right)}{p\varepsilon^2+(1-p)(1-\varepsilon)^2+\varepsilon(1-\varepsilon)},
\end{aligned}$$

Similarly, a type θ'_S seller updates her beliefs about B's signal given her own signal and B's announcement, according to:

$$\begin{aligned}
\Pr\left(\theta_B=\theta'_B|\theta_S=\theta'_S,L\right) &= \frac{\left(p(1-\varepsilon)^2+(1-p)\varepsilon^2\right)\left(\sigma'_B\right)}{\left(p(1-\varepsilon)^2+(1-p)\varepsilon^2\right)\left(\sigma'_B\right)+\varepsilon(1-\varepsilon)\left(1-\sigma''_B\right)} \\
\Pr\left(\theta_B=\theta''_B|\theta_S=\theta'_S,L\right) &= \frac{\varepsilon(1-\varepsilon)\left(1-\sigma''_B\right)}{\varepsilon(1-\varepsilon)\left(1-\sigma''_B\right)+\left(p(1-\varepsilon)^2+(1-p)\varepsilon^2\right)\left(\sigma'_B\right)}.
\end{aligned}$$

The conditional probabilities for a type θ''_S seller, are:

$$\begin{aligned}
\Pr\left(\theta_B=\theta'_B|\theta_S=\theta''_S,L\right) &= \frac{\varepsilon(1-\varepsilon)\left(\sigma'_B\right)}{\varepsilon(1-\varepsilon)\left(\sigma'_B\right)+\left(p\varepsilon^2+(1-p)(1-\varepsilon)^2\right)\left(1-\sigma''_B\right)} \\
\Pr\left(\theta_B=\theta''_B|\theta_S=\theta''_S,L\right) &= \frac{\left(p\varepsilon^2+(1-p)(1-\varepsilon)^2\right)\left(1-\sigma''_B\right)}{\left(p\varepsilon^2+(1-p)(1-\varepsilon)^2\right)\left(1-\sigma''_B\right)+\varepsilon(1-\varepsilon)\left(\sigma'_B\right)}.
\end{aligned}$$

7.2 Buyer's ex post payoffs

Suppose $\theta_B = \theta_B'$. The value to B from announcing "high" when she receives signal θ_B' is

$$\begin{aligned}
V_B\left(H|\theta_B = \theta_B'\right) &= \Pr\left(\theta_S = \theta_S'|\theta_B = \theta_B'\right)\begin{pmatrix} (E[v|\theta_B',\theta_S'] - 14) \\ +\left(E[v|\theta_B',\theta_S'] - 14\right)\end{pmatrix} \\
&\quad + \Pr\left(\theta_S = \theta_S''|\theta_B = \theta_B'\right)\begin{pmatrix} \sigma_S''\left(E[v|\theta_B',\theta_S''] - 14\right) \\ +\left(1-\sigma_S''\right)\left(E[v|\theta_B',\theta_S''] - 14\right)\end{pmatrix} \\
&= \frac{p(1-\varepsilon)^2 + (1-p)\varepsilon^2}{p(1-\varepsilon)^2 + (1-p)\varepsilon^2 + \varepsilon(1-\varepsilon)}\begin{pmatrix} \left(\frac{p(1-\varepsilon)^2}{p(1-\varepsilon)^2+(1-p)\varepsilon^2}\right)14 \\ +\left(1-\frac{p(1-\varepsilon)^2}{p(1-\varepsilon)^2+(1-p)\varepsilon^2}\right)10\end{pmatrix} \\
&\quad + \frac{\varepsilon(1-\varepsilon)}{p(1-\varepsilon)^2 + (1-p)\varepsilon^2 + \varepsilon(1-\varepsilon)}\left(p14 + (1-p)10\right) - 14.
\end{aligned}$$

The value to B from announcing "low" when she receives signal θ_B' is

$$\begin{aligned}
V_B\left(L|\theta_B = \theta_B'\right) = \Pr\left(\theta_S = \theta_S'|\theta_B = \theta_B'\right)&\begin{pmatrix} \left(1-\sigma_S'\right)\begin{pmatrix} \Pr\left(v = 14|\theta_B',\theta_S'\right)(14-9-6) \\ +\Pr\left(v = 10|\theta_B',\theta_S'\right)(10-9-5)\end{pmatrix} \\ +\sigma_S'\left(E[v|\theta_B',\theta_S'] - 10\right)\end{pmatrix} \\
+ \Pr\left(\theta_S = \theta_S''|\theta_B = \theta_B'\right)&\begin{pmatrix} \sigma_S''\begin{pmatrix} \Pr\left(v = 14|\theta_B',\theta_S''\right)(14-9-6) \\ +\Pr\left(v = 10|\theta_B',\theta_S''\right)(10-9-5)\end{pmatrix} \\ +\left(1-\sigma_S''\right)\left(E[v|\theta_B',\theta_S''] - 10\right)\end{pmatrix} \\
= \frac{p(1-\varepsilon)^2 + (1-p)\varepsilon^2}{p(1-\varepsilon)^2 + (1-p)\varepsilon^2 + \varepsilon(1-\varepsilon)}&\begin{pmatrix} \left(1-\sigma_S'\right)\begin{pmatrix} \left(\frac{p(1-\varepsilon)^2}{p(1-\varepsilon)^2+(1-p)\varepsilon^2}\right)(14-9-6) \\ +\left(1-\frac{p(1-\varepsilon)^2}{p(1-\varepsilon)^2+(1-p)\varepsilon^2}\right)(10-9-5)\end{pmatrix} \\ +\sigma_S'\begin{pmatrix} \left(\frac{p(1-\varepsilon)^2}{p(1-\varepsilon)^2+(1-p)\varepsilon^2}\right)14 \\ +\left(1-\frac{p(1-\varepsilon)^2}{p(1-\varepsilon)^2+(1-p)\varepsilon^2}\right)10 - 10\end{pmatrix}\end{pmatrix} \\
+ \frac{\varepsilon(1-\varepsilon)}{p(1-\varepsilon)^2 + (1-p)\varepsilon^2 + \varepsilon(1-\varepsilon)}&\begin{pmatrix} \sigma_S''\left(p(14-9-6) + (1-p)(10-9-5)\right) \\ +\left(1-\sigma_S''\right)\left(p14 + (1-p)10 - 10\right)\end{pmatrix}.
\end{aligned}$$

To see where the payoffs come from recall that if B announces "high" then the mechanism specifies that she gets the good for 14. If she announces low and S does not challenge she gets the good for 10. If S does challenge then we assume that the true state of the good is revealed to both parties and we are therefore back in the complete information setting[13].

When $\theta_B = \theta''_B$ we have

$$\begin{aligned}
V_B\left(H|\theta_B = \theta''_B\right) &= \Pr\left(\theta_S = \theta'_S|\theta_B = \theta''_B\right)\begin{pmatrix} E[v|\theta''_B, \theta'_S] - 14 \\ +E[v|\theta''_B, \theta'_S] - 14 \end{pmatrix} \\
&\quad + \Pr\left(\theta_S = \theta''_S|\theta_B = \theta''_B\right)\begin{pmatrix} E[v|\theta''_B, \theta''_S] - 14 \\ +E[v|\theta''_B, \theta''_S] - 14 \end{pmatrix} \\
&= \frac{\varepsilon\left(1-\varepsilon\right)}{p\varepsilon^2 + (1-p)(1-\varepsilon)^2 + \varepsilon(1-\varepsilon)}\left(p14 + (1-p)10\right) \\
&\quad + \frac{p\varepsilon^2 + (1-p)(1-\varepsilon)^2}{p\varepsilon^2 + (1-p)(1-\varepsilon)^2 + \varepsilon(1-\varepsilon)}\left(p14 + (1-p)10\right) - 14,
\end{aligned}$$

and

$$V_B\left(L|\theta_B = \theta''_B\right) = \Pr\left(\theta_S = \theta'_S|\theta_B = \theta''_B\right)\begin{pmatrix} \left(1-\sigma'_S\right)\begin{pmatrix} \Pr\left(v = 14|\theta''_B, \theta'_S\right)(14-9-6) \\ +\Pr\left(v = 10|\theta''_B, \theta'_S\right)(10-9-5) \end{pmatrix} \\ +\sigma'_S E[v|\theta''_B, \theta'_S] - 10 \end{pmatrix}$$

$$+\Pr\left(\theta_S = \theta''_S|\theta_B = \theta''_B\right)\begin{pmatrix} \sigma''_S\begin{pmatrix} \Pr\left(v = 14|\theta''_B, \theta''_S\right)(14-9-6) \\ +\Pr\left(v = 10|\theta''_B, \theta''_S\right)(10-9-5) \end{pmatrix} \\ +\left(1-\sigma''_S\right)E[v|\theta''_B, \theta''_S] - 10 \end{pmatrix}$$

[13] This could be modified so that at the bargaining stage–in the spirit of Myerson (1984)–each player has a 50% chance of making a take-it-or-leave-it offer, using the information she has at that time. If B gets to make the offer she always offers zero, and if S gets to make the offer she offers a price equal to the posterior expectation of the value of the good conditional on her signal θ_S.

$$= \frac{\varepsilon\left(1-\varepsilon\right)}{p\varepsilon^2+(1-p)(1-\varepsilon)^2+\varepsilon(1-\varepsilon)} \left(\begin{array}{c} \left(1-\sigma_S'\right)\left(p(14-9-6)+(1-p)(10-9-5)\right) \\ +\sigma_S'\left(p14+\left(1-p\right)10\right)-10 \end{array}\right)$$

$$+\frac{p\varepsilon^2+(1-p)(1-\varepsilon)^2}{p\varepsilon^2+(1-p)(1-\varepsilon)^2+\varepsilon(1-\varepsilon)} \left(\begin{array}{c} \sigma_S'' \left(\begin{array}{c} \left(\frac{p\varepsilon^2}{p\varepsilon^2+(1-p)(1-\varepsilon)^2}\right)(14-9-6) \\ +\left(1-\frac{p\varepsilon^2}{p\varepsilon^2+(1-p)(1-\varepsilon)^2}\right)(10-9-5 \end{array}\right) \\ +\left(1-\sigma_S''\right) \left(\begin{array}{c} \left(\frac{p\varepsilon^2}{p\varepsilon^2+(1-p)(1-\varepsilon)^2}\right)14 \\ +\left(1-\frac{p\varepsilon^2}{p\varepsilon^2+(1-p)(1-\varepsilon)^2}\right) \end{array}\right)10 \end{array}\right).$$

7.3 Seller's ex post payoffs

The payoff to player S conditional on $\theta_S = \theta_S'$ and B announcing "high" is

$$V_S\left(\theta_S=\theta_S', H\right) = V_S\left(\theta_S=\theta_S'', H\right) = 14.$$

since the mechanism specifies that B gets the good for 14 when she announces "high".

In the equilibria that we consider in the text, the buyer is either exactly or approximately truthful, so there is positive probability that the buyer announces "low", and we can thus compute conditional payoffs on this event using Bayes rule.

The payoff for player S conditional on challenging when $\theta_S = \theta_S'$ and B announcing "low" is

$$V_S\left(C|\theta_S=\theta_S', L\right) = \Pr\left(\theta_B=\theta_B'|\theta_S=\theta_S', L\right) \left(\left(\begin{array}{c} \Pr\left(v=10|\theta_B',\theta_S'\right)(5-9) \\ +\Pr\left(v=14|\theta_B',\theta_S'\right)(9+6) \end{array}\right)\right)$$

$$+\Pr\left(\theta_B=\theta_B''|\theta_S=\theta_S', L\right) \left(\left(\begin{array}{c} \Pr\left(v=10|\theta_B'',\theta_S'\right)(5-9) \\ +\Pr\left(v=14|\theta_B'',\theta_S'\right)(9+6) \end{array}\right)\right)$$

$$= \frac{\left(p(1-\varepsilon)^2+(1-p)\varepsilon^2\right)\left(\sigma_B'\right)}{\left(p(1-\varepsilon)^2+(1-p)\varepsilon^2\right)\left(\sigma_B'\right)+\varepsilon(1-\varepsilon)\left(1-\sigma_B''\right)} \left(\left(\begin{array}{c} \left(1-\frac{p(1-\varepsilon)^2}{p(1-\varepsilon)^2+(1-p)\varepsilon^2}\right)(5-9) \\ +\left(\frac{p(1-\varepsilon)^2}{p(1-\varepsilon)^2+(1-p)\varepsilon^2}\right)(9+6) \end{array}\right)\right)$$

$$+\frac{\varepsilon(1-\varepsilon)\left(1-\sigma_B''\right)}{\varepsilon(1-\varepsilon)\left(1-\sigma_B''\right)+\left(p(1-\varepsilon)^2+(1-p)\varepsilon^2\right)\left(\sigma_B'\right)}\left(\left(\left(1-p\right)(5-9)+p\,(9+6)\right)\right).$$

The payoff for player S conditional on not challenging when $\theta_S = \theta'_S$ and B announcing "low" is

$$V_S\left(DC|\theta_S = \theta'_S, L\right) = 10$$

The payoff for player S conditional on challenging when $\theta_S = \theta''_S$ and B announces "low" is

$$\begin{aligned}
V_S\left(C|\theta_S = \theta''_S, L\right) &= \Pr\left(\theta_B = \theta'_B|\theta_S = \theta''_S, L\right)\left(\left(\begin{array}{c} \Pr\left(v = 10|\theta'_B, \theta''_S\right)(5-9) \\ +\Pr\left(v = 14|\theta'_B, \theta''_S\right)(9+6) \end{array}\right)\right) \\
&+\Pr\left(\theta_B = \theta''_B|\theta_S = \theta''_S, L\right)\left(\left(\begin{array}{c} \Pr\left(v = 10|\theta''_B, \theta''_S\right)(5-9) \\ +\Pr\left(v = 14|\theta''_B, \theta''_S\right)(9+6) \end{array}\right)\right) \\
&= \frac{\varepsilon(1-\varepsilon)\left(\sigma'_B\right)}{\varepsilon(1-\varepsilon)\left(\sigma'_B\right) + \left(p\varepsilon^2 + (1-p)(1-\varepsilon)^2\right)\left(1-\sigma''_B\right)}\left(\left(\left(1-p\right)(5-9) + p\left(9+6\right)\right)\right) \\
&+\frac{\left(p\varepsilon^2 + (1-p)(1-\varepsilon)^2\right)\left(1-\sigma''_B\right)}{\left(p\varepsilon^2 + (1-p)(1-\varepsilon)^2\right)\left(1-\sigma''_B\right) + \varepsilon(1-\varepsilon)\left(\sigma'_B\right)}\left(\left(\begin{array}{c} \left(1 - \frac{p\varepsilon^2}{p\varepsilon^2+(1-p)(1-\varepsilon)^2}\right)(5-9) \\ +\left(\frac{p\varepsilon^2}{p\varepsilon^2+(1-p)(1-\varepsilon)^2}\right)(9+6) \end{array}\right)\right).
\end{aligned}$$

The payoff for player S conditional on not challenging when $\theta_S = \theta''_S$ and B announces "low" is

$$V_S\left(DC|\theta_S = \theta''_S, L\right) = 10$$

7.4 Proof of Theorem 2

Suppose, by way of contradiction, that as $\varepsilon \to 0$, we have $\sigma^j_j \to 1$ and $\mu^j_j \to 1$. Now consider player 2's decision whether or not to challenge at stage 1.2, when player 1 announces something other than θ^j_1. By Bayes Rule, player 2's posterior belief that player 1 saw signal θ^j_1 given that player 2 saw signal θ^j_1 and that player 1 announced something other than θ^j_1 is

$$
\begin{aligned}
\delta(\varepsilon) &\equiv \Pr\left(\theta_1=\theta_1^j\middle|\,\theta_2=\theta_2^j,\hat{\theta}_1=\theta_1^k\right)=\frac{\Pr\left(\theta_1=\theta_1^j,\theta_2=\theta_2^j,\hat{\theta}_1=\theta_1^k\right)}{\Pr\left(\theta_2=\theta_2^j,\hat{\theta}_1=\theta_1^k\right)}\\
&= \frac{\Pr\left(\hat{\omega}_1=\omega_1^k\middle|\,\theta_1=\theta_1^j,\theta_2=\theta_2^j\right)\Pr\left(\theta_1=\theta_1^j,\theta_2=\theta_2^j\right)}{\sum_{\ell=1}^n\Pr\left(\hat{\omega}_1=\omega_1^k\middle|\,\theta_2=\theta_2^j,\theta_1=\theta_1^\ell\right)\Pr\left(\theta_1=\theta_1^\ell,\theta_2=\theta_2^j\right)}\\
&= \frac{\Pr\left(\hat{\omega}_1=\omega_1^k\middle|\,\theta_1=\theta_1^j,\theta_2=\theta_2^j\right)\Pr\left(\theta_1=\theta_1^j,\theta_2=\theta_2^j\right)}{\sum_{\ell=1}^n\Pr\left(\hat{\omega}_1=\omega_1^k\middle|\,\theta_2=\theta_2^j,\theta_1=\theta_1^\ell\right)\Pr\left(\theta_1=\theta_1^\ell,\theta_2=\theta_2^j\right)}\\
&= \frac{\sigma_j^k\Pr\left(\theta_1=\theta_1^j,\theta_2=\theta_2^j\right)}{\sum_{\ell=1}^n\Pr\left(\hat{\omega}_1=\omega_1^k\middle|\,\theta_2=\theta_2^j,\theta_1=\theta_1^\ell\right)\Pr\left(\theta_1=\theta_1^\ell,\theta_2=\theta_2^j\right)}\\
&= \frac{\sigma_j^k\Pr\left(\theta_1=\theta_1^j,\theta_2=\theta_2^j\right)}{\sum_{\ell=1}^n\Pr\left(\hat{\omega}_1=\omega_1^k\middle|\,\theta_1=\theta_1^\ell\right)\Pr\left(\theta_1=\theta_1^\ell,\theta_2=\theta_2^j\right)}\\
&= \frac{\sigma_j^k\left[\frac{1}{n}\left((1-\varepsilon)^2+(n-1)\left(\frac{\varepsilon}{n-1}\right)^2\right)\right]}{\begin{array}{c}\sigma_j^k\left[\frac{1}{n}\left((1-\varepsilon)^2+(n-1)\left(\frac{\varepsilon}{n-1}\right)^2\right)\right]\\ +\sum_{\ell\neq j}\sigma_\ell^k\left[\frac{1}{n}\left((1-\varepsilon)\frac{\varepsilon}{n-1}+\frac{\varepsilon}{n-1}(1-\varepsilon)+(n-2)\left(\frac{\varepsilon}{n-1}\right)^2\right)\right]\end{array}}.
\end{aligned}
$$

Also, let

$$
\begin{aligned}
\alpha_k(\varepsilon) &= \Pr\left(\omega_1=\omega_1^k|\theta_1=\theta_1^j,\theta_2=\theta_2^j\right),\ \text{for } k\neq j\\
&= 1-\frac{(1-\varepsilon)^2}{(1-\varepsilon)^2+(n-1)\frac{\varepsilon^2}{(n-1)^2}}.
\end{aligned}
$$

Finally let $\alpha(\varepsilon)=\sum_{k\neq j}\alpha_k(\varepsilon)$. Note that if player 1 indeed saw signal θ_1^j then at stage 1.3 with probability $1-\alpha(\varepsilon)$ she will choose $\{y;t_y+\Delta\}$ and with probability $\alpha(\varepsilon)$ she will choose $\{x;t_x+\Delta\}$. Under the former choice player 2 receives a transfer of $t_y+\Delta$ and under the latter choice she receives a transfer of $t_x-\Delta$.

The payoff to player 2 from challenging is therefore

$$\begin{aligned} V_2^C &= \delta\left(\varepsilon\right)\left[\begin{array}{c} \alpha\left(\varepsilon\right)\left(\frac{1}{n}\sum_{i=m}^{n}\left(u_2\left(x,\omega_2^m\right)\right)+t_x-\Delta\right) \\ +\left(1-\alpha\left(\varepsilon\right)\left(\frac{1}{n}\sum_{i=m}^{n}u_2\left(y,\omega_2^m\right)+t_y+\Delta\right)\right) \end{array}\right] \\ &\quad +\sum_{z\neq j}\Pr\left(\theta_1=\theta_1^z|\theta_2=\theta_2^j,\hat{\theta}_1=\theta_1^k\right) \\ &\quad \cdot\left(\begin{array}{c} \Pr\left(\omega_1=\omega_1^z|\theta_1=\theta_1^z,\theta_2=\theta_2^j\right)\left(\frac{1}{n}\sum_{i=1}^{n}u_2\left(y,\omega_2^i\right)+t_y+\Delta\right) \\ +\left(1-\Pr\left(\omega_1=\omega_1^z|\theta_1=\theta_1^z,\theta_2=\theta_2^j\right)\right)\left(\frac{1}{n}\sum_{i=1}^{n}\left(u_2\left(x,\omega_2^i\right)\right)+t_x-\Delta\right) \end{array}\right) \end{aligned}$$

Note that as $\varepsilon\to 0$, $\alpha\left(\varepsilon\right)\to 1$, and that given the supposition that $\sigma_j^j\to 1$ as $\varepsilon\to 0$ we have $\delta\left(\varepsilon\right)\to 0$ as $\varepsilon\to 0$, unless σ_j^k is of the order of ε. If it is then for a mixed strategy equilibrium to exist requires ρ_j^j and ρ_k^j such that player 1 is indifferent between announcing truthfully and not. This requires $\rho_j^j\neq 1$, which is a contradiction. For if $\rho_j^j=1$ (i.e. player 2 announces truthfully) then player 1's preference is accepted, but then player 1 plainly cannot be indifferent for all social choice functions $f=\left(D,T_1,T_2\right)$.

We thus return to the case where $\delta\left(\varepsilon\right)\to 0$ as $\varepsilon\to 0$, and note that

$$\begin{aligned} & \Pr\left(\omega_1=\omega_1^z\middle|\,\theta_1=\theta_1^z,\theta_2=\theta_2^j\right) \\ = & \frac{\Pr\left(\theta_1=\theta_1^z,\theta_2=\theta_2^j,\omega_1=\omega_1^z\right)}{\Pr\left(\theta_1=\theta_1^z,\theta_2=\theta_2^j\right)} \\ = & \frac{\Pr\left(\theta_1=\theta_1^z,\theta_2=\theta_2^j\middle|\,\omega_1=\omega_1^z\right)\Pr\left(\omega_1=\omega_1^z\right)}{\Pr\left(\theta_1=\theta_1^z,\theta_2=\theta_2^j\right)} \\ = & \frac{\frac{1}{n}\left(1-\varepsilon\right)\frac{\varepsilon}{n-1}}{\frac{1}{n}\left(2\left(1-\varepsilon\right)\frac{\varepsilon}{n-1}+\left(n-2\right)\left(\frac{\varepsilon}{n-1}\right)^2\right)} \\ = & \frac{1-\varepsilon}{2\left(1-\varepsilon\right)+\frac{n-2}{n-1}\varepsilon}, \end{aligned}$$

where the third equality holds by conditional independence of θ_1 and θ_2, and the fourth equality is derived as follows.

$$\begin{aligned}
\Pr\left(\theta_1=\theta_1^z,\theta_2=\theta_2^j\right) &= \sum_{k=1}^{n}\Pr\left(\theta_1=\theta_1^z,\theta_2=\theta_2^j\middle|\,\omega_1=\omega_1^k\right)\Pr\left(\omega_1=\omega_1^k\right)\\
&= \frac{1}{n}\sum_{k=1}^{n}\Pr\left(\theta_1=\theta_1^z\middle|\,\omega_1=\omega_1^k\right)\Pr\left(\theta_2=\theta_2^j\middle|\,\omega_1=\omega_1^k\right)\\
&= \frac{1}{n}\left(\begin{array}{c}\sum_{k\in\{\ell,j\}}\Pr\left(\theta_1=\theta_1^z\middle|\,\omega_1=\omega_1^k\right)\Pr\left(\theta_2=\theta_2^j\middle|\,\omega_1=\omega_1^k\right)\\ +\sum_{k\notin\{\ell,j\}}\Pr\left(\theta_1=\theta_1^z\middle|\,\omega_1=\omega_1^k\right)\Pr\left(\theta_2=\theta_2^j\middle|\,\omega_1=\omega_1^k\right)\end{array}\right)\\
&= \frac{1}{n}\left((1-\varepsilon)\frac{\varepsilon}{n-1}+\frac{\varepsilon}{n-1}(1-\varepsilon)+(n-2)\left(\frac{\varepsilon}{n-1}\right)^2\right)\\
&= \frac{1}{n}\left(2(1-\varepsilon)\frac{\varepsilon}{n-1}+(n-2)\left(\frac{\varepsilon}{n-1}\right)^2\right),
\end{aligned}$$

so that

$$\lim_{\varepsilon\to 0}\Pr\left(\omega_1=\omega_1^z\middle|\,\theta_1=\theta_1^z,\theta_2=\theta_2^j\right)=\lim_{\varepsilon\to 0}\frac{1-\varepsilon}{2(1-\varepsilon)+\frac{n-2}{n-1}\varepsilon}=\frac{1}{2}.$$

Therefore the payoff as $\varepsilon\to 0$ to player 2 from challenging is

$$\left(\begin{array}{c}\frac{1}{2}\left(\frac{1}{n}\sum_{i=i}^{n}u_2\left(y,\omega_2^i\right)+t_y+\Delta\right)\\ +\frac{1}{2}\left(\frac{1}{n}\sum_{i=i}^{n}\left(u_2\left(x,\omega_2^i\right)\right)+t_x-\Delta\right)\end{array}\right).$$

Note that the Δs cancel out which means we can no longer conclude that player 2 will be willing to challenge for all social choice functions f. That is, there exists an f such that the payoff from challenging is smaller than the payoff from not challenging, that being

$$\frac{1}{n}\sum_{i=1}^{n}\left(u_2\left(D\left(\hat{\omega}_1,\omega_2^i\right),\omega_2^i\right)+t_2\right).$$

Thus, player 2 will not necessarily challenge if she sees signal θ_2^j and player 1 announces $\omega_1^k, k\neq j$.

Now consider other signals that player 2 could observe. Note that by the construction of the signal structure

$$\Pr\left(\theta_1 = \theta_1^j | \theta_2 = \theta_2^k, \hat{\theta}_1 = \theta_1^k\right), k \neq j = \frac{1}{n-1} \Pr\left(\theta_1 = \theta_1^j | \theta_2 \neq \theta_2^j, \hat{\omega}_1 = \omega_1^k\right),$$

which goes to zero as $\varepsilon \to 0$. Applying the same reasoning as above player 2 will not challenge in this case either.

Now let us consider player 1's choice when $\theta_1 = \theta_1^j$. Given that player 2 will not challenge when $\varepsilon \to 0$, we have for ε sufficiently small that the payoff to announcing $\hat{\theta}_1 = \theta_1^j$ is

$$V_1^j = \frac{1}{n}\left(\sum_{i=1}^{n} u_1\left(D\left(\omega_1^j, \omega_2^i\right), \omega_2^i\right) - t_1^j\right).$$

The payoff to announcing some other state $\hat{\theta}_1 = \theta_1^k, k \neq j$ is

$$V_1^k = \frac{1}{n}\left(\sum_{i=i}^{n} u_1\left(D\left(\omega_1^k, \omega_2^i\right), \omega_2^i\right) - t_1^k\right).$$

But there clearly exist social choice functions $f = (D, T_1, T_2)$ such that $V_1^k > V_1^j$, and without further restrictions on preferences we cannot rule out that these social choice functions also lead player 2 not to challenge at stage 1.2.

Identical reasoning establishes a contradiction for $\rho_j^j \to 1$ and $\rho_k^j \to 0$ for all $k \neq j$ in phase 2 of the mechanism where the players' roles are reversed.

Table 1: Signal Structure

	θ_1^1, θ_2^1	θ_1^1, θ_2^2	$\cdots$	θ_1^1, θ_2^n	$\theta_1^2\theta_2^1$	$\theta_1^2\theta_2^2$	$\cdots$	$\theta_1^2\theta_2^n$	$\cdots$	$\theta_1^n\theta_2^1$	$\theta_1^n\theta_2^2$	$\cdots$	$\theta_1^n\theta_2^n$
ω_1	$(1-\varepsilon)^2$	$\frac{\varepsilon}{n-1}(1-\varepsilon)$	$\cdots$	$\frac{\varepsilon}{n-1}(1-\varepsilon)$	$\frac{\varepsilon}{n-1}(1-\varepsilon)$	$\left(\frac{\varepsilon}{n-1}\right)^2$	$\cdots$				$\frac{\varepsilon}{n-1}(1-\varepsilon)$		
ω_2	$\left(\frac{\varepsilon}{n-1}\right)^2$	$\frac{\varepsilon}{n-1}(1-\varepsilon)$		$\left(\frac{\varepsilon}{n-1}\right)^2$	$\frac{\varepsilon}{n-1}(1-\varepsilon)$	$(1-\varepsilon)^2$	$\cdots$	$\frac{\varepsilon}{n-1}(1-\varepsilon)$			$\ddots$		
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\cdots$		$\ddots$	
ω_n	$\left(\frac{\varepsilon}{n-1}\right)^2$	$\left(\frac{\varepsilon}{n-1}\right)^2$	$\cdots$	$\frac{\varepsilon}{n-1}(1-\varepsilon)$	$\left(\frac{\varepsilon}{n-1}\right)^2$	$\left(\frac{\varepsilon}{n-1}\right)^2$	$\cdots$	$\frac{\varepsilon}{n-1}(1-\varepsilon)$					$(1-\varepsilon)^2$