# Subgame Perfect Implementation With Almost Perfect Information

Philippe Aghion, Drew Fudenberg and Richard Holden*

December 6, 2007

**Abstract**

The theory of incomplete contracts has been recently questioned using or extending the subgame perfect implementation approach of Moore and Repullo (1988). We consider the robustness of this mechanism to the introduction of small amounts of asymmetric information. Our main result is that the mechanism may not yield (even approximately) truthful revelation as the amount of asymmetric information goes to zero.

*Aghion: Harvard University, Department of Economics. email: paghion@fas.harvard.edu. Fudenberg: Harvard University, Department of Economics. email: dfudenberg@harvard.edu. Holden: Massachusetts Institute of Technology, Sloan School of Management. E52-410, 50 Memorial Drive Cambridge MA 02142. email: rholden@mit.edu. We are grateful to Mathias Dewatripont, Bob Gibbons, Oliver Hart, Philippe Jehiel, John Moore, Roger Myerson, Andrew Postlewaite, Jean Tirole, Ivan Werning and Muhamet Yildiz for helpful discussions and comments. Michael Powell provided excellent research assistance.

# 1 Introduction

The distinction between *observable* and *verifiable* information or the notion of ex ante non-describability of states of nature, which underlie the incomplete contracts theory of Grossman and Hart (1986) and Hart and Moore (1990)–have been recently 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, and our main result is that the MR mechanism may not yield even approximately truthful revelation as the amount of informational asymmetry goes to zero.

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)–and based itself on the mechanism in section 5 of Moore and Repullo (1988)–to illustrate our 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 so 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 is that the simple subgame-perfect implementation mechanism à la MR for this example, does not yield approximately truthful revelation as the

[1] For example, see Aghion-Dewatripont-Rey (1999) and more recently Maskin-Tirole (1999a, 1999b).
[2] See also Maskin and Tirole (1999b).
[3] Whereas straight Nash implementation (see Maskin 1977, 1999) does not guarantee uniqueness.

correlation between the private signals and the true value of the good becomes increasingly perfect. [4]

The intuition for this result is that the 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.

More specifically, in our modification of the Hart-Moore-Repullo example, we suppose that 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, say the Buyer, to make an announcement about the value of the good, and then the Seller is permitted to either challenge or not challenge this announcement. There are a series of other steps involved if the Seller challenges. Obviously, under perfect information, the Buyer's announcement contains no information which the Seller did not have. However, when each player receives a private signal about the value of the good, the Buyer's announcement *does* contain information–namely about her own signal of the good's valuation. The Seller will condition her belief both upon her signal and the announcement made by the Buyer, and the resulting Bayesian updating is what causes the mechanism to break down.

An important part of the appeal of the subgame perfect implementation approach of Moore and Repullo (1988) is that, unlike Nash implementation, it yields a unique equilibrium.

[4] Of course our result leaves open the question of whether more complicated mechanisms, for example with multiple rounds of messages or with some trembling that takes account of the correlation of signals, might permit approximate efficiency. However, these more complicated mechanisms will be less appealing as they impose additional complexity or informational burden on the mechanism designer. We return to this issue in Section 4.

It also does not require the monotonicity axiom of Maskin (1977, 1999) and thus a larger set of social choice functions are implementable. The fragility we identify here is a consequence of the dynamic nature of the mechanism. Of course the difficulty we identify is overcome by moving back from subgame perfect implementation to Nash implementation. This, however, comes at the cost of losing the uniqueness of equilibrium.

While this Hart-Moore-Repullo example highlights in the simplest possible setting the fragility of subgame perfect implementation mechanisms to perturbations of the information structure, the example is sufficiently simple that a 2-stage mechanism in which each player only acts once can achieve approximate efficiency, as we indicate in the last part of Section 2[5]. However, this latter mechanism is itself not robust to small perturbations of the signal structure. Based on this observation, in Section 3 we extend our analysis and result to 3-stage mechanisms in a more general setting with $n$ states of nature and transferable utility[6].

In addition to the above references, 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. Unlike us, these papers consider static implementation games with commitment.

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 mechanisms with transferable utility, and develops an example to illustrate this result. Section 4 concludes.

[5] We are grateful to Andrew Postlewaite and John Moore who each provided us with such a mechanism.

[6] Also in the setting of the example, a mechanism with asymmetric fines can yield truthful revelation. But this too does not work in the general Moore-Repullo environment, as we show in Section 3.

# 2 A Hart-Moore 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 $\psi$ 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: $\theta'$ or $\theta''$. 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 $\theta'_B$. 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 $\theta''_B$ if she announces low. S now believes that there is a high 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 $\theta'_B$. 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.

There could, in principle, exist mixed strategy equilibria which yield approximately truthful revelation as $\varepsilon$ becomes small. Proposition 1 below shows that this is not the case. More specifically, suppose that conditional on observing signal $\theta'_B$ B announces "high" with probability $1 - \sigma'_B$ and "low" with probability $\sigma'_B$, where $\sigma_B \in [0, 1]$. For signal $\theta''_B$ , the

corresponding mixing probabilities are denoted by $\sigma''_B$ and $1-\sigma''_B$. These are summarized in the following table.

| | High | Low |
|---|---|---|
| $\theta'_B$ | $1-\sigma'_B$ | $\sigma'_B$ |
| $\theta''_B$ | $\sigma''_B$ | $1-\sigma''_B$ |

The corresponding mixing probabilities for player S are

| | Challenge | Don't Challenge |
|---|---|---|
| $\theta'_S$ | $1-\sigma'_S$ | $\sigma'_S$ |
| $\theta''_S$ | $\sigma''_S$ | $1-\sigma''_S$ |

### 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 $E_\varepsilon$ of the above imperfect information game as $\varepsilon \to 0$. More specifically:

**Proposition 1** *For any fine $F$ there is no sequence of equilibrium strategies $\sigma_B, \sigma_S$ such that $\sigma'_B \to 0$ and $\sigma''_B \to 0$.*

**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 we remove this restriction in Appendix 2 below.

Now, let us reason by way of contradiction, and suppose that $\varepsilon \to 0$, we have $\sigma'_B \to 0$ and $\sigma''_B \to 0$. Now, let us look at the seller's decision whether or not to challenge the buyer when $\theta_S = \theta'_S$ and the buyer announces "low". From Appendix 1, together with $(\sigma'_B, \sigma''_B) \to (0,0)$, we have 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 (see Appendix)

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

and

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

Given that $(\sigma'_B, \sigma''_B) \to (0,0)$, we thus have: $\delta(\varepsilon) \to 0$ and $\alpha(\varepsilon) \to 0$ when $\varepsilon \to 0$. This in turn implies that 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, given $(\sigma'_B, \sigma''_B) \to (0,0)$, $S$ 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)

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

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, given that $(\sigma'_B, \sigma''_B) \to (0,0)$, we have $m(\varepsilon) \to 0$ and $n(\varepsilon) \to 1$ when $\varepsilon \to 0$. Thus again, in the limit, challenging yields strictly less than (it yields $-4$) $V_S\left(DC | \theta_S = \theta''_S, L\right) = 10$. It follows that if $(\sigma'_B, \sigma''_B) \to (0,0)$, then necessarily $(\sigma'_S, \sigma''_S) \to (1,0)$ in equilibrium when $\varepsilon \to 0$.

But now let us examine the buyers's choice when $\theta_B = \theta'_B$. Given that $(\sigma'_S, \sigma''_S) \to (1,0)$ when $\varepsilon \to 0$, we have, for $\varepsilon$ sufficiently small:

$$\begin{aligned} V_B\left(H | \theta_B = \theta'_B\right) &\simeq \gamma(\varepsilon)[\beta(\varepsilon)14 + (1-\beta(\varepsilon))10] \\ &\quad + (1-\gamma(\varepsilon))[\frac{1}{2}14 + \frac{1}{2}10] - 14 \end{aligned}$$

and

$$\begin{aligned} V_B\left(L | \theta_B = \theta'_B\right) &\simeq \gamma(\varepsilon)[\beta(\varepsilon)14 + (1-\beta(\varepsilon))10] \\ &\quad + (1-\gamma(\varepsilon))[\frac{1}{2}14 + \frac{1}{2}10] - 10, \end{aligned}$$

where (see Appendix)

$$\gamma(\varepsilon) = \frac{\frac{1}{2}(1-\varepsilon)^2 + \frac{1}{2}\varepsilon^2}{\frac{1}{2}(1-\varepsilon)^2 + \frac{1}{2}\varepsilon^2 + \varepsilon(1-\varepsilon)}$$

and

$$\beta(\varepsilon) = \frac{\frac{1}{2}(1-\varepsilon)^2}{\frac{1}{2}(1-\varepsilon)^2 + \frac{1}{2}\varepsilon^2}$$

both converge to one as $\varepsilon \to 0$. Thus

$$V_B\left(H | \theta_B = \theta'_B\right) < V_B\left(L | \theta_B = \theta'_B\right)$$

as $\varepsilon \to 0$, which yields the desired contradiction and therefore establishes the result. ■

### 2.4 Discussion of the example

In appendix 2 we show that the 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 $F = 15$ if B subsequently accepts). Again, this is shown in appendix 2.

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.

## 3 A more general mechanism

### 3.1 From 2- to 3-stage mechanisms

So far, we have used a simple example of a 3-stage mechanism in order to highlight our driving intuition. While this example was subject to robustness issues, the existing implementation

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

literature suggests that these do not carry over to $q$-stage mechanisms with $q > 2$. In this section, we provide a general result on 3-stage mechanisms with transferable utility.

More specifically, the results of Fudenberg, Kreps and Levine (1988) imply that sequential equilibrium *is* robust to a range of perturbations if each player acts only once. Moreover, actions that are guaranteed to be on the equilibrium path do not count for this definition of "act". For example, if each player acts 10 times, and at the first 9 all choices have positive probability then we should also expect robustness. Although these results are not directly applicable to this setting they are suggestive of a possible reason for the fragility of 3-stage mechanisms, but not 2-stage mechanisms. The reason that they don't apply directly is that they are concerned with situations where one starts with a given "physical game" and then constructs "elaborations" of it where players have private information but the same sets of physical actions. In contrast, the setting considered here involves the mechanism designer getting to add moves to the game once the payoff perturbations are decided.

The underlying cause of fragility of subgame perfect implementation results is that the small amount of uncertainty at the interim stage, when players have observed their signals but not yet played the game, loomed large ex post once a player has sent a message. However, the way a player responds to a deviation from her opponent depends on how she expects her to play subsequently. And if there is no future action from the opponent (as it is the case in 2-stage mechanisms) then such considerations are rendered moot.

Now, 2-stage mechanisms are special in several respects. In particular, there require strong restrictions required on the preferences of the players (Moore (1992), theorem 3.3). Leading cases which satisfy the conditions are where only one player's preferences are state dependent, or where the players' preferences are perfectly correlated. For example, in the above Hart-Moore example, only B's preferences were state dependent, and so a 2-stage mechanism could work. This is restrictive, including for the setting considered in Maskin and Tirole's irrelevance theorem. In many–if not most–incomplete contracts settings of interest both parties preferences depend on the state and their preferences are not perfectly aligned.

## 3.2 Outline of the Moore-Repullo mechanism

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.

Let $\Omega$ be the (finite) set of possible states of nature[8]. 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 $\omega$s are common knowledge among each other (but not "publicly" known in the sense that the third party introduced below does not know the agents $\omega$s).

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 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 $\omega_1$ as we now outline.

1. Agent 1 announces a preference $\omega_1$, and we proceed to stage 2.

[8] 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.

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

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

and

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

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 $2\Delta$). 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 $\omega_2$. We use the notation stage 1.2, for example, to refer to phase 1, stage 2.

**Theorem 1 (Moore-Repullo)** *Suppose that the two agents' $\omega$s are common knowledge between them, and $\Delta \gg 0$ is sufficiently large. Then any $f$ can be implemented as the unique subgame perfect equilibrium of the MR mechanism.*

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 $\Delta$ 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 $\Delta$.

### 3.3 Perturbing the information structure: Our main result

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 $\theta_1^j$ given state $\omega_1^j$ given is $1-\varepsilon$, and the probability of each signal $\theta_1^j$ 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 $\omega_2^j$ given signal $\eta_2^j$ is $1-\varepsilon$, and the probability of each state $k \neq j$ conditional on signal $\eta_2^j$ 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[9].

Denote the probability that agent 1 announces $\theta_1^j$ conditional on seeing signal $\theta_1^k$ as $\sigma_k^j$. Similarly let the probability the agent 2 announces $\phi_j$ (at stage 2) conditional on observing signal $\theta_1^k$ be $\mu_k^j$. 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 $\theta_2^j$ conditional on seeing signal $\theta_2^k$ is $\rho_k^j$ and the probability the agent 1 announces $\phi_j$ (at stage 2) conditional on observing signal $\theta_2^k$ is $\tau_k^j$.

[9] 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.

**Theorem 2** *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 $\sigma_j^j \to 1, \rho_j^j \to 1$ and $\sigma_k^j \to 0, \rho_k^j \to 0$ for all $k \neq j$.*

**Proof.** See appendix 3. ■

**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 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 $\Delta$, 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 $\Delta$ 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 $\theta_1^j$ the payoff as $\varepsilon \to 0$ to player 2 from challenging is now

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

By making $\Delta_2$ large relative to $\Delta_1$ 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.

## 3.4 An example

We conclude this section by providing an example which illustrates two points: one, that asymmetric fines do not help matters, and two that there are very natural social choice functions in simple settings which cannot be implemented in the setting with imperfect

information[10]. 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

| | $\theta_1'\theta_2'$ | $\theta_1'\theta_2''$ | $\theta_1''\theta_2'$ | $\theta_1''\theta_2''$ |
|---|---|---|---|---|
| $\beta_i^H$ | $\frac{1}{2}(1-\varepsilon)^2$ | $\frac{1}{2}(1-\varepsilon)\varepsilon$ | $\frac{1}{2}\varepsilon(1-\varepsilon)$ | $\frac{1}{2}\varepsilon^2$ |
| $\beta_i^L$ | $\frac{1}{2}\varepsilon^2$ | $\frac{1}{2}(1-\varepsilon)\varepsilon$ | $\frac{1}{2}\varepsilon(1-\varepsilon)$ | $\frac{1}{2}(1-\varepsilon)^2$ |

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, $\beta_1$. Let the probability that agent 1 announces $\beta^L$ conditional on seeing signal $\theta_1'$ as $\sigma_1'$ and the probability that she announces $\beta^H$ conditional on seeing signal $\theta_1''$ as $\sigma_1''$. 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 $\varepsilon \to 0$.

The MR mechanism for this phase involves agent 1 announcing $\beta_1$ 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 $\beta_1$. If agent 2 challenges then agent 1 pays $\Delta_1$ to the third party and then agent 1 chooses between the social choice functions

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

and

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

[10]This is adapted from Bolton and Dewatripont (2005), pp558-559.

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 $(\sigma_1', \sigma_1'') \to (0,0)$. The payoff to agent 2 from challenging given that she observed signal $\theta_2'$ is

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

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 5.1 of the appendix for these calculations) and hence $\lim_{\varepsilon \to 0} \Pr\left(\theta_1 = \theta_1'|\theta_2 = \theta_2', \beta_1^L\right) = 0$, 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 $\beta^H$.

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

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

When $\theta_2=\theta_2''$ agent 2 will not challenge an announcement of $\beta_1^L$ (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 $\beta_1^H$. 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 $\lim_{\varepsilon\to0}\Pr\left(\theta_1=\theta_1'|\theta_2=\theta_2',\beta_1^H\right)=1$, 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 $(d = Y, t_Y - \Delta_1, -t_Y + \Delta_2)$[11]. 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 Conclusion

In this paper, we have used a particular deviation from common knowledge–that of common $p$-belief. An alternative approach would be to allow for differences in beliefs about the $k$th order in the hierarchy of beliefs. This is, in some sense, a more permissive perturbation than common $p$-belief. The results of Weinstein and Yildiz (2007) imply that there exists a perturbation of beliefs above the $k$th order–for $k$ arbitrarily high–such that any equilibrium in the set of interim correlated rationalizable equilibria can be "knocked out". In other words, without full knowledge of the complete hierarchy of beliefs one cannot make predictions which are any stronger than what is implied by rationalizability. Since both subgame

[11] 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.

perfect equilibrium and Nash equilibrium are stricter concepts than interim correlated rationalizability. the implication is that subgame perfect mechanisms are not robust to such $k$th order belief perturbations; furthermore, neither is Nash implementation.

Our analysis in this paper may provide foundations for the Grossman-Hart model of vertical integration. To see this, let us introduce a stage prior to the mechanism considered above 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

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

The first-order condition is

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

Under the mechanism considered above the Seller solves the following problem for $\varepsilon$ small

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

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

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

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).

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

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

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

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

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

[9] Fudenberg, D., D.K Levine, and E. Maskin (1991), "Balanced Budget Mechansims for Advsere Selection Problems," *mimeo.*

[10] 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.

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

[12] Hart, O. and J. Moore (2003), "Some (Crude) Foundations for Incomplete Contracts," mimeo, Harvard University.

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

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

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

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

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

[18] 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.

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

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

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

# 5 Appendix 1: 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 $1/2$.*

## 5.1 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 $\theta'_S$ 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 $\theta''_S$ seller, are:

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

## 5.2 Buyer's ex post payoffs

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

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

The value to B from announcing "low" when she receives signal $\theta'_B$ is

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

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[12].

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

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

[12]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 $\theta_S$.

and

$$V_B\left(L|\theta_B=\theta_B''\right)=\Pr\left(\theta_S=\theta_S'|\theta_B=\theta_B''\right)\left(\begin{array}{c}\left(1-\sigma_S'\right)\left(\begin{array}{c}\Pr\left(v=14|\theta_B'',\theta_S'\right)(14-9-6)\\+\Pr\left(v=10|\theta_B'',\theta_S'\right)(10-9-5)\end{array}\right)\\+\sigma_S'E[v|\theta_B'',\theta_S']-10\end{array}\right)$$

$$+\Pr\left(\theta_S=\theta_S''|\theta_B=\theta_B''\right)\left(\begin{array}{c}\sigma_S''\left(\begin{array}{c}\Pr\left(v=14|\theta_B'',\theta_S''\right)(14-9-6)\\+\Pr\left(v=10|\theta_B'',\theta_S''\right)(10-9-5)\end{array}\right)\\+\left(1-\sigma_S''\right)E[v|\theta_B'',\theta_S'']-10\end{array}\right)$$

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

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

## 5.3 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".

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

is

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

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

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

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

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

$$\begin{aligned}V_S\left(DC|\theta_S=\theta'_S,L\right) &= \Pr\left(\theta_B=\theta'_B|\theta_S=\theta'_S,L\right)(10)+\Pr\left(\theta_B=\theta''_B|\theta_S=\theta'_S,L\right)(10)\\ &= \frac{\left(p(1-\varepsilon)^2+(1-p)\varepsilon^2\right)\left(\sigma'_B\right)}{\left(p(1-\varepsilon)^2+(1-p)\varepsilon^2\right)\left(\sigma'_B\right)+\varepsilon(1-\varepsilon)\left(1-\sigma''_B\right)}10\\ &\quad+\frac{\varepsilon(1-\varepsilon)\left(1-\sigma''_B\right)}{\varepsilon(1-\varepsilon)\left(1-\sigma''_B\right)+\left(p(1-\varepsilon)^2+(1-p)\varepsilon^2\right)\left(\sigma'_B\right)}10.\end{aligned}$$

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

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

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

$$\begin{aligned}V_S\left(DC|\theta_S=\theta''_S,L\right) &= \Pr\left(\theta_B=\theta'_B|\theta_S=\theta''_S,L\right)(10)+\Pr\left(\theta_B=\theta''_B|\theta_S=\theta''_S,L\right)(10)\\ &= \frac{\varepsilon(1-\varepsilon)\left(\sigma'_B\right)}{\varepsilon(1-\varepsilon)\left(\sigma'_B\right)+\left(p\varepsilon^2+(1-p)(1-\varepsilon)^2\right)\left(1-\sigma''_B\right)}10\\ &\quad+\frac{\left(p\varepsilon^2+(1-p)(1-\varepsilon)^2\right)\left(1-\sigma''_B\right)}{\left(p\varepsilon^2+(1-p)(1-\varepsilon)^2\right)\left(1-\sigma''_B\right)+\varepsilon(1-\varepsilon)\left(\sigma'_B\right)}10.\end{aligned}$$

# 6 Appendix 2: Proposition 1 for general fines

In the proof of Proposition 1 presented in the text, we restricted the fine $F$ to be equal to 9. Now, we shall remove that restriction and allow for any fine $F$. We shall also allow the price at which B is offered the good after a challenge from S to be $P$, rather than simply 6. Note that S's valuations when the fine is $F$ are

$$\begin{aligned}V_S\left(C|\theta_S=\theta'_S,L\right) &= \delta(\varepsilon)[\alpha(\varepsilon)(5-F)+(1-\alpha(\varepsilon))(6+F)]\\ &\quad+(1-\delta(\varepsilon))[(1-p)(5-F)+(6+F)p],\end{aligned}$$

and

$$\begin{aligned} V_S\left(C|\theta_S=\theta''_S,L\right) &= m(\varepsilon)[(1-p)(5-F)+p(6+F)] \\ &\quad +(1-m(\varepsilon))[n(\varepsilon)(5-F)+(1-n(\varepsilon))(6+F)], \end{aligned}$$

where $\delta(\varepsilon),\alpha(\varepsilon),m(\varepsilon),n(\varepsilon)$ are the same ex post probabilities as above.

Thus, as $\varepsilon\to 0$ we have:

$$V_S\left(C|\theta_S=\theta'_S,L\right)\to(1-p)(5-F)+(6+F)p.$$

For $F$ sufficiently small, $V_S\left(C|\theta_S=\theta'_S,L\right)<V_S\left(DC|\theta_S=\theta'_S,L\right)=10$. For $p=1/2$ this conclusion is independent of $F$ and Proposition 1 holds.

If $p>1/2$ then if $F$ is large enough then $V_S\left(C|\theta_S=\theta'_S,L\right)>V_S\left(DC|\theta_S=\theta'_S,L\right)$. The critical value of $F$ for which the seller is indifferent between challenging and not challenging given $\theta_S=\theta'_S$ and B announcing "low", is given by:

$$\bar{F}=\frac{5-p}{2p-1}.$$

Whenever $F>\bar{F}$, S will challenge conditional on $\theta_S=\theta'_S$ and B announcing "low". But as $\varepsilon\to 0$ we also have

$$V_S\left(C|\theta_S=\theta''_S,L\right)\to(5-F),$$

which, for $F>\bar{F}$, is smaller than the seller's payoff from not challenging, namely 10. So for $F>\bar{F}$ S will only challenge when she receives signal $\theta'_S$. Thus when $p>1/2$ there exists an $F$ such that $(\sigma'_B,\sigma''_B)\to(0,0)$.

If $p=1/2$ but the fines at stages 3a and 3d are different then there exist $F_1,F_2$ such that $(\sigma'_B,\sigma''_B)\to(0,0)$. In such a case we have, as $\varepsilon\to 0$

$$V_S\left(C|\theta_S=\theta'_S,L\right)\to(1-\frac{1}{2})(5-F_1)+(6+F_2)\frac{1}{2}.$$

Thus if $F_2$ is sufficiently large relative to $F_1$ then $(\sigma'_S, \sigma''_S) \to (0,0)$, since

$$V_S\left(C|\theta_S = \theta''_S, L\right) \to (5 - F_1).$$

# 7 Appendix 3: Proof of Theorem 2

Suppose by way of contradiction that $\varepsilon \to 0$, 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 $\theta_1^j$. By Bayes Rule, player 2's posterior belief that player 1 saw signal $\theta_1^j$ given that player 2 saw signal $\theta_1^j$ and that player 1 announced something other than $\theta_1^j$ 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 $\theta_1^j$ 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

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

Note that as $\varepsilon \to 0$, $\alpha(\varepsilon) \to 1$, and that given the supposition that $\sigma_j^j \to 1$ as $\varepsilon \to 0$ we have $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$. Furthermore

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

where the third equality holds by conditional independence of $\theta_1$ and $\theta_2$, 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 $\Delta$s 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 $\theta_2^j$ 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 $\varepsilon$ 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.

This establishes the result.

**Table 1: Signal Structure**

| | $\theta_1^1, \theta_2^1$ | $\theta_1^1, \theta_2^2$ | $\cdots$ | $\theta_1^1, \theta_2^n$ | $\theta_1^2\theta_2^1$ | $\theta_1^2\theta_2^2$ | $\cdots$ | $\theta_1^2\theta_2^n$ | $\cdots$ | $\theta_1^n\theta_2^1$ | $\theta_1^n\theta_2^2$ | $\cdots$ | $\theta_1^n\theta_2^n$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\omega_1$ | $(1-\varepsilon)^2$ | $\frac{\varepsilon}{n-1}(1-\varepsilon)$ | $\cdots$ | $\frac{\varepsilon}{n-1}(1-\varepsilon)$ | $\frac{\varepsilon}{n-1}(1-\varepsilon)$ | $\left(\frac{\varepsilon}{n-1}\right)^2$ | $\cdots$ | | | | $\frac{\varepsilon}{n-1}(1-\varepsilon)$ | | |
| $\omega_2$ | $\left(\frac{\varepsilon}{n-1}\right)^2$ | $\frac{\varepsilon}{n-1}(1-\varepsilon)$ | | $\left(\frac{\varepsilon}{n-1}\right)^2$ | $\frac{\varepsilon}{n-1}(1-\varepsilon)$ | $(1-\varepsilon)^2$ | $\cdots$ | $\frac{\varepsilon}{n-1}(1-\varepsilon)$ | | | $\ddots$ | | |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | | $\cdots$ | | $\ddots$ | |
| $\omega_n$ | $\left(\frac{\varepsilon}{n-1}\right)^2$ | $\left(\frac{\varepsilon}{n-1}\right)^2$ | $\cdots$ | $\frac{\varepsilon}{n-1}(1-\varepsilon)$ | $\left(\frac{\varepsilon}{n-1}\right)^2$ | $\left(\frac{\varepsilon}{n-1}\right)^2$ | $\cdots$ | $\frac{\varepsilon}{n-1}(1-\varepsilon)$ | | | | | $(1-\varepsilon)^2$ |