NUS RMI Working Paper Series – No. 2019-02

Getting Dynamic Implementation to Work

Yi-Chun CHEN, Richard HOLDEN, Takashi KUNIMOTO, Yifei SUN and Tom WILKENING

18 September 2019

NUS Risk Management Institute
21 HENG MUI KENG TERRACE, #04-03 I3 BUILDING, SINGAPORE 119613

www.rmi.nus.edu.sg/research/rmi-working-paper-series

Getting Dynamic Implementation to Work*

Yi-Chun Chen† Richard Holden‡ Takashi Kunimoto§ Yifei Sun¶
Tom Wilkening‖

September 18, 2019

Abstract

We develop a new class of two-stage dynamic mechanisms, which fully implement any social choice function under initial rationalizability in complete information environments. We show theoretically that our mechanism is robust to small amounts of incomplete information about the state of nature. We also highlight the robustness of the mechanism to a wide variety of reasoning processes and behavioral assumptions. We show experimentally that the mechanism has good performance in inducing truth-telling in both complete and incomplete information environments and that it outperforms subgame-perfect implementation mechanisms based on the design of Moore and Repullo (1988).

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

*We thank Sandro Brusco, Yeon-Koo Che, Navin Kartik, Siqi Pan, Patrick Rey, Steve Williams, and Jun Xiao for helpful comments and discussions, and seminar participants at Osaka University, Cardiff Business School, Monash, NYU-Shanghai, UTS, UWA, the 10th The Annual Organizational Economics Workshop at Sydney, the 18th Annual SAET Conference, and the 6th Xiamen University International Workshop on Experimental Economics. We gratefully acknowledge the financial support of the Australian Research Council including ARC Future Fellowship FT130101159 (Holden) and ARC Discovery Early Career Research Award DE140101014 (Wilkening).

†Department of Economics and Risk Management Institute, National University of Singapore, Singapore 117570, ecsycc@nus.edu.sg

‡School of Economics, UNSW Sydney Business School

§School of Economics, Singapore Management University, Singapore, 178903, tkunimoto@smu.edu.sg

¶School of International Trade and Economics, University of International Business and Economics, Beijing 100029, sunyifei@uibe.edu.cn

‖Department of Economics, University of Melbourne

1 Introduction

In an instantly classic paper, Maskin (1977, 1999) asked what social objectives can be implemented in a decentralized environment that respects the individual incentives of participants. Maskin showed that with a suitably constructed game form one can implement a class of social choice functions — so-called "monotonic" SCFs — in Nash equilibrium. Monotonicity is, however, somewhat restrictive. In particular, it does not allow for SCFs with distributional considerations. Since then, there has been substantial interest in using extensive form mechanisms, as they hold the prospect of using refinements of Nash equilibrium (such as subgame perfection) to implement non-monotonic SCFs.

Moore and Repullo (1988) illustrate the potential of extensive form mechanisms by showing that one can implement any social choice function—Maskin monotonic or not—using a suitably constructed three-stage mechanism. However, subsequent work has raised concerns about the sensitivity of their solution concept to common knowledge assumptions regarding rationality, payoffs, or preferences. For instance: Fudenberg et al. (1988) and Dekel and Fudenberg (1990) show that refinements of Nash equilibria may not be robust to the introduction of a small number of "crazy" types and thus may not be a good prediction of actual behavior. Aghion et al. (2012) and Aghion et al. (2018) show that extensive-form mechanisms are not robust to small deviations from common knowledge about the state of nature[1], while Fehr et al. (2017) show that heterogeneity in reciprocal preferences can cause subgame-perfect equilibrium mechanisms to break down.

A central characteristic of all extensive-form mechanisms that are based on subgame perfection is that deviations are always considered to be "one-shot deviations in behavior" that do not shatter the faith players have in the subsequent behavior of the deviating player. This faith is unwarranted (and in fact contrary to Bayes Law) when the assumptions of common knowledge of rationality, payoffs or preferences are relaxed. In such situations, belief updating occurs along the dimension of uncertainty leading to equilibria that may be far away from the intended equilibrium even when uncertainty is small.

This paper takes the perspective that all real-world economic environments have noise. Thus, from a practical perspective, subgame perfection is a non-robust solution concept to use for implementation. To identify a more robust solution concept, we focus squarely on environments with noise and explore dynamic implementation both theoretically and experimentally when imposing less stringent assumptions on how beliefs evolve. By relaxing beliefs, we can tackle the issue of updating in noisy environments head-on in both the

[1]See also Monderer and Samet (1989), Kajii and Morris (1997) for concerns of robustness to perturbations in normal form games.

construction of new mechanism and the solution concept we employ.

Following Ben-Porath (1997) and Dekel and Siniscalchi (2013), we use the notion of *initial rationalizability* as our solution concept. Like rationalizability in normal-form games, this solution concept iteratively deletes strategies that are not best replies. However, unlike backward induction, it requires that there be rationality and common beliefs of rationality only at the beginning of the game and makes no assumption about how beliefs evolve after zero probability events occur. Accommodating any belief revision assumption at any subsequent stages of the game when a zero-probability event occurs, initial rationalizability is the weakest rationalizability concept among all extensive-form games. Hence, implementation under initial rationalizability is the most robust notion of implementation among existing concepts in dynamic mechanisms.

Part one of our paper provides very permissive implementation results when using initial rationalizability as a solution concept. Before getting into the details, we want to be clear from the outset about the domain of problems in which our results apply. First, we consider environments where monetary transfers among the players are available and all players have quasilinear utilities in money. We focus on this class of environments because most of the settings in the applications of mechanism design are in economies with money. Second, we employ stochastic mechanisms in which lotteries are explicitly used. Therefore, we assume that players are probabilistically sophisticated in the sense of Machina and Schmeidler (1992). Third, we focus on private-value environments. That is, each player's utility depends only upon his/her own payoff type as well as the lottery chosen and his/her monetary payment.[2]

Within the domain described above, we show that any social choice function is implementable in initial rationalizable messages by a simple two-stage *Simultaneous Report* (SR) mechanism. The SR mechanism combines a coordination game with arbitration clauses that are triggered in the event of disagreement. In the first stage, players are arranged in a circle and report on their payoff type and the payoff type of their predecessor in the circle. A players self-report is *consistent* if it matches the report made by his successor and is *inconsistent* otherwise. If all self-reports are consistent, we use these reports to implement the social choice function. If, however, there are any inconsistent reports, all the individuals who make an inconsistent report are immediately fined and are asked to make a second report.

We use one of the second reports to select a lottery from a set of pre-specified lotteries, and use the lottery to determine the outcome. The set of lotteries are constructed so that it is a dominant strategy for a probability sophisticated maximizer to make a truthful report. We can therefore use the second report as a part of a test to determine whether the successor

[2]This is without loss of generality in the complete information case.

was lying in the previous stage. We do this by comparing each second report with the initial report of the successor. We reward the successor with a bonus if the two reports match and punish him or her with a fine if they differ. The bonuses and fines can always be set to induce truthful reporting by the successor in the first stage without requiring money from an outside source. This in turn induces the self-reporting individual to make truthful first-stage reports.

In contrast to the canonical three-stage mechanisms, the SR mechanism that we develop is robust to departures from the common knowledge assumption. Our notion of robustness, which we call "robustness to private-value perturbations," demands that a mechanism implement the desired social choice function both under complete information, and "almost" implement it in nearby environments where there is a small amount of incomplete information about the state of nature. Specifically, in such nearby environments, even conditional on the opponents' signals and types, each player's signal remains almost accurate in identifying his/her own type.[3] We prove that in the SR mechanism, any sequence of initially rationalizable message (e.g., sequential equilibrium) profiles under incomplete information converges to the truth-telling profile as the amount of incomplete information goes to zero. That is, any social choice function is robustly implementable under private-value perturbations.[4]

Having developed a mechanism with promising robustness features, part two of our paper uses experiments to assess the performance of the mechanism both in an environment with complete information and in an environment with noise. The setting we consider is identical to the one studied in Aghion et al. (2018) but with a private-value perturbation. Specifically, a buyer is to receive a buyer-specific good of either high or low quality. Before learning the value of the good, the buyer and seller would like to write a contract where the buyer pays a high price if the good is of high quality and a low price if the good is of low quality. However, the quality of the good is not verifiable by a third-party court and thus a state-dependent contract cannot be directly enforced. Contracting parties must instead rely on some form of implementation mechanism.

The SR mechanism we consider is a simple two-stage mechanism where the buyer and seller report the quality of the good in the first stage. If the reports coincide, we use them to set a report specific price. However, if the reports differ, the buyer is fined and enters into a second stage where he makes a second report that generates a binary lottery over outcomes. By construction, the buyer has a dominant strategy to report his value truthfully in the second-stage. Thus, in theory, we can use it to determine who has lied in the first

[3]As shown in Theorem 1 of Aghion et al. (2012), the mechanism proposed by Moore and Repullo (1988) is not robust to private-value perturbations.

[4]This result contrasts the impossibility result of robust subgame-perfect implementation due to Aghion et al. (2012) (Theorem 3) which is proved by making use of non-private-value perturbations.

stage and induce truthful reports through additional bonuses and fines.

Our experiment explores whether the SR mechanism induces truthful first-stage revelation under complete information and under a private-value perturbation and compares performance against a benchmark canonical subgame perfect implementation (SPI) mechanism that uses a nearly identical set of prices, fines, and rewards. Each session consists of a single mechanism and two information treatments: a no-noise treatment with complete information about the quality of the good, and a noise treatment where buyers receive correct information about the quality of the good 97.5 percent of the time while sellers receive correct information about the quality of the good 87.5 percent of the time. The SR mechanism we develop is predicted to induce truthful reports in both treatments. By contrast, the SPI mechanism has a unique subgame perfect equilibrium under complete information but has multiple initial rationalizable strategy profiles. Further, it is not predicted to be robust to the private-value perturbation and misreports by buyers are predicted to increase when noise is introduced.

We find experimental evidence that is largely consistent with the behavior predicted by theory. In the no-noise treatment of the SR mechanism, buyers and sellers report truthfully in the vast majority of cases: in scenarios where the quality of the good is low, buyers report truthfully in 97.7 percent of cases while sellers report truthfully in 86.2 percent of cases. In scenarios where the quality of the good is high, buyers report truthfully in 94.0 percent of cases and sellers report truthfully in 93.0 percent of cases. When noise is introduced, there is no significant change in the behavior of buyers and sellers.

By contrast, truth-telling rates are lower in the canonical SPI mechanism and the mechanism is not robust to noise. Buyers in the SPI mechanism misreport in one of the two scenarios in 25.5 percent of cases in the no-noise treatment and in 43.3 percent of cases in the noise treatment. These misreport rates are significantly higher than the rates seen in the SR mechanism (7.7 percent and 12.5 percent, respectively) and the difference in misreports between the no-noise and noise treatments of the SPI mechanism are significant. Thus, the SR mechanism strongly outperforms the SPI mechanism in complete information environments and appears robust to private-value perturbations.

Our results relate directly to the burgeoning literature on the robustness of theoretical mechanisms to small perturbations of the economic environment. This literature insists that mechanisms be robust, in the sense that a small perturbation of modeling assumptions does not lead to a large change in equilibria (see, for instance, Chung and Ely (2003) and Aghion et al. (2012)).

The title of our paper reflects our underlying goal of finding ways of operationalizing dynamic implementation for real world applications. Although mitigating hold-up in

environments where information is observable but not verifiable is the most well known application (Maskin and Tirole, 1999), there are many others. For example, we see revelation mechanisms as having important applications in developing countries where well-functioning courts and legal systems do not exist. Suitably designed dynamic implementation mechanisms combined with smart contracting protocols have the potential to expand what is contractible, and hence economic activity and gains from trade.

In designing our SR mechanism, we took into consideration a number of findings from the experimental literature on implementation. Sefton and Yavas (1996) and Katok et al. (2002) study various versions of the Abreu-Matsushima mechanisms and highlight issues that arise in mechanisms that use multiple iterations of backward induction. Discussing the search for good mechanisms for the selection of arbitrators, de Clippel et al. (2014) argue that one desiderata in the search for good mechanisms is that a "mechanism has as few stages as possible so that backward induction is relatively 'simple' to execute." By concentrating on two-stage mechanisms and using a weaker solution concept, our paper directly addresses the issues raised in these papers.

Finding auxiliary mechanisms that have good empirical properties has proven difficult even in simple environments with complete information.[5] Yet our mechanism is robust to a range of reasoning processes. In particular, it remains valid for any solution concept which is stronger than deletion of never sequential best replies followed by two rounds of deletion of strictly dominated strategies. This requirement is satisfied for almost all standard solution concepts in extensive-form games as well as some behavioral solution concepts such as the agent quantal-response equilibrium.

The remainder of the paper proceeds as follows. Section 2 contains our theoretical analysis and proves our main implementation results. Section 3 contains our experimental design and results. Section 4 contains some brief concluding remarks. Appendix A contains our theoretical proofs while Appendix B contains additional empirical analysis, figures, and experimental instructions.

[5]Much of the experimental literature on implementation has centered on either the public goods problems, Solomon's dilemma problems (Ponti et al., 2003; Giannatale and Elbittar, 2010), or hold-up problems. In the context of public goods, Chen and Plott (1996), Chen and Tang (1998), and Healy (2006) study learning dynamics in public good provision mechanisms. Andreoni and Varian (1999), Falkinger et al. (2000), and Chen and Gazzale (2004) study two-stage compensation mechanisms that build on work from Moore and Repullo (1988), while Harstad and Marrese (1981, 1982), Attiyeh et al. (2000), Arifovic and Ledyard (2004), and Bracht et al. (2008) study the voluntary contribution game, Groves–Ledyard, and Falkinger mechanisms respectively. In relation to the hold-up problem, Aghion et al. (2012) draws attention to the issue of information perturbations, while Fehr et al. (2017) draws attention to the issue of reciprocity. Hoppe and Schmitz (2011) study "option contracts" developed in Nöldeke and Schmidt (1995) in a one-sided setting that allows for renegotiation and highlight how attempts at renegotiation are not always successful.

2 The Theory

In this section, we first define the solution concept of *initial rationalizability* and argue that it is substantially more permissive than subgame-perfect equilibrium. We then formally construct a two-stage mechanism—the "Simultaneous Report" (SR) mechanism—in a quasilinear setting and show that it can implement any SCF in initial rationalizable strategy profiles. As a result, implementation by the SR mechanism is not sensitive to belief updating regarding other players' preferences, payoffs, or rationality.

We then show that both the implementation and the truth-telling equilibrium in the SR mechanism are robust to introducing a small amount of incomplete information. More precisely, we show that for any private-value perturbations (see Definition 4), truth-telling remains the unique initial rationalizable strategies for the SR mechanism. Finally, we highlight how the mechanism is robust to a wide variety of reasoning properties and behavioural assumptions.

2.1 The Environment

Consider a finite set of players $\mathcal{I} = \{1, \ldots, I\}$ with $I \geq 2$ located on a circle. Call player $i-1$ (resp. player $i+1$) the predecessor (resp. the successor) of player i. In particular, the successor of player I is player 1. The set of pure social alternatives is denoted by A, and $\Delta(A)$ denotes the set of all lotteries over A with countable supports. We write a for a generic alternative in A and l for a generic lottery in $\Delta(A)$.

Each player i is endowed with a payoff type θ_i which belongs to a finite set Θ_i. Each payoff type θ_i identifies a bounded utility function mapping each lottery-transfer pair (l, τ_i) in $\Delta(A) \times \mathbb{R}$ to a quasilinear utility $u_i(l, \theta_i) + \tau_i$. That is, players' values are *private*. We assume that $u_i(\cdot, \theta_i)$ admits the expected utility representation. Finally, assume that any two distinct types θ_i and θ_i' induce different preference orders over $\Delta(A) \times \mathbb{R}$, i.e., $u_i(a, \theta_i) \neq u_i(a, \theta_i')$ for some $a \in A$.

Let $\Theta \equiv \times_{i \in \mathcal{I}} \Theta_i$ be the set of type profiles or *states*. We consider a *planner* who aims to implement a *social choice function* $f : \Theta \to \Delta(A)$. We start with the complete-information environment, i.e., the true type profile $\theta \in \Theta$ is commonly known to the players but unknown to the planner. We note that the private-value assumption entails no loss of generality when we assume complete information. In Section 2.4, we will turn to study the robustness of our result in an incomplete-information environment where this common knowledge assumption is perturbed.

We will only consider finite two-stage mechanisms throughout the paper. This suffices for our purpose since the SR mechanism which we are about to define has only two stages. In

Stage 1, each player i chooses one message m_i^1 from a finite set M_i^1. Denote by $M^1 \equiv \times_{i\in\mathcal{I}} M_i^1$ the set of Stage 1 message profiles. In Stage 2, after observing the Stage 1 message profile $m^1 \in M^1$, each player i chooses a message m_i^2 from another finite set $M_i^2(m^1)$. Again, write $M^2(m^1) \equiv \times_{i\in\mathcal{I}} M_i^2(m^1)$ for the set of Stage 2 message profiles following m^1. Formally, a two-stage mechanism can be written as a two-stage game form $\Gamma = (\mathcal{H}, (M_i)_{i\in\mathcal{I}}, \mathcal{Z}, g, (\tau_i)_{i\in\mathcal{I}})$ where (1) $M_i = M_i^1 \times (\times_{m^1\in M^1} M_i^2(m^1))$; (2) $\mathcal{H} = \{\varnothing\} \cup M^1$ is the set of non-terminal histories; (3) $\mathcal{Z} = \{(m^1, m^2) : m^1 \in M^1, m^2 \in M^2(m^1)\}$ is the set of terminal histories; (4) g is the outcome function that maps each terminal history to a lottery in $\Delta(A)$; and (5) τ_i is the transfer rule that maps each terminal history to a transfer to player i.

Let $\Gamma(\theta)$ denote the two-stage game associated with Γ at state θ. A *message* (a *pure strategy*) is a pair (m_i^1, m_i^2) such that $m_i^1 \in M_i^1$ and $m_i^2 \in \times_{m^1\in M^1} M_i^2(m^1)$. For each $m \in M$, let $z(m)$ be the unique terminal history induced by m, i.e., $z(m) = (m^1, m^2(m^1))$.

2.2 Solution Concept and Implementation

We now define the solution concept of *initial rationalizability*. Consider the two-stage game $\Gamma(\theta)$ induced by a mechanism Γ at state θ. Conditional on history $h \in \mathcal{H}$, player i's payoff from a message profile m is given by

$$v_i(m, \theta_i|\varnothing) \equiv u_i(g(z(m)), \theta_i) + \tau_i(z(m)). \tag{1}$$

Moreover, for each $\tilde{m}^1 \in M^1$,

$$v_i\left(m, \theta_i|\tilde{m}^1\right) \equiv u_i\left(g(\tilde{m}^1, m^2\left(\tilde{m}^1\right)), \theta_i\right) + \tau_i\left(\tilde{m}^1, m^2\left(\tilde{m}^1\right)\right). \tag{2}$$

In order to analyze each player's reasoning about other players' messages during the entire course of play of the game, we model players' conditional beliefs by means of *a conditional probability system* (CPS). Following Dekel and Siniscalchi (2015), we formulate the notion of CPS as follows.

Definition 1 *Fix a measurable space (Ω, Σ) and a countable collection $\mathcal{B} \subset \Sigma$. A conditional probability system, or CPS, is a map $\mu : \Sigma \times \mathcal{B} \to [0,1]$ such that:*

1. *For each $B \in \mathcal{B}$, $\mu(\cdot|B) \in \Delta(\Omega)$ and $\mu(B|B) = 1$.*
2. *If $A \in \Sigma$ and $B, C \in \mathcal{B}$ with $B \subset C$, then $\mu(A|C) = \mu(A|B) \cdot \mu(B|C)$.*

In the current section, we set $\Omega = M_{-i}$ and $\mathcal{B}$ to be the collection of all nonempty subsets of M_{-i}. That is, a CPS μ_i specifies for each nonempty subset of M_{-i} a probability distribution over M_{-i} such that Bayes' rule (i.e., condition (2) of Definition 1) applies whenever possible. Let $M_{-i}(h) \subset M_{-i}$ be the set of message profiles of player i's opponent that are consistent with history h. Hence, $M_{-i}(\varnothing) = M_{-i}$ and for each $m^1 \in M^1$, we have $M_{-i}(m^1) = \left\{\bar{m}_{-i} \in M_{-i} : \left(m_i^1, \bar{m}_{-i}^1\right) = m^1\right\}$. Conditional on history $h \in \mathcal{H}$, reporting message m_i, and holding CPS μ_i, player i receives the expected payoff:

$$V_i(m_i, \theta_i, \mu_i | h) = \sum_{m_{-i}} v_i\left(m_i, m_{-i}, \theta_i | h\right) \mu_i\left[m_{-i} | M_{-i}(h)\right].$$

A message m_i is a *sequential best response* to CPS μ_i for player i who has type θ_i if, for every history h, we have

$$V_i(m_i, \theta_i, \mu_i | h) \geq V_i(m_i', \theta_i, \mu_i | h), \forall m_i' \in M_i.$$

We now define initial rationalizability:

Definition 2 (Initial Rationalizability) *Let $\Gamma(\theta)$ be a two-stage game. For every player $i \in I$, let $R_{i,0}^{\Gamma(\theta)} = M_i$. Inductively, for every integer $k \geq 1$, let $R_{i,k}^{\Gamma(\theta)}$ be the set of messages $m_i \in M_i$ that are sequential best replies to some CPS μ_i such that $\mu_i\left(R_{-i,k-1}^{\Gamma(\theta)} \middle| M_{-i}\right) = 1$. Finally, the set of **initially rationalizable** messages for player i is $R_i^{\Gamma(\theta)} = \bigcap_{k=1}^{\infty} R_{i,k}^{\Gamma(\theta)}$.*

The solution concept is arguably the weakest among standard notions of equilibrium/rationalizability which impose sequential rationality (see Dekel and Siniscalchi (2015) for more discussion). In particular, only beliefs at the beginning of the game (i.e., $\mu_i(\cdot|M_{-i})$) are restricted. In other words, a player can hold an arbitrary updated belief about his/her opponents once being surprised. For instance, at a history precluded by his/her opponents' rational moves, the player can simply cease believing that his/her opponents are rational. The feature sharply contrasts subgame-perfect equilibrium where the opponents' irrational moves are always regarded as "one shot" and never upset a player's belief in his opponents' rationality in their subsequent moves. In particular, the Moore-Repullo (MR) mechanism implements any SCF in subgame-perfect equilibrium. However, it fails to implement it when a player gives up the belief that his/her opponents will behave rationally upon seeing a history precluded by his/her opponents' rational moves.[6]

We now define our notion of implementability to be used later:

[6]We will revisit this point in Section 3.3 when we discuss our experimental design and hypotheses.

Definition 3 *A social choice function f is **implementable in initial rationalizable messages** if there exists a mechanism Γ such that, for any state $\theta \in \Theta$, we have $g(z(m)) = f(\theta)$ and $\tau_i(z(m)) = 0$ for every message profile $m \in R^{\Gamma(\theta)}$ and for every player i.*

Since we only consider finite mechanisms throughout the paper, we always have $R^{\Gamma(\theta)} \neq \varnothing$ and hence we omit the existential requirement from Definition 3.

2.3 The SR Mechanism

We start by providing a verbal description of the SR mechanism. The SR mechanism is a finite two-stage mechanism which proceeds as follows. In the first stage, each player i announces simultaneously his/her own type as well as the type of player $i-1$. If player i's announcement about his/her own type coincides with his/her successor's announcement of player i's type, player i's announcement is said to be *consistent*. If every player's announcement is consistent, then we implement the social outcome prescribed by the consistent profile. Otherwise, all players who make an inconsistent announcement pay a large penalty and enter the second stage. In the second stage, these players make an announcement of his/her type and with equal probability, one of such players, say player i^*, is picked and a lottery which is pre-assigned to the type announced is implemented. Finally, for each player i who made a second report, player $i+1$ is imposed a large reward if his/her announcement of player i's type coincides with player i's second announcement; otherwise, he/she pays a large penalty.

We now proceed to the formal details. It will become clear from the construction of the SR mechanism that we do not need the full force of the complete information assumption. Indeed, it suffices to assume that each player's type is known by at least two players and ask each player to report the type profile of all players which he/she knows in Stage 1.

2.3.1 Message Space

First, we specify the message space.

Stage 1: Each player i is asked to report his/her own type and his predecessor's type, namely,

$$M_i^1 = \Theta_i \times \Theta_{i-1}.$$

A generic element in M_i^1 is denoted as $m_i^1 = (\hat{\theta}_i^i, \hat{\theta}_{i-1}^i)$.

Stage 2: Let $\mathcal{I}^*(m^1) \equiv \{i \in \mathcal{I} : \ \hat{\theta}_i^i \neq \hat{\theta}_i^{i+1}\}$ be the set of players who make an inconsistent announcement at history m^1. For $m^1 = (\hat{\theta}_i^i, \hat{\theta}_{i-1}^i)_{i \in \mathcal{I}}$, each player $i \in \mathcal{I}^*(m^1)$ is asked to

report his/her own type, that is,

$$M_i^2\left(m^1\right) = \begin{cases} \Theta_i, & \text{if } \hat{\theta}_i^i \neq \hat{\theta}_i^{i+1}; \\ \varnothing, & \text{if } \hat{\theta}_i^i = \hat{\theta}_i^{i+1}. \end{cases}$$

A generic element in M_i^2 is denoted as $m_i^2 = \tilde{\theta}_i$.

2.3.2 Outcome Function

Next we turn to the specification of the outcome function. Recall our assumption that two distinct types θ_i and θ_i' induce different preference orders over $\Delta(A) \times \mathbb{R}$. With this assumption, we can construct the dictator lotteries by invoking the following result due to Abreu and Matsushima (1992).

Lemma 1 *For each player $i \in \mathcal{I}$, there exists a function $l_i : \Theta_i \to \Delta(A) \times \mathbb{R}$ such that*

$$u_i\left(l_i\left(\theta_i\right), \theta_i\right) > u_i\left(l_i\left(\theta_i'\right), \theta_i\right), \text{ for any } \theta_i, \theta_i' \in \Theta_i \text{ with } \theta_i \neq \theta_i'. \tag{3}$$

Second, we specify the outcome function. If all players' announcements in the first stage are consistent, then the planner implements $f(\hat{\theta})$ where $\hat{\theta} \equiv (\hat{\theta}_i^i)_{i \in \mathcal{I}}$. Otherwise, the planner randomly selects a player i^* from the set $\{i \in \mathcal{I}^*(m^1)\}$ with equal probability. The planner then implements $l_{i^*}(\tilde{\theta}_{i^*})$.

2.3.3 Transfers

We now define the transfer rule. Transfers are incurred only when some player's first announcement is inconsistent. In particular, we impose the following rules:

- Each player $i \in \mathcal{I}^*(m^1)$ pays a penalty T;
- Player $i+1$ gets the incentive transfer:

$$T_{i+1}(\hat{\theta}_i^{i+1}, \tilde{\theta}_i) = \begin{cases} T, & \text{if } \hat{\theta}_i^{i+1} = \tilde{\theta}_i; \\ -T, & \text{if } \hat{\theta}_i^{i+1} \neq \tilde{\theta}_i. \end{cases}$$

- We choose $T > D$ where[7]

$$D = \sup_{i,a,a',\theta_i} \left|u_i(a, \theta_i) - u_i(a', \theta_i)\right|.$$

[7]Recall that we assume that Θ_i is finite and hence D is bounded.

In words, each player $i \in \mathcal{I}^*(m^1)$ is penalized by T for making an inconsistent announcement of his/her own type. Moreover, for each $i \in \mathcal{I}^*(m^1)$, player $i+1$ is rewarded by T, if his/her Stage 1 announcement of player i's type coincides with player i's Stage 2 announcement; otherwise, player $i+1$ is penalized by T.

We now prove the following permissive result for implementation in initial rationalizable messages via the SR mechanism.

Theorem 1 *Any social choice function is implementable in initial rationalizable messages by the SR mechanism.*

Proof. See Appendix A.1. ■

We further elaborate on the feature of the SR mechanism. The second stage in the SR mechanism is constructed by first choosing a set of lotteries, one for each type of each player, according to which it is the unique optimal choice for each player to truthfully report his/her own type. Since the outcome in the second stage is based solely on the second reports of the party and the dictator lotteries were constructed prior to play, information or belief about the other player's type plays no role in the choice made at the second stage. This feature ensures that the mechanism is insensitive to the way in which players update their beliefs about other players. As a result, the SR mechanism is less susceptible to relaxations of common knowledge assumptions on rationality, information, and preferences.

The truthful report in the second stage plays the same role as a behavioral anchor in the level$-k$ model (see, e.g., Crawford and Iriberri (2007) and de Clippel et al. (2018)). At the first stage, once everyone knows that telling the truth is the unique optimal choice for any active player in the second stage, truth-telling must also be the uniquely optimal action for the successor(s) of each active player and hence the optimal action of each player.

Recall that the active player's re-evaluation of his/her opponent's rationality based on his opponent's deviation is irrelevant because the opponent has no opportunity to make a move. These features carry on even when we perturb the original information structure slightly, as long as each (active) player's own signal is more informative over his/her own payoff types than the other players' signals/payoff types. This feature also explains how we obtain information robustness of the SR mechanism in the next section.

2.4 Robustness to Information Perturbations

We now formulate the second robustness property of the SR mechanism. Suppose that the players do not observe the state directly but are informed of the state via some noisy signal. Following Aghion et al. (2012), we set the space of signals as $S_i = \Theta$. A signal profile is an

element $s = (s_1, ..., s_n) \in S = \times_{i \in I} S_i$. Let s_i^θ denote the signal in S_i which corresponds to θ. Also denote by s^θ the signal profile such that $s_i = s_i^\theta$ for all $i \in \mathcal{I}$.

Suppose that the state and signals are jointly distributed according to a prior distribution $\pi \in \Delta(\Theta \times S)$. A prior π^{CI} is said to be a *complete information* prior if $\pi^{\text{CI}}(\theta, s) = 0$ whenever $s \neq s^\theta$. We assume that for each $i \in \mathcal{I}$ and $\theta \in \Theta$, the marginal distribution of π on the signal space places strictly positive weight on every signal profile (i.e., $\text{marg}_S \pi(s) > 0$ for every $s \in S$) so that Bayes's rule is always well defined. For each π, we write $\pi(\cdot|s_i)$ (resp. $\pi(\cdot|s, \theta_{-i})$) for the probability measure over $\Theta \times S_{-i}$ (resp. Θ_i) conditional on s_i (resp. s).

Let $\mathcal{P}$ denote the set of priors over $\Theta \times S$. Let $\mathcal{P}$ be endowed with the following metric $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$:

$$d(\pi, \pi') = \max_{(\theta,s) \in \Theta \times S} |\pi(\theta, s) - \pi'(\theta, s)|, \forall \pi, \pi' \in \mathcal{P}.$$

We consider the following class of information perturbations.

Definition 4 *A sequence of priors $\{\pi^k\}$ is a **private-value perturbation** to π^{CI} (which we denote as $\pi^k \to \pi^{CI}$) if $d\left(\pi^k, \pi^{CI}\right) \to 0$ and for all $i \in \mathcal{I}$, $\theta \in \Theta$, and $s_{-i} \in S_{-i}$, we have*

$$marg_{\Theta_i} \pi^k \left[\theta_i | s_i^\theta, s_{-i}, \theta_{-i}\right] \to 1 \text{ as } k \to \infty.$$

Private-value perturbation requires that conditional on the opponent's signal and payoff types, player i's signal be asymptotically accurate in identifying his/her own type. Indeed, Theorems 1 and 2 of Aghion et al. (2012) both invoke private-value perturbations in proving the nonrobustness of the MR mechanism. To wit, when players' values are private, it is natural to assume that a player's own signal is more informative over their own payoff types than others' signals/payoff types.

One special case of private-value perturbations is that player i knows precisely his/her own type θ_i (e.g., Bergemann and Morris (2005)) even with information perturbations and only entertains a small amount of uncertainty about his/her opponents' types. This amounts to assuming that $\pi^k \left[\theta_i | s_i^\theta\right] = 1$ for every k and also makes the sequence of priors $\{\pi^k\}$ a private-value perturbation, as long as $d\left(\pi^k, \pi^{\text{CI}}\right) \to 0$.

We now adapt our definitions of mechanisms and solution concepts to the incomplete-information setup. We denote by $\Gamma(\pi)$ the incomplete information game induced by a two-stage mechanism Γ under prior π. Here, a CPS μ_i specifies for each nonempty subset E of $\Theta \times S_{-i} \times M_{-i}$ a distribution $\mu_i[\cdot|E]$ over $\Theta \times S_{-i} \times M_{-i}$ with the property that Bayes'

rule applies whenever possible.[8]

Now conditional on history $h \in \mathcal{H}$, using message m_i and holding CPS μ_i, player i's expected payoff is computed as follows:

$$V_i(m_i, s_i, \mu_i|h) = \sum_{\theta, s_{-i}, m_{-i}} v_i\left(m_i, m_{-i}, \theta_i|h\right) \mu_i\left[(\theta, s_{-i}, m_{-i})|\Theta \times S_{-i} \times M_{-i}(h)\right].$$

Similar to the case of complete information, we say that a message m_i is a *sequential best response* to CPS μ_i for player i with signal s_i if, for every $h \in \mathcal{H}$, we have

$$V_i(m_i, s_i, \mu_i|h) \geq V_i(m_i', s_i, \mu_i|h), \forall m_i' \in M_i.$$

We say that a CPS μ_i is *consistent with* s_i if $\text{marg}_{\Theta \times S_{-i}} \mu_i\left(\cdot|\Theta \times S_{-i} \times M_{-i}(h)\right) = \pi\left(\cdot|s_i\right)$ for the following two cases: (1) initial history $h = \varnothing$; (2) every history $h \neq \varnothing$ such that

$$\mu_i\left(\Theta \times S_{-i} \times M_{-i}(h)|\Theta \times S_{-i} \times M_{-i}\right) = 0.$$

First, the belief over $\Theta \times S_{-i}$ induced by a CPS at initial history is required to be consistent with player i's prior belief $\pi\left(\cdot|s_i\right)$. Second, if history h is assigned probability zero at the beginning of the game according to the CPS μ_i, then conditional on history h, μ_i must maintain the initial belief about states and signal profiles of the opponents given by $\pi\left(\cdot|s_i\right)$.

The following two definitions are the counterparts of Definitions 2 and 3 in the almost complete information environments. We first define the solution concept of initial rationalizability under incomplete information.

Definition 5 (Initial Rationalizability under Incomplete Information) *Let $\Gamma(\pi)$ be a two-stage game. The set of initial rationalizable messages of player i with signal s_i is defined as $R_i\left(s_i|\Gamma(\pi)\right) = \bigcap_{k=1}^{\infty} R_{i,k}\left(s_i|\Gamma(\pi)\right)$, where $R_{i,0}\left(s_i|\Gamma(\pi)\right) = M_i$ and, inductively, for every integer $k \geq 1$,*

$$R_{i,k}\left(s_i|\Gamma(\pi)\right) = \left\{ m_i \in M_i \left| \begin{array}{l} \text{there exists CPS } \mu_i \text{ over } \Theta \times S_{-i} \times M_{-i} \text{ such that} \\ (1)\ \mu_i\left[(\theta, s_{-i}, m_{-i})\left|\Theta \times S_{-i} \times M_{-i}\right.\right] > 0 \\ \Rightarrow\ m_{-i} \in R_{-i,k-1}\left(s_{-i}|\Gamma(\pi)\right); \\ (2)\ m_i \text{ is a sequential best response to } \mu_i; \text{ and} \\ (3)\ \mu_i \text{ is consistent with } s_i. \end{array} \right. \right\}.$$

As we did under complete information, we only consider finite mechanisms through-

[8] Using the formal notation introduced for CPSs in Definition 1, we set $\Omega = \Theta \times S_{-i} \times M_{-i}$ and let $\mathcal{B}$ be the collection of all nonempty subsets of Ω.

out the paper. So, we have $R(s^\theta|\Gamma(\pi)) \neq \varnothing$. The following is the definition of robust implementation we adopt.

Definition 6 *A social choice function f is **robustly implementable in initial rationalizable strategies** if there exists a mechanism $\Gamma = (M, g)$ such that for any state $\theta \in \Theta$, any signal profile $s^\theta \in S$, any private-value perturbation $\{\pi^k\}$ to π^{CI}, and any sequence of message profiles $\{m^k\}_{k=1}^\infty$ with $m^k \in R(s^\theta|\Gamma(\{\pi^k\}))$ for each k, we have $g(z(m^k)) \to f(\theta)$ as $k \to \infty$ and $\tau_i\left(z\left(m^k\right)\right) = 0$ for each k.*

Within the class of private-value perturbations, our robustness notion is formulated with the permissive solution concept of initial rationalizability. Specifically, we allow each player's CPS to have any degree of correlations among player's strategies, other players' signals, and the payoff type profiles. Hence, our robust implementation result holds even with the stronger yet more standard solution concept of sequential equilibrium (defined in the online appendix of Aghion et al. (2012)) is used.[9] We are now ready to state our robust implementation result regarding the SR mechanism that we previously defined in Section 2.3.

Theorem 2 *Any social choice function is robustly implementable in initial rationalizable strategies by the SR mechanism.*

Proof. See Appendix A.2. ■

Note that Theorem 2 is consistent with the impossibility result of robust subgame-perfect implementation proved in Theorem 3 of Aghion et al. (2012). Indeed, Theorem 3 of Aghion et al. (2012) invoke perturbations which are not private-value. Our Theorem 2 shows that it is actually necessary for Aghion et al. (2012) to invoke these non-private-value perturbations, as we have obtained a possibility result with respect to private-value perturbations.

2.5 Additional Robustness Results and Discussion

2.5.1 Retaliatory Preferences

Fehr et al. (2017) consider an implementation problem where players care about not only material payoffs but also "psychological" payoffs obtained from retaliating against perceived unkind acts. Studying a Subgame Perfect Implementation (SPI) mechanism, they show that

[9] In fact, if we adopt a solution concept where it is commonly believed that each player's strategy only depends on his own signal but not on the payoff type profile, we can further weaken Definition 4 in requiring only $\text{marg}_{\Theta_i}\pi^k\left[\theta_i|s_i^\theta, s_{-i}\right] \to 1$.

when retaliatory types are heterogenous and private, retaliation behaviors may lead to an implementation failure where fines occur with positive probability in a mixed strategy equilibrium. This equilibrium is based on heterogeneity in buyer reciprocity and belief updating: highly reciprocal buyers who enter into the arbitration stage may choose to retaliate against his/her opponent's challenge by not validating his/her challenge even at the cost of a material loss. Fear of such retaliation can cause sellers to avoid the arbitration system if they believe that a buyer will retaliate. However, this reluctance will lead even non-reciprocal buyers to lie and pretend to be reciprocal leading to an equilibrium where buyers lie with positive probability and sellers challenge with positive probability.

Due to belief updating, implementation failure can occur in the SPI mechanism even when the number of reciprocal buyers is (vanishingly) small. By contrast, in the SR mechanism, the seller takes no actions after observing the buyer's report, and so there is no scope for belief updating about the retaliatory type of the buyer. This implies that if just over half the buyers are non-reciprocal and the most reciprocal buyer is still 'reasonable' in the sense that he or she is unwilling to pay more than a dollar to destroy a dollar of another player's money when maximally aggrieved, the majority of players will be truthful in the second stage and our implementation results hold.

In other psychological environments, the SR mechanism will at least partially implement a social choice function if (i) the SR mechanism (fully) implements the social choice function in a psychological environment where all agents are non-reciprocal and (ii) no retaliatory motives exist after the realization of the state but before the players first reports. Thus, for a large class of psychological environments, reciprocity will at worse cause the SR mechanism to act as a coordination game with a salient truth-telling equilibrium.[10]

2.5.2 Agent Quantal Response Equilibrium

The SR mechanism is robust to alternative reasoning processes and behavioral assumptions. In particular, the SR mechanism implements the SCF after (1) deleting strategies that violate sequential rationality; and (2) deleting strictly dominated strategies for two rounds. As discussed in greater detail in the proof of Theorem 1, we choose the dictator lotteries $l_{i^*}(\cdot)$, the incentive transfers T, and the arbitration fee T so that (1') sequential rationality

[10]Note also that the SPI mechanism gives a "challenged" player a behavioral motive to retaliate against the "challenger" because the challenger must actively force the challenged into arbitration and this causes him or her to be fined. In contrast, in the SR mechanism the successor in the circle is just a truth-teller expecting the predecessor to also be truthful. Thus, it is unclear whether the predecessor should wish to retaliate against a truthful report when such reports are not intended to cause harm. If it is assumed that players report truthfully in the second stage if their successor reported truthfully in the first stage, then the mechanism is fully robust to reciprocity if all players are reasonable.

ensures that player i^* will truthfully announce his/her type in the second stage; (2') the first-round deletion of strictly dominated strategies ensures that each player i wants to match his/her first stage report on the type of player $i-1$ with the second stage report chosen by player $i-1$; (2") the second-round deletion of strictly dominated strategies ensures that each player i wants to match his/her first stage report on his/her own type with the first stage report chosen by player $i+1$. Consequently, our result remains valid for any solution concept which is stronger than deletion of never sequential best replies followed by two rounds of deletion of strictly dominated strategies. This is a remarkably weak requirement. For instance, it is satisfied almost all standard solution concepts in extensive-form games as well as some behavioral solution concepts such as the *agent quantal response equilibrium* proposed by McKelvey and Palfrey (1998), provided that the noise parameter is sufficiently small.

The SR mechanism relies on strict inequalities at all stages of the game and is thus remarkably robust to level-k reasoning. When play is anchored by truth-telling level-0 types, all types are truthful. If we allow for any level-0 play, then the SR mechanism is level-3 implementable.[11] These results are consistent with de Clippel et al. (2018), who show that a slight weakening of standard strict incentive constraints are necessary for level-k implementation.

3 The Experiment

Part one of this paper suggests that the SR mechanism is robust to private value perturbations and may be robust to a variety of alternative reasoning processes and behavioral assumptions. In this section, we study the empirical properties of the mechanism using laboratory experiments in a two-person environment where buyers and sellers seek to implement a state-dependent contract with observable but non-verifiable information. This environment was selected because it is the underlying contracting problem needed to resolve two-sided holdup and it is an environment where other mechanisms based on subgame perfection have performed poorly.

To concentrate directly on the robustness properties of the SR mechanism, we consider behavior in both a complete information environment and an environment with a private-value perturbation. We further benchmark behavior against a canonical SPI mechanism that is not predicted to be robust to private-value perturbations.

[11] To see this, note that in the SR mechanism, level-1 players will always report truthfully in the second stage of the mechanism. Thus, a level-2 player will make a truthful report regarding their predecessors type and will also be truthful in the second stage. It follows that a level-3 player will best respond to a level-2 type by making truthful reports at all stages.

3.1 Environment and Mechanisms

Our experimental environment is based on Aghion et al. (2018), which borrows its setup from Hart and Moore (2003). In each of 20 periods, a buyer and a seller are matched and the seller is randomly assigned one of two sealed containers with equal probability. One container is worth 70 Experimental Currency Units (ECU) to the buyer and the other container is worth 20 ECU.[12]

Each container has two compartments: a buyer's compartment and a seller's compartment. Each compartment is filled with red and blue balls whose composition changes by information treatment:

1. **No-Noise Treatment:** In the no-noise treatment, both the buyer's compartment and the seller's compartment of the container worth 70 ECU is filled with 40 red balls and 0 blue balls. Likewise, both the buyer's compartment and the seller's compartment of the container worth 20 ECU is filled with 40 blue balls and 0 red balls.

2. **Noise Treatment:** In the noise treatment, the buyer's compartment of the container worth 70 ECU is filled with 39 red balls and 1 blue ball. The seller's compartment of the container worth 70 ECU is filled with 35 red balls and 5 blue balls. Similarly, the buyer's compartment of the container worth 20 ECU is filled with 39 blue balls and 1 red ball. The seller's compartment of the container worth 20 ECU is filled with 35 blue balls and 5 red balls.

The two parties in a group do not initially know which container has been allocated to the seller. However, over the course of a period, the buyer privately observes a ball drawn from the buyer's compartment of the container and the seller privately observes a ball drawn from the seller's compartment. These "signals" provide complete information about the container being traded and the signal observed by their matched partner in the no-noise treatment. In the noise treatment, the buyer's signal is more accurate in identifying the value of the container than the seller's signal: the buyer will receive the correct signal 97.5 percent of case while the seller will receive the correct signal 87.5 percent of the time. The signals will coincide 85.3 percent of the time. Throughout the rest of the paper we refer to the red signal as the **high signal** and the blue signal as the **low signal**.

In each period, the buyer and seller have the task of trading the container using either a Simultaneous Report (**SR**) mechanism or a Subgame-Perfect Implementation mechanism (**SPI**). Both mechanisms use near identical price schedules, bonuses, and fines to implement a

[12]The exchange rate of ECU to Australian dollars was a rate of 2 ECU = 1 AUD. As discussed below, we randomly paid two periods: one from periods 1-10 and one from periods 11-20.

state-contingent trading scheme that (under complete information) trades containers worth 20 ECU at a price of 10 ECU and containers worth 70 ECU at a price of 35 ECU. The mechanisms are implemented as follows:

The Simultaneous-Report mechanism: In treatments using the Simultaneous-Report mechanism, each period is composed of four stages: a report stage, a signal stage, a verification stage, and an arbitration stage. In the report stage, both the buyer and the seller are asked to privately report the value of the container under two scenarios: the scenario where he or she observes the high signal and the scenario where he or she observes the low signal. The buyer and seller may report a **high value** of 70 or a **low value** of 20 in each scenario.

In the signal stage, the buyer privately draws a ball from the buyer's compartment of the container. After observing the signal, the buyer makes a formal report which corresponds to his or her decision in the report stage for that signal. The seller also privately draws a ball from the seller's compartment and also makes a formal report that corresponds with his or her decision in the report stage.

Following the signal stage, each party is made aware of the formal report made by their matched party, but not their matched partner's signal nor their strategy. Thus, the strategy method that we employ generates a compete set of reports in each period but does not affect the information observed in the mechanism itself. Obtaining a complete panel of first-stage reports improves our ability to control for heterogeneity across individuals. It also reduces variation across periods that is driven by the random assignment of containers and signals to different buyers and sellers.[13]

In the verification stage, the formal report of the buyer and seller are compared with one another. If the formal reports coincide, the two parties trade at a price equal to one half of the reported value (i.e., 35 after a high value report and 10 after a low value report). If the reports do not coincide, the buyer pays an arbitration fee of 40 ECU and enters into the arbitration stage.

In the arbitration stage, the buyer is asked to make a second report. As shown in Table 1, the buyer may report a value of 0, 20, or 70.[14] We use the second report along with a fair six-sided die to determine whether trade occurs and the price.[15]. If the second report of the

[13]Note that we do not employ the strategy method for internal nodes of the experiment (e.g., the arbitration stage of the SR mechanism) because the sequential nature of the mechanism is important.

[14]Theoretically, the second stage only requires reports of 20 and 70 for the mechanism to work. We included the additional possibility of reporting zero so that we could distinguish between misreports in the second stage that were designed to minimize the probability of trade and those that were designed to intentionally match the misreport of one's trading partner.

[15]Note that in the current treatments, the dice is mapped into a simple binary lottery. We use the dice description as it is easier to extend to other environments with more outcome states.

buyer matches the first-stage report of the seller, the seller is rewarded a bonus of 40 ECU in addition to her earnings from trade. In other cases, the seller also pays a fine of 40 ECU.

At the end of the period, the true value of the container is revealed. If trade occurs in the period, the profits of the buyer and seller are given by:

$$\begin{aligned} \pi_B &= Value - Price - BuyerArbitrationFee \\ \pi_S &= Price + SellerBonus - SellerArbitrationFee. \end{aligned}$$

If trade does not occur, the container is destroyed. However, both parties must still pay their arbitration fees.

Table 1: Trade Prices if Buyer Enters Arbitration

Buyer's Secondary Report	Outcome if Dice Roll is a 1-3	Outcome if Dice Roll is a 4-6
0	No Trade	No Trade
20	Trade at 10	No Trade
70	Trade at 10	Trade at 35

The Subgame-Perfect Implementation mechanism: In order to benchmark the performance of the mechanism, we also conducted sessions using a three-stage Subgame-Perfect Implementation mechanism based on Moore & Repullo (1988). In treatments using the **SPI** mechanism, we elicit a report for the buyer in both the scenario where he receives the high signal and the scenario where he receives the low signal using the strategy method discussed above. The buyer may make a high or low report in each scenario. We then draw a signal for the buyer from the buyer's compartment and make a formal report to the seller that corresponds with the buyer's decision.

The seller in the mechanism next draws a signal from the seller's compartment and is informed of the formal report of the buyer. The seller next has the option to "call" or "not call" the arbitrator.[16] If the arbitrator is not called, the parties trade at a price equal to one half of the reported value. If the seller calls the arbitrator, the buyer is fined 40 ECU and enters into the arbitration stage.

In the arbitration stage, the buyer is given a counter-offer equal to 35 if he reported a value of 20 and 85 if he reported a value of 70. The buyer may accept or reject the counter offer. If the counter offer is accepted, trade occurs and the seller receives a bonus of 40 ECU. If the counter offer is rejected, no trade occurs and the seller is fined 40 ECU.

[16]Since the seller's arbitration decision is in the second stage of the mechanism, we do not use the strategy method when eliciting this actions. Note, however, that our main interest is on the buyer side where misreports are predicted to emerge when noise is introduced.

3.2 Protocol

Our experiment utilized a 2x2 design in which we generated within-subject variation in noise and between-subject variation in mechanism. Within a session, subjects played ten periods of the no-noise treatment followed by ten periods of the noise treatment using a single mechanism. All sessions consisted of exactly 20 participants who were evenly divided between buyers and sellers at the beginning of the experiments. Buyers and sellers were matched with each other at most once in each of the two information treatments.

All of the experiments were run in the Experimental Economics Laboratory at the University of Melbourne in March of 2019. The experiments were conducted using the programming language z-Tree (Fischbacher, 2007). A total of 12 sessions were run: 6 sessions using the SR mechanism and 6 sessions using the SPI mechanism. All of the 240 participants were undergraduate students at the university and were invited from a pool of more than 6000 volunteers using ORSEE (Greiner, 2015).

Upon arrival at the laboratory, participants were randomly assigned buyer and seller roles and asked to read the instructions for their assigned mechanism in the no-noise treatment. Consistent with previous implementation experiments, the instructions described the game in detail, walked through a series of examples that calculated the payoffs of both parties along the equilibrium path and along the off-equilibrium paths, and culminated in a quiz. In the quiz, the subjects were required to calculate the payoffs that each buyer and seller would receive for both potential values of the container and both on and off-equilibrium actions.[17]

After completing the instructions, we read additional oral instructions that reiterated the matching structure and the payment rules discussed below. In the oral instructions we announced that the second treatment would be identical to the first except that some of the blue balls would be moved to the container worth 70 ECU and some of the red balls would be moved to the container worth 20 ECU. Thus, subjects were informed about all aspects of the no-noise and noise treatments at the start of the experiment with the exception of the exact noise distribution.

After playing 10 periods of the no-noise treatment, we handed out a second set of instructions that discussed how the balls had been moved between the containers and compartments. Subjects were informed explicitly about the probability of all possible combinations of signals and containers in this set of instructions to reduce the computational burden.

We randomly selected one period from the no-noise treatment and one period from the

[17]While the instructions for both mechanisms were complete in describing the outcome of each set of actions in the mechanism, we did not explicitly state that the mechanism is designed to induce truthful reports for both buyers and sellers because the SPI mechanism does not have this property in the noise treatment. This precluded some forms of training that have been used in previous research such as playing against a Nash best-responding computer.

noise treatment for payment at an exchange rate of 2 ECU to $1 AUD. To avoid bankruptcies, participants received a show-up fee of $35 AUD. The average payment at the end of the experiment was $51.84 AUD. At the time of the 2019 experiments, $1 AUD ≈ $0.71 USD.

The experimental design, instructions, and analysis plan were pre-registered at open science (osf.io/p6ukx). Prior to pre-registration, we ran one pilot session of the SR mechanism and one session of the SPI mechanism to obtain a better estimate of the distribution of buyer misreports in order to perform power calculations. We do not include these pilots in the results as they were done before finalizing the analysis plan.

3.3 Hypotheses

The SR mechanism used in our experiment is designed to implement truthful reports for both buyers and sellers. Given the incentives induced by the mechanism we would predict the following pattern of behavior in the no-noise treatment under the solution concepts of subgame perfection and initial rationalizability:

Hypothesis 1 *In the no-noise treatment, the path of play under the Simultaneous-Report mechanism involves both the buyer and seller making truthful reports. If the buyer enters into arbitration, the buyer makes a truthful secondary report.*

In the noise treatment, if the buyer and seller make truthful reports, there is 14.4 percent chance that the reports will not coincide. In these cases, the buyer's signal is correct 84.8 percent of the time. Thus the expected value of the container is 62.4 if the buyer receives the high signal and 27.6 if the buyer receives the low signal. By reporting a high value of 70, the buyer has the potential to trade with the seller at a price of 35. Thus, for a very large class of risk preferences, the buyer should make a high second report of 70 after receiving the high signal and a low second report of 20 after receiving the low signal. We would thus predict the following behavior under both subgame perfection and initial rationalizability:

Hypothesis 2 *In the noise treatment, the path of play under the Simultaneous-Report mechanism involves both the buyer and seller making reports that match their signal. If the buyer enters into arbitration, the buyer makes a secondary report that matches his signal.*

Truth-telling is also the path of play in the unique subgame perfect equilibrium of the SPI mechanism in the no-noise treatment and is one of potentially many initial rationalizable strategy profiles. However, when noise is introduced, the original truth-telling equilibrium of the SPI mechanism is no longer supported by both solution concepts. Instead two types of equilibria emerge: (i) a unique mixed strategy equilibrium in which the buyer reports a low

value with the high signal with a positive probability and (ii) a continuum of pure strategy equilibria in which the buyer with the low signal reports a high value.[18] Using subgame perfection as the basis for the null hypothesis, we predict the following:

Hypothesis 3 *In the no-noise treatment, the proportion of buyer who report truthfully in the SPI mechanism will be equal to the proportion of buyer who report truthfully in the SR mechanism. In the noise treatments, the proportion of buyers who report truthfully in the SPI mechanism will be smaller than the proportion of buyers who report truthfully in the SR mechanism.*

As discussed earlier, truth-telling corresponds to the unique initial rationalizable strategy profile of the SR mechanism while it corresponds to one of potentially many initial rationalizable strategy profiles in the SPI mechanism.[19] The SR mechanism is also more robust to noise in the best response functions and less sensitive to retaliatory preferences. Thus, under initial rationalizability and for a number of alternative theoretical assumptions, we would predict that the SR treatment will have a greater level of truth-telling than the SPI mechanism in the no-noise treatment. In testing how noise influences the two mechanisms, we provide both direct proportion tests of buyer misreports using data only from the noise treatments and a difference-in-difference estimator that uses data from both information treatments to control for potential differences in the no-noise treatment. These estimates provide both an absolute difference between the two noise treatments and the relative change in buyer misreport rates that occur when noise is introduced.

3.4 Results

We will refer to a **truthful report** as a case where a buyer or seller makes a high report with a high signal or a low report after a low signal. We will say the SR mechanism **induces truth-telling first-stage strategies** if both reports by the buyer are truthful and both

[18]For the purposes of conducting power calculations, we concentrated on the mixed strategy equilibrium where buyers mix between misreporting and telling the truth with the high signal and sellers challenge a report that does not match their signal with a positive probability. This equilibrium is the only one that is not based on some form of out-of-equilibrium belief. In the mixed strategy equilibrium, buyers misreport on average 13 percent of the time. The sample size was thus selected to detect an effect size of 0.13 at a significance level of 0.05 and a power of 0.80.

[19]To see that the MR mechanism allows for multiple initial rationalizable strategy profiles in the complete information environment, consider the situation where the seller receives the high-valued container. Suppose that—counter to his signal—the buyer makes a low report and the seller is choosing whether to call the arbitrator. Under initial rationlizability, the seller may abandon the belief that the buyer is rational and could instead entertain a belief that the buyer will reject the counter offer. As such, the seller may not call the arbitrator. Hence, a situation where the buyer makes a low report and the seller does not challenge the report is consistent with initial rationalizability in the high-value case.

reports by the seller are truthful. The buyer **misreports** in a period if he reports a low signal in the high-signal scenario or the high signal in the low-signal scenario.

The seller's strategies are not fully observed in the SPI mechanism and thus we cannot directly compare the proportion of truth-telling strategies in the two mechanisms. Instead, we will compare observed actions in periods where both the buyer's signal and seller's signal match the true state. We will say that the **formal report is truthful** if a buyer's or seller's formal report matches his or her signal and the **formal report is misreported** otherwise. A group in the SR mechanism **displays truth-telling behavior** if the formal report of the buyer and seller are truthful. A group in the SPI mechanism **displays truth-telling behavior** if the buyer's formal report is truthful and the seller does not challenge.

3.4.1 The Simultaneous Report Mechanism Under Complete Information

Under hypothesis 1, our experimental design predicts that both buyers and sellers will make truthful reports in the first stage in both the high-signal scenario and the low signal scenario. The data from the no-noise treatment is largely consistent with this hypothesis.

Result 1 *In the no-noise treatment of the Simultaneous Report mechanism, buyers report truthfully in* 97.7 *percent of cases in the low-signal scenario and in* 94.0 *percent of cases in the high-signal scenario. Sellers report truthfully in* 86.2 *percent of cases in the low-signal scenario and in* 93.0 *percent of cases in the high-signal scenario. Buyers who enter arbitration report the true value in* 77.3 *percent of cases.*

Figure 1 shows the pattern of play in the no-noise treatment in sessions using the Simultaneous Report mechanism. Panel (a) shows the proportion of reports that were truthful in the report stage for both the buyers and sellers while panel (b) shows how buyers responded in the arbitration stage. The left-hand side of panel (c) shows a histogram of the aggregate number of buyer misreports over the 10 periods of the no-noise treatment. The right-hand side of panel (c) shows the aggregate number of periods where each seller made a misreport in the low-signal scenario.[20]

As seen in panel (a), misreports were rare for buyers and uncommon for sellers. Buyers told the truth 97.7 percent of the time in the low-signal scenario and 94.0 percent of the time in the high-signal scenario. Sellers told the truth 86.2 percent of the time in the low-signal scenario and 93.0 percent of the time in the high-signal scenario.[21]

[20]Seller lies in the low-signal scenario were found to be prevalent in a SPI mechanism tested in Aghion et al. (2018). Thus we included information on seller reports in the pre-analysis plan.

[21]As shown in the appendix, there are no apparent time trends in the behavior of buyers in the SR mechanism. However, sellers appear to be learning to tell the truth in the low-signal scenario through

Aggregating the behavior of the buyer and seller, the simultaneous report mechanism induces truth-telling first-stage strategies in 74.7 percent of groups. The buyer's formal report was truthful in 96.0 percent of cases while the seller's formal report was truthful in 90.3 percent of cases. This led to 86.8 percent of groups displaying truth-telling behavior. As such, we observe only 58 cases where the buyer enters into the second stage after following his signal and 21 cases where the buyer enters the second stage after lying. As seen in panel (b), buyers report truthfully 84.5 percent of the time if arbitration is due to a seller misreport and in 57.1 percent of the time if arbitration is due to the buyers own misreport. Based on the distribution of buyer secondary reports, sellers who tell the truth earn an average of 21.6 ECU more than sellers who misreport in the low-signal scenario and 17.2 ECU more than sellers who misreport in the high-signal scenario. Buyers earn 13.6 ECU more for telling the truth in the low-signal scenario and 33.3 ECU more in the high-signal scenario. Thus the mechanism generates strong incentives for truthful reporting for both parties.

Finally, as seen in panel (c), the majority of buyers and sellers are truthful in all 10 periods and there are no individuals who misreport in every period. This implies that (i) the mechanism induces truth-telling strategies for the majority of individuals from the very start of the experiment and (ii) there is little heterogeneity in strategy. On the buyer side, 45 out of 60 buyers are truthful in every period and an additional 9 buyers make one or two misreports. On the seller side, 32 sellers never make a misreport with a low-signal and an additional 15 make one or two misreports.

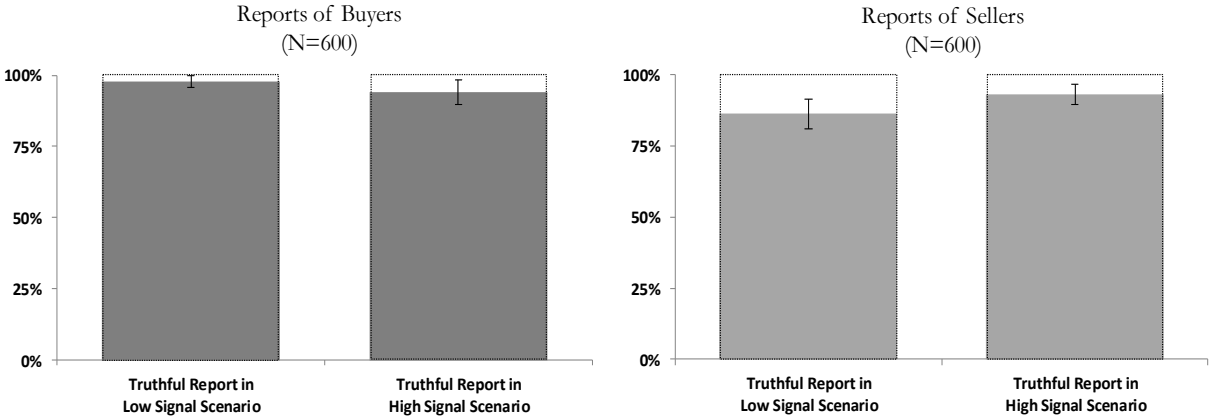
Taken together, behaviour in the no-noise treatment of the SR mechanism is strongly consistent with predicted behavior. Buyers report truthfully in over 90% of cases for both scenarios and seller's converge to similar behavior. Buyers also report truthfully in the majority of cases where they enter arbitration and buyers and sellers have strong incentives to report truthfully in the first stage given the empirical distribution of the data. Finally, there is no apparent heterogeneity in strategy across subjects.

3.4.2 The Simultaneous Report Mechanism Under a Private-Value Perturbation

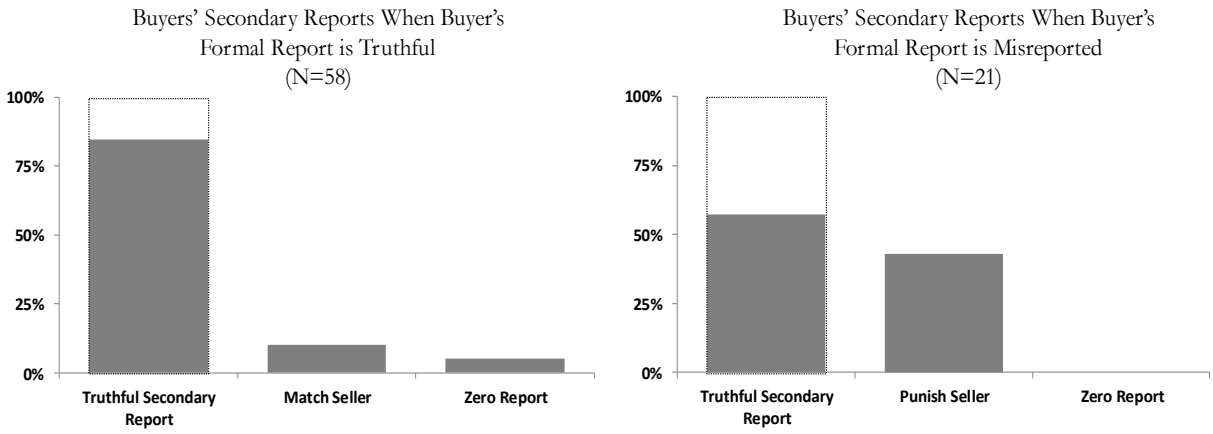
Our second hypothesis predicts that in the noise treatment, both buyers and sellers will continue to make truthful reports in the first stage in both the scenario where they receive a high signal and the scenario where they receive a low signal. We also predict that buyers

experience: A seller who lies in the low-signal scenario in one period lies in the same scenario only 26.1 percent of the time in the next period if this scenario arises and they are able to observe the buyers response. Since sellers rarely switch from a truthful strategy to a lying strategy, the aggregate truth-telling rate for this scenario increases from 82.3 percent in periods 1-5 to 90.0 percent in periods 6-10.

(a) Proportion of Truthful Reports in Report Stage



(b) Reports in Second Stage



(c) Aggregate Number of Misreports by Buyers and Sellers

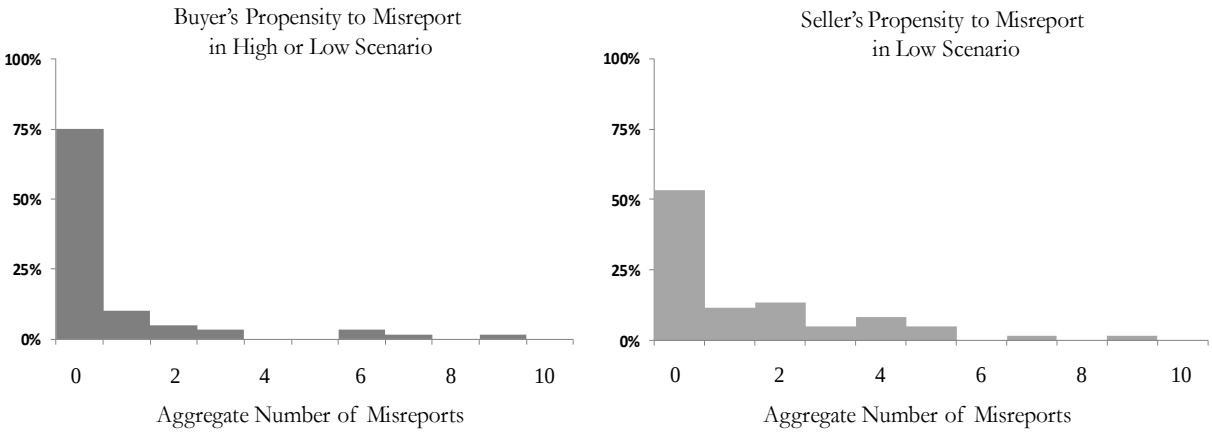


Figure 1: Pattern of Play in No-Noise Treatment of Simultaneous Report Mechanism

will continue to report truthfully in the second stage. The data is largely consistent with these predictions:

Result 2 *In the noise treatment of the Simultaneous Report, buyers report truthfully in* 96.5 *percent of cases in the low-signal scenario and in* 90.0 *percent of cases in the high-signal scenario. Sellers report truthfully in* 82.8 *percent of cases in the low-signal scenario and in* 92.8 *percent of cases in the high-signal scenario. Buyers who enter arbitration report the true value in* 71.3 *percent of cases.*

We also would predict that there is no statistical difference in behavior between behavior in the noise and no-noise treatment of the SR mechanism.

Result 3 *Comparing behavior in the no-noise and noise treatments with the Simultaneous Report mechanism, there is no significant difference in (i) the proportion of buyers who misreport, (ii) the proportion of sellers who misreport, or (iii) the proportion of groups that exhibit truth-telling first-stage strategies.*

Figure 2 shows the pattern of play in the noise treatment of the SR mechanism and is directly comparable to Figure 1. As can be seen in the left-hand side of panel (a), the frequency of truthful reports by buyers remains high with the buyer reporting truthfully in 96.5 percent of cases with the low signal and in 90.0 percent of cases after the high signal. The proportion of misreports made by buyers in the noise treatment is not significantly different from the proportion of misreports made in the no-noise treatment in a simple regression where buyer misreports is regressed on the noise treatment dummy (p-value $= 0.12$).[22]

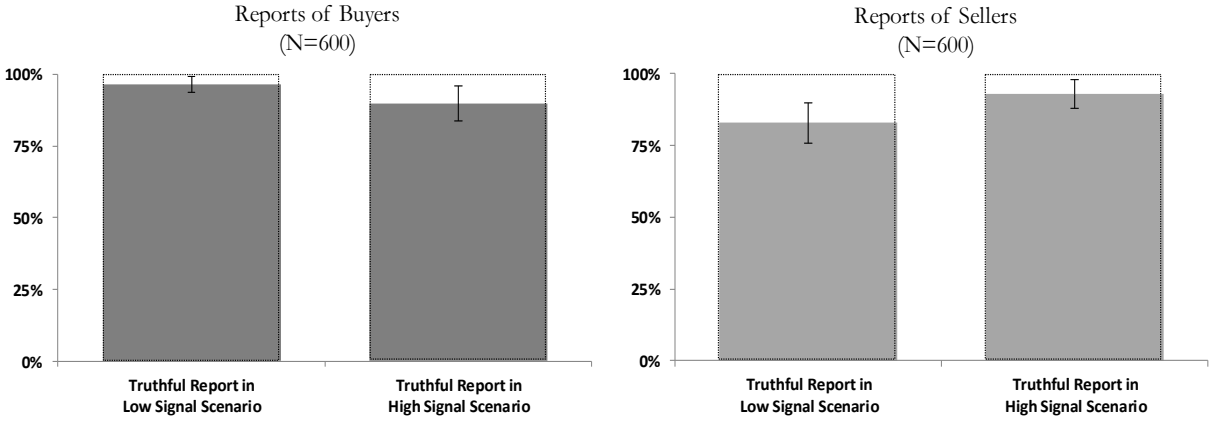
Sellers report truthfully in 82.8 percent of cases after the high signal and 92.8 percent of cases after the low signal. The proportion of seller misreports in the low-signal scenario is not significantly different from the no-noise treatment using a simple regression where seller misreports in the low-signal scenario are regressed on the noise treatment dummy (p-value $= 0.354$).

As seen in panel (b), buyers continue to follow their signal when they enter into the arbitration stage. In cases where the buyer's formal report was truthful, the buyers second report followed his original signal in 96 out of 126 cases. In cases where the buyer's formal report was misreported, the buyer's second report is truthful in 16 out of 31 cases.

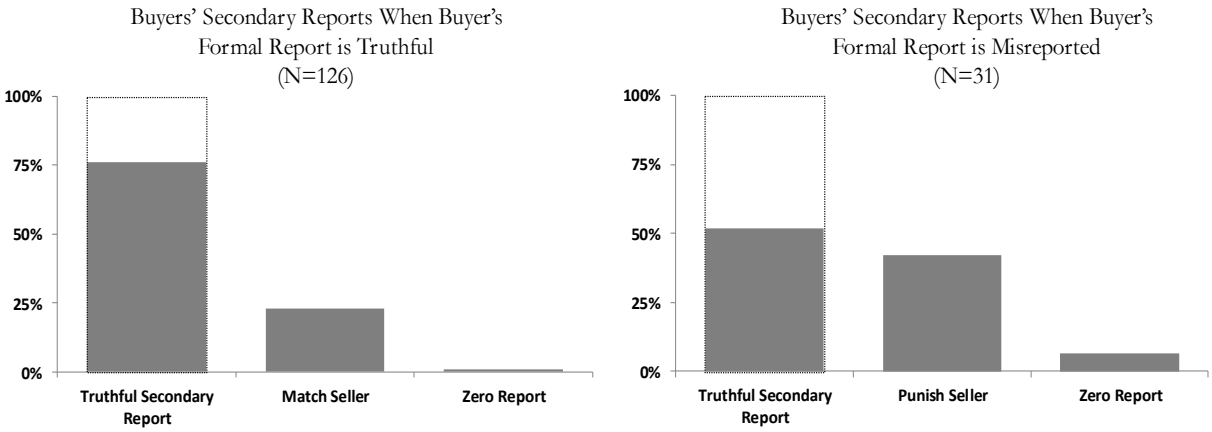
Finally, as seen in panel (c), the majority of buyers make truthful reports in both the high and low signal scenario and the majority of sellers make truthful reports in the low-signal scenario. The distribution of buyer misreports in the noise treatment is not

[22]Unless otherwise specified, regressions on buyers behavior is clustered by buyer, regressions on sellers behavior is clustered by seller, and regressions on group outcomes are clustered by both buyers and sellers.

(a) Proportion of Truthful Reports in Report Stage



(b) Reports in Second Stage



(c) Aggregate Number of Misreports by Buyers and Sellers

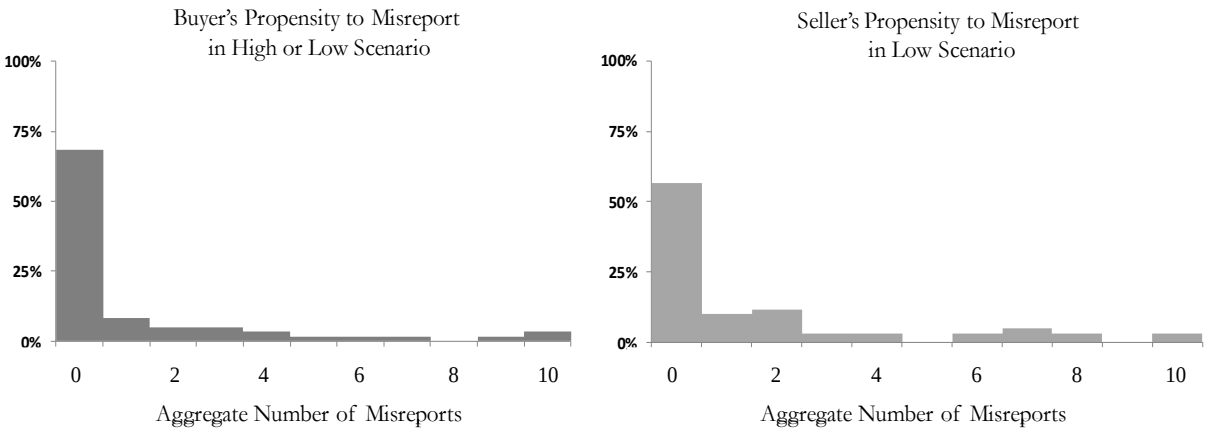


Figure 2: Pattern of Play in Noise Treatment of Simultaneous Report Mechanism

significantly different in the distribution of misreports in the no-noise treatment using a Wilcoxon Sign-Ranked test (p-value = 0.14). Likewise, there is no significant difference in the distribution of seller challenges with the low signal (p-value = 0.65).

At the aggregate level, 68.5 percent of groups exhibit first-stage truth-telling strategies in the noise treatment. This is not significantly different from the proportion of groups exhibiting first-stage truth-telling strategies in the no-noise treatment when a variable that is one if a group exhibits first-stage truth-telling strategies and zero otherwise is regressed on the treatment variable (p-value = 0.15).

3.5 The Relative Performance of the Simultaneous Report Mechanism

Thus far we have shown that the SR mechanism is effective at inducing truthful reports and leads to the efficient outcome in the majority of cases under complete information and under a private value perturbation. We now compare the mechanism to the SPI mechanism, which is predicted to generate buyer misreports under a private value perturbation.

Result 4 *In the no-noise treatment, buyers are significantly more likely to make a misreport in the SPI mechanism than in the SR mechanism. The introduction of noise leads to a significant increase in misreports in the SPI mechanism but does not increase misreports in the SR mechanism.*

Figure 3 compares the proportion of periods in which the buyer misreports his signal in the no-noise treatments (left) and noise treatments (right). The error bars are 95% confidence intervals. As can be seen in the left hand side, buyers misreport in 7.7 percent of cases in the no-noise treatment with the SR mechanism and in 25.5 percent of cases in the no-noise treatment with the SPI mechanism.[23] This difference is significant in a simple regression where buyer lies is regressed on the SPI treatment dummy with data restricted to the no-noise treatments (p-value < 0.01).[24]

As can be seen in the right hand side, buyers misreport in 12.5 percent of cases in the noise treatment with the SR mechanism and in 43.3 percent of cases with the SPI

[23]The large number of buyer lies in the high-signal scenario is consistent with the alternative equilibrium outlined in Section 3.3 where a seller believes that the buyer may not be rational after observing the buyer report a low value with the high signal.

[24]An extended analysis of the SPI treatment is provided in the appendix. As seen there, we also perform a Mann-Whitney-Wilcoxen tests on the distribution of buyer reports across the treatments at the buyer level and the session level. The null hypothesis of these tests are rejected at the .01 level for all comparisons between the SR and SPI treatments and between the noise and no-noise treatments of the SPI mechanism.

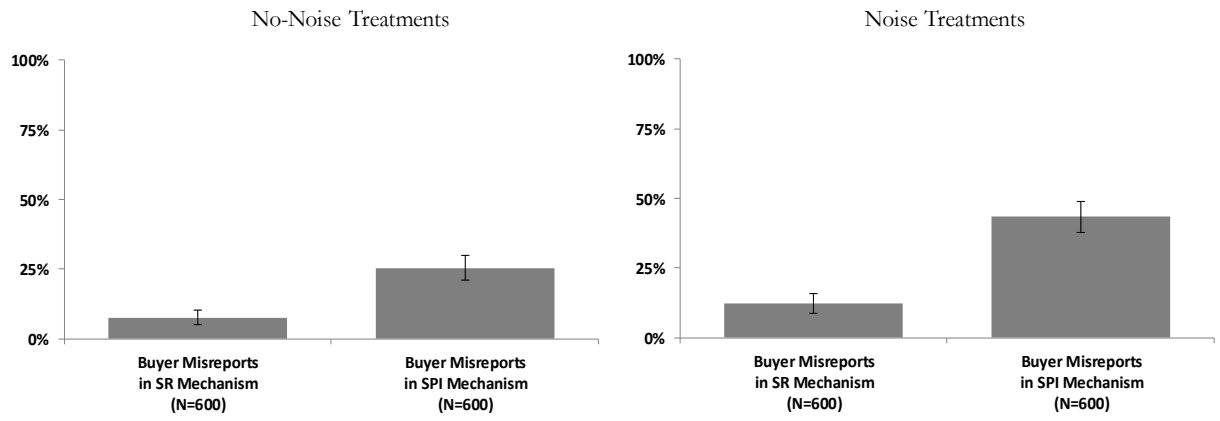


Figure 3: Proportion of Buyer Misreports in the SR and SPI Mechanisms in the No-Noise and Noise Treatments

mechanism. Consistent with the theoretical predictions, there are significantly more lies in the SPI mechanism in the noise treatment than the SR mechanism (p-value < 0.01).

Finally, using the no-noise treatment as a baseline, there is an 4.8 percent increase in buyer misreports when noise is introduced in the SR mechanism and an 18.8 percent increase in buyer lies when noise is introduced in the SPI mechanism. Using a simple difference-in-difference estimator where buyer lies is regressed on the SPI treatment variable, the noise treatment variable, and the interaction of noise and the SPI treatment variable, the interaction term is significantly different from zero (p-value $= 0.03$). Thus, the introduction of noise increases the number of lies both in absolute terms and when only considering the relative change in misreport rates that occur when noise is introduced.

At the aggregate level, groups display truth-telling behavior in 78.3 percent of cases in the no-noise treatment with the SPI mechanism. This is significantly different from the 86.3 percent of groups that display truth-telling behavior in the no-noise treatment of the SR mechanism when a variable that is one if a group displays truth-telling behavior and zero otherwise is regressed on the SPI mechanism treatment variable (p-value $= 0.02$). Groups display truth-telling behavior in 72.4 percent of cases in the noise treatment with the SPI mechanism. This is not significantly lower than in the no-noise treatment with the SPI mechanism (p-value $= 0.06$). However, it is significantly lower than the 82.5 percent of groups that display truth-telling behavior in the noise treatment of the SR mechanism (p-value $= 0.01$).

4 Conclusion

The question of what social objectives can be achieved in decentralized environments is a fundamental one, and one that is germane to a wide class of problems. Beginning with Maskin (1977, 1999), implementation theory has been remarkably successful in establishing strong positive results pertaining to this question.

Extensive-form mechanisms have been utilized to obtain particularly striking results, such as in Moore and Repullo (1988) who show that any SCF can be implemented as the unique subgame perfect equilibrium of a suitably constructed multi-stage mechanism in "economic environments".[25]

However, there is also a long tradition in game theory (see, for instance: Fudenberg et al. (1988), Monderer and Samet (1989), Dekel and Fudenberg (1990) and Kajii and Morris (1997)) of skepticism about the robustness of refinements of Nash equilibrium to small perturbations of the environment. Aghion et al. (2012) raise these types of concerns in the context of implementation theory, and Fehr et al. (2017) and Aghion et al. (2018) illustrate them as a practical matter in laboratory settings.

The key issue is that extensive-form mechanisms give rise to consideration of how beliefs evolve when unexpected play occurs. These considerations drive the non-robustness of mechanisms that use refinements of Nash equilibrium as a solution concept.

Our contribution in this paper is to articulate a mechanism that is robust theoretically and experimentally to these considerations about the evolution of beliefs during play. Our *Simultaneous Report* mechanism fully implements any social choice function under initial rationalizability in complete information environments. This solution concept iteratively deletes strategies that are not best replies, but only mandates rationality and common beliefs at the beginning of the game. Crucially, it makes no assumption about how beliefs evolve after zero probability events occur. This makes it the weakest rationalizability concept for extensive-form games.

As a theoretical matter, our mechanism is robust to small amounts of incomplete information about the state of nature. We also highlight the robustness of the mechanism to a wide variety of reasoning processes and behavioral assumptions.

Our mechanism performs very well experimentally. Truth-telling rates are high for both buyers and sellers in both an environment with complete information and one with a private value perturbation. The mechanism also outperforms a canonical SPI mechanism that uses a near identical price schedule and fines and bonuses of the same size.

In general, one would expect that when mechanisms work well, economic and other

[25] i.e. with transferable utility or with at least one divisible private good.

activity would be mediated by contract. When mechanisms do not work well, one would expect authority, in one form or another, to play a larger role. This has clear implications for the theory of the firm, but also for other settings where interactions can be structured. The organization of the political process is a leading example of such a setting, as are "vertical legal relationships", such as between different courts or tiers of government.

These political and legal environments may well be more complicated than the simple revelation game studied in our experiments. Understanding the efficacy of our SR mechanism—or a suitably adapted variant—in these richer environments may be a fruitful direction for further work.

References

Abreu, D. and H. Matsushima (1992): "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica*, 60, 993–1008.

Aghion, P., E. Fehr, R. Holden, and T. Wilkening (2018): "The Role of Bounded Rationality and Imperfect Information in Subgame Perfect Implementation — An Empirical Investigation," *Journal of the Eurpoean Economic Association*, 16, 232–274.

Aghion, P., D. Fudenberg, R. Holden, T. Kunimoto, and O. Tercieux (2012): "Subgame-Perfect Implementation Under Information Perturbations," *Quarterly Journal of Economics*, 1843–1881.

Andreoni, J. and H. Varian (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. and J. Ledyard (2004): "Scaling up Learning Models in Public Good Games," *Journal of Public Economic Theory*, 6, 203—238.

Attiyeh, G., R. Franciosi, and R. M. Isaac (2000): "Experiments with the Pivot Process for Providing Public Goods," *Public Choice*, 102, 95—114.

Ben-Porath, E. (1997): "Rationality, Nash equilibrium and backwards induction in perfect-information games," *The Review of Economic Studies*, 64, 23–46.

Bergemann, D. and S. Morris (2005): "Robust mechanism design," *Econometrica*, 73, 1771–1813.

Bracht, J., C. Figuières, and M. Ratto (2008): "Relative performance of two simple incentive mechanisms in a public goods experiment," *Journal of Public Economics*, 92, 54 – 90.

Chen, Y. and R. Gazzale (2004): "When Does Learning in Games Generate Convergence to Nash Equilibria? The Role of Supermodularity in an Experimental Setting," *American Economic Review*, 94, 1505–1535.

Chen, Y. and C. Plott (1996): "The Groves–Ledyard Mechanism: An Experimental Study of Institutional Design," *Journal of Public Economics*, 59, 335–364.

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

Chung, K.-S. and J. C. Ely (2003): "Implementation with Near-Complete Information," *Econometrica*, 71, 857–871.

Crawford, V. P. and N. Iriberri (2007): "Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions?" *Econometrica*, 75, 1721–1770.

de Clippel, G., K. Eliaz, and B. Knight (2014): "On the Selection of Arbitrators," *American Economic Review*, 104, 3434–3458.

de Clippel, G., R. Saran, and R. Serrano (2018): "Level-k mechanism design," *The Review of Economic Studies*.

Dekel, E. and D. Fudenberg (1990): "Rational behavior with payoff uncertainty," *Journal of Economic Theory*, 52, 243–267.

Dekel, E. and M. Siniscalchi (2013): "Epistemic game theory," Tech. rep., Mimeo.

——— (2015): "Epistemic Game Theory," *Handbook of Game Theory with Economic Applications*, 4, 619–702.

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

Fehr, E., M. Powell, and T. Wilkening (2017): "Behavioral Constraints on the Design of Subgame-Perfect Implementation Mechanisms," Mimeo.

Fischbacher, U. (2007): "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments," *Experimental Economics*, 10, 171–178.

Fudenberg, D., D. M. Kreps, and D. K. Levine (1988): "On the Robustness of Equilibrium Refinements," *Journal of Economic Theory*, 44, 354 – 380.

Giannatale, S. D. and A. Elbittar (2010): "King Solomon's Dilemma: An Experimental Study on Implementation," Working Paper 477, CIDE.

Greiner, B. (2015): "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE," *Journal of the Economic Science Association*, 1, 114–125.

Harstad, R. M. and M. Marrese (1981): "Implementation of Mechanism by Processes: Public Good Allocation Experiments," *Journal of Economic Behavior & Organization*, 2, 129–151.

——— (1982): "Behavioral explanations of efficient public good allocations," *Journal of Public Economics*, 19, 367–383.

Hart, O. and J. Moore (2003): "Some (Crude) Foundations of incomplete contracts," *Mimeo*.

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

Hoppe, E. I. and P. W. Schmitz (2011): "Can Contracts Solve the Hold-Up Problem? Experimental Evidence," *Games and Economic Behavior*, 73, 186–199.

Kajii, A. and S. Morris (1997): "The robustness of equilibria to incomplete information," *Econometrica*, 65, 1283–1309.

Katok, E., M. Sefton, and A. Yavas (2002): "Implementation by Iterative Dominance and Backward Induction: An Experimental Comparison," *Journal of Economic Theory*, 104, 89–103.

Machina, M. J. and D. Schmeidler (1992): "A More Robust Definition of Subjective Probability," *Econometrica*, 60, 745–780.

Maskin, E. (1977, 1999): "Nash Equilibrium and Welfare Optimality," *Review of Economic Studies*, 66, 39–56.

Maskin, E. and J. Tirole (1999): "Unforeseen contingencies and incomplete contracts," *The Review of Economic Studies*, 66, 83–114.

McKelvey, R. D. and T. R. Palfrey (1998): "Quantal Response Equilibria for Extensive Form Games," *Experimental Economics*, 1, 9–41.

Monderer, D. and D. Samet (1989): "Approximating common knowledge with common beliefs," *Games and Economic Behavior*, 1, 170–190.

Moore, J. and R. Repullo (1988): "Subgame Perfect Implementation," *Econometrica*, 56, 1191–1220.

Nöldeke, G. and K. Schmidt (1995): "Option Contracts and Renegotiation: A Solution to the Hold-Up Problem," *RAND Journal of Economics*, 26, 163–179.

Ponti, G., A. Gantner, D. López-Pintado, and R. Mongtgomery (2003): "Solomon's Dilemma: An Experimental Study on Dynamic Implementation," *Review of Economic Design*, 8, 217–239.

Sefton, M. and A. Yavas (1996): "Abreu—Matsushima mechanisms: experimental evidence," *Games and Economic Behavior*, 16, 280–302.

Table of Contents: Appendix

Appendix A: Theory
A.1: Proof of Theorem 1
A.2: Proof of Theorem 2

Appendix B: Additional Analyses and Treatments
B.1: The SPI Treatment
B.2: Additional Figures Comparing the SR and SPI Mechanisms
B.3: Pre-Analysis Plan
B.4: Insructions

Appendix A: Theory

A.1 Proof of Theorem 1

Let θ be the true state. We prove Theorem 1 in the following three steps.

A.1.1 Truth-Telling Condition

Claim 1 *If* $m_i \in R_{i,1}^{\Gamma(\theta)}$ *and* $i \in \mathcal{I}^*(m^1)$, *then* $m_i^2(m^1) = \theta_i$.

Proof. Let m^1 be a message profile realized at Stage 1 such that $\mathcal{I}^*(m^1) \neq \varnothing$. First, for every $i \in \mathcal{I}^*(m^1)$, $l_i(m_i^2)$ is implemented with probability $1/|\mathcal{I}^*(m^1)|$. Second, m_i^2 determines the outcome only when $l_i(m_i^2)$ is chosen. Hence, by Lemma 1, $m_i^2(m^1) = \theta_i$ is the unique best response conditional on m^1. ■

A.1.2 Inter-Stage Coordination Condition

Claim 2 *If* $m_i \in R_{i,2}^{\Gamma(\theta)}$, *then* $m_i^1 = (\hat{\theta}_i^i, \theta_{i-1})$ *for some* $\hat{\theta}_i^i \in \Theta_i$, *i.e., player* i *must report the type of player* $i-1$ *truthfully at Stage 1.*

Proof. Since $m_i \in R_{i,2}^{\Gamma(\theta)}$, we know that m_i is a sequential best reply to some CPS μ_i such that $\mu_i(R_{-i,1}^{\Gamma(\theta)} | M_{-i}) = 1$. We fix such μ_i and $m^1 \in M^1$ as a message profile chosen at the first stage. By Claim 1, it follows that for each $j \in \mathcal{I}$,

$$\text{marg}_{M_j}\mu_i\left(m_j^2\left(m^1\right) = \theta_j | M_{-i}\right) = 1 \text{ if } j \in \mathcal{I}^*\left(m^1\right).$$

Fix an arbitrary message profile $m_{-i} \in R_{-i,1}^{\Gamma(\theta)}$. In what follows, we can assume that each player, who is called upon in Stage 2, always announces his/her true type. No matter how player i chooses $\hat{\theta}_{i-1}^i$, player i's resulting payoff difference from altering the outcome is bounded from above by D.

We shall show that against any message profile $m_{-i} \in R_{-i,1}^{\Gamma(\theta)}$ of player i's opponents, reporting $m_i^1 = (\hat{\theta}_i^i, \theta_{i-1})$ in Stage 1 is strictly better for player i than reporting $m_i^1 = (\hat{\theta}_i^i, \hat{\theta}_{i-1}^i)$ with $\hat{\theta}_{i-1}^i \neq \theta_{i-1}$. More specifically, we establish this claim by considering the extra transfers associated with different choices player i might make in the following two cases.

Case 1. $\hat{\theta}_{i-1}^{i-1} \neq \theta_{i-1}$.

For player i, reporting $\hat{\theta}_{i-1}^i \neq \theta_{i-1}$ will result in either the penalty T (if $\hat{\theta}_{i-1}^i \neq \hat{\theta}_{i-1}^{i-1}$) or no transfer (if $\hat{\theta}_{i-1}^i = \hat{\theta}_{i-1}^{i-1}$), while reporting $\hat{\theta}_{i-1}^i = \theta_{i-1}$ will result in the reward T. Thus, the transfer gain from reporting $\hat{\theta}_{i-1}^i = \theta_{i-1}$ relative to $\hat{\theta}_{i-1}^i \neq \theta_{i-1}$ is at least T. Since

$T > D$, reporting $\hat{\theta}^i_{i-1} = \theta_{i-1}$ in the first stage is strictly better for player i than reporting $\hat{\theta}^i_{i-1} \neq \theta_{i-1}$.

Case 2. $\hat{\theta}^{i-1}_{i-1} = \theta_{i-1}$.

For player i, reporting $\hat{\theta}^i_{i-1} \neq \theta_{i-1}$ will result in the penalty T, while reporting $\hat{\theta}^i_{i-1} = \theta_{i-1}$ will not induce any transfer. Thus, the transfer gain from reporting $\hat{\theta}^i_{i-1} = \theta_{i-1}$ relative to $\hat{\theta}^i_{i-1} \neq \theta_{i-1}$ is T. Again, since $T > D$, reporting $\hat{\theta}^i_{i-1} = \theta_{i-1}$ in Stage 1 is strictly better for player i than reporting $\hat{\theta}^i_{i-1} \neq \theta_{i-1}$.

Thus, in both cases, it is strictly better for player i to report θ_{i-1} in the first stage than to report any $\hat{\theta}^i_{i-1} \neq \theta_{i-1}$. We conclude that against any $m_{-i} \in R^{\Gamma(\theta)}_{-i,1}$, reporting $(\hat{\theta}^i_i, \hat{\theta}^i_{i-1})$ with $\hat{\theta}^i_{i-1} \neq \theta_{i-1}$ is strictly dominated by $(\hat{\theta}^i_i, \theta_{i-1})$. Hence, player i reports the type of player $i-1$ truthfully in the first stage, i.e., $m^1_i = (\hat{\theta}^i_i, \theta_{i-1})$ for every $m_i \in R^{\Gamma(\theta)}_{i,2}$. ■

A.1.3 Within-Stage Coordination Condition

Claim 3 *If $m_i \in R^{\Gamma(\theta)}_{i,3}$, then $m^1_i = (\theta_i, \theta_{i-1})$.*

Proof. Let $m_i \in R^{\Gamma(\theta)}_{i,3}$. Then, we know that m_i is a best reply to some CPS μ_i such that $\mu_i(R^{\Gamma(\theta)}_{-i,2}|M_{-i}) = 1$. We fix such μ_i. By Claim 2, μ_i has the following property:

$$\mu_i\left(m^1_{-i}|M_{-i}\right) > 0 \Rightarrow m^1_{i+1} = (\hat{\theta}^{i+1}_{i+1}, \theta_i) \text{ for some } \hat{\theta}^{i+1}_{i+1} \in \Theta_{i+1}.$$

That is, we know that player $i+1$ makes a truthful announcement about player i's type in the first stage. Hence, if player i misreports his/her own type by announcing some $\hat{\theta}^i_i \neq \theta_i$, he/she will be penalized by T. Since $T > D$, player i's unique best response is to truthfully announce his/her own type in the first stage. Hence, every player i will truthfully report his/her type at the first stage, i.e., $\hat{\theta}^i_i = \theta_i$. Combining this with Claim 2, we conclude that $m^1_i = (\theta_i, \theta_{i-1})$. ■

A.2 Proof of Theorem 2

To prove Theorem 2, we continue to use the same SR mechanism which we defined in Section 2.3. Similar to the proof of Theorem 1, we prove Theorem 2 by establishing three conditions. In the proof, let $\{\pi^k\}$ denote a private-value perturbation to π^{CI}.

A.2.1 Truth-Telling Condition

Claim 4 *Let $m^1 \in M^1$. For every $i \in \mathcal{I}^*(m^1)$, we have that $m^2_i(m^1) = \theta_i$ for any $m_i \in R_{i,1}\left(s^\theta_i|\Gamma\left(\pi^k\right)\right)$ and any k sufficiently large.*

Proof. Fix $k \geq 1$, player $i \in \mathcal{I}$, and a message $m_i \in R_{i,1}\left(s_i^\theta | \Gamma\left(\pi^k\right)\right)$. Observe that m_i is a sequential best response against some CPS $\mu_{i,k}$ consistent with signal s_i^θ. We fix such μ_i. Consider any $m^1 \in M^1$ such that $i \in \mathcal{I}^*\left(m^1\right)$. Conditional on m^1, only the Stage 2 message m_i^2 matters for player i's payoff; moreover, m_i^2 matters only when $l_i(m_i^2)$ is chosen by the mechanism. By Lemma 1, $m_i^2\left(m^1\right) = \theta_i$ is the unique sequential best reply against $\mu_{i,k}$ for player i with signal s_i^θ, as long as

$$\text{marg}_{\Theta_i}\mu_{i,k}\left(\theta_i | M_{-i}\left(m^1\right)\right) \to 1 \text{ as } k \to \infty,$$

which actually follows from Bayes' rule. Specifically, we write $\mu_{i,k}(\cdot)$ for $\mu_{i,k}(\cdot|\varnothing)$ and compute the following:

$$\begin{aligned}
& \text{marg}_{\Theta_i}\mu_{i,k}\left(\theta_i | M_{-i}\left(m^1\right)\right) \\
= & \sum_{\theta_{-i}, s_{-i}} \sum_{\bar{m}_{-i} \in M_{-i}(m^1)} \mu_{i,k}\left(\theta_i, \theta_{-i}, s_{-i}, \bar{m}_{-i} | M_{-i}\left(m^1\right)\right) \\
= & \frac{\sum_{\theta_{-i}, s_{-i}} \sum_{\bar{m}_{-i} \in M_{-i}(m^1)} \mu_{i,k}\left(\theta_i, \theta_{-i}, s_{-i}, \bar{m}_{-i}\right)}{\sum_{\theta_i', \theta_{-i}, s_{-i}} \sum_{\bar{m}_{-i} \in M_{-i}(m^1)} \mu_{i,k}\left(\theta_i', \theta_{-i}, s_{-i}, \bar{m}_{-i}\right)} \\
= & \frac{\sum_{\theta_{-i}, s_{-i}} \sum_{\bar{m}_{-i} \in M_{-i}(m^1)} \mu_{i,k}\left(\bar{m}_{-i} | \theta_i, \theta_{-i}, s_{-i}, s_i^\theta\right) \pi^k\left(\theta_i, \theta_{-i}, s_{-i} | s_i^\theta\right)}{\sum_{\theta_i', \theta_{-i}, s_{-i}} \sum_{\bar{m}_{-i} \in M_{-i}(m^1)} \mu_{i,k}\left(\bar{m}_{-i} | \theta_i', \theta_{-i}, s_{-i}, s_i^\theta\right) \pi^k\left(\theta_i', \theta_{-i}, s_{-i} | s_i^\theta\right)} \\
= & \frac{\sum_{\theta_{-i}, s_{-i}} \sum_{\bar{m}_{-i} \in M_{-i}(m^1)} \mu_{i,k}\left(\bar{m}_{-i} | \theta_i, \theta_{-i}, s_{-i}, s_i^\theta\right) \pi^k\left(\theta_i | s_i^\theta, s_{-i}, \theta_{-i}\right) \pi^k\left(\theta_i, \theta_{-i}, s_{-i} | s_i^\theta\right)}{\sum_{\theta_i', \theta_{-i}, s_{-i}} \sum_{\bar{m}_{-i} \in M_{-i}(m^1)} \mu_{i,k}\left(\bar{m}_{-i} | \theta_i', \theta_{-i}, s_{-i} s_i^\theta\right) \pi^k\left(\theta_i' | s_i^\theta, s_{-i}, \theta_{-i}\right) \pi^k\left(\theta_i', \theta_{-i}, s_{-i} | s_i^\theta\right)}, \quad (4)
\end{aligned}$$

where the second equality follows from Condition (2) in Definition 1 and the third equality follows from the consistency of $\mu_{i,k}(\cdot)$ with s_i^θ. Finally, the definition of private-value perturbations implies that

$$\text{marg}_{\Theta_i}\pi^k\left[\theta_i | s_i^\theta, s_{-i}, \theta_{-i}\right] \to 1 \text{ as } k \to \infty \text{ for any } s_{-i},\ \theta_{-i}.$$

Hence, it follows from (4) that $\text{marg}_{\Theta_i}\mu_{i,k}\left(\theta_i | M_{-i}\left(m^1\right)\right) \to 1$ as $k \to \infty$. ■

A.2.2 Inter-Stage Coordination Condition

Claim 5 *For any $i \in \mathcal{I}$, k sufficiently large, and $m_i \in R_{i,2}\left(s_i^\theta | \Gamma\left(\pi^k\right)\right)$, we have $m_i^1 = (\hat{\theta}_i^i, \theta_{i-1})$ for some $\hat{\theta}_i^i \in \Theta$, i.e., player i must report the type of player $i-1$ truthfully at Stage 1.*

Proof. First, by Claim 4, if there exists $m^1 \in M^1$ such that $i-1$ in $\mathcal{I}^*\left(m^1\right)$, then player $i-1$ will report θ_{i-1} truthfully, as long as he/she plays $m_{i-1} \in R_{i-1,1}\left(s_{i-1}^\theta | \Gamma\left(\pi^k\right)\right)$ for k

large enough. Moreover, since Claim 2 holds under complete information, player i is strictly better off by reporting his/her predecessor's true type (i.e., $\hat{\theta}_{i-1}^{i} = \theta_{i-1}$) than telling a lie. This strict truth-telling incentive remains the same under perturbations $\{\pi^k\}$ (with k large), so long as player $i-1$ reports type θ_{i-1} with probability close to one in Stage 2. Since $\pi^k \to \pi^{\text{CI}}$, which implies $\pi^k\left(\theta, s_{-i}^{\theta} | s_i^{\theta}\right) \to 1$ as $k \to \infty$, player i of type s_i^{θ} believes with probability close to one that player $i-1$ also receives s_{i-1}^{θ} for any k large enough. Hence, player i reports the type of player $i-1$ truthfully in the first stage, i.e., $m_i^1 = (\hat{\theta}_i^i, \theta_{i-1})$ for any $m_i \in R_{i,2}\left(s_i^{\theta} | \Gamma\left(\pi^k\right)\right)$ and any sufficiently large k. ■

A.2.3 Within-Stage Coordination Condition

Claim 6 *For any $i \in \mathcal{I}$, k sufficiently large, and $m_i \in R_{i,3}\left(s_i^{\theta} | \Gamma\left(\pi^k\right)\right)$, we have $m_i^1 = (\theta_i, \theta_{i-1})$.*

Proof. Since Claim 3 holds under complete information, player i finds it strictly better to report his/her own type at Stage 1 rather than to tell a lie about it. This strict better reply of telling his/her true type as opposed to misreporting his/her type remains the same under the perturbation $\{\pi^k\}$ (with k large), so long as player $i+1$ reports player i's type θ_i truthfully with probability close to one at Stage 1. Therefore, it follows that $m_i^1 = (\theta_i, \theta_{i-1})$ for any $m_i \in R_{i,3}\left(s_i^{\theta} | \Gamma\left(\pi^k\right)\right)$ and any k sufficiently large. ■

B Additional Analyses and Treatments

B.1 The SPI Mechanism

In this appendix, we report on the behavior of buyers and sellers in treatments that use the SPI mechanism. Recall from Section 3.3 that the SPI mechanism is predicted to have a unique truth-telling equilibrium in the no-noise environment under subgame perfection, but that there are many initial rationalizable strategy profiles. We find:

Result B.1 *In the no-noise treatment with the SPI mechanism, buyers misreport the value of their good with the high signal in* 22.5 *percent of cases. Sellers challenge buyer's who misreport a high signal in* 78 *percent of cases and the buyer rejects a legitimate challenge in* 37.7 *percent of cases.*

Figure B.4 displays the patterns of play we observed in the first ten periods of the experiment. The left column examines play in the low-signal scenario, and the right column examines play in the high-signal scenario. Panel (a) summarizes the buyers' announcement decisions, Panel (b) summarizes the sellers' challenge decisions for different announcements, and Panel (c) summarizes the buyers' decisions to accept or reject counter offers.

Panel (a) shows that buyers are almost always truthful in the low-signal scenario. However, buyers misreport in the high-signal scenario in 22.5 percent of cases. This misreport rate is very similar to the long-run lie rate observed in Aghion et al. (2018), which study a similar mechanism and environment in experiments that lasted between 10 and 40 periods.

The left hand side of panel (b) shows that sellers mistakenly challenge a low report with a low signal in 11.0 percent of cases. This rate of false challenges is not significantly different to the proportion of sellers who misreport in the low-signal scenario in the SR mechanism in a simple regression that regresses misreporting behavior on the SPI treatment (p-value $= 0.43$).[26] Data for this test includes all observations from the low-signal scenario of the SR mechanism, but uses only the observations in the SPI mechanism where the low-signal is observed and the buyer has reported a low value because the seller's challenge behavior is not observed in the other cases.

The right hand side of panel (b) shows that sellers challenge a misreport in the high-signal scenario in only 78 percent of cases. Thus while appropriate challenges occur in the majority of cases, at least some sellers are reluctant to challenge. As seen in panel (c), buyers accept the counteroffer after an appropriate challenge in only 62 percent of cases and retaliate against appropriate challenges in 38 percent of cases.

[26] This regression was not in the pre-analysis plan and has been added based on the relatively high misreport rate of sellers observed in the SR mechanism.

We now turn to the SPI mechanism in the noise treatment:

Result B.2 *The introduction of noise leads to a significant increase of misreports by buyers with the high signal and a significant decrease in the proportion of challenges made by sellers who have high signals and observe a low report.*

Figure B.5 shows the path of play in the noise treatment with the SPI mechanism and is directly comparable with figure B.4 above. As seen in panel (a), buyers lie in 40 percent of observations in the high-signal scenario but in only 6.3 percent of observations in the low-signal scenario. The misreport rate in the high-signal scenario is significantly higher in the noise treatment than the no-noise treatment in a simple regression that regresses buyer misreports in the high-signal scenario on the noise treatment dummy (p-value $< .01$).

Panel (b) shows that sellers challenge in only 62.4 percent of cases when they have the high signal and the buyer has made a low announcement. This challenge rate is significantly lower than in the no-noise treatment in a simple regression that regresses seller challenges on the noise treatment dummy using data from observations where the seller has the high signal and receives a low report (p-value $< .034$).[27] Finally, panel (c) shows that buyers accept a counteroffer in 78 percent of cases when they misreport their value, have the high signal, and are challenged. This is slightly higher than in the no-noise baseline treatment, but the difference is not significant (p-value $= 0.059$).

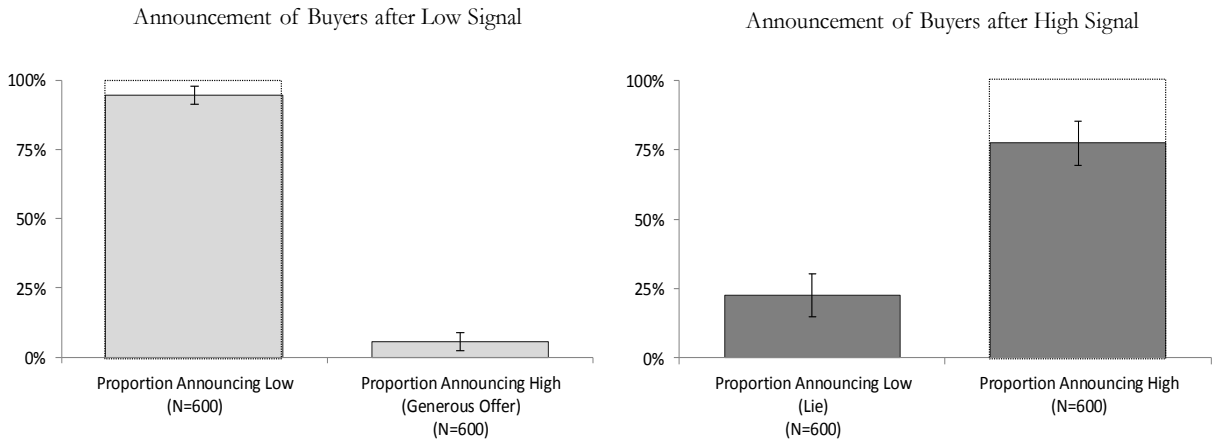
In both the no-noise treatment and noise treatment, buyers have a higher expected value for telling the truth in the high-signal scenario (31.8 in the no-noise treatment and 29.3 in the noise treatment) than they are expected to receive by lying and accepting all challenges (9.3 in the no-noise treatment and 20.35 in the noise treatment). Thus, we might expect to see truth-telling rates increase over time. Panel (a) of Figure B.6 tracks the proportion of truthful announcements in the high-signal scenario over time. This data is overlayed with the predictions and 95% confidence intervals from a simple linear random effects regression that regresses the reporting decision on the period. While there appears to be a small decrease in misreports over time, the time series is not significant in either random effects regression at the .05 level (No-noise treatment: p-value $= .059$; noise treatment: p-value $= .121$).

Finally, panel (c) of figure B.6 shows the distribution of buyer misreports in the no-noise and noise treatments of the SPI mechanism. While truth-telling is the modal action in the no-noise treatment, behavior here is heterogeneous with a small number of buyers misreporting in every period. When noise is introduced, the distribution of lies shifts to the right and the distribution becomes bimodal, with some buyers lying in every period.[28]

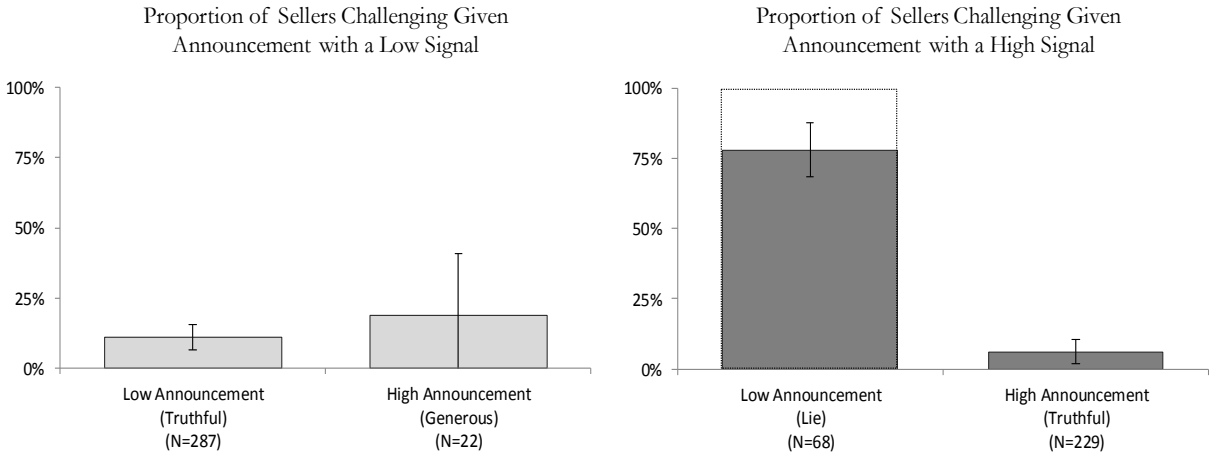
[27] This regression was not in the pre-analysis plan, but the result is consistent with theory.

[28] Sell also figure B.8 which shows buyer reports in both treatments simultaneously.

(a) Announcements of Buyers



(b) Challenges of Sellers



(c) Acceptances of Counter-Offers by Buyers

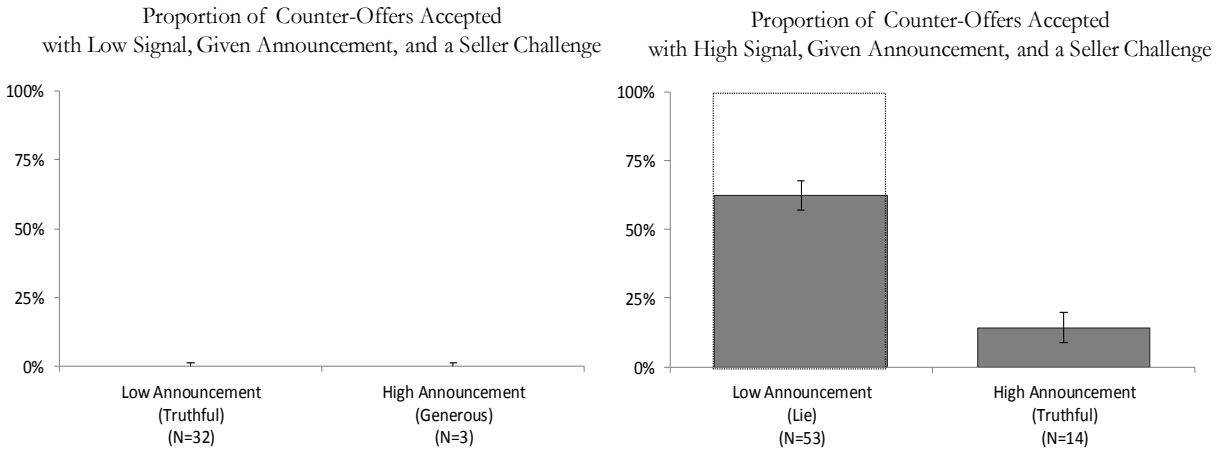
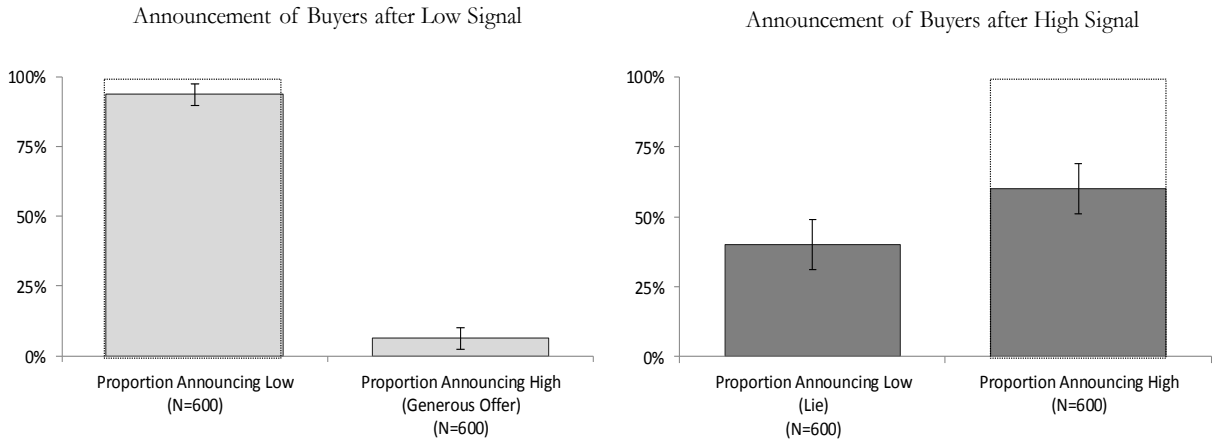
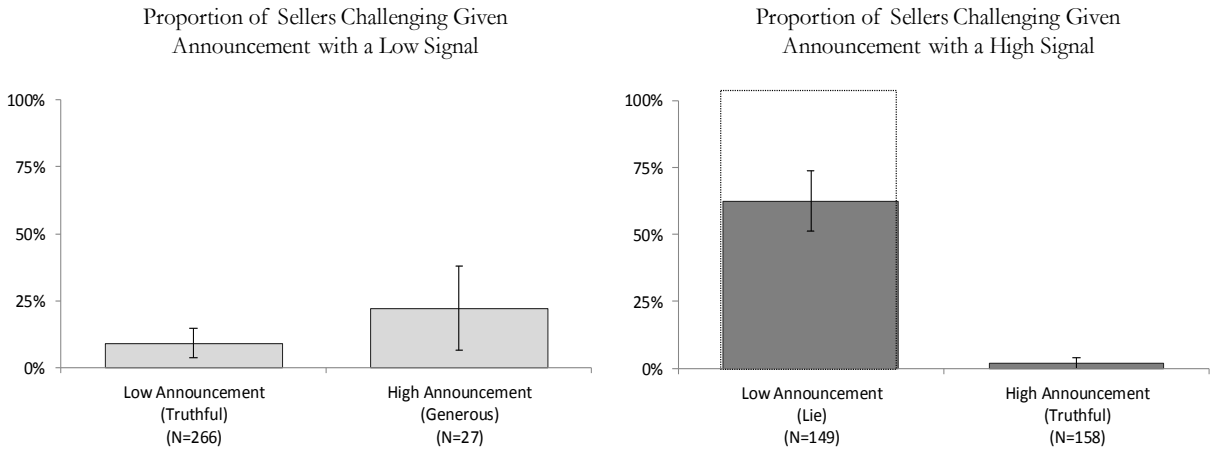


Figure B.4: Path of Play in No Noise Treatment with the SPI mechanism

(a) Announcements of Buyers



(b) Challenges of Sellers



(c) Acceptances of Counter-Offers by Buyers

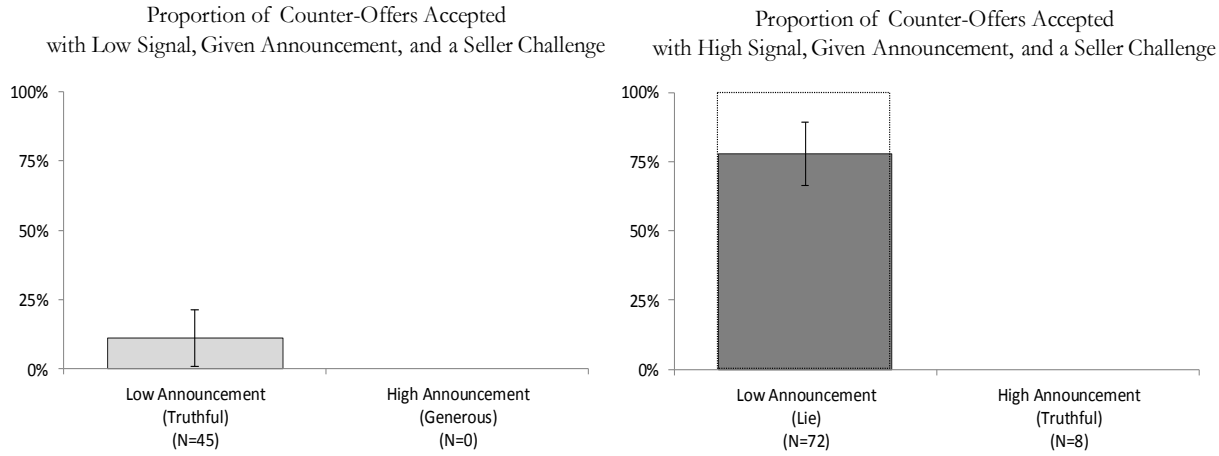
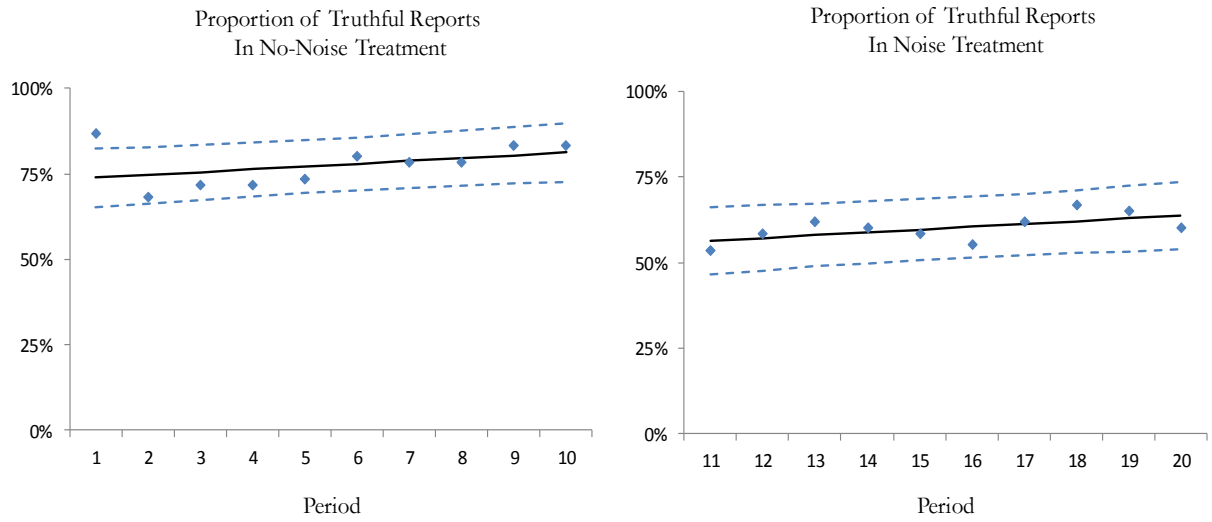
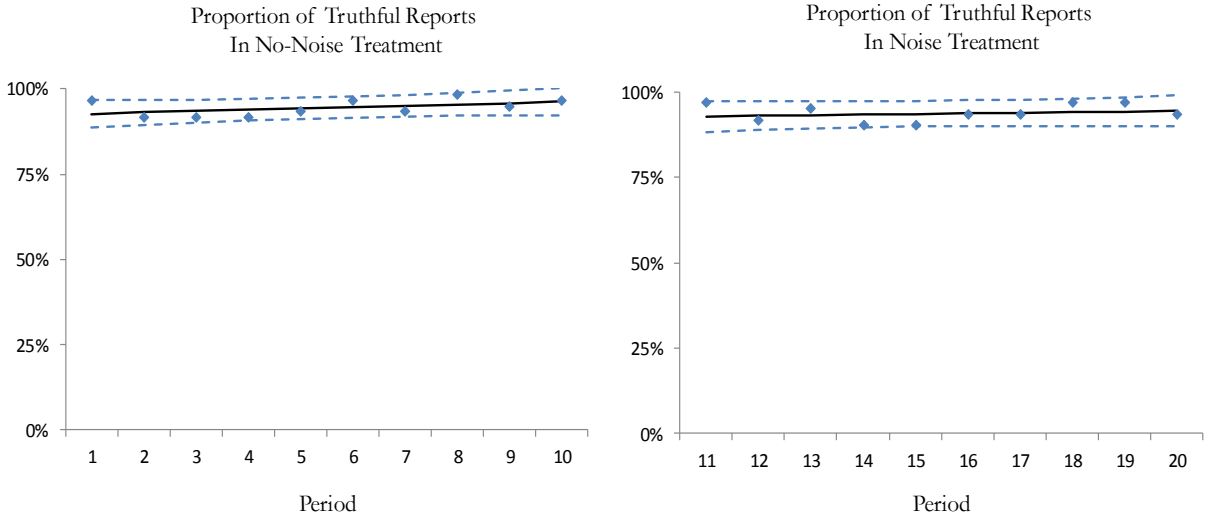


Figure B.5: Path of Play in Noise Treatment with the SPI mechanism

(a) Proportion of Truthful Reports by Buyers in High-Signal Scenario of SPI Mechanism



(b) Proportion of Truthful Reports by Buyers in Low-Signal Scenario of SPI Mechanism



(c) Aggregate Number of Misreports by Buyers in SPI Mechanism

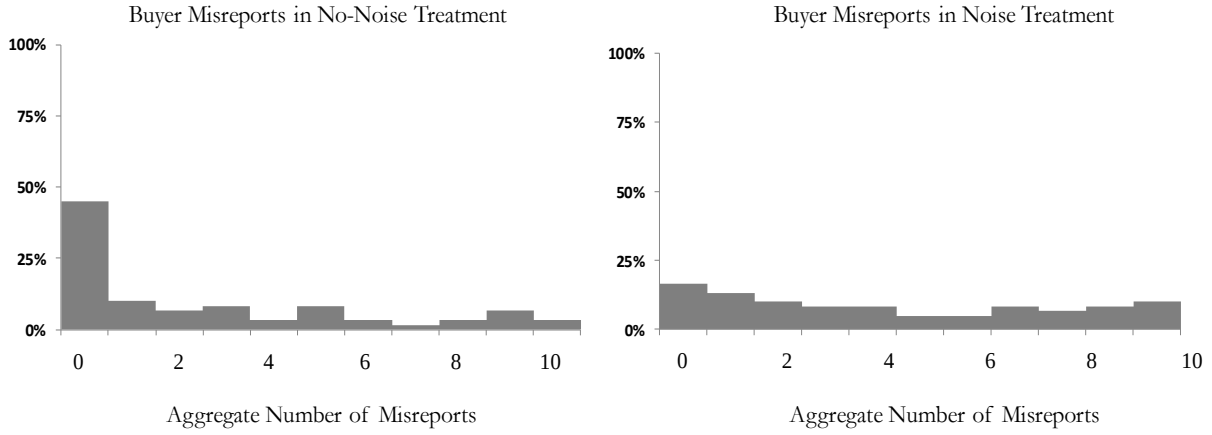


Figure B.6: Evolution and Distribution of Buyer Misreports in SPI Mechanism

B.2 Additional Figures: Comparison of misreports in the SR and SPI mechanisms

This appendix provides additional analysis that compares the behavior of buyers in each of the four treatments. Table B.1 reports coefficients from OLS regressions with buyer lies as the left hand side, and the treatments on the right. Column (1) includes data only from the noise treatments while column (2) includes all four treatments. Column (3) is a random effects regression and clusters data at the session level. Note that while the $*$'s represent two-sided significance levels, the predicted difference between the SR and SPI mechanisms in the noise treatments is one-sided in the pre-analysis plan.

Table B.2 provides non-parametric comparisons of the treatments with data aggregated at either the individual buyer level or the session level while Figure B.7 shows buyer misreports over time in all four treatments. The prediction and confidence intervals that has been overlayed on the data in Figure B.7 are from simple random effects regressions with a treatment-specific linear time trend.

Figure B.8 shows the aggregate number of misreports made by buyers in both the No-Noise and Noise treatments of both mechanisms. As seen in panel (a), 34 out of 60 buyers reported truthfully in all 20 periods of sessions using the SR mechanism. Lies are more prevalent in the SPI mechanism, and 47 percent of buyers lie more frequently in the noise treatment than the no-noise treatment while only 25 percent of buyers lie less.

Finally figures B.9 and B.10 shows the proportion of truthful reports for buyers and sellers in the SR mechanism over time. The prediction and confidence intervals that has been overlayed on the data are from simple random effects regressions with a treatment-specific linear time trend. The only time trend in these graphs that is significant is the time series for sellers in the no-noise treatment in the low-signal scenario. Recall that by the construction of the mechanism, buyers are always punished if they enter the arbitration stage of the mechanism while sellers may be rewarded or punished based on the actions of the buyer. If a seller is uncertain about the incentives generated in the mechanism, they may experiment with lies until they are able to observe how the buyers behave. Such experiment is apparent in an ex-post analysis of the data: A seller who lies in the low-signal scenario in period t lies in the next period 76.5 percent of the time if the high-signal scenario occurred and the repercussions of the lie are not observable. By contrast, if a seller lies in the low-signal scenario and the low-signal scenario occurs, sellers lie only 26.1 percent of the time in the next period. Thus, the time trend appears to be based on learning that occurs only in cases where this scenario is played out and the seller observes the second report of the buyer.

	(1)	(2)	(3)
SPI Treatment	0.308 ***	0.178 ***	0.178 ***
	(0.055)	(0.048)	(0.044)
Noise Treatment		0.048	0.048
		(0.030)	(0.043)
SPI x Noise Treatment		0.130 **	0.130 *
		(0.058)	(0.079)
Constant	0.125 ***	0.077 ***	0.770 ***
	(0.032)	(0.024)	(0.017)
R^2	0.118	0.156	0.110
Observations	1200	2400	2400

Dependent varaible is 1 if the buyer lies by announcing low with a high signal and 0 otherwise. Regression (1) is a linear probability model that includes data only from the moise treatments. Regression (2) and (3) are linear probability model that includes data from all four treatments. Regressions (1) and (2) are clustered at the buyer level. Regression (3) uses individual level random effects and is clustered at the sesssion level. *,**,*** denote two-tailed siginificance at the 10%, 5% and 1% respectively.

Table B.1: Buyer misreports in the SR and SPI mechanisms.

Non-Parametric Tests of Buyer Lies	**Unit of Observation**	**p-value**
No-Noise Treatment vs Noise Treatment in SPI Mechanism (Wilcoxen sign-rank test)	Individual Level	0.0030
	Session Level	0.0464
No-Noise Treatment vs Noise Treatment in SR Mechanism (Wilcoxen sign-rank test)	Individual Level	0.1421
	Session Level	0.4630
SR vs SPI Mechanisms in No-Noise Treatents (Mann-Whitney Wilcoxen test)	Individual Level	0.0000
	Session Level	0.0039
SR vs SPI Mechanisms in Noise Treatments (Mann-Whitney Wilcoxen test)	Individual Level	0.0001
	Session Level	0.0039

Non-parametric comparaison of treatments. p-values shown for two-sided version of each test.

Table B.2: Non-parametric tests comparing the proportion of misreports made by buyers.

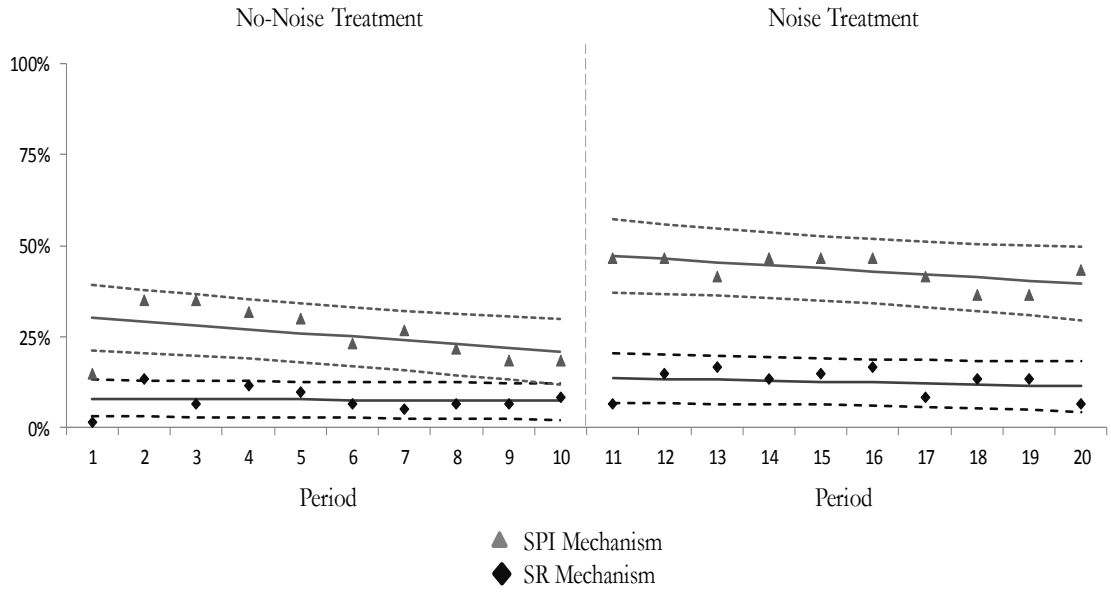
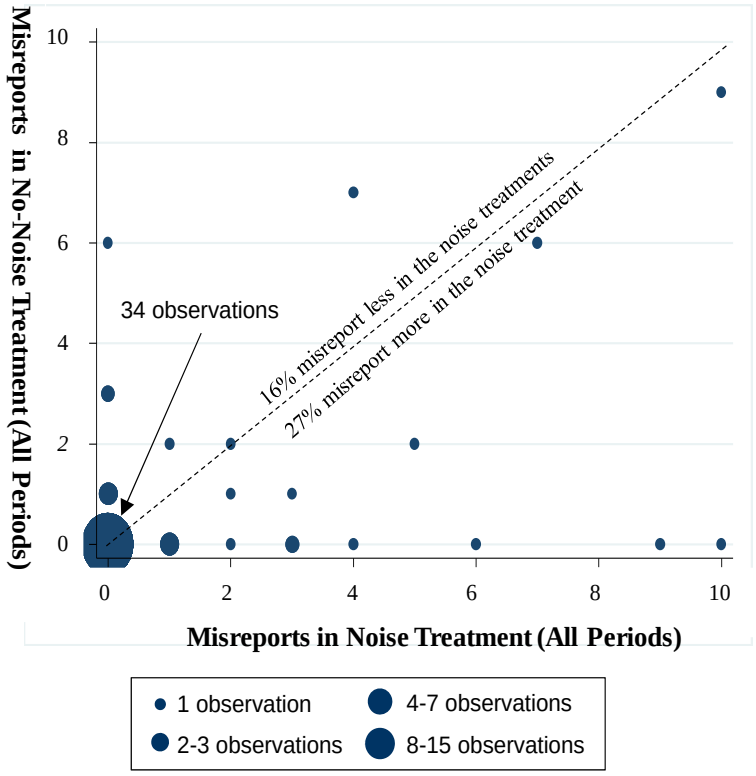


Figure B.7: Buyer misreports over time. 95 percent confidence intervals constructed from random effects regression with a treatment-specific linear time trend.

(a) Aggregate Number of Buyer Misreports in SR Mechanism (N=60)



(b) Aggregate Number of Buyer Misreports in SPI Mechanism (N=60)

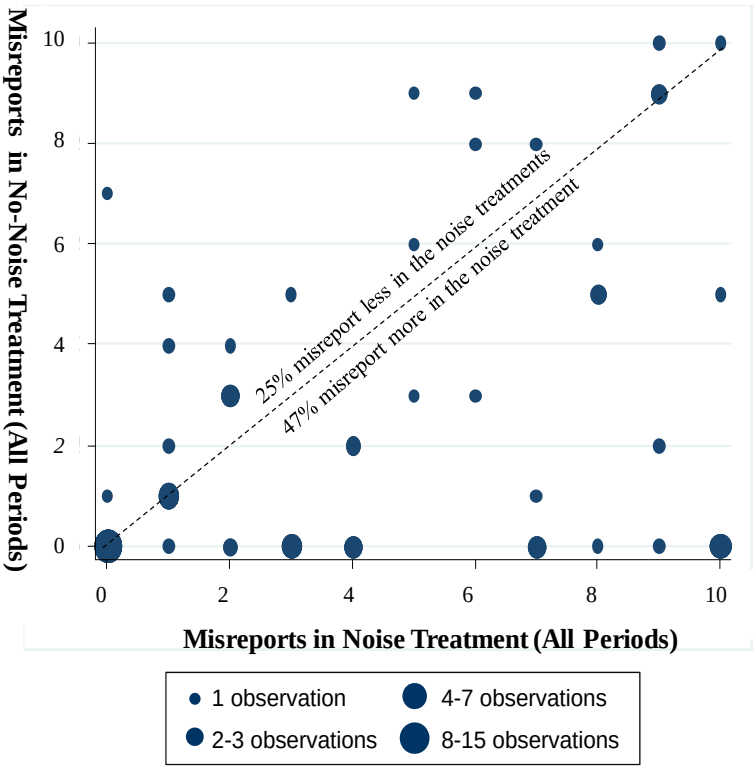
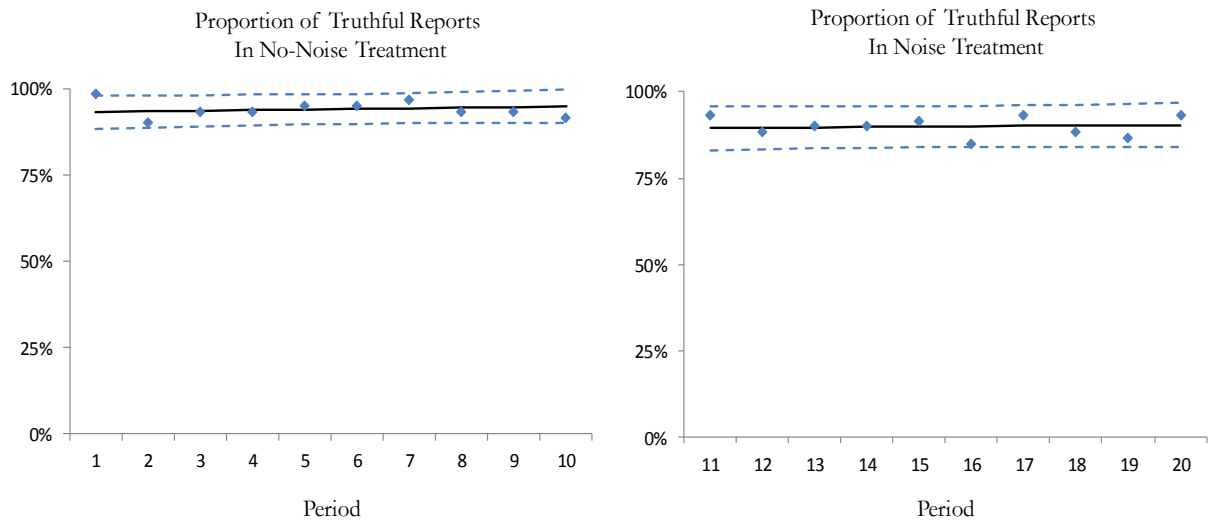


Figure B.8: Aggregate number of misreports made by buyers in both the No-Noise and Noise treatments of the SR and SPI mechanism

(a) Proportion of Truthful Reports by Buyers in High-Signal Scenario of SR Mechanism



(b) Proportion of Truthful Reports by Buyers in Low-Signal Scenario of SR Mechanism

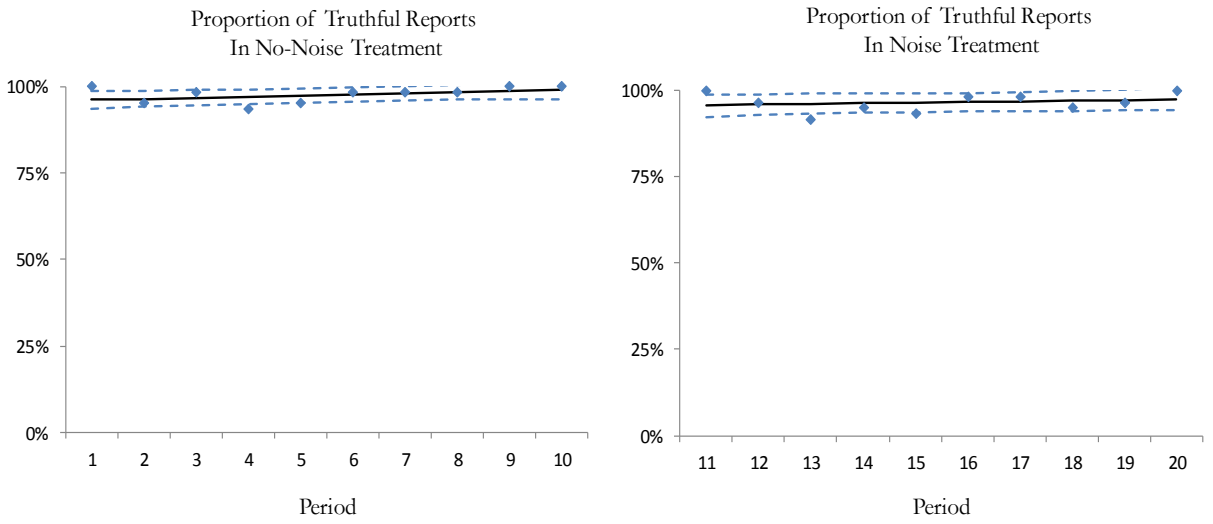
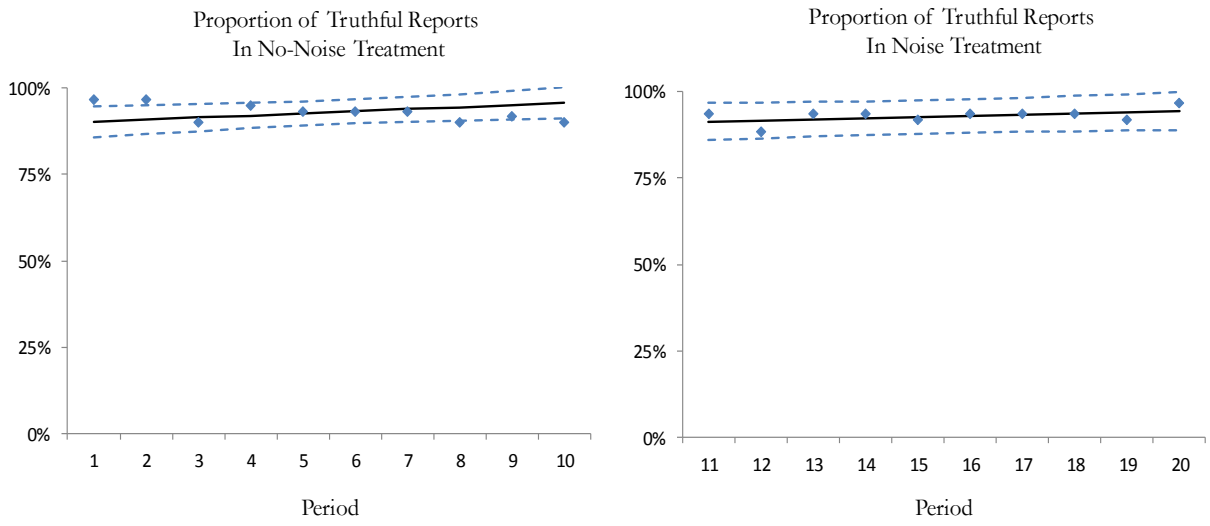


Figure B.9: Evolution of Buyer Misreports in No-Noise and Noise Treatments of Simultaneous Report Mechanism

(a) Proportion of Truthful Reports by Sellers in High-Signal Scenario of SR Mechanism



(b) Proportion of Truthful Reports by Sellers in Low-Signal Scenario of SR Mechanism

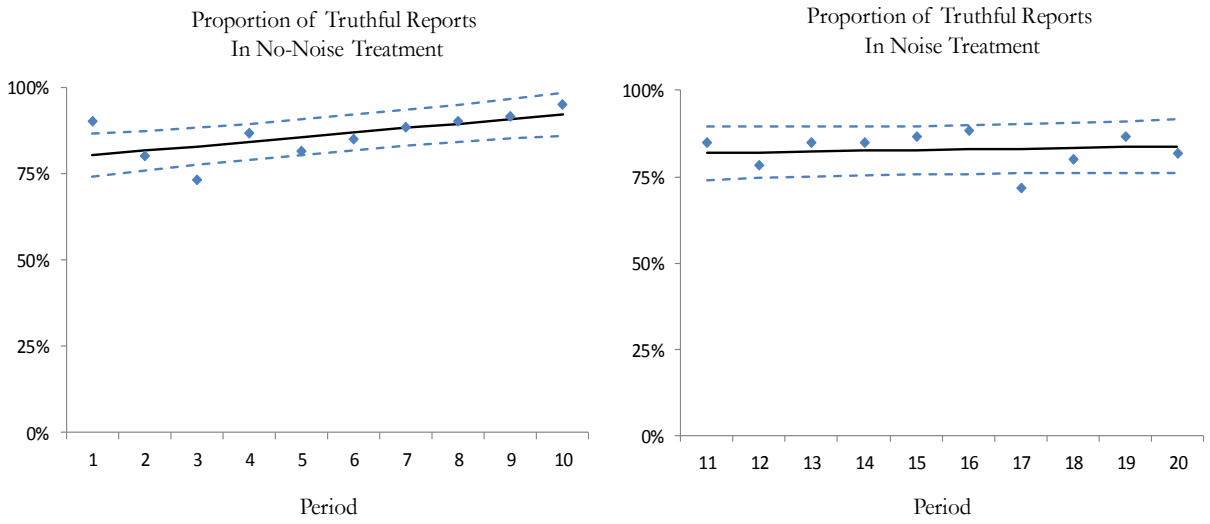


Figure B.10: Evolution of Seller Misreports in No-Noise and Noise Treatments of Simultaneous Report Mechanism

B.3 Pre-Analysis Plan

Both the experimental design and analysis plan were pre-registered at open science prior to the initial experiment. The registration can be found at osf.io/p6ukx.

We pre-registered the design, experimental hypotheses, and analysis plan. All statistics and figures in the pre-analysis plan have been included in the main document or the appendix. Based on the initial analysis, we also included the following in the appendix: (1) a short analysis of seller misreports in the SPI mechanism, (2) session level clustered analysis of the main treatment effects, and (3) time series graphs of all treatments.

B.4 Instructions

A full set of instructions are available with the registration at osf.io/p6ukx. For convenience, we have included a copy of the buyer's instructions in the SR mechanism. The order of events in a session were as follows:

1. After being randomly assigned computers, subjects were given the first set of written instructions and were asked to read them at their own pace and complete the quiz.
2. After all quizzes were checked, the first verbal summary was read allowed.
3. Subjects then played the first ten periods of the experiment.
4. We handed out the written instructions for phase 2 and gave the subjects four minutes to read them before beginning the second verbal summary.
5. The second verbal summary was read aloud.

Instructions for Buyer:

Welcome to the experiment. Throughout this experiment, you will be the potential buyer of a container. Other participants in the experiment have been assigned to the role of sellers. If you read the instructions carefully, depending on your decisions and the decisions of others, you can earn a significant amount of money. You will receive this money privately, in cash, after the experiment.

If you have any questions before or during the experiment, please do not talk out loud. Raise your hand and an experimenter will assist you.

During the experiment we shall not speak of Dollars, but of Experimental Currency Units (ECU). Your entire earnings will be calculated in ECUs. At the end of the experiment we will randomly select two of the periods for payment – one from periods 1-10 and one from periods 11-20. The ECUs you earn in these periods will be converted to Dollars at the rate of

2 ECU= $1,

and will be immediately paid to you in cash. In addition, if you stay to the end of the experiment, we will give you an additional one-off payment of **$35** for your participation.

I. Overview of the experiment

This experiment is broken up into twenty periods. In each period, you will be matched with a seller and have the task of deciding on a price at which to trade a container. If there is disagreement about the price, an arbitrator may be called. The role of the arbitrator is played by the computer whose actions will be described below.

The experiment is divided into two phases of 10 periods. In the first 10 periods, you will be randomly matched with a <u>different</u> seller in each period. Thus, the seller that you interact with in this period will be different from the seller you interact with in future periods. As discussed below, we will randomly select two periods for payments. One of the payment periods will be from the first 10 periods and one will be from the second 10 periods.

Instructions for the second phase will be handed out at the end of period 10. The parameters for the second part of the experiment have been predetermined and thus your choices in the first 10 periods have no bearing on your role or potential choices in the second half of the experiment.

Each period of the experiment will be divided into two parts: The Allocation of Containers and the Report Game.

Part I: The Allocation of Containers
Part II: The Report Game

Please turn the page for details about Part I.

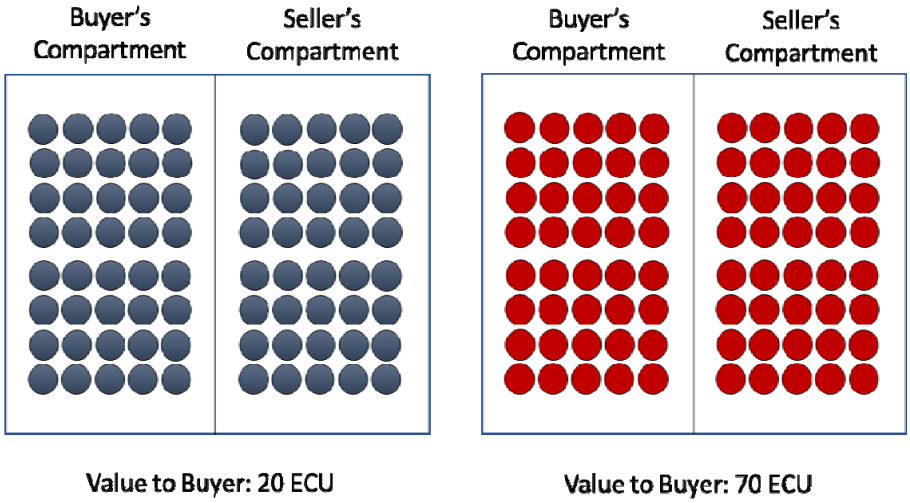
Part I: The Allocation of Containers

In each period, the computer will randomly select one of two possible containers and give it to the seller. One container is worth 70 ECU to you while the other container is worth 20 ECU. Each container is equally likely to be selected.

Each of the two containers has two compartments that are filled with red and blue balls. We will refer to the left compartment as the "buyer's compartment" and the right compartment as the "seller's compartment."

1. The buyer's compartment and the seller's compartment of the container worth 20 ECU is filled with 40 blue balls.

2. The buyer's compartment and the seller's compartment of the container worth 70 ECU is filled with 40 red balls.



Both you and your matched partner will not initially know which container has been given to the seller while trading. However, one of the balls in the buyer's compartment will be randomly drawn and secretly shown to you. One of the balls in the seller's compartment will be drawn from the same container and secretly shown to the seller.

In the first 10 periods, since the container worth 70 ECU is filled with only red balls and the container worth 20 ECU is filled with only blue balls, the color of the ball will perfectly inform you about the container being traded. If you see a red ball, you have a 100% chance to be trading a container worth 70 ECU. If you see a blue ball, you have a 100% chance of trading a container worth 20 ECU.

The arbitrator, which is played by the computer, does not get to draw a ball from a container. Thus, the arbitrator cannot determine which container you are trading and will only take specific actions based on your actions in the next part of the experiment.

Part II: The Report Game

Part II of the experiment is divided into four stages:

Stage 1: The Report Stage:

In the report stage, you will be asked to report the value of the container under two scenarios: the scenario where you observe a red ball and the scenario where you observe a blue ball. For each scenario, you may report a value of 20 or a value of 70. Your reports on the two scenarios may be different or the same.

Stage 2: The Signal Stage:

In the signal stage, you will randomly draw a ball out of the buyer's compartment of the container. After observing the color of the signal, you will make a report which corresponds to your decision in the report stage for that color signal.

Example 1: *In the report stage, you decide to report a value of 70 in the case of receiving the red signal and a value of 20 in the case of receiving a blue signal. In the signal stage you draw a blue ball. Thus, your report will be 20.*

Stage 3: The Arbitration Stage:

In the Arbitration Stage, the seller will be informed of your report. The seller will then draw and observe a ball from the seller's compartment of the container. He or she then has the option to call or not call the arbitrator.

If the Arbitrator Is Called: If the seller chooses to call the arbitrator, you (i) pay an arbitration fee of 40 ECU and (ii) enter the Arbitration Response Stage. (See next page)

If the Arbitrator Is Not Called: If the seller choose to not call the arbitrator, you will trade at a price that is based on your report. The trade prices are as follows:

Your Report	Trade Price
20	10
70	35

Your earnings will be the value of the container minus the trade price. The seller's earnings is equal to the trade price.

Example 2*: You receive the red signal. You report a value of 70. This report results in a trade price of 35. The seller does not call the arbitrator. After opening the container, the value is 70 ECU. Your earnings would be 35 (70-35=35).*

If No Arbitrator is Called:

Your Earnings = Value - Trade Price

The seller's earnings would be 35.

If No Arbitrator is Called:

Seller's Earnings = Trade Price

Example 3: *You receive the red signal. You report a public value of 20. This report results in a public price of 10. The seller does not call the arbitrator. After opening the container, the value is 70 ECU. Your earnings would be 60 (70-10=60). The seller's earnings would be 10.*

Stage 4: The Arbitration Response Stage:

If the seller decides to call the arbitrator, the arbitrator gives you the option to purchase the container from the seller at an alternative price. This price, called the arbitrator's counter offer, is determined by your original report. The following table shows the price offered in arbitration for each of your possible reports:

Your Report	Arbitrator's Counter Offer
20	35
70	85

You have two options, you can **accept** the arbitrator's counter offer and buy the container at the arbitrator's counter offer or you can **reject** the offer and not purchase a container for the period.

Accept: If you **accept** the offer by the arbitrator, trade occurs at a price equal to the counter offer. In addition, the seller receives an arbitration bonus of 40 ECU from the arbitrator. Your earnings would be the difference between your value and the counter offer minus the 40 ECU arbitration fee. The seller's earnings would be the counter offer price plus the 40 ECU arbitration bonus.

Reject: If you **reject** the offer by the arbitrator, no trade will occur. In addition, the seller will also be charged an arbitration fee of 40 ECU.

Example 4: *You have the red signal. You report a value of 70. The seller calls the arbitrator. The arbitrator offers you the chance to buy the container at 85. You accept the arbitrator's offer*

and buy the container. If the true value of the container is 70 ECU, your earnings would be -55 (70 – 85 - 40 = – 55).

If You Accept:

Your Earnings = Value – Arbitrator's Counter Offer – Arbitration Fee

The seller's earnings would be 125 (85 + 40 = 125).

If You Accept:

Seller's Earnings = Arbitrator's Counter Offer + Arbitration Bonus

Example 5: *You have the red signal. You report a value of 70. The seller calls the arbitrator. The arbitrator offers you the chance to buy the container at 85. You reject the arbitrator's offer and do not buy the container. Your earnings would be -40.*

If You Reject:

Your Earnings = – Arbitration Fee

The seller's earnings would be -40.

If You Reject:

Seller's Earnings = – Arbitration Fee

Stage 5: Calculating Earnings

At the end of each period, we will open the container and determine whether the value of the container is 70 or 20. The earnings you receive in a period are based on your earnings from trade and any fines or bonuses that you pay or receive from the arbitrator.

Buyer Earnings = Value – Price – Arbitration Fee

The sellers earning is equal to:

Seller's Earnings = Price– Arbitration Fee + Arbitration Bonus

How will I be paid?

Your payment in this experiment is based on two things: Your participation payment and your earnings from the experiment.

1) **One-off Completion Payment:** If you stay to the end of the experiment, you will be awarded a one-off payment of **$35** for your participation.

2) **Earnings from One Period:** We will randomly select **two** of the twenty periods for payment. One of these periods will be from periods 1-10 and one of these periods will be from periods 11-20. Your payment will be the earnings in these periods plus your One-off payment.

Your Payment = Earnings from two periods + Completion Payment

---

What happens if I lose money in a period selected for payment?

Depending on your actions and the actions of other participants, your earnings in the randomly selected period may be negative. If this is the case, we will subtract this total from the one-off payment. Thus, if you lose money in a period, your earnings may be below $35.

How am I matched with Sellers?

In each of the first 10 periods you will be randomly matched with a different seller. Thus, the seller that you interact with this period will be different from the seller you are matched with in the next period.

Quiz

Quiz: Please answer the following 6 questions (you can write on these instructions). When you have completed the questions, raise your hand and a monitor will come to check your answers. When all participants have completed their instructions, the experiment will begin.

Question 1: You receive the blue signal:

What is the chance that the container is worth 20 ECU?

Question 2: You receive the blue signal. You report a value of 20:

What is the trade price if the seller does not call the arbitrator?

Upon opening the container, the value is 20. If the seller does not call the arbitrator, how much money do you earn?

If the seller does not call the arbitrator, how much money does the seller earn?

Question 3: You receive the blue signal. You report a value of 20: The seller calls in an arbitrator:

What is the arbitrator's counter offer?

If you **accept** the arbitrator's offer, how much money do you earn if the container is worth 20 ECU?

If you **reject** the arbitrator's offer, how much money do you earn?

Please turn the page to continue the quiz

Question 4: You receive the red signal:

What is the chance that the container is worth 20 ECU?

Question 2: You receive the red signal. You report a value of 20:

What is the trade price if the seller does not call the arbitrator?

Upon opening the container, the value is 70. If the seller does not call the arbitrator, how much money do you earn?

If the seller does not call in an arbitrator, how much money does the seller earn?

Question 3: You receive the red signal. You report a value of 20: The seller calls in an arbitrator:

What is the arbitrator's counter offer?

If you **accept** the arbitrator's offer, how much money do you earn if the container is worth 70 ECU?

If you **reject** the arbitrator's offer, how much money do you earn?

If you **accept** the arbitrator's offer, how much money does the seller earn?

Tables

Trade Prices If Arbitrator is Not Called

Buyer's Report	Trade Price
20	10
70	35

Counter Offer Prices if Arbitrator is Called

Buyer's Report	Arbitrator's Counter Offer
20	35
70	85

Arbitration Fees and Bonuses

- If Buyer Enters Arbitration: Buyer Pays Arbitration Fee of 40
- If Buyer Accepts the Counter Offer: Seller Receives Arbitration Bonus of 40
- If Buyer Rejects the Counter Offer: Seller Pays Arbitration Fee of 40

Earnings Calculations:

If Arbitrator is Not Called:

Buyer's Earnings = Value - Trade Price

Seller's Earnings = Trade Price

If Arbitrator is Called and Counter Offer is Accepted:

Buyer's Earnings = Value – Arbitrator's Counter Offer– Arbitration Fee

Seller's Earnings = Arbitrator's Counter Offer + Arbitration Bonus

If Arbitrator is Called and Counter Offer is Rejected:

Buyer's Earnings = – Arbitration Fee

Seller's Earnings = – Arbitration Fee

Verbal summary SR (FI first): In this experiment you will take on the roles of buyers and sellers who have the task of deciding on a price at which to trade a container. If there is disagreement about the price, an arbitrator may be called. The role of the arbitrator is played by the computer who takes specific actions based on the reports of the buyer and seller.

The experiment will be broken up into twenty periods. In each of the first ten periods you will be matched with a different person from the other side of the market. Thus the partner you are matched with in this period will be different than the one you are matched with in the next period.

Each of the first 10 periods is divided into two parts:

In part 1, the seller will be randomly given a container which is either worth 20 or 70 ECU to the buyer. Each of the two potential containers has a buyer's compartment and a seller's compartment. Each compartment is filled with 40 balls. In the first 10 periods:

1) Each compartment of the container worth 70 ECU is filled with 40 red balls and 0 blue balls
2) Each compartment of the container worth 20 ECU is filled with 40 blue balls and 0 red balls

In each period, a random ball from the buyer's compartment of the assigned container will be drawn and shown to the buyer. A random ball from the seller's compartment of the assigned container will be drawn and shown to the seller.

If you see a red ball, you have a 100% chance to be trading a container worth 70 ECU. If you see a blue ball, you have a 100% chance of trading a container worth 20 ECU. Note that since each container has only one ball color both you and your matched partner will observe the same color ball and thus know perfectly which container you are trading.

After observing a ball, called a signal in the experiment, you will then continue to part two. Part two of the experiment is divided into four stages: The report stage, the signal stage, the verification stage, and the arbitration stage.

In the report stage, both you and your matched partner will be asked to report the value of the container under the scenario of receiving the red ball and the scenario of receiving the blue ball. In the signal stage, we will draw a ball from the buyer's compartment of the container and the buyer will make a report which corresponds to his or her decision in the report stage for this color ball. We will draw a ball from the seller's compartment of the container and the seller will make a report which corresponds to his or her decision in the report stage for this color ball.

In the verification stage, we will compare the two reports. If they match, we will use them to set the trade price. If they do not match, the buyer will be charged an arbitration fee and continue to the arbitration stage. In the arbitration stage, the buyer will be asked to make a second report. We will use this second report and the roll of a six sided dice to determine the trade price. We will also compare the buyer's second report to the seller's report. The seller will receive an arbitration bonus if they match and must pay an arbitration fee if they differ.

After 10 periods, we will hand out new instructions for phase 2. Phase 2 will be identical to Phase 1 except that the composition of red and blue balls in the container may change. The parameters for the second part of the experiment have been predetermined and your choices in the first 10 periods have no bearing on your role or potential choices in phase 2.

If anyone has any questions, just raise your hand… OK, we will begin.

Periods 11-20: The Allocation of Containers

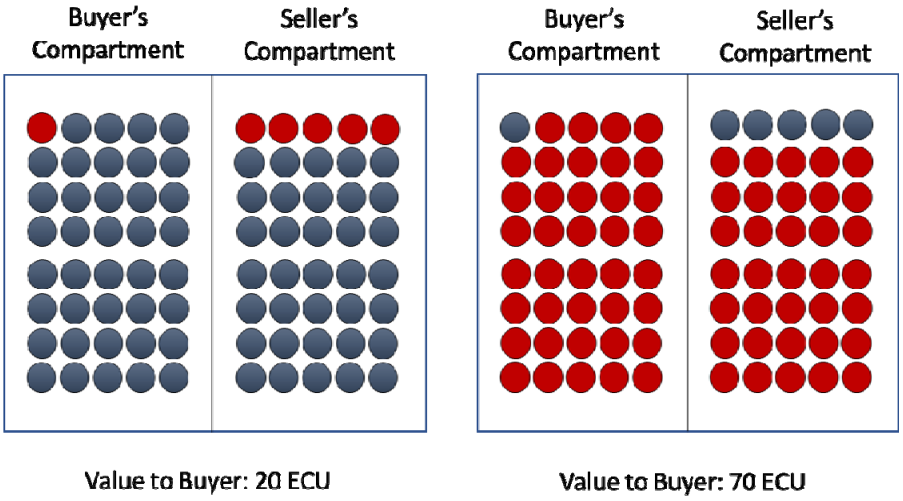
The second part of this experiment is identical to the first except that some of the red and blue balls have been switched.

As before, in each period, the computer will randomly select one of two possible containers and give it to the seller. One container is worth 70 ECU to you while the other container is worth 20 ECU. Each container is equally likely to be selected.

Each of the two containers is filled with red and blue balls. Now, however:

1. The buyer's compartment of the container worth 20 ECU is filled with 39 blue balls and 1 red ball. The seller's compartment of the container worth 20 ECU is filled with 35 blue balls and 5 red ball.

2. The buyer's compartment of the container worth 70 ECU is filled with 39 red balls and 1 blue ball. The seller's compartment of the container worth 70 ECU is filled with 35 red balls and 5 blue balls.



Both you and your matched partner will not initially know which container has been given to the seller while trading. However, one of the balls in the buyer's compartment will be randomly drawn and secretly shown to you. Likewise, one of the balls in the seller's compartment will be drawn from the same container and secretly shown to the seller.

Please Turn the Page

Unlike the first 10 periods, seeing a red or blue ball does not give you perfect information about the container being traded.

If you are a buyer and see a red ball, you have a 97.5% chance of trading a container worth 70 ECU. If you see a blue ball, you have a 2.5% chance of trading a container worth 70 ECU.

If you are a seller and see a red ball, you have a 87.5% chance of trading a container worth 70 ECU. If you see a blue ball, you have a 12.5% chance of trading a container worth 70 ECU.

The containers that you are trading contain both red and blue balls and thus there may be cases where the colour of the ball you observe is different than the colour of the ball that your matched partner observes. To help you make decisions, we have calculated the probability that different events occur.

Signal Combinations

If the container is worth 70:

- Both the buyer and the seller will receive a red signal 85.3125% of the time
- The buyer will receive a red signal and the seller will receive a blue signal 12.1875% of the time.
- The buyer will receive a blue signal and the seller will receive a red signal 2.1875% of the time
- Both the buyer and the seller will both receive a blue signal 0.3125% of the time.

If the container is worth 20:

- Both the buyer and the seller will receive a blue signal 85.3125% of the time
- The buyer will receive a blue signal and the seller will receive a red signal 12.1875% of the time.
- The buyer will receive a red signal and the seller will receive a blue signal 2.1875% of the time
- You will both receive the red signal 0. 3125% of the time.

Verbal Summary (Phase 2): The second part of the experiment is identical to the first part except that some of the blue and red balls have been swapped between the two containers.

In the container worth 70 ECU:

- The buyer's compartment is filled with 39 red balls and 1 blue ball
- The seller's compartment is filled with 35 red balls and 5 blue balls.

In the container worth 20 ECU:

- The buyer's compartment is filled with 39 blue balls and 1 red ball
- The seller's compartment is filled with 35 blue balls and 5 red balls.

Unlike the first 10 periods, seeing a red or blue ball does not give you perfect information about the container being traded.

If you are a buyer and see a red ball, you have a 97.5% chance of trading a container worth 70 ECU. If you see a blue ball, you have a 2.5% chance of trading a container worth 70 ECU.

If you are a seller and see a red ball, you have a 87.5% chance of trading a container worth 70 ECU. If you see a blue ball, you have a 12.5% chance of trading a container worth 70 ECU.

Note also that there may be cases where the colour of the ball you observe is different than the colour of the ball that your matched partner observes. We have included the probability of each of these events on the back of your new instructions.

As before, you will be matched with a different person on the other side of the market in each period. Thus the person you are matched with in this period will be different than the person you are matched with in all future periods.

If anyone has any questions, just raise your hand… OK, we will begin.