Subgame Perfect Implementation under Approximate Common Knowledge: Evidence from a Laboratory Experiment*

Philippe Aghion, Ernst Fehr, Richard Holden, Tom Wilkening

July 29, 2011

Abstract

A remarkable result, due to Moore & Repullo (1988) is that *any* social choice function can be implemented as the *unique* equilibrium of a suitably constructed dynamic game (subgame perfect implementation). Yet such mechanisms are never observed in practice, which raises the question: why not? In this paper we test such mechanisms experimentally and find that even small deviations from perfect information cause large and persistent deviations from efficiency and truth-telling. This fragility provides an explanation for why these mechanisms are not observed in practice, which in turn has implications for the debate on the foundations of incomplete contracting.

JEL Classifications:

Keywords: Implementation Theory, Incomplete Contracts, Experiments

*We thank Jacob Goeree, Oliver Hart, Mike Powell, Klaus Schmidt and seminar participants at the 2010 Asian-Pacific ESA Conference (Melbourne, Australia), Bocconi, Chicago and Stockholm for helpful comments. We gratefully acknowledge the financial support of the Australian Research Council.

1 Introduction

Nash implementation has revolutionized economic theory, in particular by providing solid foundations to the field of mechanism design, the area of microeconomics that analyzes conditions under which various social choice outcomes can be generated as (Nash) equilibria of adequately designed mechanisms. Yet, Nash implementation per se suffers from three limitations. First, it allows only a certain class of social choice rules to be implemented, those which are "Maskin Monotonic" (Maskin, 1977; Maskin, 1999). Roughly speaking, this does not permit the implementation of social choice rules that involve distributional concerns between the agents. Second, Nash implementation typically involves multiple equilibria, so that even if a desirable equilibrium exists, an undesirable one may too.[1] Finally, Nash implementation is often not renegotiation proof generating incentives for parties to deviate and renegotiate.

These three issues may in turn explain why the literature on subgame perfect implementation (see Moore and Repullo, 1988) has attracted so much attention. This literature shows that *any* social choice function can be implemented as the *unique* subgame perfect equilibrium of a suitably designed dynamic mechanism by using subgame perfection as the solution concept. A particularly successful application of Moore-Repullo is to the debate on the foundations of incomplete contracts. In their influential 1986 paper, Grossman and Hart argued that in contracting situations where states of nature are *observable* but not *verifiable,* asset ownership (or vertical integration) can help limit *ex post* hold-up and thereby encourage *ex ante* investments. However, in subsequent work, Maskin & Tirole (1999*b*)[2] used subgame perfect implementation to show that the non-verifiability of states of nature can be overcome using a 3-stage subgame perfect implementation mechanism which induces truth-telling by all parties as the unique equilibrium outcome.

One objection to such subgame implementation mechanisms, of course, is that they are hardly observed in practice. This in turn raises the question as to why one does not observe them. One answer, recently put forward by Aghion, Fudenberg,

[1] Uniqueness can be obtained through the use of so-called "integer games" whereby parties simultaneously announce an integer and the player with the largest announcement has her preferred option implemented. These have been widely criticized, particularly since the infinite strategy means that best responses are not well-defined (Jackson 1992), and for being unimportant in practice.

[2] See also Maskin & Tirole (1999*a*).

Holden, Kunimoto & Tercieux (2009), hereafter AFHKT, is that subgame perfect implementation is not robust to arbitrarily small deviations from common knowledge. Namely, they analyze the robustness of the Moore-Repullo (MR) and other extensive form mechanisms, to "common p-belief" deviations from common knowledge, which refers to situations where each player believes with probability p slightly less than 1 that the other players believes with probability slightly less than one that the state of nature is equal to a particular realization (and so on up the belief hierarchy).[3] A main result in AFHKT is that the "good" truth-telling equilibrium of the Moore-Repullo mechanism fails to exist under p-belief perturbations and that if an extensive mechanism (subgame perfect) implements a monotonic social choice function (SCF) as the unique "good" equilibrium outcome of the corresponding game under common knowledge, then a "bad" Nash equilibrium of that game also exists under an arbitrarily small p-belief perturbation.

Yet one may reasonably question the practical importance of this theoretical result. For example it could be that players mentally code small amounts of asymmetric information as *no* amount. Or perhaps participants implement the desired social choice function with probability close to one, in which case deviations in common knowledge would be negligible to the functioning of the mechanism as p-belief perturbations grow small.[4]

One may also worry that the welfare consequences of deviations from the truth-telling equilibrium may be very small. As shown by AFHKT, the failure of the Moore-Repullo mechanism to implement truth-telling is often due to the reluctance of the responding party to invoke the mechanism after an announcement that differs from the responder's signal. As the probability of this case decreases as noise decreases, the welfare consequences may be low, especially if the action space is discrete and the fine sizes are bounded.

In this paper we conduct a laboratory experiment to test the robustness of a simple Moore-Repullo mechanism to small amounts of private information. In our

[3]The notion of common p-belief perturbations has been first introduced by Monderer & Samet (1988).

[4]See, for instance Abreu & Matsushima (1992) for a discussion of such so-called "virtual" implementation. For an experimental study related to virtual implementation, see Katok, Sefton & Yavas (2002).

experiment a buyer is matched with a seller and randomly assigned one of two sealed containers. Ex ante, before the parties receive private signals about the value of the container, the probability of each container being the relevant one is 0.5. One container is worth 70 Experimental Currency Units (ECU) to the buyer while the other container is worth 20 ECU. Each container is filled with red and blue balls which signal the value of the container to the two parties. In the no-noise (or common knowledge) treatment, the container worth 70 ECU is filled with 20 red balls and zero blue balls, and conversely the container worth 20 ECU is filled with 20 blue balls and zero red balls. Deviating from common knowledge here corresponds to increasing the fraction of blue balls in the 70 ECU container and the fraction of red balls in the 20 ECU container.

The Moore-Repullo mechanism, discussed in detail in Section 2, is implemented as follows. Both the buyer and seller privately draw a ball from the assigned container with replacement. After observing a draw from the container, the buyer makes a public announcement about the container type, i.e., whether it is a 70 ECU or a 20 ECU container. Next, based on her draw and the announcement of the buyer, the seller has the option of accepting the announcement and trading at half the announced value or calling the arbitrator.

The arbitrator is played by the computer and takes specific actions based on the announcement of the buyer. These actions are common knowledge to all parties. If the arbitrator is called, the buyer is immediately fined 25 ECU and then given a counter offer based on his initial announcement that he may accept or reject. If the buyer announced a value of 70 ECU, the arbitrator gives a counter offer of 75 ECU. If the buyer announced a value of 20 ECU, the arbitrator gives a counter offer of 25 ECU. Acceptance of the counter offer leads to trade and a payment of the fine to the seller. Rejection of the counter offer results in zero payment for both parties and an additional fine of 25 ECU to the seller. Thus a rejection of the counter offer results in a payment of -25 ECU for both parties.

Our experimental design is constructed so that truth-telling is the unique subgame perfect equilibrium in the no-noise treatment while one of many potential mixed strategy equilibria, consisting of frequent lies by both buyers and sellers, is expected in each of the noise treatments. To test these predictions, each subject is exposed

to ten rounds of the no-noise treatment and ten rounds of one of two possible noise treatments. Our focus is on how even small deviations from the no-noise treatment (i.e from perfect information) impact on the efficiency and level of truth-telling induced by the Moore-Repullo mechanism.

Our conclusions can be summarized as follows. First, the introduction of noise increases the proportion of buyers who are lying by 15 to 25 percentage points. Lies in the noise treatments are persistent over time while lies in the no-noise treatment are decreasing over the ten rounds. Second, there are very few false challenges by the seller (i.e. challenges of a low buyer announcement although the seller's private signal confirms the buyer's low announcement) in the no-noise treatment whereas the proportion of sellers who are falsely challenging in the noise treatments relative to the no-noise treatment increases by 15 percentage points. Third, the subgame perfect implementation mechanism fails to induce full truth-telling for buyers even in the no-noise treatments. This failure to generate truth telling in the no-noise treatment appears to be due primarily to the buyers' fear of being challenged even after truthful announcements. Finally, the introduction of noise reduces buyer welfare by 40% and seller welfare by 2.5%. Overall, our finding suggest that small amounts of private information do indeed lead to large deviations from truth telling and significantly more lies than under complete information.

This paper relates to two strands of literature. It first contributes to the theoretical literature on mechanism design (Maskin, 1999; Moore-Repullo, 1988; Chung & Ely, 2003) by illustrating the non-robustness of the Moore-Repullo mechanism to small deviations from common knowledge.[5] Our paper also contributes to the experimental literature on contracts (Falk & Kosfeld, 2006; Fehr, Gachter & Kirchsteiger, 1997; Dufwenberg & Lundholm, 2001; Charness, Cobo-Reyes, Jimenez, Lacomba & Lagos, 2009). Here our paper contributes by providing a first attempt at testing subgame perfect implementation and the Moore-Repullo mechanism in a lab and by using lab experiments to empirically study deviations from common knowledge. It is also the first paper to demonstrate the fragility of subgame perfect implementation to small amounts of uncertainty.

[5]See also Cremer & McLean (1987), Johnson, Pratt & Zeckhauser (1990), and Fudenberg, Levine & Maskin (1991).

The remaining part of the paper is organized as follows. Section 2 provides a brief sketch of the role of common-knowledge in Moore-Repullo mechanisms. Sections 3 and 4 detail, respectively, the laboratory implementation and our main findings. Section 5 concludes by suggesting broader implications from our experiment.

2 Theoretical motivation

In this section we present a simple example which will guide our experimental design.

2.1 Common knowledge

The following example is a slight modification of one used in Aghion, Fudenberg, Holden, Tercieux and Kunimoto (2009), and based on Hart and Moore (2003).[6] 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 $V_B = \theta - p$, where θ is the value of the good and p is the price. S's payoff is just $V_S = p$.

The good can be of either high (the state is $\theta = \theta^H$) or low quality ($\theta = \theta^L$). If it is high quality then B values it at 70, and if it is low quality then B values it at 20.

Suppose that the quality θ is observable and common knowledge to both parties. Even though θ is not verifiable by a court, and therefore no initial contract between the two parties can be made credibly contingent upon θ, truthful revelation of θ by the buyer B can be achieved through the following Moore-Repullo (MR) mechanism:

1. B announces either "high" (i.e $\theta = \theta^H$) or "low" (i.e $\theta = \theta^L$). If B announces "high"and S does not "challenge" B's announcement, then B pays S a price equal to 35 and the game then ends.

2. If B announces "low"and S does not "challenge" B's announcement, then B pays a price equal to 10 and the game ends.

3. If S challenges B's announcement then:

   (a) B pays a fine of $F = 25$ to T (a third party).

[6] The original example is also reported in Aghion & Holden (2011).

(b) B is made a counter-offer for the good at a price of 75 if his announcement was "high" and a price of 25 if his announcement was "low."

(c) If B accepts the counter-offer then S receives the fine $F = 25$ from T (and also the counter-offer price from B) and the game ends.

(d) If B rejects the counter-offer then S pays $F = 25$ to T. S also gives the good to T who destroys it and the game ends.

When the true value of the good is common knowledge between B and S this mechanism yields truth-telling as the unique subgame perfect equilibrium. To see this, suppose the true valuation is 70. If B announces "high" then B pays 35 and we stop. If, however, B announces "low" then S will challenge because at stage 3a B pays 25 to T and, this cost being sunk, B will still accept the good for 25 at stage 3b (since it is worth 70). Anticipating this, S knows that if she challenges B she receives $25 + 25 = 50$, which is greater than the 10 that she would receive if she did not challenge. Moving back to stage 1, if B lies and announces "low" when the true state is θ^H, he gets $70 - 25 - 25 = 20$, whereas he gets $70 - 35 = 35$ if he tells the truth.

Thus the above mechanism yields unique implementation in subgame perfect equilibrium. That is, for any realization of θ, there is a unique subgame perfect equilibrium which yields different prices for different valuations of the good. Moreover, in each state, the unique subgame perfect equilibrium is appealing from a behavioral point of view since it consists of telling the truth. Both of these properties fail once we introduce small common p-belief perturbations.

2.2 The failure of truthtelling in (small) common p-belief perturbations

We now introduce a small common p-belief perturbation from common knowledge about the valuation θ. We assume that the players have a common prior μ that $\mu(\theta = \theta^H = 70) = .5$ and $\mu(\theta = \theta^L = 20) = .5$.[7] Each player receives an independent

[7] AFHTK consider a more general setting with arbitrary prior. However, to map closest to the experiment, we develop the theoretical part with the same values, priors, and error distributions as those used in the actual experiment in the next section.

draw from a signal structure with two possible signals: s^H or s^L, where s^H is a high signal highly correlated with θ being equal to 70, and s^L is a low signal highly correlated with θ being equal to 20. Using the notation s_B^H (resp. s_B^L) to indicate that B received the high signal s^H (resp. the low signal s^L), the following table shows the joint probability distribution ν^ε over θ, the buyer's signal s_B, and the seller's signal s_S :

ν^ε	s_B^H, s_S^H	s_B^H, s_S^L	s_B^L, s_S^H	s_B^L, s_S^L
$\theta = 70$	$\frac{1}{2}(1-\varepsilon)^2$	$\frac{1}{2}\varepsilon(1-\varepsilon)$	$\frac{1}{2}\varepsilon(1-\varepsilon)$	$\frac{1}{2}\varepsilon^2$
$\theta = 20$	$\frac{1}{2}\varepsilon^2$	$\frac{1}{2}\varepsilon(1-\varepsilon)$	$\frac{1}{2}\varepsilon(1-\varepsilon)$	$\frac{1}{2}(1-\varepsilon)^2$

As ε converges to zero, there is convergence to common knowledge about θ by the two parties (i.e., $\nu^\varepsilon \to \mu$).

First, as in AFHKT we can show that there is no equilibrium in pure strategies in which the buyer and seller always reports truthfully. To see this, suppose instead that such an equilibrium exists, and further suppose that B gets signal s_B^L, announces "low," and is challenged. Under a truth-telling equilibrium, the buyer's belief is that his signal and the seller's signal are incorrect with equal probability, and thus the expected value of the good is 45. As this is above the counter-offer price of 25, the buyer has an incentive to purchase regardless of his signal.

Anticipating the acceptance of false challenges, the seller now has an incentive to challenge even if his signal is s_S^L. It follows that there does not exist a truthfully revealing equilibrium in pure strategies. A similar pattern holds in the case of buyers who receives signal s_B^H and is considering whether to make the "high" or "low" announcement. In this case, under the truth-telling equilibrium, the seller will be unsure as to the value of the good and may not challenge the announcement if she believes the buyer will reject the counter-offer.

Second, AFHKT show that one can find a sequence of p-belief value perturbations parametrized by some noise variable ε, such that convergence to common knowledge corresponds to $\varepsilon \longrightarrow 0$, but truth-telling by the buyer (call it the "good" equilibrium) is not approximately implementable as a mixed strategy sequential equilibrium of the above MR mechanism when $\varepsilon \longrightarrow 0$.

Third, AFHKT show that one can find a sequence of p-belief value perturbations parametrized by some noise variable ε and converging to common knowledge as $\varepsilon \longrightarrow$

0, such that the above MR mechanism under these perturbations admits a "bad" sequential equilibrium in which the probability of the buyer misreporting her signal, remains bounded away from zero as $\varepsilon \longrightarrow 0$.[8]

Finally, AFHKT show that when there is a finite level of noise, the set of consistent beliefs expand markedly, and thus multiple equilibria exist. Each of these potential equilibria have a positive amount of lies by buyers and sellers.

3 The Experiment

3.1 The subgame-perfect implementation game

At the center of our experimental design is a computerized version of the Subgame-Perfect Implementation game we discussed in the previous section. In each of twenty periods, a buyer is matched with a seller and randomly assigned one of two sealed containers.[9] One container is worth 70 ECU to the buyer while the other container is worth 20 ECU.[10] Containers are selected with equal probability and both the buyer and seller do not initially know which container has been chosen while trading.

Each of the two containers is filled with red and blue balls whose composition changes by treatment:

1. **No-Noise Treatment:** In the no-noise treatment, the container worth 70 ECU is filled with 20 red balls and 0 blue balls. The container worth 20 ECU is filled with 20 blue balls and 0 red balls.

2. **5% Noise Treatment:** In the 5% noise treatment, the container worth 70 ECU is filled with 19 red balls and 1 blue ball. The container worth 20 ECU is filled with 19 blue balls and 1 red ball.

[8] More generally, AFHKT show that, given any mechanism which "subgame-perfect" implements a non-monotonic social choice function $f(\theta)$, there always exists arbitrarily small common p-belief value perturbations under which a "bad" sequential equilibrium whose outcome remains bounded away from $f(\theta)$ for at least one state of nature θ, also exists.

[9] Subjects are randomly assigned the role of a buyer or of a seller and remain in this role throughout the experiment.

[10] The experiment was conducted in experimental currency (ECU) and converted to Australian dollars at a rate of 10 ECU = 1 AUD.

3. **10% Noise Treatment:** In the 10% noise treatment, the container worth 70 ECU is filled with 18 red balls and 2 blue balls. The container worth 20 ECU is filled with 18 blue balls and 2 red balls.

At the beginning of each period, one of the balls in the container chosen is randomly drawn and secretly shown to the seller. This ball is put back into the container and a second ball is randomly drawn for the buyer but held privately to one side. These signals provide perfect information regarding the container being traded in the no-noise treatment and almost perfect information in the 5% and 10% noise treatments.

The buyer is next asked to make a public announcement concerning the value of the container for the case in which the ball drawn for him is red or blue. He may announce a value of either 70 ECU or 20 ECU in each of the two cases. After making choices for both possible signals, the color of the ball drawn is revealed to him and his declared strategy is implemented by the computer. This strategy method gives us a complete set of data in each period which precludes changes in the frequency of lies over time due to random assignment of signals to different subsets of buyers. The strategy method also allows for a complete panel of choices which improves our ability to control for heterogeneity across individuals.

The public announcement of the buyer is next seen by the seller as well as a computerized arbitrator who acts as the implementation mechanism. After observing the announcement, the seller has the option of accepting the announcement or calling the arbitrator. If the seller accepts the announcement, trade occurs at a price equal to 1/2 of the announcement. If, however, the seller elects to call the arbitrator, the buyer is immediately charged a fine of 25 ECU and the game continues on to the arbitration response stage.

In the arbitration response stage, the buyer is given a counter offer by the computerized arbitrator which is based on his initial announcement. If he announced a value of 70 ECU, the arbitrator gives a counter offer of 75 ECU. If he announced a value of 20 ECU, the arbitrator gives a counter offer of 25 ECU.

If the buyer accepts the counter offer, trade occurs at the counter offer price. In this case the seller is given the 25 ECU which was previously charged as a fine to the buyer. If, however, the buyer rejects the counter offer, no trade occurs and the seller

is also charged a fine of 25 ECU yielding a loss of 25 ECU for both parties. Note that the structure of fines ensures that under full information the subgame-perfect equilibrium is unique.[11]

In the event that trade occurs, the actual value of the container is revealed and the profits of the buyer and seller are realized based on the value of the container, the price, and any arbitration fees. The profits of each individual are calculated after each period.

In addition to action profiles of the implementation mechanism, we also elicited beliefs about the likelihood of actions of the other party. The belief elicitation was done in each period directly after the buyer or seller took their action. For a buyer, we elicited the likelihood that the seller would challenge an announcement of 20 ECU and 70 ECU in each period given his observed signal, and we did so right after the buyer made her announcement decision but before discovering the seller's action. For a seller, we asked about his beliefs right after the seller made his challenge decision. Likelihoods were recorded using a 4-point likert scale (Never/Unlikely/Likely/Always). Similarly, we asked each seller the likelihood that their challenge would be rejected given their signal and the announcement of the buyer. Beliefs were unpaid in order to prevent hedging.[12]

3.2 Experimental design and hypotheses

Our experimental design utilizes a within-subject design in which each subject is exposed to ten periods of the no-noise treatment and ten periods of one of the two noise treatments. A total of 16 sessions were run: eight with a 5% noise level and eight with a 10% noise level. We conducted half the sessions starting with the no-noise treatment and switching to the noise treatment in period 11. We reversed the order of the two treatments in the remaining sessions. Each session contained between 20 and 24 subjects who were evenly divided between buyers and sellers at the beginning of the experiment. Buyers and sellers were matched with each other at most once in

[11] The mechanism can also be made renegotiation proof by allowing for Nash bargaining in the case of disagreement and placing the fines in escrow so they cannot be recovered in cases of disagreement.

[12] In games where both beliefs and action are compensated, risk averse individuals may find it optimal to hedge risk by stating beliefs which differ from their true estimates. See Blanco, Engelmann, Koch & Normann (2008) for a discussion of hedging.

each of the two treatments.

	Treatment 1	Treatment 2	Number of Subjects
Session 1-4	No Noise	5% Noise	88
Session 5-8	5% Noise	No Noise	84
Session 9-12	No Noise	10% Noise	90
Session 13-16	10% Noise	No Noise	86

Table 1: Treatments and Observations - 10 Periods per Treatment

As with the theoretical model, our interest is in how the introduction of almost perfect information influences the efficiency and level of truth telling in the mechanism. To this end, we designed our no-noise treatment to maximize the likelihood of the mechanism functioning well. In particular, in order to minimize the importance of fairness considerations, we set the price in the absence of a challenge equal to half of the buyer's announcement and used a moderate fine to reduce potential reciprocity. We also chose a reduced message space of announcements in order to remove the potential for small lies.[13]

Relative to the no-noise treatment where truth telling is predicted, the theoretical model predicts two types of lies to become prevalent and persist over time. First, buyers who have a high signal have an incentive to pretend that they have observed the low "blue" signal and announce 20 ECU. As sellers must contend with the potential of different signals, the sellers likelihood of challenging is expected to decrease in the degree of noise. Thus we predict:

Hypothesis 1 *The likelihood that a buyer announces a low valuation with a high signal is increasing in the level of noise. The likelihood that a seller challenges a low announcement with a high signal is decreasing in the level of noise.*

Second, sellers have an incentive to falsely challenge a low announcement when noise is introduced. In cases where the seller's signal is the low "blue" signal and

[13]The effects that arise if buyers' strategy space is larger such that they can announce false, intermediate, values is studied in Fehr, Powell & Wilkening (2010). If subjects have a preference for reciprocity or fairness, the act of challenging may be perceived as a negative or unfair act, particularly when the fine size is very large. This may lead buyers to reject challenges even when they are warranted. Such reciprocity can make the mechanism break down with a larger action space since both buyers and sellers understand that the counter offer will be rejected in the future.

the buyer announces 20 ECU, sellers may still challenge. Buyers must now contend with the possibility that they had the wrong signal as the amount of noise increases. These predictions are summarized below:

Hypothesis 2 *The likelihood that a seller with a low signal challenges a low announcement ("seller lies") increases with the level of noise. The likelihood that a buyer accepts such a challenge although he received a low signal is increasing with the level of noise.*

While we expect some deviation from truth-telling in the no-noise treatment due to different initial expectations and heterogeneity in the willingness to accept gambles, it is nonetheless the case that individuals should learn about the mechanism over time and decrease lies. Thus, a priori, we would expect lies in the no-noise treatment to be decreasing over time.

By contrast, a natural equilibrium in the noise treatment is a mixed strategy equilibrium in which both buyer lies and seller false challenges are predicted to persist.[14] Thus, unlike the no-noise treatment, we predict:

Hypothesis 3 *In both the 5% and 10% noise treatments, the probability of both buyer lies and seller false challenges are bounded away from zero and persistent over time.*

3.3 Procedures

All of the experiments were run in the Experimental Economics Laboratory at the University of Melbourne in September and October of 2009. The experiments were conducted using the programming language z-Tree (Fischbacher 2007). All of the 348 participants were undergraduate students at the University, who were randomly invited from a pool of more than 2000 volunteers using ORSEE (Greiner 2004).

Upon arrival to the laboratory, participants were divided into buyers and sellers and asked to read the instructions. To be as fair as possible to the mechanism, the instructions described the game in detail, explaining each possible signal, announcement, and arbitration action profiles in order to make the payoff consequences of a

[14] As AFHKT show, the presence of noise leads the set of consistent beliefs to expand markedly, and thus to the existence of multiple equilibria.

challenge and the rejection/acceptance of a challenge transparent. The instructions then ended with a set of practice questions which tested subjects' understanding of the signal valuations and the payoff consequences of accepting or rejecting counter-offers after a lie and after a truthful announcement. Once the answers of all participants were checked, the experimenter read aloud a summary of the instructions. The purpose of the summary was to ensure that the main features of the experiment were common knowledge amongst the participants.

Subjects then participated in the main experiment which was conducted in two parts. Subjects first played 10 periods of their assigned treatment, being matched with a different partner on the other side of the market in each period. At the start of period 11, new instructions were distributed concerning the change in information structure between treatments which were read aloud. Subjects then played 10 additional periods, again matching with the same partner at most once.

Following a short questionnaire in which gender and other demographic information were recorded, payments to the subjects were made in cash based on the earnings they accumulated throughout the experiment. In addition, each subject received a show-up fee of \$10. Since payoffs during the experiment could be negative, the subjects could use the show-up fee to prevent bankruptcy during the experiment.[15] The average salient payment at the end of the experiment was \$51.10 AUD. At the time of the experiment \$1 AUD = \$0.80 USD.

4 Results

4.1 How does the absence of common knowledge affect behavior and beliefs?

Our analysis of the experimental data focuses on the theoretical predictions stated in Hypothesis 1, 2 and 3. For each result, we provide support based on descriptive statistics, figures, and statistical tests.

[15] We had no bankruptcies.

Result 1 *Consistent with hypothesis 1, the introduction of noise increases the proportion of buyers who are lying by 15 to 25 percentage points. In the no-noise treatment, the proportion of lies is decreasing over time, while in the noise treatments, the proportion of lies is relatively stable and persists at high levels. In addition, the introduction of noise decreases the sellers' propensity to challenge low announcements if they themselves receive a high signal.*

We define a buyer lie as the announcement of a low value after observing a high signal.[16] Figure 1 shows the proportion of these lies in each period divided between the 8 sessions which started with the no-noise treatment and the 8 sessions which started with one of the two noise treatments. As can be seen by studying the top left and bottom right panels, while there are lies in the no-noise treatment, they are decreasing over time. This appears to be due to learning of individuals who announce truthfully and are not challenged.[17]

By contrast, in the noise treatments there is a dramatic increase in buyer lies relative to the no-noise treatment, which is persistent over time. In the sessions that begin in the no-noise treatment, there is a 30.3 percentage point increase in buyer lies between the last period of the no-noise treatment and the first period of the noise treatment. Likewise, in the sessions that start in the noise treatment, there is a 21.2 percentage point decrease in lies between the last period of the noise treatment and the first period of the no-noise treatment. This dramatic change in the frequency of lies is highly robust, occurring in all 16 of our individual sessions.

As foreshadowed by the large difference in means and the apparent difference in the time series, the introduction of noise leads to an increase in buyer lies which is significant under a wide variety of specifications. Table 2 shows the marginal effects of the noise treatments on buyers' lies from a probit regression under three different sets of controls. As can be seen in column (1), which looks at the last 5 periods of each treatment, increasing noise from 0% to 5% leads to a 19.3 percentage point

[16] Buyers could also lie by announcing a high value after observing the low signal. However, in practice, this occurred very infrequently. Looking at all the observations in our data, this type of lie occurred in only 102 out of 3378 observations (2.93%). Given the consistency in announcements after the low signal, we find it unlikely that lies after high signals are being driven by a fundamental misunderstanding of the environment.

[17] Individual decision making in the no-noise treatment is analyzed at the end of the section.

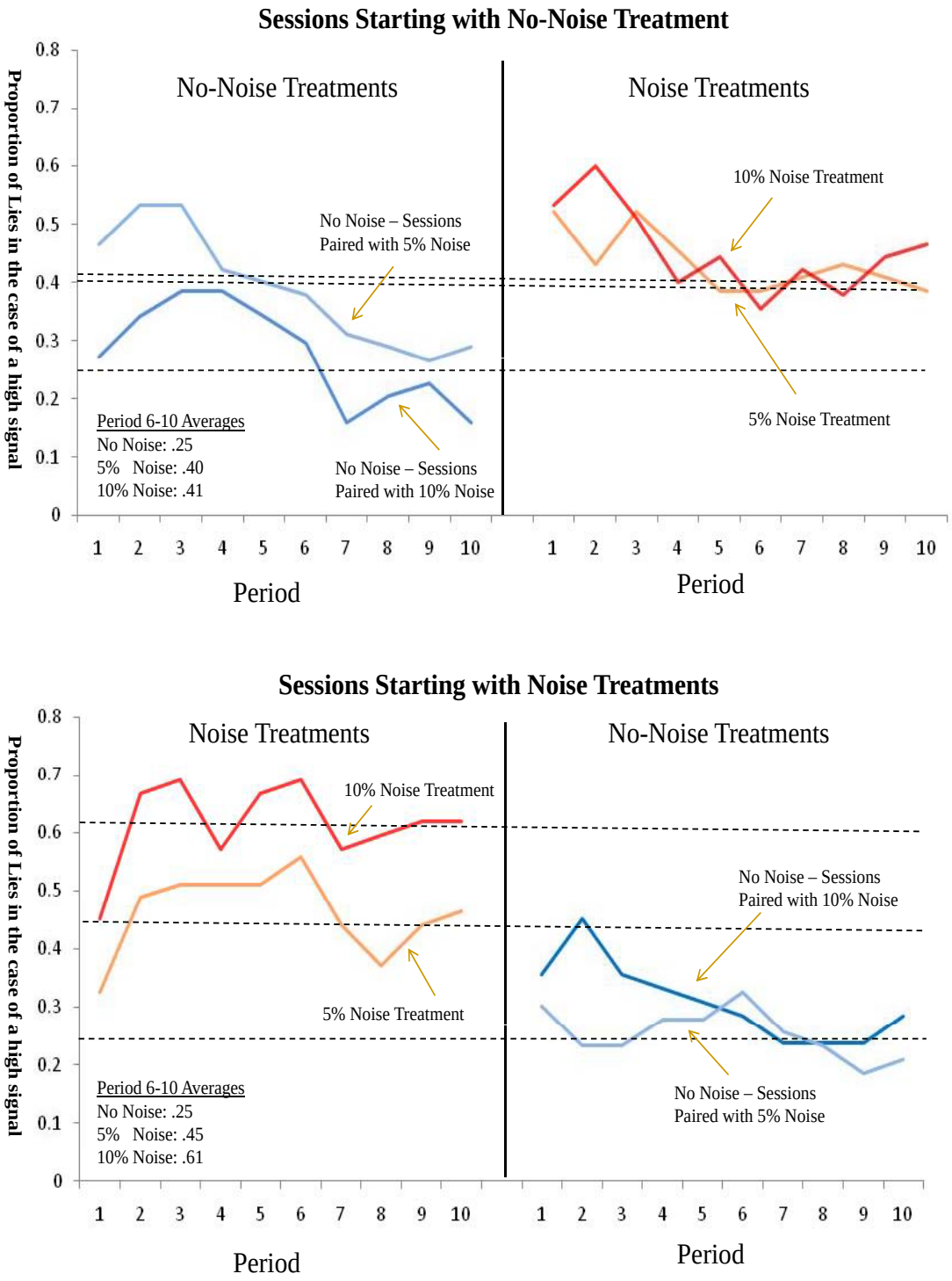


Figure 1: Proportion of Buyer Lies in the case of a high signal. The dotted lines indicated the average proportion of lies in the last 5 periods of each treatment.

increase in the proportion of lies. Similarly, increasing the noise to 10% from the baseline results in a 26.7 percentage point increase in the proportion of lies.[18]

Table 2: Lies by Buyer

	(1)	(2)	(3)
Treatment with 10% Noise	0.267***	0.130**	0.092***
	(0.048)	(0.053)	(0.102)
Treatment with 5% Noise	0.193***	0.068	0.020
	(0.045)	(0.049)	(0.095)
Period in No-Noise Treatment		–0.022***	–0.021***
		(0.004)	(0.008)
Period in Noise Treatments		–0.005	-0.001
		(0.004)	(0.007)
Period Fixed Effects?	Yes	No	No
Psuedo. R^2	.049	.033	.043
Observations	1740	3480	1740

Significance levels: * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors in parentheses, clustered by individual. Probit regression with marginal effects shown. Columns (1) and (3) includes only periods 6-10 of each treatment. Column (2) includes all periods. Column (1) includes period fixed effects for each treatment.

The regressions in columns (2) and (3) include two linear time trends: one for the no-noise treatment and one for the two noise treatments. Column (2) includes all 10 periods while column (3) includes only periods 6-10. As can be seen, there is a dramatic decrease in the proportion of lies over time in the no-noise treatment. In other words, in the no-noise treatment, people learn over time not to lie. By contrast, lies in the noise treatments appear to be more persistent, with the proportion of lies remaining relatively stable over time in both noise treatments. As seen by comparing the two columns, the reduction of lies in the no-noise treatment continues over the

[18]The treatment effects are also significant in non-parametric specifications. For example, aggregating data to the treatment level, buyers lie more in the noise treatment in 15 out of 16 sessions when all periods are used and all 16 sessions when the last 5 periods of each treatment are used. Both are significant in a Wilcoxen sign-rank test ($p < .01$). Clustered version of the rank-sum test developed by Datta & Satten (2005) yield similar results ($p < .01$).

entire sample while lies in the noise treatments stabilize over time.[19]

A different way to analyze persistence is to analyze the switching pattern of each individual buyer over time between making high and low announcements with the high (red) signal. Using observations from the last 5 periods of each session, we constructed a transition matrices P of announcements over time and then analyzed the steady state by calculating P^N as N grows large. We find that in the no-noise treatment, the steady state proportion of lies is 20%. In the 5% noise treatment, the steady state lies are 42.7% while in the 10% noise treatment, the steady state lies are 50.0%.

One might still be concerned that ten periods per treatment is not enough time for individuals to learn and that the difference in treatment is simply an artifact of slower learning in noisier environments. In the Appendix, we further study buyer lies at the individual level. We show that in the no-noise treatment, the majority of buyers are decreasing their lies over time, while in the noise treatments there are both upward and downward revision in lies over time. These findings suggest that our results would be robust to a longer sequence of periods which allowed for greater opportunities for learning.

Related to the decision of buyers to lie about their signal is the willingness of sellers to challenge when their signal indicates a high valuation. Table 3 shows the probability of a seller calling in the arbitrator given a low announcement and a high signal. As predicted, there is a significant decrease in challenges in the 5% and 10% noise treatments relative to the baseline treatment.[20]

Thus far, we have shown that, as predicted by hypothesis 1, the introduction of noise increases the buyers' probability of lying and decreases the sellers' propensity

[19] One might wonder why the treatment effects are not significant for the 5% noise treatment in column (2) and (3). This is due to the way the time trends are coded. Recall that the period variable runs from 1 to 10 in each treatment. As the time trend between the noise and no-noise treatments can differ, part of the treatment effect is subsumed in the difference in time trends. In column (2), period is between 1 and 10 and thus has a mean of 5.5. In column (3), period is between 6 and 10 and thus has a mean of 8. An alternative specification where period is centered with mean zero has treatment coefficients nearly identical to the original regression.

[20] Significance based on a Mann-Whitney test with the challenging frequency of each individual used as the variable of interest (5% noise treatment vs. baseline: z = 2.769, p-value ¡ .01; 10% noise treatment vs. baseline: z = 2.834, p-value = .01). Similar results hold for a probit regression with data clustered at the individual level.

	Challenge Probability	Number of Observations
No-Noise Treatment	95.2%	128
5% Noise Treatment	86.6%	112
10% Noise Treatment	87.5%	103

Table 3: Empirical Probability of a Challenge given a high signal and a low buyer announcement.

to challenge low announcements if their private signal is high. We next turn to our second hypothesis which predicts that the introduction of noise will also lead to false challenges by the seller when they receive the low "blue" signal.

Result 2 *Consistent with hypothesis 2, the introduction of noise substantially increases the proportion of sellers who falsely challenge. This increase persists over time.*

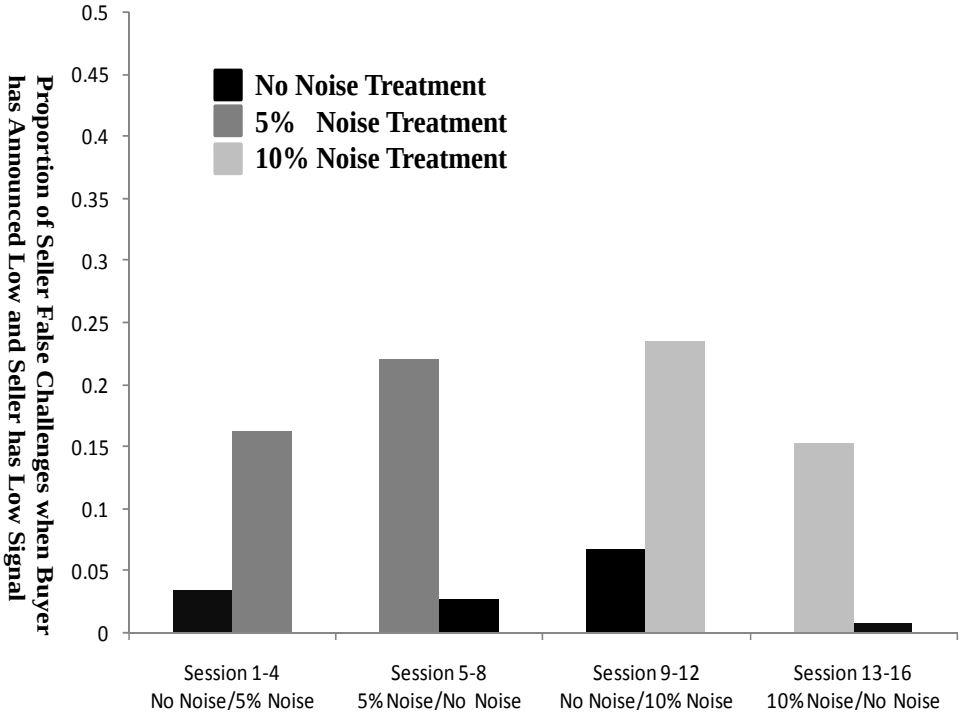
Figure 2 presents the average proportion of sellers' false challenges ("seller lies") for the last 5 periods of each treatment split into the four different session types. As can be seen by looking at the black bars, the sellers' propensity to falsely challenge in the no-noise treatment is very low. This suggests that sellers understood the structure of the game and the certainty with which their lie would be challenged by the buyers.[21]

As predicted by the theory, the introduction of noise generates a marked increase in seller initiated lies in both of the noise treatments, which is significant under both non-parametric and parametric specifications. Table 4 shows the marginal effects of the noise treatments on sellers' lies from a probit regression under the same set of controls used to analyze the buyer problem.

As can be seen in column (1), which looks at the last 5 periods of each treatment, increasing the probability of a wrong signal from 0% to 5% leads to a 20.7 percentage

[21]Looking at the beliefs data of the sellers supports this supposition. In the no-noise treatment, 64.3% of sellers with a low signal believe that the buyer will always reject a challenge while only 4.7% believe that a rejection is unlikely or will never occur. In the 5% and 10% noise treatments, only 26.7% of sellers believe they will always be challenged while 14.6% believe that a rejection is is unlikely or will never occur. These distributions are significantly different at the .01 level using a clustered version of the chi-squared test developed by Donner & Klar (2000).

Figure 2: Proportion of Lies by Sellers with Low Signal and Facing a Low Announcement - Last 5 Periods of Each Treatment



No-Noise Treatment: Mean 3.4%, 443 Obs; 5% Noise Treatment: Mean 19.5%, 200 Obs; 10% noise Treatment: Mean 19.1%, 189 Obs.

point increase in the proportion of lies under a probit specification. Similarly, increasing the noise to 10% from the baseline results in a 20.4 percentage point increase in lies.

As shown in regressions (2) and (3), which include a time trend for the no-noise treatment and a second time trend for the 5% and 10% noise treatments, false challenges in the noise treatments are stable over time while the number of false challenges decreases over time in the no-noise treatment. There is no apparent difference in time trend when restricting attention to periods 6-10, as done in regression (3), than when analyzing the full sample, as done in regression (2).[22]

Table 4: False Challenges by Sellers

	(1)	(2)	(3)
Treatment with 10% Noise	0.204***	0.105**	0.012
	(0.052)	(0.044)	(0.125)
Treatment with 5% Noise	0.206***	0.119**	0.014
	(0.047)	(0.054)	(0.128)
Period in No-Noise Treatment		–0.012***	–0.021***
		(0.003)	(0.012)
Period in Noise Treatments		–0.003	-0.004
		(0.003)	(0.007)
Period Fixed Effects?	Yes	Yes	Yes
Psuedo. R^2	.113	.085	.107
Observations	832	1614	832

Significance levels relative to 0: * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors in parentheses, clustered by individual. Probit regression with marginal effects shown. Columns (1) and (3) includes only periods 6-10 of each treatment. Column (2) includes all periods. Column (1) includes period fixed effects for each treatment.

Consistent with what theory would predict, the incentive for sellers to challenge falsely appears to stem from the belief that there exists a proportion of buyers who

[22]One may again wonder why the coefficient on the treatment effects differ across regressions (2) and (3). As before the period variable is coded from 1 to 10 in regression (2) while it is coded from 6 to 10 in regression (3). As the average period is different in these regressions, the time trends subsume part of the treatment effect. An alternative specification where period is centered with mean zero has treatment coefficients in regressions (2) and (3) nearly identical to the original regression.

will accept challenges when they have a low signal. Of the periods of the noise treatments in which a seller believed that a false challenge would never be accepted, a false challenge was initiated in only 6 of 184 (3% of the) cases. By contrast, sellers who believe that it was likely that the buyer would accept the false challenge initiated a false challenge in 73 out of 110 cases (66.3% of the time).

Table 5 reports the actual probability of a buyer accepting the counter-offer given a low signal, a low announcement, and a challenge. As can be seen, buyers increase their probability of accepting by just over 15 percentage points in the 5% noise treatment and by 33 percentage points in the 10% noise treatments.

	Acceptance Probability	**Number of Observations**
No-Noise Treatment	6.8%	15
5% Noise Treatment	23.1%	52
10% Noise Treatment	40.9%	66

Table 5: Probability of a buyer accepting a challenge with the low signal and low announcement

As foreshadowed by the marked increase in lies and the rejection of challenges, the introduction of noise leads to a marked decrease in earnings. As shown in Table 6, the buyers in the two noise treatments have significant reductions in their earnings relative to that of the no-noise treatment. The sellers, by contrast, have a very small decrease in earnings. This difference in buyer and seller outcomes is due in part to the number of buyers who appear to be minimizing potential losses in the mechanism, a phenomenon we study next.

	Buyer's Average Earnings	**Seller Average Earnings**
No-Noise Treatment	18.9	22.5
5% Noise Treatment	13.8	21
10% Noise Treatment	12.1	21.8

Table 6: Average earnings of the buyer and seller in the last 5 periods of each session. Expected earnings under the truth telling equilibrium are 22.5 respectively.

4.2 Explaining deviations from truth telling in the no-noise treatments

Thus far we have shown that p-belief perturbations from common knowledge do indeed lead to deviations from truth telling for both buyers and sellers. As predicted by the theoretical model, these deviations are both persistent across time and welfare decreasing.

A careful study of figure 1, however, reveals additional deviations from truth telling which were not predicted *ex ante* by our model. As was noted earlier and can be seen in the top-left and bottom-right panels of the figure, buyers lie roughly 25% of the time in the no-noise treatment. This result is surprising, particularly given that buyers and sellers were perfectly informed about the value of the container and the instructions and control quiz carefully explained how the mechanism functioned.

What might cause buyers to lie with such high frequency? As discussed in Bolton & Dewatripont (2005), one potential reason for the failure of subgame-perfect implementation is that individuals must place a large amount of faith in the rationality of other players. A buyer who announces truthfully must have faith that the seller will not mistakenly challenge a truthful announcement and lead to disagreement. However, if a buyer's fear of such false challenges is high enough, it may be in his best interest to minimize potential losses and adopt a maximin strategy.

In practice, it was relatively rare for sellers to falsely challenge an announcement of 70. Over the course of the experiment, a challenge of an announcement of 70 occurred in only 4.23% of observations in the no-noise treatment. Nonetheless, the belief that some sellers challenge a truthful announcement of 70 may induce the buyers to lie. The implemented mechanism implies that a challenged announcement of 70 will lead to relatively large buyer losses regardless of whether the buyer accepts or rejects the challenge. If the buyer accepts the challenge, he will earn 70 - 75 - 25 = -30; if he rejects the challenge, he will earn -25. These losses contrast sharply with the payoff the buyer can guarantee himself by lying and accepting the counter-offer. In this case the buyer earns 70 - 25 - 25 = 20 for sure.

In fact, the fear of false challenges for truthful announcements plays an important role in explaining why we observe buyers' lies in the no-noise treatments: there,

individuals who believe there is some probability of a truthful announcement being challenged are 19 percentage points more likely to lie than those who believe that truthful announcements will never be challenged.

Table 7 below extends Table 2 by including the belief that an announcement of 70 will be challenged as an additional explanatory variable in the regression. This allows comparison of the relative magnitudes of the treatment effect and the effect stemming from the fear of false challenges. As can be seen in the table, the treatment effects are highly significant, as is the likelihood that a truthful announcement will be challenged. The interaction between beliefs and the noise treatments is negative, suggesting that the fear of false challenges has a smaller impact on the buyer's probability of lying in the noise treatments than under the no-noise treatment.

Table 7: Extended Probit Regression studying buyer lies including beliefs that truthful high announcements will be challenged

	(1)	(2)	(3)
Treatment with 10% Noise	0.298***	0.180***	0.119
	(0.055)	(0.061)	(0.113)
Treatment with 5% Noise	0.234***	0.122**	0.055
	(0.048)	(0.058)	(0.104)
Belief that an announcement of	.141***	0.124***	0.140***
70 will be challenged	(0.028)	(0.024)	(0.028)
Belief x Noise Treatments	-.061***	-0.047**	-0.060**
	(0.029)	(0.024)	(0.029)
Period in No-Noise Treatment		–0.017***	–0.020***
		(0.005)	(0.009)
Period in Noise Treatments		0.013**	0.021*
		(0.006)	(0.011)
Period Fixed Effects?	Yes	No	No
Psuedo. R^2	.083	.065	.082
Observations	1740	3480	1740

Significance levels: * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors in parentheses, clustered by individual. Probit regression with marginal effects shown. Columns (1) and (3) includes only periods 6-10 of each treatment. Column (2) includes all periods. Column (1) includes period fixed effects for each treatment.

If a fear of false challenges is generating an incentive for buyers to lie, eliminating the ability of the seller to challenge high announcements should reduce the likelihood of lies in all treatments. As a further test to determine whether the fear of false challenges is generating buyer lies in the no-noise treatment, we ran four additional sessions in which buyers who announced a value of 70 could not be challenged by the seller. In two of the sessions, individuals started in the 10% noise treatment and ended in the no-noise treatment while in the other session, individuals started in the no-noise treatment and ended in the 10% noise treatment. To allow for the maximal amount of learning, we analyze the second treatment of each session.

In the sessions ending with no-noise treatments, buyer lies fell to 7.1% while those in the 10% noise treatment fell to 27%. While the gap between the noise and no-noise treatments remains large, the decrease in lies in both treatments is evidence that the fear of false challenges is a major source of lies in the mechanism.[23] This suggests that ambiguity regarding the strategy of others may have important consequences for the functioning of potential implementation mechanisms. We conclude:

Result 3 *The Subgame-Perfect Implementation mechanism fails to induce truth-telling for a subset of buyers in the no-noise treatment. This appears to be partly due to a fear that sellers will mistakenly challenge a truthful announcement.*

Perhaps surprisingly, the lack of seller false challenges in the no-noise treatment is also consistent with the fear of irrationality hypothesis. For a seller who fears irrationality on the side of the buyers, a strategy which minimizes the potential of losses is to never call the arbitrator. In this way, the buyer does not get to respond to the seller's action and thus cannot reduce the sellers payoff by rejecting a legitimate challenge. For a seller with a low signal and matched with a buyer who makes a low announcement, therefore, the only reason that a seller might challenge is to lie in an attempt to extract surplus from the buyers.

It is interesting to note that the type of sequential mechanism we tested in the above additional sessions is not capable of implementing all Social Choice Functions.

[23]The difference in treatment effects is significance at the 10% level based on a Mann-Whitney test where the lie frequency of each individual is the variable of interest: z = 1.897, p-value: .0578. Similar results hold for a probit regression with data clustered at the individual level (p=.015). Lies in the first 10 periods of each treatment follow the pattern seen in the rest of the paper with 40% lies in the 10% noise treatment and 20.5% lies in the no-noise treatment.

Moore (1992) calls mechanisms like this "simple sequential mechanisms" and provides conditions under which they can implement a desired SCF. Roughly speaking, this requires that only one part have state dependent preferences, or that preferences be perfectly correlated.

5 Concluding remarks

In this paper we conducted a laboratory experiment to test the robustness of a simple Moore-Repullo mechanism to introducing small amounts of private information. Our main findings were: first, that the introduction of noise leads to a significant increase in the proportion of buyers who are lying. Second, that lies in the noise treatments are persistent over time whereas lies in the no-noise treatments are decreasing over the successive rounds. Third, that the Moore-Repullo mechanism fails to induce truth-telling by all buyers even in the no-noise treatment, and that this appears to be (partly) due to the buyers' fear of being falsely challenged by the sellers. Overall, these findings suggest that small amounts of private information do indeed lead to large deviations from truth telling and significantly more lies than under complete information, as predicted by the theoretical analysis of Aghion, Fudenberg, Holden, Kunimoto and Tercieux (2009).

But in addition, we also found a non-negligible amount of buyers' lying in the no-noise treatment. Based on the relationship between lies and beliefs, our data suggests that individuals in all treatments are concerned about the rationality of their matched partner and often take strategies to minimize the risk imposed by the actions of others. Our results here provide initial evidence that implementation which relies heavily on the strategic actions of others may be eschewed by participants relative to simpler mechanism. Fudenberg, Kreps & Levine (1988) provide a theoretical basis for this behavior, showing that if one allows for any mapping from terminal nodes to payoffs then equilibrium refinements lose their bite. Put differently, an arbitrarily small possibility of a "crazy" type can have a large impact on the set of equilibria an analyst might expect to observe. This warrants further investigation empirically.

Our research opens two general avenues for future research. First, while we have restricted our analysis to Moore-Repullo mechanisms, it would be interesting to see

whether other classes of dynamic mechanisms are robust to similar perturbations in practice. For one sided hold-up problems, simple sales schemes appear to be significantly more robust both to imperfect information and to behavioral forces than more complicated multi-stage mechanisms which require higher degrees of rationality. In the same vein, one might use lab experiments to test the robustness of other types of implementation mechanisms, for example virtual implementation, to small deviations from standard rationality and preferences hypotheses.

A second avenue for future research would be to use similar lab experiments to analyze the effect of asset ownership on the equilibrium outcomes of perturbed extensive form mechanisms. In particular asset ownership could be naturally modeled as an outside option for the asset holder, which in turn would affect either party's incentive to report the state of nature truthfully or to challenge the other party. In particular, it would be interesting to see whether asset ownership helps achieve better equilibrium outcomes that are also robust to introducing small amounts of private information as in the above experiment.

References

Abreu, D. & Matsushima, H. (1992), 'Virtual implementation in iteratively undominated strategies', *Econometrica* **60**(5), 993–1008.

Aghion, P., Fudenberg, D., Holden, R., Kunimoto, T. & Tercieux, O. (2009), 'Subgame-perfect implementation under value perturbations', Mimeo.

Aghion, P. & Holden, R. (2011), 'Incomplete contracts and the theory of the firm: What have we learned over the past 25 years?', *Journal of Economic Perspectives* (Forthcoming, Spring).

Blanco, M., Engelmann, D., Koch, A. K. & Normann, H.-T. (2008), Belief elicitation in experiments: Is there a hedging problem?, IZA Discussion Papers 3517, Institute for the Study of Labor (IZA).

Bolton, P. & Dewatripont, M. (2005), *Contract Theory*, The MIT Press, The Massachusetts Institute of Technology.

Charness, G., Cobo-Reyes, R., Jimenez, N., Lacomba, J. A. & Lagos, F. (2009), 'The hidden costs of control: Comment', Mimeo.

Cremer, J. & McLean, R. P. (1987), 'Full extraction of the surplus in bayesian and dominant strategy auctions', *Econometrica* **56**(1247-1257).

Datta, S. & Satten, G. (2005), 'Rank-sum tests for clustered data', *Journal of the American Statistical Association* **471**(1), 908–915.

Donner, A. & Klar, N. (2000), *Design and Analysis of Cluster Randomization Trials in Health Research*, Arnold, London.

Dufwenberg, M. & Lundholm, M. (2001), 'Social norms and moral hazard', *Economic Journal* **111**(473), 506–525.

Falk, A. & Kosfeld, M. (2006), 'The hidden cost of control', *The American Economic Review* **96**(1), 1611–1630.

Fehr, E., Gachter, S. & Kirchsteiger, G. (1997), 'Reciprocity as a contract enforcement device: Experimental evidence', *Econometrica* **65**(4), 833–860.

Fehr, E., Powell, M. & Wilkening, T. (2010), 'Handing out guns at a knife fight: Behavioral limitations of the Moore-Repullo mechanism', Mimeo.

Fehr, E., Zehnder, C. & Hart, O. (2009), 'Contracts, reference points, and competition-behavioral effects of the fundamental transformation', *Journal of the European Economic Association* **7**(2-3), 561–572.

Fischbacher, U. (2007), 'z-tree: Zurich toolbox for ready-made economic experiments', *Experimental Economics* **10**(2), 171–178.

Fudenberg, D., Kreps, D. M. & Levine, D. K. (1988), 'On the robustness of equilibrium refinements', *Journal of Economic Theory* **44**(2), 354–380.

Fudenberg, D., Levine, D. K. & Maskin, E. (1991), 'Balanced-budget mechanisms for adverse selection', *Unpublished working paper* .

Greiner, B. (2004), The online recruitment system orsee 2.0 - a guide for the organization of experiments in economics, Working Paper Series in Economics 10, University of Cologne, Department of Economics.

Grossman, S. J. & Hart, O. (1986), 'A theory of vertical and lateral integration', *Journal of Political Economy* **94**(691-719).

Jackson, M. (1992), 'Implementation in undominated strategies: A look at bounded mechanisms', *Review of Economic Studies* **59**, 757–775.

Johnson, S., Pratt, J. W. & Zeckhauser, R. J. (1990), 'Efficiency despite mutually payoff-relevent private information', *Econometrica* **58**(873-900).

Katok, E., Sefton, M. & Yavas, A. (2002), 'Implementation by iterative dominance and backward induction: An experimental comparison', *Journal of Economic Theory* **104**, 89–103.

Maskin, E. (1977. Published 1999), 'Nash equilibrium and welfare optimality', *Review of Economic Studies* **66**(39-56).

Maskin, E. & Tirole, J. (1999*a*), 'Two remarks on the property-rights literature', *Review of Economic Studies* **66**(1), 139–49.

Maskin, E. & Tirole, J. (1999*b*), 'Unforseen contingencies and incomplete contracts', *Review of Economic Studies* **66**(1), 39–56.

Monderer, D. & Samet, D. (1988), 'Approximating common knowledge with common beliefs', *Games and Economic Behavior* **1**, 170–190.

Moore, J. (1992), *Advances in Economic Theory: Sixth World Congress Volume I*, Cambridge University Press, chapter Implementation, contracts, and renegotiation in environments with complete information, pp. 182–282.

Moore, J. & Repullo, R. (1988), 'Subgame perfect implementation', *Econometrica* pp. 1191–1220.

6 Appendix: Individual level analysis

In the main text, we used aggregate data to evaluate whether buyer lies in the two treatments were persistent over time. While this aggregate data shows that there is no time trend in the noise treatment and a decrease in lies in the no-noise treatment, one might be concerned that this is simply due to a difference in learning speeds between the two treatments.

As further support, this appendix shows data at the individual level. We first study the evolution of lies within the noise and no-noise treatments by dividing the treatments into two five period blocks and comparing the proportion of lies between the two blocks. We then compare lies across the noise and no-noise treatments to get a sense how individual heterogeneity is impacting our results.

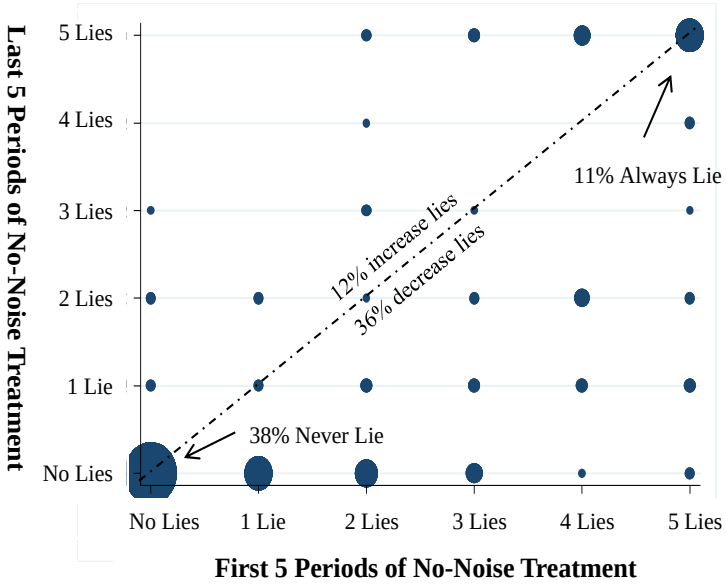
The top of Figure 3 begins our analysis by categorizing individuals based on the number of lies they make in the first five periods of the no-noise treatment relative to the last five periods of the no-noise treatment. The bubble sizes are on a logarithmic scale, normalized by the most frequent outcome, which in this case is the SPNE prediction of zero lies in both five-period blocks.

As can be seen, individuals appear to fall into three main categories. Just under 40% of individuals never lie, reflecting an immediate understanding of the mechanism. An additional 36% of individuals reduce their lies over time, suggesting that these individuals are learning to play their expected value maximizing strategy over time. Finally, the remaining 23% of individuals lie in every period or increase their lies over time.

In the main text, a Markov switching matrix was constructed which predicted 20% lies in steady state. The individual data appears to support this analysis. As 11% of individuals lie in every period, we expect very little learning from these individuals as there is no experimentation which might lead to learning and adaptation. Likewise, of the individuals who increased lying, 8% lie in all five post periods suggesting that they have stopped experimenting. We thus view it unlikely that a longer time series would fully eliminate lying in the no-noise treatment.

Paralleling the previous analysis, the bottom of Figure 3 shows the identical graph for the noise treatments. The bubbles are larger here as the most frequent outcome (zero lies) is played only 18% of the time which scales the size of the other categories. As can be seen, the evolution of lying here is more ambiguous, with the strategies of individuals shifting toward the polar cases of always lying or never lying, but with no strong directional drift. Given the lack of strong convergence in either direction, we see the conclusions drawn from the aggregate data as being reasonable.

A final consistency check is to look at individual decisions across the noise and no-noise treatments. Figure 4 categorizes individuals based on the number of lies



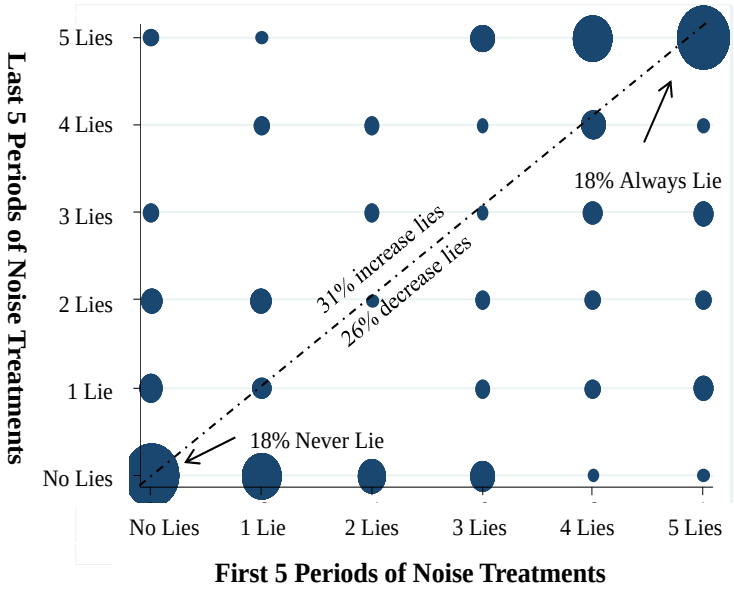


Figure 3: The evolution of lies within each treatment. As can be seen, buyer lies are decreasing over time in the No-Noise treatments while no such pattern emerges in the Noise treatments.

they make in all ten periods of the noise treatment relative to all ten periods of the no-noise treatment. Consistent with our aggregate data, there is a marked decrease in lies in the no-noise treatment, with 60% of individuals decreasing lies between the two treatments. Of the remaining individuals, there is again a subset who lies in every period who (as discussed in the main text) appear to be playing a maximin strategy, and a subset who always tell the truth. The increase in lies by a small subset of the population appears to be due to learning by individuals who started in the no-noise treatment. Splitting the sample by the ordering of treatments, 71% of individuals who lie more often in the no-noise treatment are from the sessions where the no-noise treatment was run first.

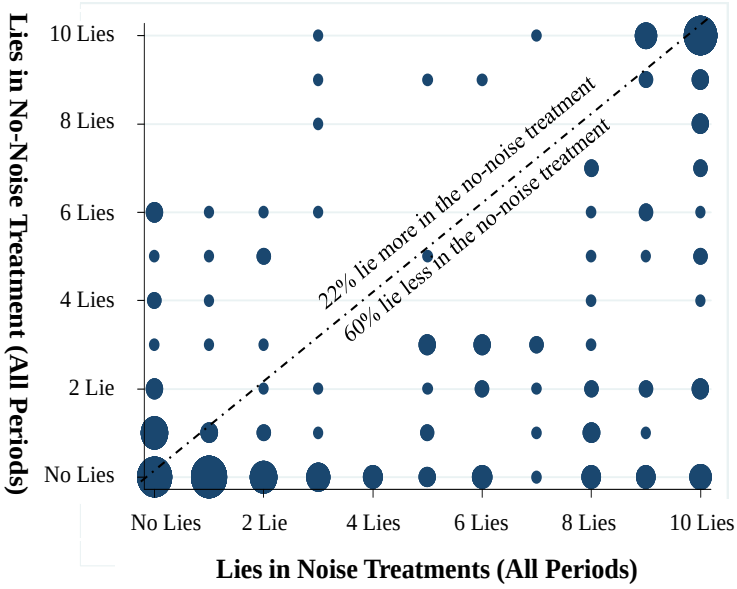


Figure 4: Lies by individual in the Noise and No-Noise treatments. While the majority of individuals decrease lies in the no-noise treatment, there exists a minority who lie in every period. These individuals appear to fear false challenges and instead adopt a maximin strategy. Of the individuals who lie more in the no-noise treatment, 71% start in the no-noise treatment. This suggests the size of this group is an artifact of learning.