Subgame Perfect Implementation under Perfect and Almost Perfect Information: An Empirical Test

Philippe Aghion, Ernst Fehr, Richard Holden, Tom Wilkening*

August 29, 2014

Abstract

In this paper we conduct a laboratory experiment to test the extent to which Moore and Repullo's subgame perfect implementation mechanism induces truth-telling in practice, both in a setting with perfect information and in a setting where buyers and sellers face a small amount of uncertainty regarding the good's value. We find that Moore-Repullo mechanisms fail to implement truth-telling in a substantial number of cases even under perfect information about the valuation of the good. This failure to implement truth telling is due to beliefs about the irrationality of one's trading partner. Deviations from truth-telling become more frequent and more persistent when agents face small amounts of uncertainty regarding the good's value. Our results suggest that both beliefs about irrational play and small amounts of uncertainty about valuations may constitute important reasons for the absence of Moore-Repullo mechanisms in practice.

Keywords: Implementation Theory, Incomplete Contracts, Experiments
JEL Classification Codes: D23, D71, D86, C92

*We owe special thanks to Michael Powell. We also thank Christopher Engel, Jacob Goeree, Oliver Hart, Martin Hellwig, Andy Postlewaite, Klaus Schmidt, Larry Samuelson, and seminar participants at the 2010 Asian-Pacific ESA Conference (Melbourne, Australia), Bocconi, Chicago Booth, Harvard, MIT, Stanford, the IIES in Stockholm, the Max Planck Institute in Bonn, and UNSW for helpful comments. We gratefully acknowledge the financial support of the Australian Research Council and the University of Melbourne Faculty of Business and Economics.

1 Introduction

Subgame Perfect Implementation has attracted much attention since it was introduced by Moore & Repullo (1988). A main reason for this success is the remarkable property that almost 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.[1] This was perceived as a substantial improvement over Nash implementation, which suffered from two main limitations: first, it would allow only a certain class of social choice rules to be implemented, those which are "Maskin Monotonic" (Maskin, 1977; Maskin, 1999); roughly speaking, Nash implementation 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.[2]

A common objection to subgame implementation mechanisms, however, is that they are hardly observed in practice. Theoretical attempts at explaining their lack of use include Fudenberg, Kreps & Levine (1988) and more recently Aghion, Fudenberg, Holden, Kunimoto & Tercieux (2012), henceforth AFHKT. In Fudenberg, Kreps & Levine, subgame perfect implementation fails due to the existence of (low-probability) "crazy types" with a systematic preference for truth-telling or for lying. In AFHKT, subgame perfect implementation is shown not to be robust to arbitrarily small deviations from common knowledge.

In this paper we use a laboratory experiment to test the extent to which the Moore-Repullo mechanism implements truth-telling in practice, both in a setting with perfect information and in a setting where buyers and sellers do not share common knowledge about the good's valuations. We perform three treatments: one with complete information about the value of the good (we refer to it as the no-noise treatment); one with 5% incomplete information (i.e. traders receive information about the good's valuation that is correct 95% of the time); and one with 10% incomplete information (traders receive correct information 90% of the time).

Our first main result is that the Moore-Repullo mechanism fails to induce truth-telling even under complete information. Although the under-reporting of the goods value decreases over time, a substantial share of under-reporting prevails throughout all periods. We perform various tests and additional experiments to try and understand this implementation failure,

[1] Subgame perfect implementation also assumes that individuals are sequentially rational and that transfers of any size are allowed.

[2] 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.

and we find that under-reporting under complete information is largely driven by buyers' beliefs that even the announcement of a high valuation will be challenged by the seller with positive probability, whereas falsely announcing a low valuation will not always be challenged by the seller. In fact we find that buyers are overly pessimistic about being challenged by sellers after the truthful announcement of a high valuation (i.e. their subjective probabilities of being challenged after announcing a high valuation are higher than the actual fraction of sellers who actually challenge a high valuation announcement) and buyers are overly optimistic about not being challenged by sellers after the false announcement of a low valuation (i.e. their subjective probabilities of not being challenged after falsely announcing a low valuation are higher than the actual fraction of sellers who actually do not challenge the announcement of a low valuation). As it turns out, eliminating the ability for sellers to challenge high announcements dramatically reduces the extent of buyers' lies. This suggests that a significant amount of truth-telling under complete information may be driven by lying aversion rather than economic incentives for truth-telling.

Our second main finding is that the introduction of incomplete information in the 5% noise and the 10% noise treatments causes a large increase in the under-reporting of the goods value. As shown in a follow-up treatment, there is a significant increase in lies even when incomplete information is reduced to 1%. The introduction of incomplete information also causes sellers to challenge truthful low announcements. These false challenges are supported by the sellers' beliefs that there is a chance that buyers will accept the false challenge. Both under-reports of the goods value by buyers and false challenges by the seller are significantly larger than the no-noise treatment in follow-up experiments where the ability for sellers to challenge high announcements is eliminated.

This paper relates to several strands of literature. It first contributes to the literature on mechanism design and more specifically on subgame perfect implementation (Maskin, 1999; Moore & Repullo, 1988; Maskin & Tirole, 1999; Chung & Ely, 2003) by pointing at two main sources for the failure of Moore-Repullo mechanisms: namely, players' beliefs about the possibility of irrational challenges by other players and (small) deviations from common knowledge. In particular we show that beliefs about the irrationality of the trading partner undermines the Moore-Repullo mechanism even in the case of perfect information about the good's value. This in turn suggests that future work should concentrate on the design and examination of mechanisms that are robust to deviations from perfect information and perfect rationality.

Second, our paper contributes 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 in-

tegration) can help limit *ex post* hold-up and thereby encourage *ex ante* investments (see Grossman and Hart (1986)). However, in subsequent work, Maskin and Tirole (1999a, 1999b) 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. Our paper sheds light on why such mechanisms are not observed in practice, which in turn can explain why vertical integration or control allocation matter.[3]

The remaining part of the paper is organized as follows. Section 2 presents the simple model which guides our experimental design. Section 3 describes the experiment and hypotheses. Section 4 presents the experimental results under complete and incomplete information. And Section 5 concludes by suggesting broader implications from our experiment and avenues for future research.

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, Kunimoto & Tercieux (2012), and based on Hart & Moore (2003).[4] 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

[3] Other papers cut from old intro we should probably cite: The systematic under estimation of the rationality of others is similar to results in Huck & Weizsäcker (2002) who find that beliefs about the play of others are distorted toward the uniform prior. Our results that many individuals tell the truth when their monetary gain from truth telling is negative is related to the literature on lying aversion (Gneezy, 2005; Sanchez-Pages & Vorsatz, 2007; Ederer & Fehr, 2009).

Our paper also contributes to the experimental literature on incentives, contracts, and implementation (Falk & Kosfeld, 2006; Fehr, Gachter & Kirchsteiger, 1997; Dufwenberg & Lundholm, 2001; Charness, Cobo-Reyes, Jimenez, Lacomba & Lagos, 2009; Sefton & Yavas (1996); Katok, Sefton & Yavas, 2002; Cabrales, Charness & Corchon, 2003).

An extensive experimental literature also exists looking at efficiency of implementation mechanisms in the public goods provision problem. Chen & Plott (1996), Chen & Tang (1998), and Healy (2006) highlight the importance of supermodularity in Groves–Ledyard mechanisms in improving public goods provision. Andreoni & Varian (1999) and Falkinger, Fehr, Gachter, & Winter-Ebmer (2000) study two-stage compensation mechanisms that build on work from Moore & Repullo (1988), while Harstad & Marese (1981, 1982), Attiyeh, Franciosi, & Isaac (2000), Arifovic & Ledyard (2004), and Bracht, Figuieres, & Ratto (2008) study the VCG, Groves–Ledyard, and Falkinger mechanisms respectively. Masuda, Okano & Saijo (2013) study approval mechanisms and like our study emphasize the need for implementation mechanisms to be robust to multiple reasoning processes and behavioral assumptions.

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

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. Before θ is realized both parties would prefer to trade at a price $p(\theta) = \frac{\theta}{2}$. This price always ensures that trade occurs when it is efficient and splits the surplus evenly between the buyer and the seller in all states of the world so that inequity aversion does not influence the desire for trade.

The value θ is *observable* and common knowledge to both parties but *non-verifiable* by a court. The assumption that the value θ is non-verifiable implies that no contract can be written that is credibly contingent on θ. However, truthful revelation of θ can be achieved through the following Moore-Repullo (MR) mechanism which can indirectly generate the desired price schedule:

1. B announces either "high" or "low". If "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).

   (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. The logic of this equilibrium is that the initial-prices, counter-offer prices, and fines are constructed so that if B and S are commonly known to be sequentially rational, B only has an incentive to announce "high" if $\theta = \theta^H$ and "low" if $\theta = \theta^L$. For this to be true, the mechanism must satisfy three conditions.

(i) **Counter-Offer Condition.** B must prefer to accept any counter offer for which he has announced "low" when $\theta = \theta^H$. B must prefer to reject any counter offer for which he has announced "low" when $\theta = \theta^L$ or for which he announced "high".

(ii) **Appropriate-Challenge Condition.** S must prefer to challenge an announcements of "low" when $\theta = \theta^H$ and must prefer not to challenge an announcement of "low" when $\theta = \theta^L$. S must prefer to never challenge "high".

(iii) **Truth-Telling Condition.** B must prefer to announce "low" if $\theta = \theta^L$ and "high" if $\theta = \theta^H$.

We refer to a challenge of "low" when $\theta = \theta^H$ as an **appropriate challenge**. The Counter-Offer Condition requires that after an appropriate challenge the counter offer price is below the value of the good so that B has a pecuniary incentive to accept the counter offer. Since the counter offer price after "low" is 25 this requirement is met. The Counter-Offer Condition also requires that after any other challenge the counter offer price is above the value of the good so that B has a pecuniary interest to reject the counter offer. Since the counter offer prices for "low" is 25 and the counter-offer price for "high" is 75, this second requirement is met.

As prices and counter-offers are constructed to satisfy the Counter-Offer Condition, B will reject counter-offers following inappropriate challenges and will accept counter-offers following appropriate challenges. This implies the Appropriate-Challenge Condition is satisfied if S has an incentive to challenge only in cases when B will accept such a challenge (i.e., when B announces "low" when $\theta = \theta^H$). This condition is satisfied since the counter-offer price of challenging a "low" announcement (25) plus the fine (25) exceeds the price that occurs if the announcement is not challenged (10).

Finally, for the Truth-Telling Condition to be satisfied, B must prefer to announce "low" if $\theta = \theta^L$ and "high" if $\theta = \theta^H$. Since the price paid by announcing "high" is higher than the price paid by announcing "low" and an appropriate challenge never occurs when $\theta = \theta^L$, B never has an incentive to overreport his value by announcing "high" when $\theta = \theta^L$. Further, B will always be challenged for announcing "low" when $\theta = \theta^H$. Adding the counter-offer price and the fine, a buyer's total payment if he **lies** by announcing "low" when $\theta = \theta^H$ is 50. As the price paid for announcing "high" is 35 and lower than the total payments from lying, B has no incentive to underreport and announces truthfully when $\theta = \theta^H$ as well.

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 Truth Telling Under (Small) Informational Perturbations

We now introduce a small common p-belief perturbation from common knowledge about the valuation θ. We assume (i) the players have a common prior μ, (ii) $\mu(\theta = \theta^H = 70) = .5$, and (iii) $\mu(\theta = \theta^L = 20) = .5$.[5] Each player receives an independent draw from a signal structure with two possible signals: s^H or s^L, where s^H is a high signal where θ equals 70 with probability $1 - \epsilon$, and s^L is a low signal where θ is equal to 20 with probability $1 - \epsilon$. We use 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).

First, as in AFHKT, we can show that there is no equilibrium in pure strategies in which the buyer and seller always report 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 challenges with a low signal and "low" announcement, 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. For slight changes in the environment, a similar pattern can hold in the case of a buyer 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.[6]

AFHKT further show that when introducing even a small level of noise, the set of consistent beliefs expands markedly, which gives rise to equilibria that involve a positive amount

[5] AFHTK consider a more general setting with an 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.

[6] These arguments extend to the limiting case where the value perturbations get very small. AFKHT show that one can find a sequence of p-belief value perturbations parameterized 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 $\varepsilon \longrightarrow 0$. They also show that one can find a sequence of p-belief value perturbations parameterized 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 $\varepsilon \longrightarrow 0$. 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.

of lies by buyers and/or false challenges by sellers. We consider particular deviations from perfect information and derive the corresponding mixed strategy Perfect Bayesian Equilibria (PBEs) in the experimental design section.

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.[7] One container is worth 70 ECU to the buyer while the other container is worth 20 ECU.[8] 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.

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 assigned to the buyer 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

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

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

and almost perfect information in the 5% and 10% noise treatments.[9]

Before the buyer knows the color of his ball he is 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 announcement 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.[10]

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 fines. The profits of each individual are calculated after each period.

In addition to action profiles of the implementation mechanism, we also elicited beliefs

[9] In the control quiz, subjects are asked to calculate the likelihood of the other party having the same color ball as them in each treatment. For the no-noise treatment we announce in the verbal summary that "if you see a red ball, you know with 100% certainty that your matched partner has also seen a red ball. Likewise, if you see a blue ball, you know with 100% certainty that your matched partner has also seen a blue ball." For the noise treatments we announce the probability that both parties observe the same signal.

[10] We ran two pilot sessions without the strategy method. The lying rates in these pilot sessions were similar to those reported in the results section.

[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.

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.

Our choice of unpaid beliefs in our main experiments was due to subgame perfection playing an important role in the theoretical equilibrium of the mechanism. From a design standpoint, we wanted to have a full set of belief information including beliefs about future counterfactual actions. In order to elicit these beliefs in an incentive compatible way, we would have had to use the strategy method for eliciting the challenge decisions of the seller and the acceptance and rejection decisions of the buyer. As the strategy method turns an extensive form game into a normal form game, we were averse to using the strategy method at interior nodes of our original experiments.[12] We discuss two additional experiments later in the paper: one where we elicited beliefs using an incentive-compatible mechanism developed by Karni (2009) and one where we elicited no beliefs to test for potential priming effects.

3.2 Experimental Design and Protocols

Our experimental design utilizes a within-subjects design in which each subject is exposed to 10 periods of the no-noise treatment and 10 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 each of the two treatments.

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

[12] We also felt that explaining an additional belief elicitation mechanism would take attention away from the main experiment. Further, in games where both beliefs and actions 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 (2010) for a discussion of hedging.

	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

than 3000 volunteers using ORSEE (Greiner 2004). An additional 340 participants were recruited in follow-up sessions conducted in 2010 and 2013.

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 challenge and the rejection/acceptance of a challenge transparent. The instructions also included a summary table which showed the payoff consequences of each combination of container value, announcements, challenges, and responses to challenges for both the buyer and the seller. 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 with an exchange rate of 10 ECU to \$1 AUD. 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.[13] The average salient payment at the end of the experiment was \$51.10 AUD. At the

[13]While we had no bankruptcies in the experiment, there is a potential that the description of bankruptcy rules could prime individuals to be more loss averse in the experiment. To test for this, we ran additional treatments where we paid for a single period and increased the show-up fee to \$35 to cover the worst outcome. We find no significant difference in our results. The robustness test of our experimental protocol

time of the 2009 experiments $1 AUD = $0.80 USD.

3.3 Hypotheses

3.3.1 The No-Noise Treatment

The Moore-Repullo mechanism used in our experiment is designed to implement truthful announcements and efficient trade. Our predictions in the no-noise treatment are as follows:

Hypothesis 1 *In the no-noise treatment buyers truthfully announce their signals and sellers do not challenge these announcements.*

As discussed in the theoretical section, hypothesis 1 is based on three conditions that must be satisfied in order for the mechanism to function: the counter-offer condition, the appropriate-challenge condition and the truth-telling condition. Each of these conditions has implicit assumptions about how individuals behave and require at least some consistency between an individual's beliefs and the actions of other individuals at later stages of the game. We briefly discuss some of the potential issues that might cause the conditions underlying the mechanism to be violated.

The counter-offer condition requires that a buyer who is appropriately challenged is willing to accept the counter offer instead of rejecting it. If individuals care only about their own payoffs in the mechanism, as is assumed by theory, this condition is satisfied for any counter-offer price that is below the value of the good in the high state and above the value of the good in the low state. Given our two-state design, any counter-offer price between 20 and 70 would thus suffice in creating a pecuniary incentive to accept appropriate challenges.

As discussed in detail in Fehr, Powell & Wilkening (2014), there is strong evidence of non-pecuniary benefits for rejecting an appropriate challenge when individuals are negatively reciprocal. A buyer who lies and is challenged suffers a pecuniary reduction in his income that unambiguously reduces his utility relative to what he would receive if he had not been challenged. If buyers view this reduction in their payment as an unkind act they may retaliate against sellers by rejecting appropriate counter offers. This implies that for the mechanism to function properly, the monetary gain from accepting the counter offer must exceed the combined pecuniary and non-pecuniary values of rejecting.

In our experiment we were particularly concerned with the counter-offer condition and used parameters that both maximized the net pecuniary value from accepting the counter offer and minimized the non-pecuniary value for rejecting. To maximize the pecuniary value, we set the counter offer price at 25 so that the buyer's return for accepting the challenge (45

are discussed in section 4.4.

ECU) is very large. We also chose a relatively low fine as the desire to retaliate is likely to be influenced by (1) the amount of money lost by being challenged and (2) the amount of the seller's payoffs that can be destroyed by rejecting. On net, a buyer who retaliates after a low announcement must prefer the payoffs of $\{-25, -25\}$ for the Buyer and Seller over payoffs of $\{20, 50\}$. Equivalently, he must be willing to destroy \$.60 of his own money to destroy \$1.00 of the seller's money after a low announcement and a challenge. This is much larger than the levels of reciprocity estimated from Fehr et al. (2014) and seen in standard ultimatum games.

Moving up to the next stage of the game, the appropriate-challenge condition requires that sellers make appropriate challenges but not inappropriate challenges. For this condition to hold it must be that individuals have beliefs about the actions of the buyer that lead to the desired challenge decisions.

While subgame perfection assumes that the beliefs of individuals are consistent with the actions other individuals make in later stages of the game, there are reasons to believe that forming consistent beliefs is particularly difficult in the acceptance and rejection stage. When the counter-offer condition and the appropriate-challenge condition are met, a buyer who announces "low" in the high state is deviating in a way that reduces his material payoffs relative to a truthful announcement. For beliefs to be fully consistent with the SPNE, the seller must believe that such a buyer will come back to his senses and accept the counteroffer at the next stage. This consequence of the "one-shot deviation" principle is likely to be violated if lies are correlated across choices as would be the case if (for instance) lies were generated by confusion.[14]

In order to maximize the incentive of sellers to make challenges across a large range of potential beliefs, we chose to pass the fine F to the seller in the case that the counter offer is accepted. Given a belief $\rho_s^{L|H}$ that a buyer will reject a counter offer after a low announcement in the high state, a sellers expected value for challenging is $50(1-\rho_s^{L|H})-25\rho_s^{L|H}$. Comparing this to the return of 10 that the seller could guarantee by not challenging, the seller has a pecuniary incentive to challenge if:

$$50(1-\rho_s^{L|H})-25\rho_s^{L|H} > 10 \tag{1}$$

which is satisfied when $\rho_s^{L|H} < .533$. Under risk neutrality, this implies that the seller has an incentive to challenge even if she believes a buyer who lies will randomly accept or reject counter offers after a lie.

[14]See Bolton & Dewatripont (2005) for a general discussion of this issue in subgame-perfect implementation.

Finally, for the truth-telling condition to hold, it must be that a buyer, given his beliefs about the actions of the seller, has an incentive to make a truthful announcement rather than a lie. For the truth-telling condition to hold, both the buyer's belief about the likelihood of being challenged after a lie and the likelihood of being challenged after a truthful announcement guide his decision. Given belief $\rho_b^{L|H}$ that the buyer will be challenged after a low announcement in the high state, a buyer who will accept the counter offer receives a pecuniary utility of lying of $60(1-\rho_b^{L|H})+20(\rho_b^{L|H})$. Likewise, given belief $\rho_b^{H|H}$ that a truthful announcement will be challenged in the high state, the pecuniary utility of a truthful announcement is $35(1-\rho_b^{H|H})-25(\rho_b^{H|H})$. For the buyer to have a pecuniary value for truthful announcement in the high state, it must be the case that:

$$35(1-\rho_b^{H|H})-25(\rho_b^{H|H})>60(1-\rho_b^{L|H})+20(\rho_b^{L|H}) \tag{2}$$

or

$$\frac{2}{3}\rho_b^{L|H}>\frac{5}{12}+\rho_b^{H|H}. \tag{3}$$

Note that if $\rho_b^{H|H}=0$, this requirement would be satisfied for $\rho_b^{L|H}>\frac{5}{8}$.

Informed by the discussion above, we parameterized the model with an eye toward making each of the intermediate conditions as slack as possible. In places where parameters affected multiple constraints simultaneously (such as the fine size or counter-offer price), we erred toward ensuring that the counter-offer condition was satisfied as this condition feeds into the other two conditions. We also set the price in the absence of a challenge equal to half of the buyer's announcement in order to minimize the importance of fairness considerations and make the subgame perfect equilibrium salient.

3.3.2 The Noise Treatments

As soon as one introduces noise in agents' information about the state of nature (i.e about the valuation of the good to be traded), the truth-telling equilibrium vanishes and pure and mixed strategy equilibria arise in which either (i) the buyer makes announcements which are different than his signal and/or (ii) the seller challenges announcements which are the same as her signal. This section discusses these equilibria and shows that while point predictions are sensitive to preferences and the equilibrium selected by participants, the introduction of noise leads to an increase in combined lies by buyers and sellers.

We begin by discussing a pure strategy sequential equilibrium that exists in the model. For any amount of noise, one can sustain the following "bad" (sequential) equilibrium with the appropriate sequence of beliefs: B announces low (i.e a value of 20 ECU) at stage 1

regardless of his signal, S never challenges at stage 2, and (off-equilibrium) B always rejects a counter offer made at stage 3 if that stage were to be reached.

More specifically, this equilibrium can be sustained as a sequential equilibrium with the buyer's (off-equilibrium) belief that the true state is low ($\theta = \theta^L$) when he is challenged and the arbitrator's counteroffer is made. To establish sequential rationality, we proceed by backward induction. At Stage 3, regardless of his signal, B believes with probability one that the state is θ^L. Accepting S's offer at a price of 25 (resp. 75) leads to a payoff of $20-25-25=-30$ (resp. $20-25-75=-80$) whereas rejecting it leads to a payoff of -25. Thus, it is optimal for B to reject the offer. Moving back to Stage 2, if S chooses "Challenge," S anticipates that her offer will be rejected by B at stage 3, and thus anticipates that, as ε goes to zero, the payoff is approximately equal to -25 if her signal is high and to -25 if the signal is low. On the contrary, if S chooses "No Challenge," S guarantees a payoff of 10. Thus, regardless of her signal, it is optimal for S not to challenge. Moving back to Stage 1, suppose first that B receives the high signal s_B^H. Then, as ε becomes small, B believes with high probability that the true state is θ^H so that his expected payoff from announcing "low" is close to $70-10=60$, greater than 35, which B obtains when announcing "high." Thus, it is optimal for B to announce "low." A similar reasoning applies if B receives the low signal s_B^L. Finally, consistency of beliefs follows by identical arguments to those in AFHKT (footnote 13). Thus, the above is indeed a sequential equilibrium.

A second pure strategy (sequential) equilibrium can be sustained where the buyer always announces high regardless of his signal. In this equilibrium, the buyer's (off-equilibrium) belief is that the true state is high with probability .1 at stage 2 when he receives the low signal, announces a low valuation, and is challenged. The expected value for accepting the challenge is $.9\times-5+.1\times45=0$. Thus, he is indifferent between accepting and rejecting the challenge. If at stage 1 the buyer believes that the seller will always challenge, the expected value of this sequence of play is -25. The buyer can do strictly better by announcing a high value with the low signal and thereby guarantee himself a return of $.9\times-15+.1\times35=-10$.

In addition to the "bad" pure strategy equilibria described above, the noise treatments also generate a mixed strategy equilibrium where the buyer announces his signal truthfully and the seller who has a low signal and observes a low announcement mixes between challenging and not challenging. We call the case where the seller deviates by challenging a low announcement with a low signal a **false challenge**. A buyer in this equilibrium who has followed his signal and announced low at Stage 1, and then has been challenged at Stage 2, mixes at Stage 3 between accepting the challenge and rejecting it.

If the buyer has non-standard preferences, the act of being challenged may be perceived by the buyer as an "unkind" act since his payoffs in the resulting subgame are strictly

below what he would have received if he were not challenged. A reciprocal buyer may gain additional utility from rejecting the seller's challenge and retaliating against this unkind act. When the buyer is prone to moderate levels of reciprocity, we show in the appendix that retaliation (and fear of retaliation) can lead to an alternative mixed strategy equilibrium in which the seller never challenges with a low signal and mixing occurs on the buyer side. In this equilibrium, the buyer who has a high signal mixes between making a high and low announcement and the seller who has a high signal and observes a low announcement mixes between challenging and not challenging.[15]

While different equilibria and assumptions about preferences lead to slightly different point predictions as noise increases, a universal property of all these equilibria is that total lies by buyers and false challenges by sellers increase as noise increases. We thus view an overall increase in buyer and seller lies to be the most robust prediction of our model.

Hypothesis 2 *The likelihood that a seller with a low signal challenges a low announcement 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.*

Hypothesis 3 *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.*

4 Experimental Results

We describe the results of the experiment in this section. Section 4.1 uses the data from the no-noise treatments to study Hypothesis 1. Section 4.2 uses data on beliefs and from a number of additional experiments to interpret some of the results from 4.1. Section 4.3 uses data from both the no-noise and noise treatments to study Hypotheses 2 and 3.

We call a draw of a red ball the **high signal**, a draw of a blue ball the **low signal**, an announcement of 70 a **high announcement** and an announcement of 20 a **low announcement**. As before, we define a **lie** as an announcement by B of a low value after observing a high signal. We define an **appropriate challenge** as a challenge by S of a low

[15] Structural estimates of reciprocity done for Fehr, Powell, and Wilkening (2014) predict that individuals are willing to sacrifice between \$.25 and \$.4 to destroy \$1 of wealth of the seller after a legitimate challenge in a similar subgame-perfect implementation mechanism. Given these levels of reciprocity, the mixed strategy would predict that the buyer lies 9.2% of the time in the 5% noise treatment and 19.2% of the time in the 10% noise treatment. With no reciprocity, the model predicts that the seller false challenges 28.5% of the time in the 5% noise treatment and 62.5% of the time in the 10% noise treatment. Power calculations for the experiment were based on the mixed strategy equilibrium with reciprocity.

announcement with the high signal, an **inappropriate challenge** as a challenge by S of a high announcement with the high signal, and a **false challenge** as a challenge by S of a low announcement with the low signal.

4.1 The Mechanism Under Perfect Information

Under Hypothesis 1, our experimental design predicts that in the no-noise treatment, the counter-offer condition, appropriate-challenge condition, and truth-telling condition will hold. These conditions imply that B will always tell the truth, S will make only appropriate challenges, and B will accept counter offers if and only if they result from an appropriate challenge. The data from the no-noise treatment provides support for only two of these conditions.

Result 1 *(a) In a majority of cases B's accept counter offers after appropriate challenges and reject counter offers after false challenges. (b) Further, S's make appropriate challenges the majority of the time and make inappropriate challenges and false challenges very infrequently. (c) However, a significant proportion of B's lie by making a low announcement with the high signal. Thus the mechanism does not fully induce truth-telling in the no-noise treatment.*

Figure 1 displays the patterns of play we observed in the no-noise treatment of the experiment. The left column examines play when an individual receives a low signal while the right side examines play when an individual receives a high signal. Panel (a) summarizes B's announcement decision, Panel (b) summarizes S's challenge decision, and Panel (c) summarizes B's decision to accept or reject counter offers. The error bars show 95% confidence intervals of each proportion with standard errors clustered at the individual level.

Panel (a) shows that after a low signal, 97.2% of individuals announce that the value is low. By contrast, after a high signal, 30.8% deviate from the theoretical prediction of Hypothesis 1 and lie. We discuss this deviation from truth-telling in greater detail below after detailing play in the other stages of the game.

Panel (b) shows the proportion of announcements that are challenged after each combinations of announcement and signal. As can be seen, a low announcement with a low signal is challenged only 4.1% of the time while a high announcement with a high signal is challenged only 4.8% of the time. This implies that inappropriate challenges rarely occur in the data. By contrast, S's challenge a low announcement with a high signal 93.4% of the time implying that S's almost always make appropriate challenges.

Finally, Panel (c) shows the proportion of counter-offers that are accepted for each combination of announcement and signal. B's always reject challenges after truthful announcements and almost always accept challenges after a lie.

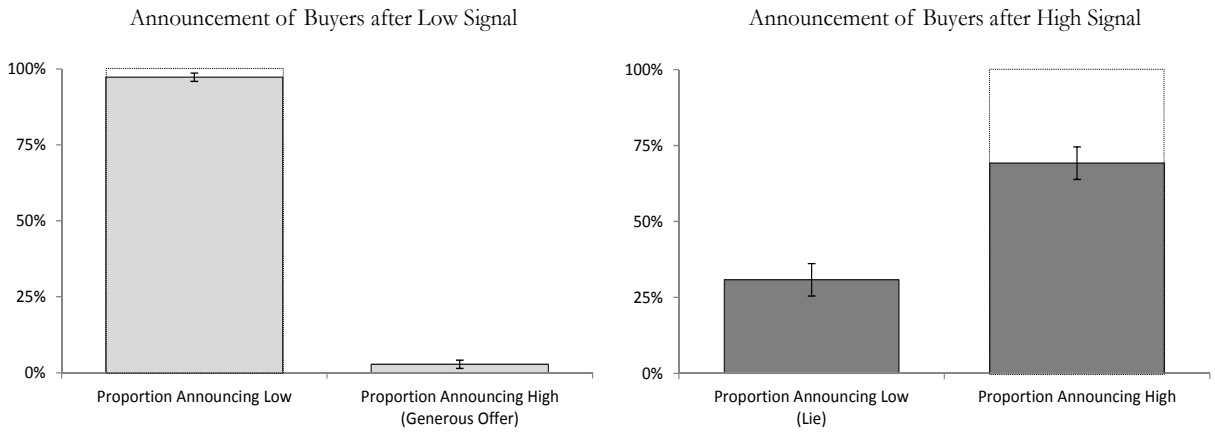
While there are small deviations from the theoretical predictions of the model in the challenge stage and counter-offer stage, these deviations tend to vanish over time. Panel (a) of Figure 2 tracks the proportion of truthful announcements that are challenged in each period. This data is overlayed with the predictions and 95% confidence intervals from a simple linear random effects regression that regresses the challenge decision on the period. As can be seen, challenges of truthful announcements are diminishing and the proportion of truthful announcements that are challenged is not significantly different from the theoretical prediction of 0% by period 10. Similarly, as seen on the right side of Panel (b), challenges of lies are increasing over time and the proportion of lies is not significantly different to the theoretical prediction of 100% by period 10. Taken together, the data strongly supports the appropriate-challenge condition.

Panel (c) of Figure 2 tracks the proportion of counter-offers that are accepted after a lie over time using the same construction of the prediction line and 95% confidence intervals as in the previous panels. While some B's initially reject counter offers, the proportion of counter-offers being accepted increases over time and is not significantly different to the theoretical prediction by period 10. Thus, there is strong evidence that the counter-offer condition is met in the data.

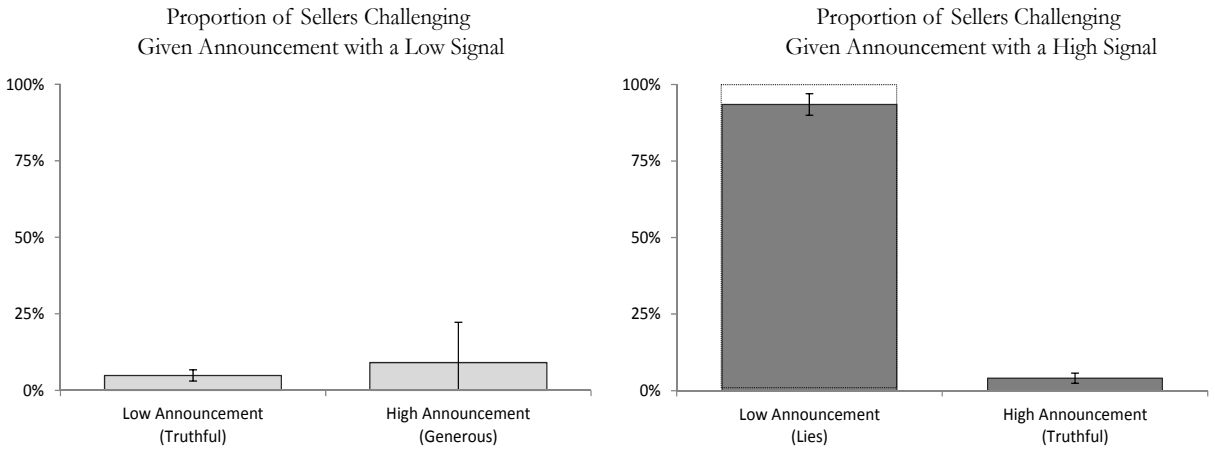
Given that the appropriate-challenge condition and the counter-offer condition hold, B's have pecuniary incentives to announce truthfully by construction of the mechanism. Thus, we might expect that small lies converge to zero over time. Figure 3 shows that this is not the case. As can be seen in Panel (a), the proportion of B's who are lying is indeed decreasing over time. However, this proportion is above 20% and significantly different from the theoretical prediction of zero even in period 10.

Panel (b) shows a histogram of buyer lie rates in the no-noise treatment using all periods. As can be seen, 38% of buyers never lie in the no-noise treatment while 11% of individuals lie in every period. This bimodal distribution becomes more pronounced over time: in a restricted sample of the last five periods of the treatment, 61% of buyers never lie while 17% lie in each period. Thus, while many individuals stop lying over time a significant subset of individuals do not stop lying. We explore why these individuals may find it in their interest to lie in the next section.

(a) Announcements of Buyers



(b) Challenges of Sellers



(a) Acceptances of Counter-Offers by Buyers

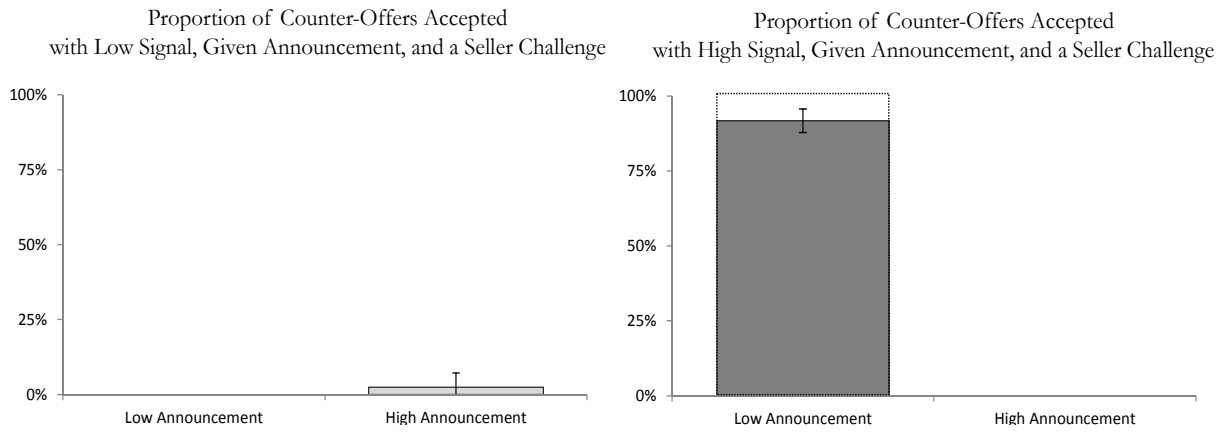
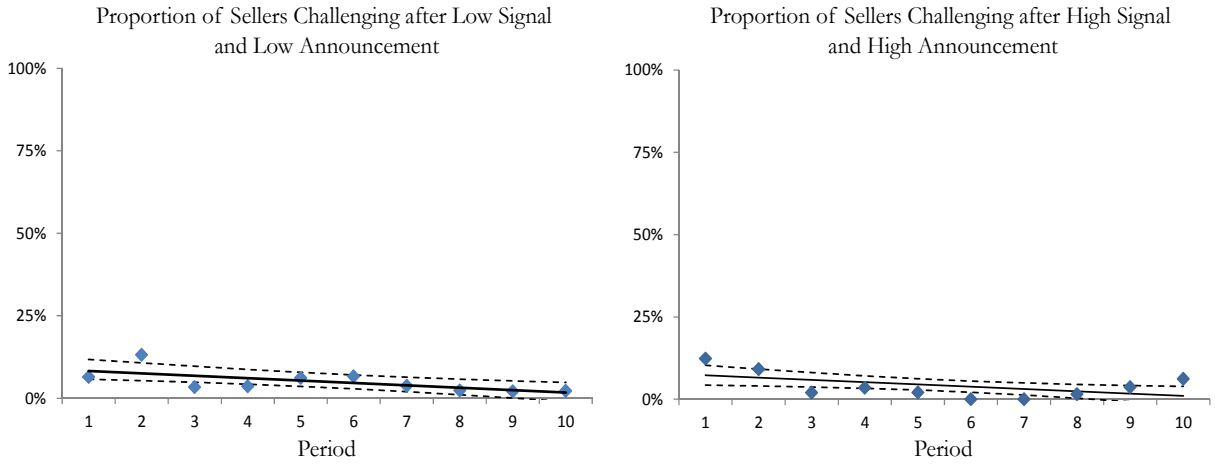
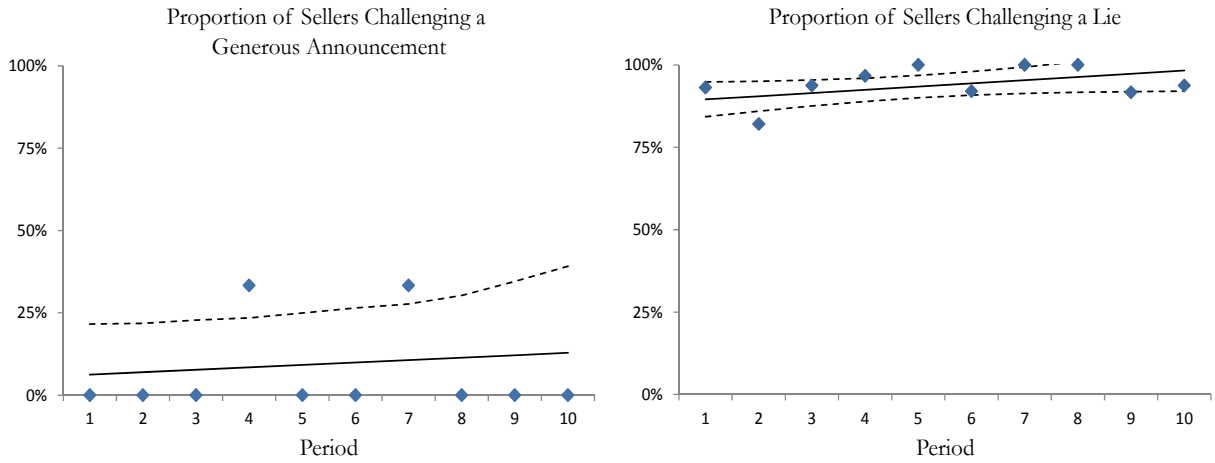


Figure 1: Pattern of Play in No-Noise Treatment

(a) Challenges of Truthful Announcements by Seller over Time



(b) Challenges of Generous Offers and Lies by Sellers over Time



(c) Acceptances of Counter-Offers by Buyers over Time

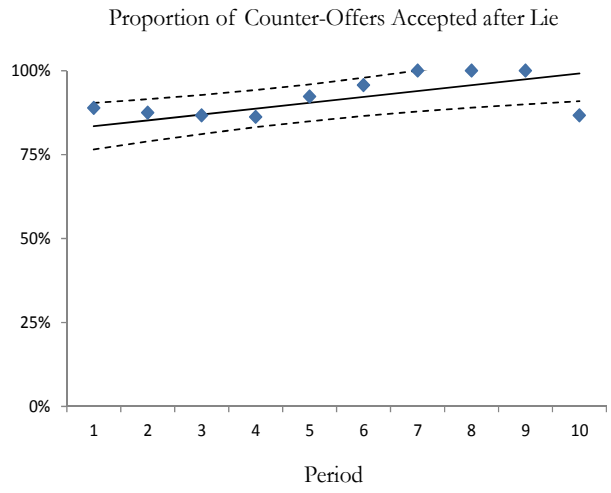


Figure 2: Evolution of Play in Challenge Stage and Counter-Offer Stage of No-Noise Treatment

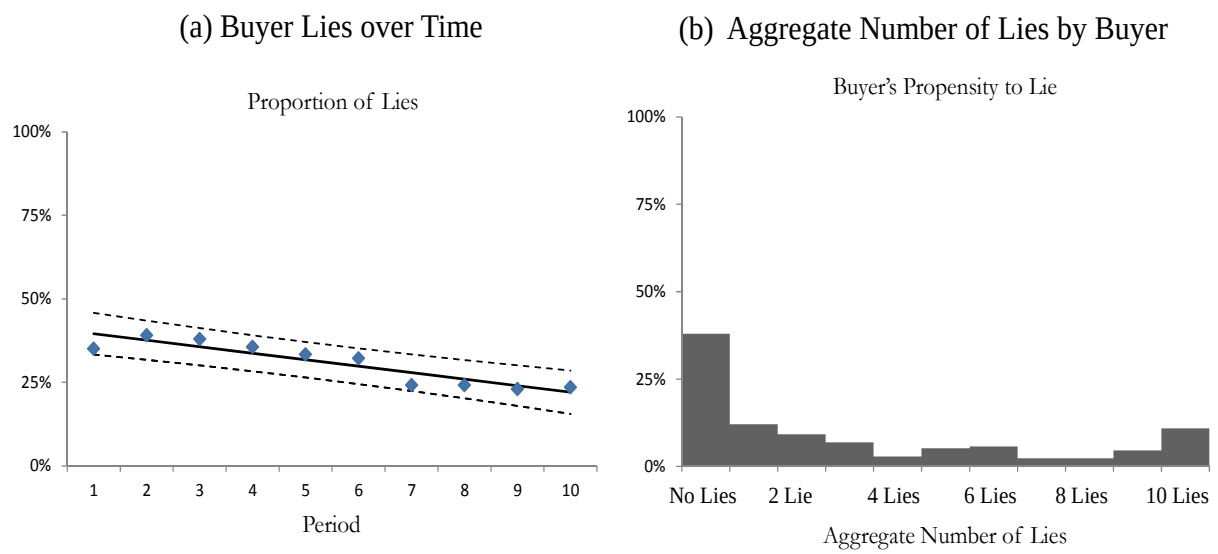


Figure 3: Evolution and Distribution of Lies in Announcement Stage of No-Noise Treatment

4.2 Understanding Deviations from Truth Telling in the No-Noise Treatment

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. B's who announce truthfully must have faith that S's will not make an inappropriate challenge. However, if a B's fear of such an inappropriate challenge is high enough, it may be in his best interest to adopt a strategy that minimizes his potential losses.

In practice, it is relatively rare for S's to make an inappropriate challenge. As was seen in the last section, high announcements were challenged in only 4.8% of observations. Nonetheless, the belief that some S's challenge a truthful high announcement may induce B's to lie. The implemented mechanism implies that a challenged high announcement will lead to relatively large losses for B regardless of whether B accepts or rejects the challenge. If B 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 of 20 that B can guarantee himself by lying, being challenged by S, and accepting the counter-offer.

Looking at the beliefs data of B, it appears that the fear of inappropriate challenges is indeed an important determinant of lies. Table 2 reports the results of regression analysis where the dependent variable is 1 if B lies after the high signal and 0 if B makes a truthful announcement. This variable is regressed on the belief that a lie will be challenged and the belief that a truthful announcement will be challenged. To allow for potential non-linearities in the belief data we treat B's beliefs as categorical data and split the 4-point likert scale into

a series of dummy variables. We use the category "Never" as the omitted dummy category. Column (1) reports the results of a simple linear probability model with errors clustered at the individual level. Column (2) reports the results of a fixed effects regression with both time and individual level fixed effects.

As can be seen in column (1), B's belief about the likelihood that he will be challenged after a truthful announcement is a good predictor of his likelihood of making a small lie. B's are 39.7 percentage points more likely to lie if they believe that a truthful announcement is "Likely" to be challenged relative to an individual who believes a truthful announcement will "Never" be challenged. The probability of making a small lie is increasing as an individual's belief becomes more pessimistic suggesting a monotonic relationship between beliefs and lies.

The fixed effects regression in column (2) controls for variation between individuals and estimates the impact of beliefs using within subject variation. As can be seen in Figure 4, which shows the evolution of beliefs over time, the proportion of individuals who believe that truthful announcements will "Never" be challenged and lies will "Always" be challenged is increasing. Thus, the estimates in the fixed effect regression can be interpreted as the change in an individual's propensity to lie as his beliefs shift toward the Nash Equilibrium predictions. As with the aggregate regressions, individuals are more likely to make truthful announcements as they become less pessimistic about truthful announcements being challenged.

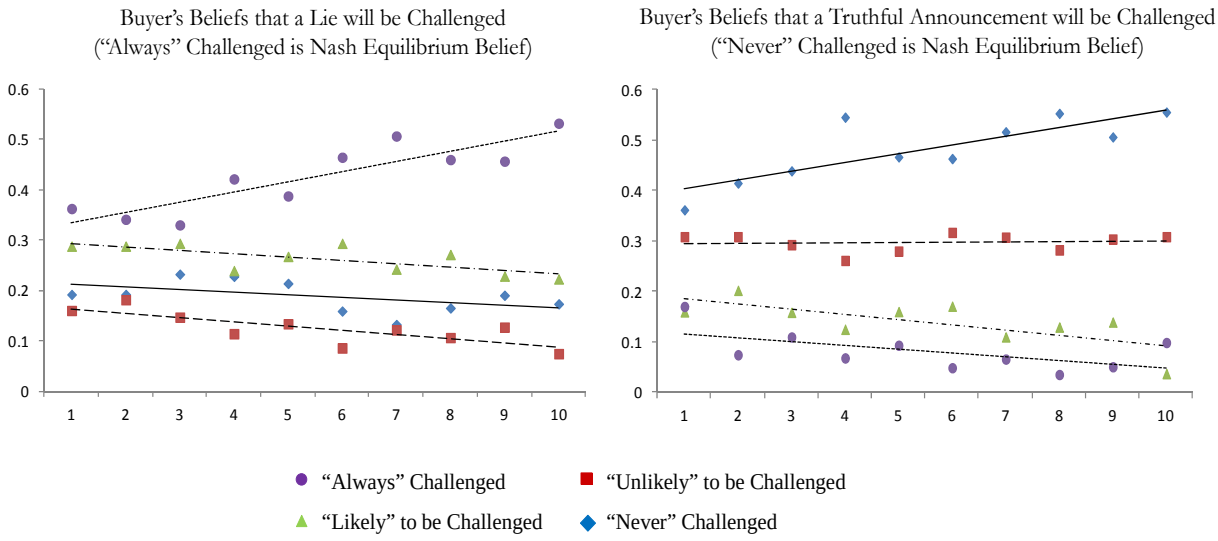


Figure 4: Evolution of Buyer Beliefs of being challenged after a high signal over time

Table 2: Probit Regression of Small Lies by Buyers

Buyers Belief that Seller Will Challenge a Low Announcement with High Signal	**(1)**	**(2)**
"Unlikely"	-0.027	-0.170 ***
	(0.089)	(0.064)
"Likely"	-0.040	-0.024
	(0.071)	(0.064)
"Always"	-0.127 *	-0.113 *
	(0.066)	(0.059)
Buyers Belief that Seller will Challenge a High Announcement with High Signal		
"Unlikely"	0.065	0.025
	(0.051)	(0.044)
"Likely"	0.397 ***	0.186 ***
	(0.070)	(0.055)
"Always"	0.590 ***	0.234 ***
	(0.074)	(0.063)
Constant	0.249 ***	0.325 ***
	(0.060)	(0.049)
Individual Fixed Effects	No	Yes
Time Fixed Effects	No	Yes
R^2	0.203	0.156
Observations	851	851

Dependent varaible is 1 if the buyer lies by announcing low with a high signal and 0 otherwise. The omitted category is Seller "Never" Challenges. Regression (1) is a linear probability model with errors clustered by individual. Regression (2) is a fixed effect regression with both time and individual fixed effects. *,**,*** denote siginificance at the 10%, 5% and 1% respectively.

4.2.1 Using incentive-compatible beliefs to better understand buyer lies

To explore further the way in which beliefs may be guiding lies in the no-noise treatment we ran an additional experiment in which we elicited probabilistic beliefs of being challenged using an incentive-compatible elicitation mechanism developed in Karni (2009).[16] In this follow-up treatment, we restricted attention to only the no-noise treatment and ran additional periods to study convergence. We ran two sessions with 30 periods and two sessions with 40 periods with random matching across periods. A total of 90 individuals participated in the experiment. The details of this elicitation mechanism can be found in the appendix.[17]

Result 2 *The majority of B's have pessimistic beliefs about being challenged after a truthful announcement of 70. The majority of B's have optimistic beliefs about being challenged after a lie of 20.*

Figure 5 compares the empirical challenge probability of S's to B's belief of being challenged. Both the means and 95 percent confidence intervals shown are calculated from individual averages. As can be seen on the right hand side of the figure, buyers in the environment are strongly pessimistic about the likelihood of being challenged after a truthful high announcement. While the empirical probability of being challenged is 9.1%, the average belief is 30.4%. This pessimism is prevalent across the population, with 80.1% of individuals having pessimistic beliefs about being challenged relative to the empirical distribution. The difference of beliefs and the empirical distribution is significant in both a simple t-test ($t = -5.379$, p-value $< .01$) and a Mann-Whitney-Wilcoxen test ($z = -5.125$, p-value $< .01$).[18]

Vice versa, buyers in the environment are optimistic about the likelihood of being challenged after a lie with a high signal. While S's challenge 85.0% of the time after a lie (a

[16] Akin to a standard BDM mechanism (Becker, DeGroot & Marschak, 1964), the belief elicitation mechanism gives B a dominant strategy to announce his true beliefs by using B's reported belief to assign him to one of two lotteries — one that is contingent on S's challenge decision and one that is independent of this decision — across a set of binary lottery pairs. We randomly select one of these lottery pairs to be played so that beliefs impact the assignment of B to a lottery but not the explicit characteristics of this lottery. We use the strategy method in this follow up experiment for S's challenge decisions as we want to elicit incentive-compatible beliefs from B about the likelihood of being challenged after a truthful announcement and after a lie. To do so we need to know S's challenge decision for both announcements. See the appendix for full details.

[17] As we were concerned with potential hedging, the follow-up experiment paid only for one period of the experiment and only for the announcement game or the belief elicitation game. There was a 50% chance that the announcement game would be paid and a 50% chance that one announcement-signal combination of the belief elicitation game would be paid. We set the exchange rate for this experiment at 1 ECU = \$1 AUD to maintain comparable incentives between experiments and paid a show-up fee of \$35 to avoid bankruptcies.

[18] Observations are an individual buyer's average belief and an individual S's average challenge rate over all periods.

15.0% deviation from the Nash Equilibrium), the average belief is 58.7% (a 41.3% deviation from the Nash Equilibrium). This optimism is again prevalent across the population, with 76.7% of individuals having optimistic beliefs about being challenged relative to the empirical distribution. The difference between beliefs and the empirical distribution is again significant (t-test: $t = 4.703$, p-value $< .01$; Mann-Whitney-Wilcoxen test: $z = 5.56$, p-value $< .01$).

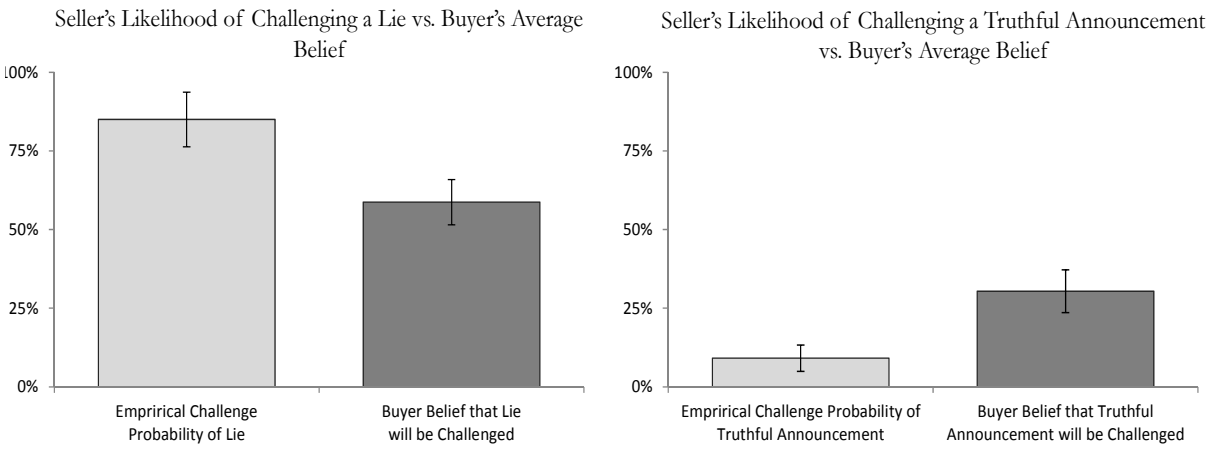


Figure 5: Buyer Beliefs About the Probability of Challenge Relative to Empirical Challenge Probabilities of Sellers

Given the optimistic beliefs about outcomes after a lie and pessimistic beliefs about outcomes after truthful announcing, a natural hypothesis is that B's may believe that they are monetarily better off lying than telling the truth. To test this hypothesis, we construct the expected value of lying and telling the truth after a high signal if B responds optimally to a challenge and has truthfully reported his beliefs. We next take the difference between these expected values to estimate the expected monetary gain from truth telling.

Result 3 *The majority of B's believe they have a higher expected value lying than they do making a truthful announcement. B's with more optimistic beliefs about being challenged after a lie and more pessimistic beliefs about being challenged after a truthful announcement are more likely to lie.*

Figure 6 shows the cumulative density function of the expected gain from truth telling split between observations where an individuals is lying (N = 543) and observations where an

individual is telling the truth (N = 491).[19] As can be seen, the empirical CDF of the expected monetary gain from truth telling for individuals who tell the truth first order stochastic dominates the CDF for individuals who lie, indicating that heterogeneity in beliefs is an important factor in the decision to announce truthfully.[20] For both distributions, however, the proportion of individuals where the expected monetary gain from truth telling is negative is large, with 79.2% (72.7%) of observations where the buyer lies (tells the truth) falling into this category.

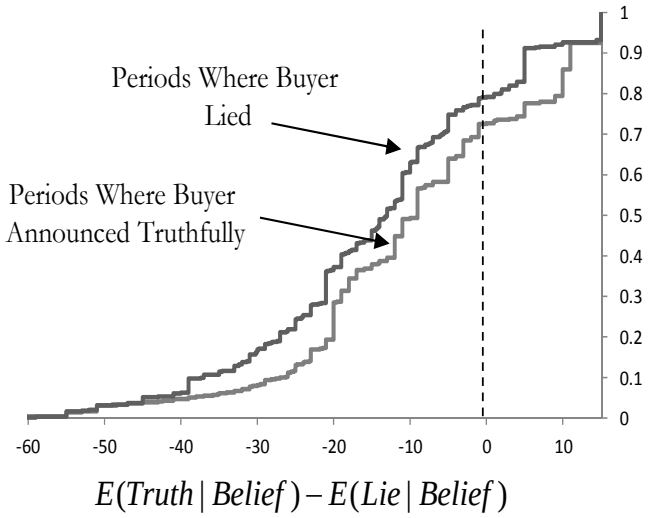


Figure 6: Cumulative Density Function of Expected Value of Telling the Truth split between observations where the buyer is lying (dark grey) and telling the truth (light grey).

One potential reason for the high level of pessimism seen in B's beliefs about being inappropriately challenged is that at least a subset of individuals are choosing announcement strategies that limit their ability to learn over time. 28.9% of individuals lie in each of the last 10 periods of the session and in at least 90% of periods overall. These individuals account for 60.7% of overall lies and 71.7% of lies that occur in the last 10 periods. As a B who lies in each period gets no new information about the likelihood of being challenged after a truthful announcement, the data suggests that alternative self-confirming equilibrium may be being selected in the game instead of the predicted subgame-perfect equilibrium as a result of B's

[19] We restrict attention to observations where (i) an individual believed that announcing low with a high signal had a higher chance of being challenged than announcing high with a low signal and (ii) the individual announced low with a low signal. The proportion of individuals where the expected monetary gain from truth-telling is negative in the full sample is 83.6%

[20] These distributions are significantly different in a bootstrapped version of the Mann-Whitney-Wilcoxen test where we randomly sampled a single period from each buyer in each iteration. p-value $< .01$. See Datta & Satten (2005) for a discussion.

actions and initial beliefs.[21]

Overall, our data suggests something of a paradox in the functioning of the Moore-Repullo mechanism. While the mechanism was designed to induce truth-telling using pecuniary incentives, most individuals who are truthful are distrustful of their partner and believe that such actions will lead to monetary loss. Truthful announcements are therefore being supported not by pecuniary incentives, but instead by non-pecuniary ones.

4.3 The Mechanism Under Almost-Perfect Information

Thus far we have seen that under perfect information the counter-offer condition and the appropriate-challenge condition hold but that there is a subset of individuals who fear that truthful announcements will be challenged and lie in order to mitigate this risk. Our theory model predicts that as we increase the amount of noise, two new violations will occur. First, as described in Hypothesis 2, S's are predicted to challenge low announcements with low signals (what we call a false challenge) and B's are predicted to accept some of these challenges. Second, as described in Hypothesis 3, B's with high signals are predicted to lie with greater frequency and S's are predicted to mix between accepting and rejecting these lies. We find support for most of these theoretical predictions:

Result 4 *The introduction of noise leads to false challenges by B's and acceptance of counter-offers by B's who made a low announcement with the low signals. The introduction of noise also leads to an increase in lies by B's and a small but insignificant decrease in challenges of low announcements by S's with a high signal.*

Figure 7 shows the proportion of S's who make a false challenges and the proportion of B's who accept a challenge with a low signal after a low announcement in each of the three treatments. The error bars show 95% confidence intervals of each proportion with standard errors clustered at the individual level. As can be seen on the left hand side, while there are very few false challenges in the no-noise treatment, the proportion of false challenges increases to 20.7% in the 5% noise treatment and 18.8% in the 10% noise treatment. Both noise treatments have significantly more false challenges than in their respective no-noise treatment based on a linear regression with errors clustered at the individual level (p-value $< .01$ in both cases).

[21] See Fudenberg, Kreps & Levine (1988), Fudenberg and Levine (1993) and Kalai and Lehrer (1993) for a discussion of self-confirming equilibrium. Notice that in our context, the consistent self-confirming equilibrium where B's always lie is a Nash Equilibrium, just not the subgame-perfect equilibrium that we are trying to implement.

As can be seen in the right hand side, B's are also much more likely to accept a challenge with a low signal and a low announcement in the noise treatments than in the no-noise treatment. While B's accepted a challenge after a low announcement and a low signal in only 2.4% of observations in the no-noise treatment, they accepted 27.7% of such challenges in the 5% noise treatment and 30.2% of such challenges in the 10% noise treatment. Both noise treatments have significantly more acceptances of challenges after a low announcement and a low signal than in their respective no-noise treatment based on a linear regression with errors clustered at the individual level (p-value $< .01$ in both cases).

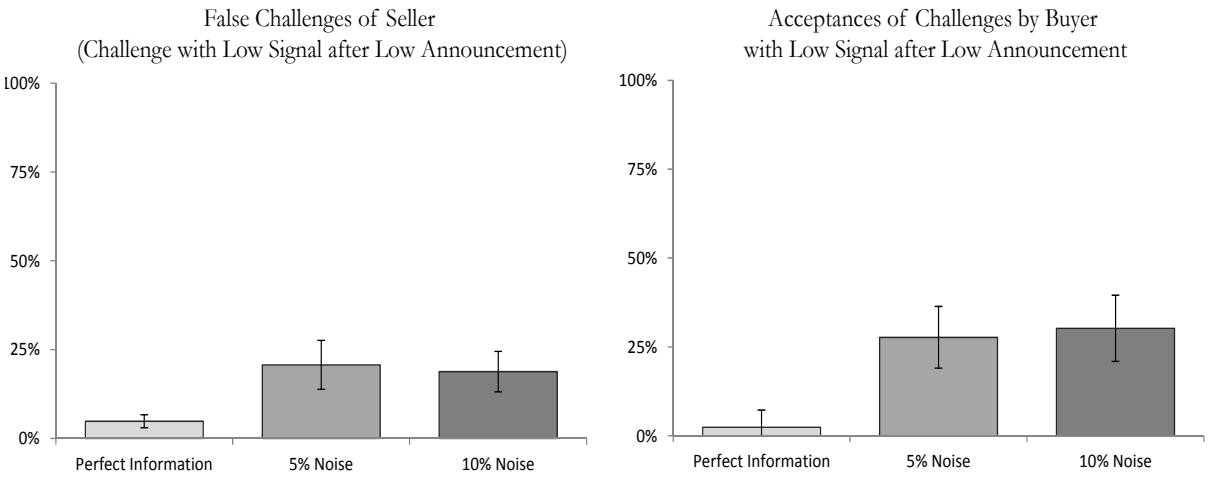


Figure 7: Seller's False Challenges and Buyer's Probability of Accepting a Challenge with Low Signal after a Low Announcement

Figure 8 shows the proportion of B's with a high signal who lie and the proportion of S's with a high signal who challenge after a low announcement across the three treatments. The error bars show 95% confidence intervals of each proportion with standard errors clustered at the individual level. As can be seen on the left hand side, B's lie in 45.9% of cases in the 5% noise treatment and in 52.2% of cases in the 10% noise treatment. Both of these lie rates are significantly higher than the no-noise treatment, where lies occur in 30.8% of cases (p-value $< .01$ in both cases). As can be seen on the right hand side of the figure, there is a small but not significant decrease in the challenges of low announcements with the high signal between the no-noise treatment, where 93.4% of cases were challenged, to the 5% and 10% noise treatments where the proportion of cases challenged were 85.2% and 88.6% respectively (10% noise treatments: p-value $= .351$; 5% noise treatment: p-value $= .147$). All three challenge rates are high, however, indicating that it would not be in B's interest to lie.[22]

[22] As shown in the next section, some of B's lies in the noise treatment can be attributed to individuals who fear an inappropriate challenge and will always accept a counter offer. Thus, it isn't too surprising that

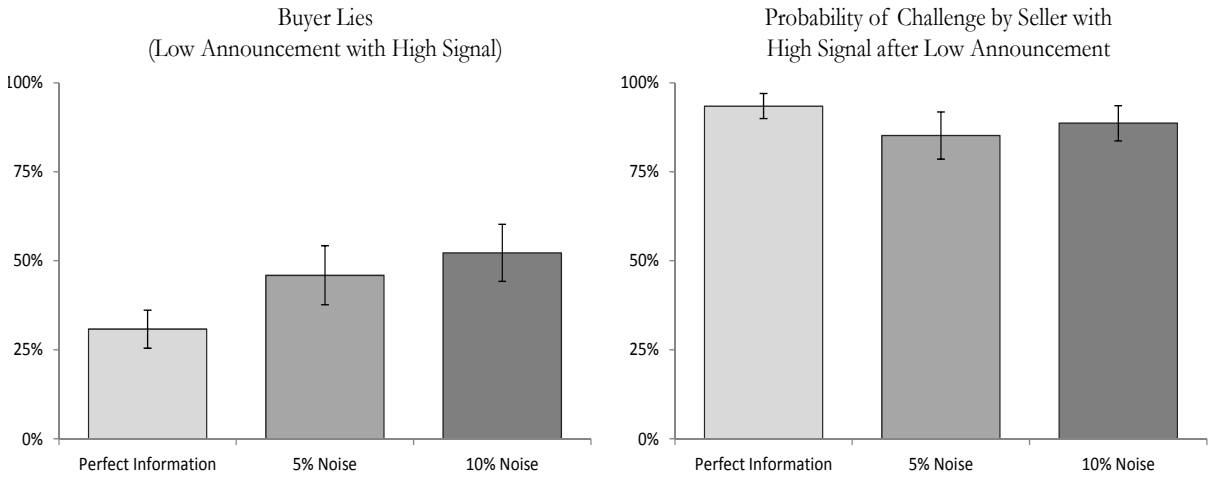


Figure 8: Buyer Lies and Seller's Probability of Challenging with High Signal after a Low Announcement

As foreshadowed by the increase in lies and the increase in false challenges that were subsequently rejected, the introduction of noise leads to a marked decrease in earnings. However, this decrease in earnings is asymmetric. As shown in Table 3, B's in the two noise treatments have significant reductions in their earnings relative to that of the no-noise treatment. S's, by contrast, have a very small decrease in earnings. This difference in the outcomes of B's and S's is due to the fact that B's who lie are frequently challenged and accept the counter offer in over 90% of these cases. Thus, they effectively are transferring money to the S's every time they lie.

	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 3: 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.

4.3.1 A direct test of the fear of false challenge hypothesis in the noise and no-noise treatments

As with the no-noise data, a likely reason that the lie rates are so high in the noise treatments is that some B's fear that truthful announcements will be challenged. Figure 9 shows the aggregate number of lies that each individual made in the no-noise treatment paired with

S's challenge rates are quite high in all treatments.

the aggregate number of lies they made in the noise treatment. While there is a significant increase in lies in the noise treatment for 60% of B's, there is also a subset of individuals who lied in every period of both the noise and no-noise treatment.

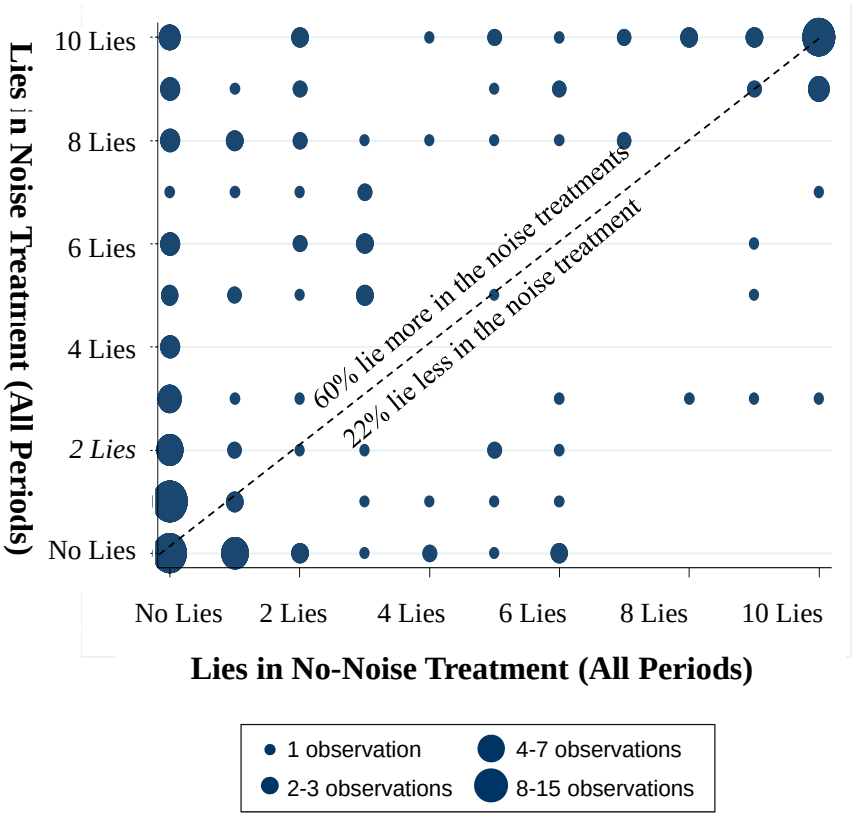


Figure 9: Aggregate Lies in the Noise and No-Noise Treatments

If the belief that truthful announcements will be challenged is the main driver of lies in the no-noise treatment and also drives a subset of lies in the noise treatment, then eliminating the potential of such challenges should increase the likelihood of truth telling in both treatments. We test this hypothesis by running four additional sessions with 10 periods of the no-noise treatment and 10 periods of the 10% noise treatment where we eliminated the ability for S to challenge an individual who makes a high announcement. Two of the sessions 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. This "no-inappropriate challenge" mechanism is expected to increase the expected gain from truth telling in both the noise and the no-noise treatments. We expect, therefore, that a large proportion of lies will decrease in this treatment relative to the baseline but that the gap between the no-noise and noise treatments will remain. A total of 82 individuals participated in these additional experiments.

Result 5 *Eliminating the ability of S to challenge high announcements dramatically reduces B's lies in both the no-noise treatment and the noise treatment. The introduction of noise leads to an increase in B's lies in both the baseline mechanism and the new mechanism.*

Figure 10 shows the proportion of lies in the original sessions with 10% noise and the new sessions using the no-inappropriate challenge mechanism. The error bars show 95% confidence intervals with standard errors clustered at the individual level. As can be seen, lies in both the noise treatment and the no-noise treatment decrease with the no-inappropriate challenge mechanism as we would expect if pessimistic beliefs about being challenged after a truthful announcement is a major contributor to lying.[23] This decrease in lies is particularly pronounced when comparing the second treatment in each session, where buyer lies fell to only 7.1% in the no-noise treatment and 27.0% in the 10% noise treatment.

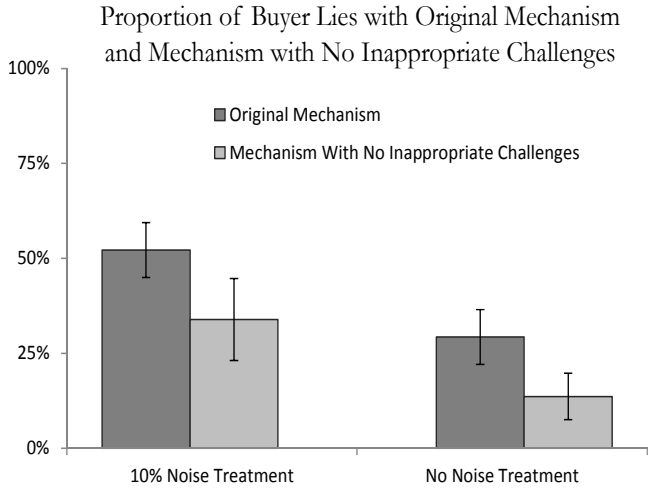


Figure 10: Frequency of Buyer Lies with original mechanism and alternative simple mechanism where high announcements cannot be challenged.

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. Moore (1992) calls mechanisms like this "simple sequential mechanisms" and provides conditions under which they can implement a desired social choice function. Roughly speaking, this requires that only one party has state dependent preferences, or that preferences are perfectly correlated.

[23]The difference in lie frequency between the original mechanism and the no false challenge mechanism 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).

4.4 Additional robustness tests

In the additional experiments where we elicited incentive-compatible beliefs, the overall lying rate was 46.3%. This lying rate was significantly larger than in our original experiments where lies occurred in 30.8% of observations (Mann-Whitney Wilcoxen Test: p-value $< .01$).

A likely explanation for the difference in lying rates between the two treatments is that the experiments with incentive-compatible beliefs used the strategy method. As discussed in the design section, this concern was the primary motivation for avoiding incentive-compatible belief elicitation methods in our original experiments and for adopting a simpler 4-point likert scale.

An alternative possibility, however, is that eliciting beliefs itself encourages B's to lie. This would be the case, for instance, if B's are primed to believe that S's will challenge them by being asked about this possibility. As this priming effect could affect our original experiment and partially explain buyer lies, we ran 4 additional sessions using 10 periods of the 5% noise treatment followed by 10 periods of the no-noise treatment where we did not elicit beliefs. We followed the payment protocol used in the incentive-compatible belief elicitation experiments where we pay a \$35 show-up fee and select a single period for payment, as the explicit discussion of bankruptcy in our original experiments may also be leading B's to mitigate risk through lies.[24] A total of 88 individuals participated in these additional experiment.

Result 6 *Lying rates in sessions without belief elicitation are not significantly different to their corresponding sessions in the original experiments.*

Figure 11 shows the distribution of buyer lies in the no-noise treatment of our original experiment and the experiment without beliefs. As can be seen, the data for both treatments has the same bimodal shape with a large portion of individuals never lying and a small fraction of individuals lying in every period. A Mann-Whitney-Wilcoxen test of the distribution of lies is not significant (p-value = .3303). There is also no significant difference in the distribution of lies in the noise treatments (p-value = .3930).

4.4.1 How small is small?

While we chose the levels of 5% and 10% noise based in order to have enough power to differentiate between treatments, the AFHKT suggests that very small levels of noise can lead to a breakdown of the mechanism. To study whether deviations from perfect information

[24]The additional experiments also included a longer set of quiz questions that checked understanding in all subgames. This was suggested by a reviewer as a way to improve common knowledge.

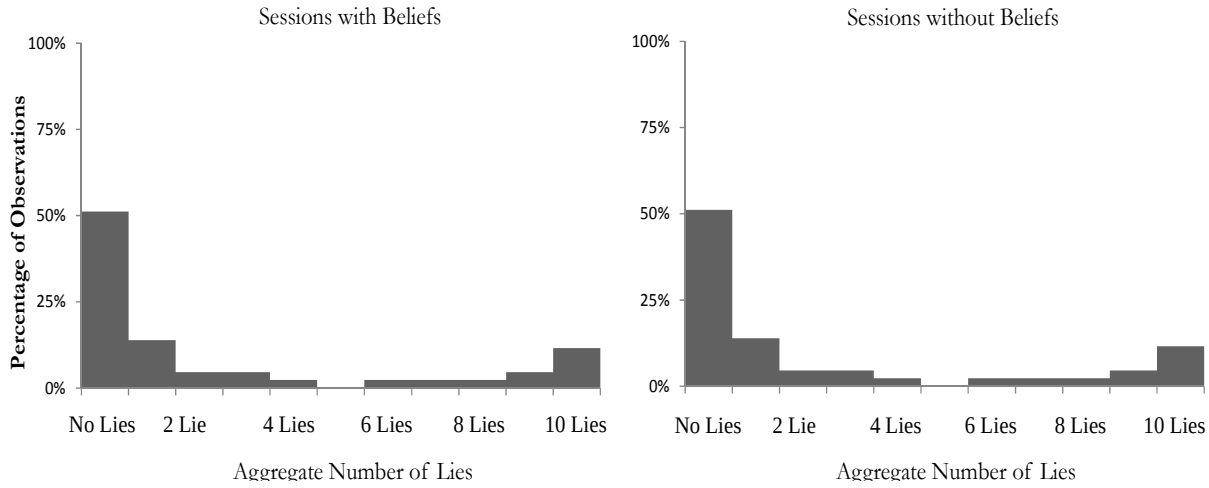


Figure 11: Distribution of Lies in Sessions With and Without Belief Elicitation

impact the distribution of lies even for very small levels of noise, we ran an additional four sessions without beliefs where we started with 10 periods of a 1% noise treatment and ended with a no-noise treatment. A total of 82 individuals participated in these additional experiments.

Result 7 *Even a very small perturbation in common knowledge leads to an increase in lies relative to the no-noise treatment.*

Figure 12 shows the proportion of buyer lies and seller false challenges in the noise treatment with 95% confidence intervals clustered at the individual level. The dotted lines in each figure show the proportion of buyer lies and seller false challenges in the subsequent no-noise treatment.

As can be seen in the left hand panel, the proportion of buyer lies observed in the 1% noise sessions is similar to the proportion of lies observed in the original 5% noise sessions and the 5% noise sessions without beliefs. There is no significant difference between this treatment and the other treatments based on a linear regression where buyer lies are regressed on the treatment dummy for the 5% noise sessions with beliefs (t = .76, p-value = .449) and the 5% noise sessions without beliefs (t = -.09, p-value = .927). Each of the noise treatments also has significantly more lies than in their paired no-noise treatment in a linear regression with errors clustered at the individual level (p-value $<$ .01 in all cases).

As can be seen in the right hand panel, the proportion of seller false challenges is smaller in the 1% noise session than in the other two treatments with lies occurring 10.3% of the time. Using the same specification as above, the proportion of false challenges in the 1% noise sessions is significantly less than the number in the 5% noise sessions without beliefs (t = 2.00, p-value = .047) and close to the significance threshold in the sessions with beliefs

(t = .192, p-value = .057). However, there is no significant difference in the proportion of seller false challenges between the 1% noise treatment and its corresponding no-noise treatments in a linear regression with errors clustered at the individual level (t = 1.48, p-value = .147).

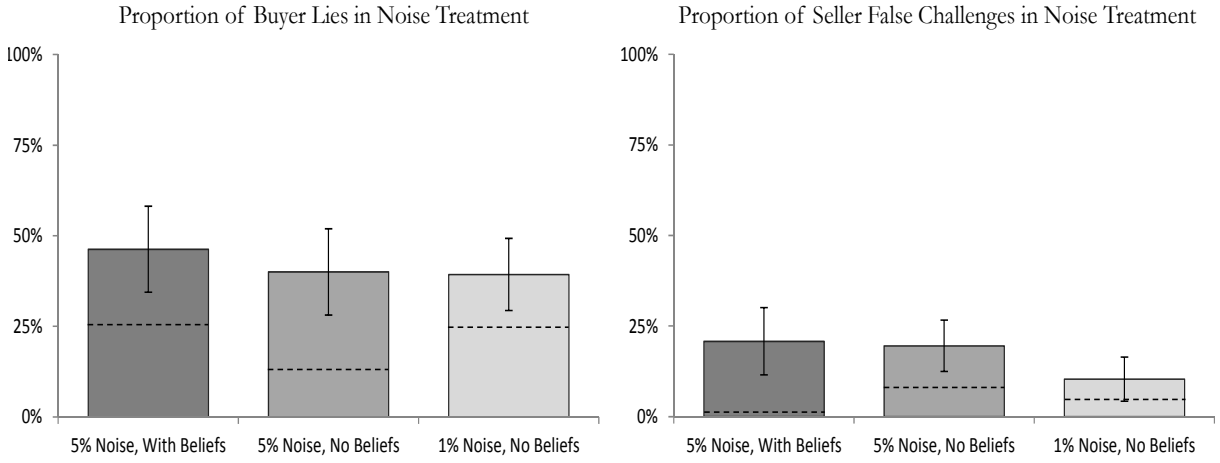


Figure 12: Difference in Lies between Noise and No-Noise Treatments

Taken together, while there is a small reduction in seller false challenges when noise rates decline, the large number of buyer lies in the 1% noise treatment illustrates that even small departures from common knowledge has a significant impact on the willingness of individuals to report truthfully. Our results thus illustrate the non-robustness of the Moore-Repullo mechanism to small amounts of noise.

5 Conclusion

In this paper we conducted a laboratory experiment to test the extent to which Moore and Repullo's subgame perfect implementation mechanism induces truth-telling in practice, both in a setting with perfect information and in a setting where buyers and sellers do not share common knowledge about the good's value. Our first finding is that even in the no-noise treatment, where no lies are predicted in equilibrium, buyers lie by announcing a low value with a high signal roughly 25% of the time. Our data suggests that in all treatments a substantial proportion of these lies are driven by pessimism about being inappropriately challenged after a high announcement. This pessimism is strong enough that a large majority of individuals who are telling the truth believe they would be better off lying and suggests that the mechanism is being supported in part by non-pecuniary incentives for telling the truth.

Our second main finding is that the introduction of noise leads to an increase in buyers' lies and sellers' false challenges in a way consistent with the analysis in AFHKT. The introduction of noise increases the proportion of buyers who announce a low value with a high signal by 15 to 25 percentage points; these lies are persistent and do not diminish with experience. Similarly, the proportion of sellers who falsely challenge in the noise treatments increases by 15 percentage points relative to the no-noise treatment. These deviations exist even when the level of noise is reduced to a very small 1% level.

If we adjust the Moore-Repullo mechanism by ruling out such false challenges, buyers lying rate in the no-noise treatment decreases by 15.6 percentage points. Likewise, the institutional removal of such false challenges also decreases the lying rate in the noise treatments significantly. However, in the noise treatments this deviation from the Moore-Repullo mechanism does not solve the lying problem. Even if the fear of false challenges of high announcements is ruled out a lying rate of 27% prevails in the 10% noise treatment, which indicates the pervasive influence of uncertainty regarding the good's value on lying behavior.

Our findings suggest important avenues for future research. First, the fact that individuals are willing to sacrifice their material well being to tell the truth suggests that preference for honesty should help implementation.[25] Second, in view of the empirical relevance of common knowledge, it also is important to design mechanisms that are robust to at least small amounts of imperfect information about the good's value. Third, it would be interesting to know (theoretically and empirically) how the introduction of asset ownership affects the functioning of 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 good's value truthfully or to challenge the other party. 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. Finally, similar experiments could be used to test the robustness of other implementation mechanisms, starting with virtual implementation. Overall, our analysis and findings in this paper raise a number of exciting issues to be tackled by future research.

References

Aghion, P., Fudenberg, D., Holden, R., Kunimoto, T. & Tercieux, O. (2012), 'Subgame-perfect implementation under value perturbations', *Quarterly Journal of Economics*

[25] Current research by Holden, Kartik, and Tercieux (2014) suggests that when individuals have a known preference for honesty, full implementation can be achieved with simple mechanisms requiring only two rounds of iterated deletion of strictly dominated strategies.

127(4), 1843–1881.

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* **25**(2), 181–197.

Andreoni, J. & Varian, H. (1999), 'Pre–play contracting in the prisoners' dilemma', *Proceedings of the National Acadamy of Science of the United States of America* **96**, 10933–10938.

Arifovic, J. & Ledyard, J. (2004), 'Scaling up learning models in public good games', *Journal of Public Economic Theory* **6**(2), 203–238.

Attiyeh, G., Franciosi, R. & Isaac, R. (2000), 'Experiments with the pivot process for providing public goods', *Public Choice* **102**(1-2), 95–114.

Blanco, M., Engelmann, D., Koch, A. K. & Normann, H.-T. (2010), 'Belief elicitation in experiments: Is there a hedging problem?', *Experimental Economics* **25**(4), 412–438.

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

Bracht, J., Figuires, C. & Ratto, M. (2008), 'Relative performance of two simple incentive mechanisms in a public goods experiment', *Journal of Public Economics* **92**(12), 54 – 90.

Cabrales, A., Charness, G. & Corchon, L. (2003), 'An experiment on Nash implmentation', *Journal of Economic Behavior & Organization* **51**, 161–193.

Charness, G., Cobo-Reyes, R., Jimenez, N., Lacomba, J. A. & Lagos, F. (2012), 'The hidden advantage of delegation: Pareto improvements in a gift exchange game', *American Economic Review* **105**(5), 2358–79.

Chen, Y. (1996), 'The Groves–Ledyard mechanism: an experimental study of institutional design', *Journal of Public Economics* **59**(3), 335–364.

Chen, Y. & Plott, C. (1996), 'The groves–ledyard mechanism: An experimental study of institutional design', *Journal of Public Economics* **59**(3), 335–364.

Chen, Y. & Tang, F. (1998), 'Learning and incentive-compatible mechanisms for public goods provision: an experimental study', *Journal of Political Economics* **106**(3), 633–662.

Dufwenberg, M. & Kirchsteiger, G. (2004), 'A theory of sequential rationality', *Games and Economic Behavior* **47**, 268–298.

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

Ederer, F. & Fehr, E. (2009), Deception and incentives: How dishonesty undermines effort provision, Working paper.

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

Falkinger, J., Fehr, E., Gächter, S. & Winter-Ebrner, R. (2000), 'A simple mechanism for the efficient provision of public goods: experimental evidence', *American Economic Review* **90**(1), 247–264.

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. (2014), 'Handing out guns at a knife fight: Behavioral limitations of subgame perfect implementation', 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. (1993), 'Self–confirming equilibrium', *Econometrica* **61**(3), 523 – 545.

Gneezy, U. (2002), 'Deception: The role of consequences', *American Economic Review* **95**(1), 384 – 394.

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.

Harstad, R. M. & Marese, M. (1981), 'Implementation of mechanism by processes: public good allocation experiments', *Journal of Economic Behavior & Organization* **2**(2), 129–151.

Hart, O. & Moore, J. (2003), 'Some (crude) foundations for incomplete contracts', Mimeo.

Healy, P. J. (2006), 'Learning dynamics for mechanism design: An experimental comparison of public goods mechanisms', *Journal of Economic Theory* **129**(1), 114 – 149.

Huck, S. & Weizsäcker, G. (2002), 'Do players correctly estimate what others do?: Evidence of conservatism in beliefs', *Journal of Economic Behavior & Organization* **47**(1), 71 – 85.

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

Kalai, E. & Lehrer, E. (1993), 'Rational learning leads to nash equilibrium', *Econometrica* **61**(5), 1019 – 1045.

Karni, E. (2009), 'A mechanism for eliciting probabilities', *Econometrica* **77**(2), 603–606.

Kartik, N., Tercieux, O. & Holden, R. (2014), 'Simple mechanisms and preferences for honesty', *Games and Economic Behavior* **83**(C), 284 – 290.

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**(1), 39–56.

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

Masuda, T., Okano, Y. & Saijo, T. (2014), 'The minimum approval mechanism implements the efficient public good allocation theoretically and experimentally', *Games and Economic Behavior* **83**(1), 73–85.

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* **56**(5), 1191–1220.

Sanchez-Pages, S. & Vorsatz, M. (2007), 'An experimental study of truth-telling in a sender-receiver game', *Games and Economic Behavior* **61**(1), 86 – 112.

Sefton, M. & Yavas, A. (1996), 'Abreu-matsushima mechanisms: experimental evidence', *Games and Economic Behavior* **16**(2), 280–302.

Appendix A: Point Predictions of the Mixed Strategy Equilibrium

In this section, we derive the point predictions of the mixed strategy equilibrium of our game for each of our three noise treatments. We begin with the standard model where all participants are risk neutral and selfish. We then show how the point predictions of the model change when buyers receive positive utility for rejecting a challenge of the seller. This alternative model is discussed in detail in Fehr, Powell, and Wilkening (2010) where a sequential reciprocity equilibrium in the spirit of Dufwenberg & Kirchsteiger (2004) is developed.

The mixed strategy equilibrium with selfish risk-neutral buyers

As in the main text, let the true valuation of the good be $\theta \in \{\theta^H = 70, \theta^L = 20\}$, with both states being equally likely. Let each player receive one of two possible signals, s^H and s^L, where s^H is a high signal correlated with θ being equal to 70, and where s^L is a low signal 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
$\theta = 70$	$\frac{1}{2}(1-\varepsilon)^2$	$\frac{1}{2}\varepsilon(1-\varepsilon)$	$\frac{1}{2}\varepsilon(1-\varepsilon)$	$\frac{1}{2}\varepsilon^2$
$\theta = 20$	$\frac{1}{2}\varepsilon^2$	$\frac{1}{2}\varepsilon(1-\varepsilon)$	$\frac{1}{2}\varepsilon(1-\varepsilon)$	$\frac{1}{2}(1-\varepsilon)^2$

For a given noise level ε, an action profile of a buyer consists of a probability of announcing low after observing each signal and a probability of rejecting the challenge given a signal and an announcement. Denote L^H as the probability of making a *low* announcement after observing a high signal and L^L as the probability of making a low announcement after a low signal. Further, let $R^{a_B|s_B}$ be the probability that the buyer rejects a challenge given his own announcement $a_B \in \{L, H\}$, his own signal $s_B = \{L, H\}$ and a challenge by the seller.

An action profile of the seller consists of a probability of challenging an announcement of the buyer for each potential announcement and signal. Let $C^{a_B|s_S}$ be the probability that the seller challenges given signal $s_S \in \{L, H\}$ and an observed announcement of the buyer $a_B = \{L, H\}$.

While there are 10 potential mixing probabilities to specify in an equilibrium, we can use some of the structure of the mechanism to rule out mixing on some action sets. A buyer who announces high and is challenged faces a price of 75 which is above his actual value

of the good regardless of the state. Thus the buyer will always reject arbitration if he has announced high and $R^{H|L} = R^{H|L} = 1$. This also implies that the seller will never call the arbitrator if the buyer announces high, and thus $C^{H|L} = C^{H|H} = 0$. Further, a buyer who has a high signal and announces low will update his belief about the quality of the good based on the probability that the seller challenges. However, for any equilibrium where the seller challenges with positive probability, the most pessimistic posterior the buyer can have after being challenged is that the state is low with probability 1/2. As the counteroffer price is 25 and the buyer's expected value for accepting the good with this belief is 45, the buyer will always accept the counteroffer, and thus $R^{L|H} = 0$. Finally, the best a buyer can do with a low signal if he always announces high is to receive 35 with probability ε and -15 with probability $1 - \varepsilon$. If in equilibrium the buyer earns more than $35\varepsilon - 15(1 - \varepsilon)$ for a low announcement, it will be the case that $L^L = 1$.[26]

Taking as given the actions of buyers and sellers in the six states specified above, the mixed strategy equilibria are based on (i) the proportion of times a buyer announces low given a high signal, L^H, (ii) the challenge probabilities given a low announcement, $C^{L|L}$ and $C^{L|H}$, and (iii) the probability that the buyer rejects a challenge given a low signal, a low announcement, and a challenge, $R^{L|L}$. These four mixing probabilities form the basis of all PBE where all stages of the subgame are reached and beliefs of both parties are consistent with the action profiles of the other party.

Given that beliefs of all parties must be consistent with their actions, a necessary condition for the mixed strategy equilibrium is that each individual is indifferent between each of their actions given the mixing probabilities of the other parties. These indifference conditions generate four linear constraints on the four mixing probabilities of the buyer and seller and generate a four-by-four linear system which derives unique point predictions. The construction of each linear constraint is as follows:

(1) Buyer's indifference between announcing low and high with a high signal: For the buyer to be indifferent between announcing high and low, the expected value of these announcements must be equal when aggregated over all potential states of nature.

Panel (a) of Figure 13 shows the four potential states of nature where the buyer can have a high signal after nature draws the true value of the container and (conditional) signals for the buyer and seller. For each state, the expected value of each potential announcement is shown as a function of the challenge probabilities of the seller. For example, as seen on the far left of the figure, with probability $\frac{1}{2}\varepsilon(1 - \varepsilon)$, the buyer receives the high signal, the seller receives the low signal, and the true state of nature is low. If in this state the buyer

[26] We argue in the main text that there is a pure strategy equilibrium where $L^L = 0$ and challenges never occur.

announces low, he will not be challenged $1-C^{L|L}$ percent of the time and be challenged $C^{L|L}$ percent of the time. As he has the high signal, he will always accept the counteroffer and thus these two outcomes yield values of $20 - P_{20} = 10$ and $20 - F - P_A = -30$ respectively. If, on the other hand, the buyer announces high, he will never be challenged (since $C^{H|L} = 0$) and receive $20 - P_{70} = -15$ for sure.

Taking into account the probability of each one of these potential states and the state's outcome, a buyer is indifferent between a high and low announcement if:

$$\psi(\varepsilon)C^{L|H} + \delta(\varepsilon)C^{L|L} = \frac{P_{70} - P_{20}}{F + P_A - P_{20}}, \tag{4}$$

Where $\psi(\varepsilon) = \varepsilon^2 + (1-\varepsilon)^2$ is the probability that the signals are the same for a given ε and $\delta(\varepsilon) = 2\varepsilon(1-\varepsilon)$ is the probability that they are different.

(2) Buyer's indifference between accepting and rejecting a challenge with a low signal and low announcement: In an equilibrium in which the seller is mixing between challenging and not challenging a low announcement with a low signal, it must be the case that the buyer is also indifferent between rejecting and accepting such a challenge. Panel (b) of Figure 13 shows the probability of reaching this acceptance and rejection as a function of the signals and the challenge probabilities of the seller and under the assumption that $L^L = 1$. Taking into account the probability of each of these potential states and the state's outcome, a buyer is indifferent between rejecting and accepting the challenge if:

$$C^{L|L} - \tau(\varepsilon)C^{L|H} = 0, \tag{5}$$

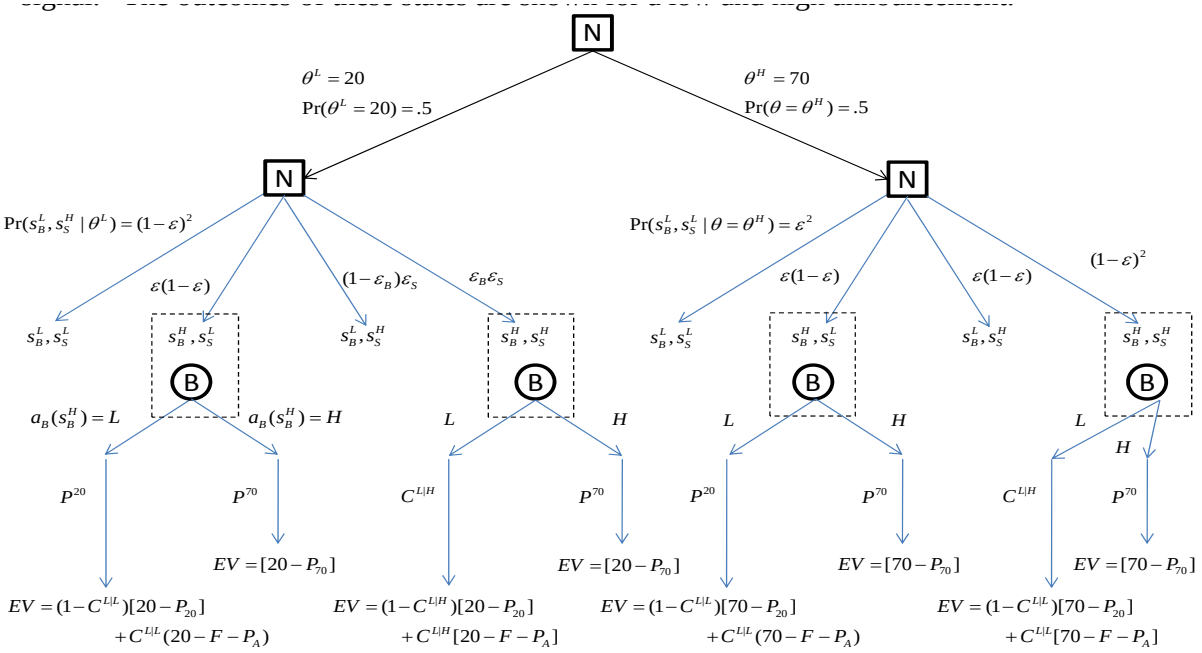
where

$$\tau(\varepsilon) = -\frac{\varepsilon(1-\varepsilon)[70 - P_A] + (1-\varepsilon)\epsilon[20 - P_A]}{\varepsilon^2[70 - P_A] + (1-\varepsilon)^2[20 - P_A]} \tag{6}$$

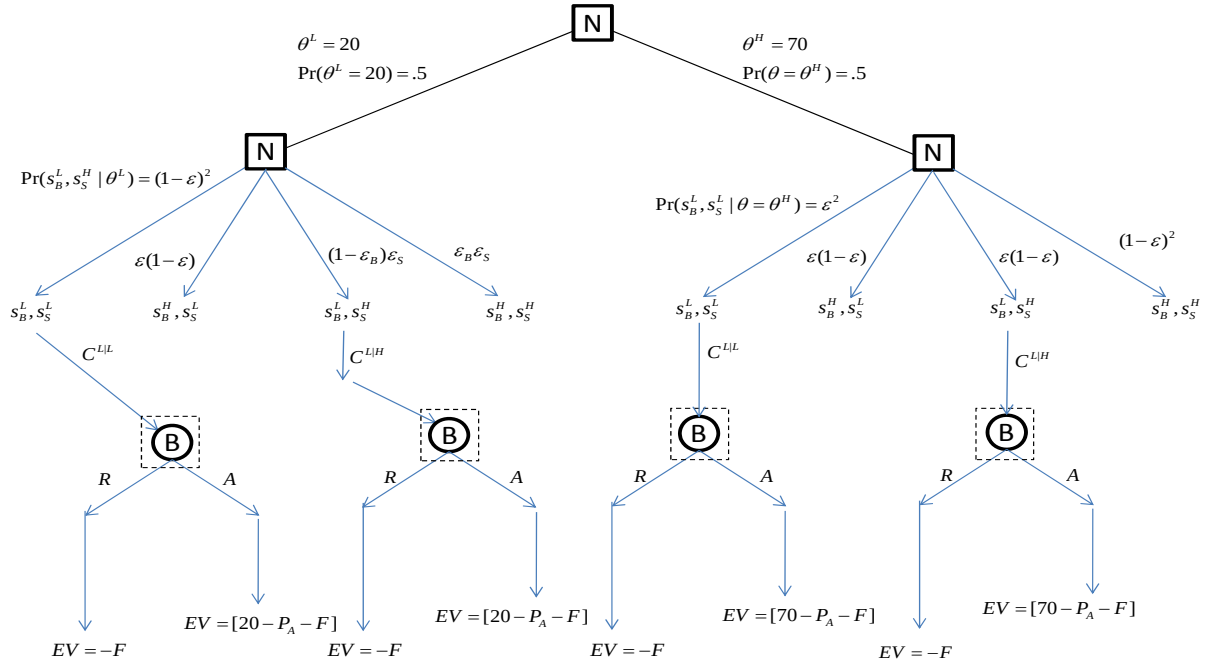
is the ratio of expected outcomes when the two parties have opposite signals relative to when they have the same signal. Note that $\tau(\varepsilon)$ is positive for all ϵ we consider since the denominator is negative.

Seller's indifference between challenging and not challenging after a low signal: As with the buyer, the seller's indifference for challenging after a low and high signal are based on the two mixing probabilities of the buyer. Panel (a) of Figure 14 shows the expected value for challenging and not challenging for states of the world where the seller has a high signal and observes a low announcement. The likelihood of reaching each of these potential states is based on the likelihoods that the buyer will make a low announcement with each signal (L^H and $L^L = 1$) while the expected value of challenging is based on the likelihood that the buyer will accept this challenge ($R^{L|L}$ and $R^{L|H} = 1$). A seller is indifferent to challenging

Figure 13: States Contributing to the Decision of the Buyer to Lie and Reject a Potentially False Challenge



(b) The four potential states which contribute to a buyer's decision to accept or reject a potentially false challenge. The outcomes of these states are shown for a rejected and accepted counteroffers.



and not challenging with the high signal if:

$$-L^H + \frac{\delta(\varepsilon)}{\psi(\varepsilon)} \frac{P_A + 2F}{P_A + F - P_{20}} R^{L|L} = \frac{\delta(\varepsilon)}{\psi(\varepsilon)} \tag{7}$$

where, as before $\psi(\varepsilon) = \varepsilon^2 + (1-\varepsilon)^2$ is the probability that the signals are the same for a given ε and $\delta(\varepsilon) = 2\varepsilon(1-\varepsilon)$.

Seller's indifference between challenging and not challenging after a high signal: Panel (b) of figure 14 shows the expected value for challenging and not challenging for states of the world where the seller has a low signal and observes a high announcement. As before, the seller's likelihood of reaching each potential state depends on L^L while the expected value within these nodes depends on $R^{L|L}$. A seller is indifferent to lying and not lying if:

$$-L^H + \frac{\psi(\varepsilon)}{\delta(\varepsilon)} \frac{P_A + 2F}{P_A + F - P_{20}} R^{L|L} = \frac{\psi(\varepsilon)}{\delta(\varepsilon)}. \tag{8}$$

Note that this is identical to the seller's indifference condition for challenging with the low signal except that the ratio of states is inverted.

Given the four indifference conditions, the point predictions of the model come from solving the four-by-four system of simultaneous equations. The solution to this system is as follows:

Result 8 *With selfish agents, the mixed strategy equilibrium with* $\varepsilon = .05$ *is* $L^H = 0$, $R^{L|L} = .53333$, $C^{L|H} = .66$, *and* $C^{L|L} = .285$. *The mixed strategy equilibrium with* $\varepsilon = .1$ *is* $L^H = 0$, $R^{L|L} = .53333$, $C^{L|H} = .625$, *and* $C^{L|L} = .625$.

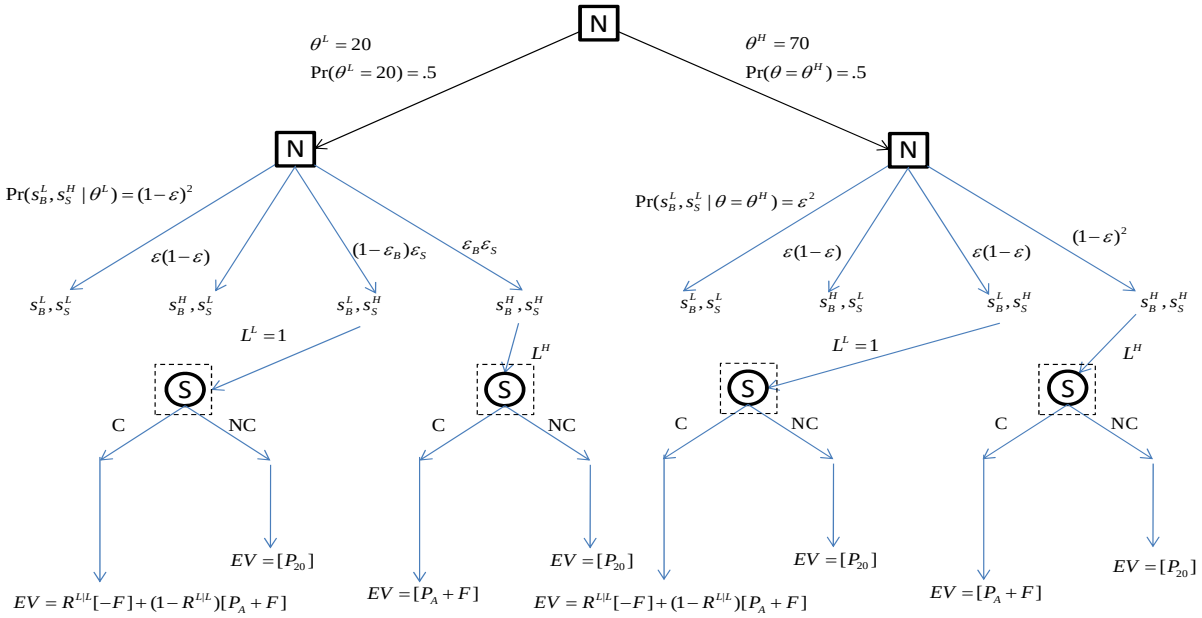
The surprising restriction that $L^H = 0$ is due to the fact that the seller must be indifferent to mixing in the case of a high and low signal. In the next section, we show that when buyers have negative reciprocity and wish to retaliate against the seller for a challenge, the buyer may strictly prefer to reject after a low signal and a low announcement. This (likely) scenario eliminates seller false challenges (i.e. sets $C^{L|L}$ to zero) and instead leads to buyer challenges.

5.1 Point Predictions with Negative Reciprocity

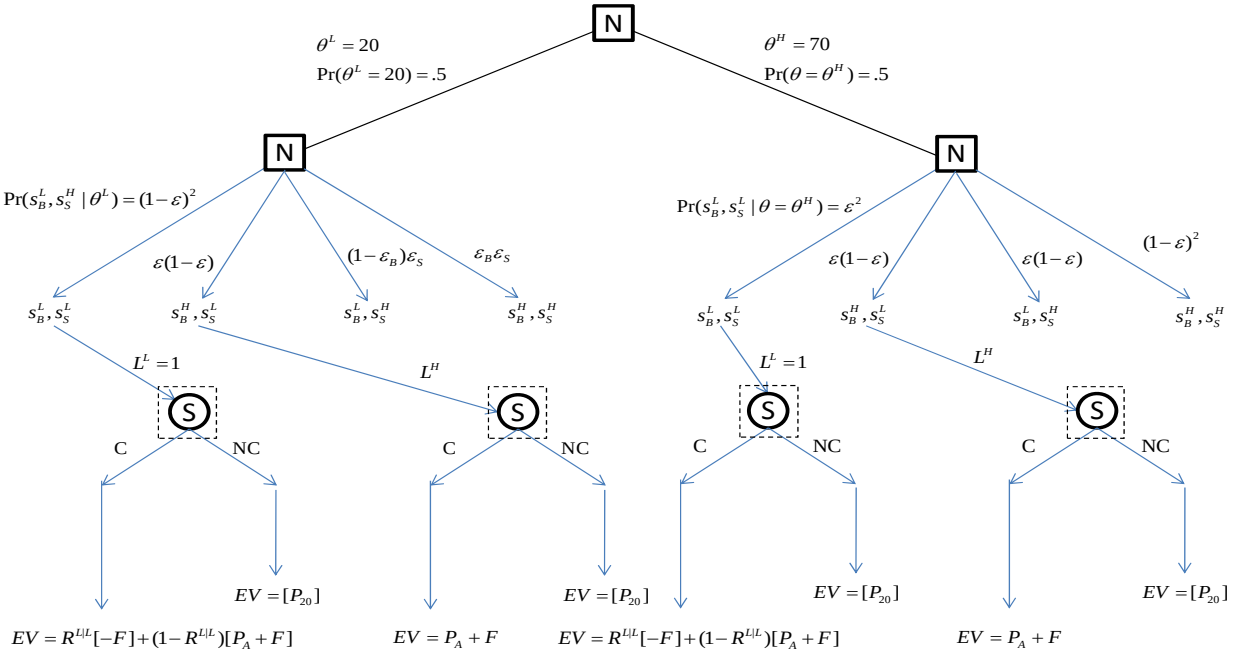
In FPW, we show that buyers who view a seller's challenge as an unkind act may retaliate against the sellers by rejecting appropriate counteroffers where the expected value for accepting the counteroffer is small. While the parametrization of the Moore-Repullo mechanism is set to make such retaliation unlikely without noise, there is no easy way to avoid retaliation from affecting the point predictions of the mixed strategy equilibrium where, by construction, the buyer is indifferent between accepting and rejecting a challenge. In this section, we

Figure 14: States Contributing to the Decision of the Seller to Challenge with a High and Low Signal

challenge



(a) (b) The four states which contribute to a seller's decision to challenge a low announcement when observing a low signal. The outcomes of these states are shown in the case of a challenge and no challenge



discuss how the point predictions of the model changes when buyers become more reciprocal and the level of buyer reciprocity is common knowledge.

Following Dufenberg & Kirchsteiger (2004), FPW shows that in a psychological games framework, a challenge by the seller is always seen by the buyer as an unkind act. Buyers who are prone to negative reciprocity may gain a "psychological" payoff by reducing the payoff of the seller and rejecting the counteroffer. Rather than reconstructing the entire arguments of these previous works, we use this insight in a reduced form way. Let the utility of a buyer who rejects the sellers counteroffer be $-F + 75\rho$, where ρ is the additional utility the buyer receives from rejecting the sellers counteroffer and reducing the seller's payoff by 75. Note that since 75ρ is the amount of money that the buyer is willing to leave on the table to reject a counteroffer of the seller, ρ can be thought of as the amount $\$x$ that the buyer is willing to forego to punish the seller by \$1.

Figure 15 maps the point predictions of the four mixing probabilities as ρ increases. As can be seen by looking at the point prediction of $C^{L|L}$, negative reciprocity initially reduces the proportion of false challenges. Such reductions in the likelihood of challenges increase the utility of the buyer for accepting the counteroffer and offset the utility from retaliation. As ρ increases, there exists a cutoff point for which the buyer will reject the counteroffer even if the seller never lies. At this point $C^{L|L} = 0$ and mixing occurs only over L^H and $C^{L|H}$. For very high ρ, outside the range of parameters seen in FPW, negative reciprocity can lead to the buyer rejecting challenges even when he has the high signal.

Based on the levels of negative reciprocity estimated in FPW, we expect ρ to range between .2 and .4. At these parameter values, mixing typically occurs over L^H and $C^{L|H}$. Note, however, that for all reciprocity levels below $\rho = .6$, the overall amount of buyer lies (L^H) plus seller false challenges ($C^{L|L}$) is unambiguously increasing in noise.

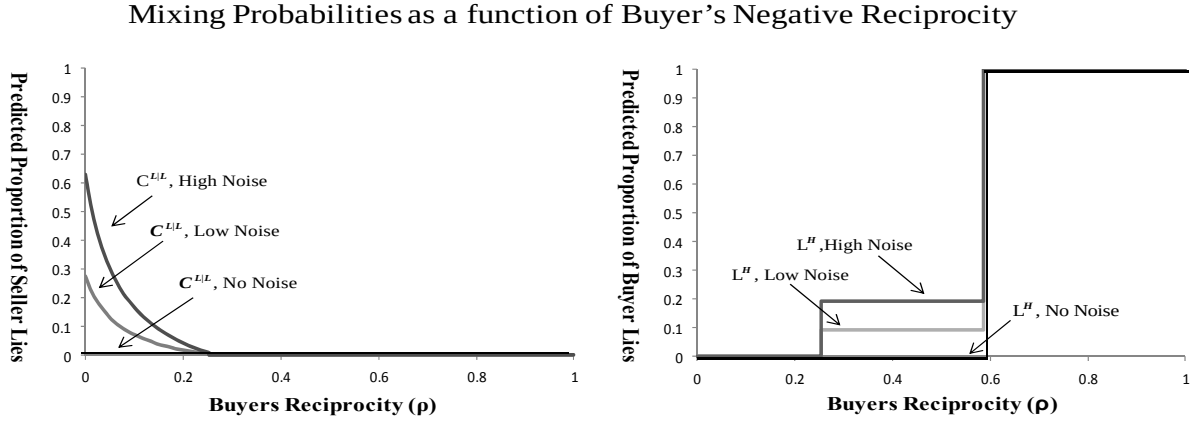


Figure 15: Point predictions for seller false challenges and buyer lies as a function of buyer reciprocity