Getting Dynamic Implementation to Work*

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

August 17, 2017

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. Unlike alternative three-stage mechanisms, our mechanism is predicted to be robust to small amounts of incomplete information about the state of nature and to moderate levels of reciprocity. In experiments, we find that the mechanisms can induce efficient investment in a two-sided hold-up problem with ex-ante investment and performs better than both a three-stage mechanism introduced by Moore and Repullo (1988) and a one-stage mechanism introduced by Kartik, Tercieux and Holden (2014). We also show that the mechanism can be made renegotiation proof if the agents are strictly risk averse and highlight the robustness of the mechanism to a wide variety of reasoning processes and behavioral assumptions.

Keywords: Implementation Theory, Incomplete Contracts, Experiments

JEL Codes: D71, D86, C92

---

*We thank Yeon-Koo Che, Navin Kartik, Patrick Rey, and Steve Williams for helpful comments and discussions, and seminar participants at Monash, UTS, the 10th The Annual Organizational Economics Workshop at Sydney, 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, 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) (AFHKT) and Aghion et al. (2017) show that extensive-form mechanisms are not robust to small deviations from common knowledge about the state of nature[1], while Fehr et al. (2014) (FPW) 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, updating occurs along the dimension of uncertainty leading to equilibria that may be far away from the intended equilibrium even when uncertainty is small.

The purpose of this paper is to explore dynamic implementation both theoretically and experimentally when imposing less stringent assumptions on how beliefs evolve. Following Ben-Porath (1997) and Dekel and Siniscalchi (2013), we use the notion of *initial rational-*

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

izability 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 only at the beginning of the game and makes no assumption about how beliefs evolve after zero probability events. 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.

We begin our paper by developing a two-stage mechanism that implements the first-best under initial rationalizability in the two-sided hold-up problem with pure cooperative investments.[2] Borrowing from Che and Hausch (1999), we consider an environment where a buyer and seller are interested in trading a relationship-specific "widget". Prior to production, each party may make a costly investment to increase the joint surplus from trade. Investments by the buyer reduce the production cost of the widget for the seller, while investments by the seller increase the value of the widget for the buyer. Investments, costs, and values are common knowledge among the trading parties, but they are not verifiable by a third-party such as a court. This implies that the two parties cannot write an enforceable contract that conditions payments on investment, value, or cost and hence, the ex ante investments are prone to holdup and will be below the first-best levels.

While investment is not verifiable by a third-party, reports are. Thus, the two parties can, in principle, write a contract that specifies trade prices as a function of reports made by the two parties. If both parties always tell the truth in equilibrium, then their reports can be used to set prices that promote efficient investment.

The *Simultaneous Report* mechanism (SR) that we develop combines a coordination game with arbitration clauses that are triggered in the event of disagreement. In the first stage of the mechanism both parties simultaneously report the cost and value. If both the value and the cost reports of the two parties coincide, trade occurs at a price that is based on the mutually reported value and cost information. If, however, there is a disagreement, one of the parties is immediately fined and enters an arbitration stage where they are asked

[2]As discussed in Che and Hausch (1999), the pure cooperative case is one where investment improves the outcome for the other party but offers no (or negative) direct benefits to the investor. See Chung (1991), Aghion et al. (1994), and Nöldeke and Schmidt (1995) for option contracts that can solve the hold-up problem under the alternative selfish investment case where investment yields direct benefits.

to make a second report.

Similar to the BDM mechanism (Becker et al., 1964) that is commonly used in experimental work to elicit beliefs, we construct a set of pre-specified lotteries in the arbitration stage such that it is a dominant strategy for an expected-utility maximizer to make a truthful second report. We use the second report, along with lotteries, to determine whether trade occurs and at what price. We also use the second report as a part of a test to determine whether the counter-party was lying in the previous stage. We do this by comparing the second report of the party in arbitration with the initial report of the party not in arbitration. We reward the counter-party with a bonus if the two reports match and punish them with a fine if they differ.

Implementation requires that we induce truth-telling in the first stage of the game and avoid the arbitration stages, which result in fines and inefficient trades. We show that the SR mechanism can accomplish this under initial rationalizability and requires only the deletion of never sequential best replies followed by two rounds of deletion of strictly dominated strategies.

Having illustrated the theoretical potential of the SR mechanism in the two-sided hold-up context, we then use experiments to study actual behavior in the mechanism. In our experiments, subjects first choose investment levels and then are exogenously entered into the SR mechanism. Thus, for the mechanism to be deemed a success it must not only produce truthful reports but also induce first-best investment levels for both parties.[3]

We find experimental evidence that is largely consistent with the behavior predicted by our theory. In the first 10 periods of the experiment where the mechanism is exogenously imposed, buyers make truthful first-stage value and cost reports in 92.6 percent of cases. Likewise, sellers make truthful first-stage value and cost reports in 91.7 percent of cases. Buyers choose the optimal level of investment in 89.6 percent of cases while sellers choose the optimal level of investment in 83.3 percent of cases. In aggregate, 87.1 percent of dyads improve their performance relative to the theoretical no-mechanism benchmark and 72.9 percent of dyads exhibit first-best investments and truth-telling behavior.[4]

[3]We will discuss the debate about the foundations of the incomplete contracting literature pioneered by Grossman and Hart (1986) and Hart and Moore (1990) below. But at this point it is worth noting that the "mechanism critique" leveled by Maskin and Tirole (1999) involves using an appropriately crafted mechanism to induce first-best ex ante investments. If such a mechanism exists, it would render asset ownership irrelevant.

[4]The only stage that does not confirm strongly to the theoretical prediction is the second report stage

A criticism often leveled against implementation mechanisms is that they are not observed in practice. This suggests that parties may be reluctant to use these mechanisms if given the choice. In a second block of 10 periods we add an opt-in stage where both parties have the option to eliminate the SR mechanism and trade at a fixed price. We find that both buyers and sellers are willing to use the mechanism and that opt-in rates are above 75 percent for both parties. Groups that opt into the mechanism behave very closely to theory with 90.5 percent of dyads achieving the first best.

In order to benchmark the performance of the mechanism, we also compare efficiency of the mechanism to a baseline treatment where the trade price is fixed and two other mechanisms that are predicted to induce the first-best under alternative equilibrium concepts: a three-stage mechanism based on Moore and Repullo (1988) and a one-stage mechanism proposed by Kartik et al. (2014). We find that our SR mechanism is 19.8 percent more efficient than the fixed price mechanism, 35 percent more efficient than the mechanism based on Moore and Repullo (1988) and 62 percent more efficient than the mechanism proposed by Kartik et al. (2014). However, relative to its theoretical benchmark, there is some efficiency losses due to fines.

Having developed a two-stage mechanism that has both good theoretical and experimental properties for the two-sided hold up problem, we next ask whether these mechanisms can be used in a more general class of problems. Part two of our paper provides very permissive implementation results when using initial rationalizability as a solution concept. We show that two-stage mechanisms similar to our SR mechanism can be constructed with a unique truth-telling sequential equilibrium in pure strategies that is robust to any "private-value perturbation." 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 agents have preferences that can be represented by a von Neumann-Morgenstern expected utility func-

where a reasonably large portion of subjects match the false report of their partner rather than making a truthful report. Despite this deviation, truth-telling continues to be a best response to the empirical distribution of second-stage reports. We discuss below how truth-telling remains a best-response to heterogeneity in preferences.

tion.[5] Third, we focus on private-value environments. That is, each player's utility depends only upon her own payoff type as well as the lottery chosen and her monetary payment.[6]

Our notion of robustness, which we call "private-value robustness", involves the proposed (finite) mechanism implementing the desired social choice function both under complete information, and "almost" implementing it in nearby environments where there is a small amount of incomplete information about the state of nature. That is, any sequence of sequential equilibria under incomplete information converges to the unique complete information equilibrium as the amount of incomplete information goes to zero. We prove that, for two or more players, any social choice function is robustly implementable under private-value perturbations by a finite mechanism, and that the unique sequential equilibrium of the mechanism is truth-telling in pure strategies.

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), Aghion et al. (2012), and Kartik et al. (2014)).

In delineating the class of environments where implementation mechanisms can achieve first-best investment levels, we also contribute to the debate on the foundations of incomplete contracts. The Maskin-Tirole critique suggested that asset ownership–which plays a central role in Grossman-Hart-Moore Property Rights Theory–may be irrelevant. AFHKT demonstrates that the mechanism proposed in Maskin-Tirole is not robust to small amounts of incomplete information, but in some sense their result is "too strong" in that it shows that *all* sequential mechanisms are non-robust. Yet we observe certain simple mechanisms in practice: e.g. so-called "cut and choose" mechanisms. This raises squarely the question of what is the class of environments where mechanisms are applicable, and in what environments do mechanisms break down. In the former we would expect contracts to play a more prominent role in governing economic activity, and in the latter we would expect asset ownership and residual control rights to be relatively more important.

In designing our SR mechanism, we relied on a number of findings from the experimental literature on implementation. Sefton and Yavas (1996) and Katok et al. (2002)

[5] We discuss how our mechanism can accommodate more general preferences over lotteries in section 5.5.2.
[6] This is without loss of generality in the complete information case.

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.

An extensive experimental literature studies the efficiency of implementation mechanisms in public goods provision problems[7], Solomon's dilemma problems (Ponti et al., 2003; Giannatale and Elbittar, 2010), and the one-sided hold-up problem. In the hold-up context, 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. The authors argue that contracts may act as reference points which in turn makes renegotiation costly. We explore mechanisms that can also solve the hold-up problem in a two-sided settings with cooperative investments. As discussed in Section 5.5.3, our mechanisms can be made robust to renegotiation if the parties are strictly risk averse.

We also build on Fehr et al. (2014) and Aghion et al. (2017) who document how deviations in assumption of common knowledge of rationality, payoffs and preferences cause three-stage mechanisms proposed in the literature to break down. Responding to Aghion et al. (2017), the mechanism we propose is robust to private-value perturbations. Responding to Fehr et al. (2014), truth-telling remains a retaliation-proof equilibrium in our mechanism and is the unique equilibrium if subjects perceive truthful first-stage announcements as a neutral or kind act. Truth-telling is also a best response when subjects believe their match partners best respond with noise and in cognitive hierarchy models where truthful announcements are perceived as focal for level-0 types. Finally, truth-telling is the unique consistent self-confirming equilibrium. Our mechanism is thus robust to multiple reasoning processes and behavioral assumptions. This desiderata for mechanism selection is suggested in Masuda et al. (2014) who study implementation mechanisms for public goods provision.

[7]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.

The remainder of the paper proceeds as follows. Section 2 introduces the SR mechanism and highlights its properties in the setting we subsequently use in our experiments. Section 3 outlines the experimental setup, while Section 4 reports the results of the experiments. Section 5 contains our theoretical analysis and proves our main implementation result. Section 6 contains some brief concluding remarks.

2 An illustration in the bilateral trade setup

In this section, we illustrate how the **Simultaneous Report (SR)** mechanism can be used to induce first-best investment in the bilateral trade setup that we use in our main treatment.

Following the work of Che and Hausch (1999), our experiments consider a two-sided hold-up environment with pure cooperative investments. In this environment, a seller can produce a non-divisible object for a buyer. The object has no outside option value to the seller, but the seller's production costs can be saved if the object is not produced.

Prior to bargaining over the production and exchange of the object, both the buyer and seller have the opportunity to make relationship-specific investments. The seller can choose an investment level $e_S \in \{0, 25, 75\}$ to increase the value of the final good for the buyer. Investment is privately costly to the seller but increases the value of the good to the buyer, which is denoted by $v(e_S)$. We assume that $v(0) = 200$, $v(25) = 250$, and $v(75) = 320$. Based on these values and investment costs, a seller investment of 75 is efficient.

Similarly, the buyer can choose an investment level $e_B \in \{0, 25, 75\}$ to reduce the production cost for the seller. Denoting the seller's production cost as $c(e_B)$, we assume that $c(0) = 130$, $c(25) = 80$, and $c(75) = 10$. A buyer investment of 75 is efficient.

We assume that both the buyer's value and the seller's cost are *observable* to both parties but *non-verifiable* by a court. These assumptions imply that while the true cost and true value is common knowledge, it is impossible to write an enforceable contract contingent on c and v. Without a contract, bargaining over the trade price, p, occurs after investments are made resulting in the potential of hold-up of both the buyer and the seller. To highlight this holdup problem, suppose first that the buyer has all the bargaining power and makes a take-it-or-leave-it offer to the seller, resulting in a trade price of $p = c(e_B)$. Since the trade price does not depend on the seller's investment choice, the seller has no incentive to choose high investment even though doing so would be socially efficient. Likewise, suppose that the

seller makes a take-it-or-leave-it offer to the buyer, resulting in a trade price of $p = v(e_S)$. As this trade price does not depend on the buyer's investment choice, the buyer has no incentive to choose high investment. Consequently, both parties would prefer a trade price that is sensitive to both v and c.

We now show that it is possible to construct a contract that is based solely on publicly observable reports that can generate the price schedule given in Table 1 using our SR mechanism. This price schedule yields first-best investment for both parties and the mechanism is based on reports that can be verified by the court.

Table 1: Price Schedule

$p(v,c)$	$c = 130$	$c = 80$	$c = 10$
$v = 200$	165	115	45
$v = 250$	215	165	95
$v = 320$	285	235	165

2.1 The SR mechanism

The SR mechanism is comprised of up to two stages: a report stage and an arbitration stage. In the report stage the buyer and the seller are asked to simultaneously report a value-cost pair. Denote the buyer's first-stage reports by $(\hat{v}^B, \hat{c}^B)$ and the seller's first-stage reports by $(\hat{v}^S, \hat{c}^S)$. We distinguish two situations:

- If both the buyer and the seller report the same pair $(\hat{v}, \hat{c})$, then they trade the object according to a pre-specified **price schedule** $p(\hat{v}, \hat{c})$, which is identical to the one in Table 1;

- Otherwise, one of the following three cases applies:

  - If there is a discrepancy *only* in the reported values, i.e., $\hat{v}^B \neq \hat{v}^S$ and $\hat{c}^B = \hat{c}^S$, the buyer will be fined an **arbitration fee** F by the arbitrator and the buyer will enter the arbitration stage. The seller is considered the outside party.

  - If there is a discrepancy *only* in the reported costs, i.e., $\hat{v}^B = \hat{v}^S$ and $\hat{c}^B \neq \hat{c}^S$, the seller will be fined an **arbitration fee** F by the arbitrator. Then, the seller will enter the arbitration stage. The buyer is considered the outside party.

– If there are discrepancies in both the reported values and the reported costs, , i.e., $\hat{v}^B \neq \hat{v}^S$ and $\hat{c}^B \neq \hat{c}^S$, both the buyer and the seller will be fined **arbitration fee** F by the arbitrator. Then, each player will enter the arbitration stage with 50% chance.

- If the buyer enters the arbitration stage, the buyer will be asked to make a second report on his own value, $\tilde{v}$. Based on the second-stage report $\tilde{v}$, a **dictator lottery** $l(\tilde{v})$ will be implemented. In addition, the seller will get an **incentive transfer** $T_S\left(\hat{v}^S, \tilde{v}\right)$ based on the seller's first-stage value report and the buyer's second-stage report.

- If the seller enters the arbitration stage, the seller will be asked to make a second report on his own cost, $\tilde{c}$. Based on the second-stage report $\tilde{c}$, a **dictator lottery** $l(\tilde{c})$ will be implemented. In addition, the buyer will get an **incentive transfer** $T_B\left(\hat{c}^B, \tilde{c}\right)$ based on the buyer's first-stage cost report and the seller's second-stage report.

The dictator lotteries, incentive transfers, and arbitration fee are constructed so that if the buyer and seller are sequentially rational and have mutual knowledge of sequential rationality, they will make their first-stage reports truthfully, i.e., $\hat{v}^B = \hat{v}^S = v$ and $\hat{c}^B = \hat{c}^S = c$ and the arbitration stage will never occur. To achieve the goal, the mechanism must satisfy the following three conditions:

1. **Arbitration Stage Truth-Telling Condition.** Whenever the buyer or seller enters the arbitration stage, he/she will report the truth, i.e., $\tilde{v} = v$ and $\tilde{c} = c$.

2. **Inter-stage Coordination Condition**. When the seller reports costs truthfully in the second stage the buyer will report costs truthfully in the first stage, i.e., $\hat{c}^B = \tilde{c} = c$. When the buyer reports value truthfully in the second stage the seller reports value truthfully in the first stage, i.e., $\hat{v}^S = \tilde{v} = v$.

3. **Within-stage Coordination Condition.** When the seller reports value truthfully in the first stage, the buyer will report value truthfully in the first stage, i.e., $\hat{v}^B = \hat{v}^S = v$. When the buyer reports cost truthfully in the first stage, the seller reports cost truthfully in the first stage, i.e., $\hat{c}^S = \hat{c}^B = c$.

To achieve the three conditions, we employ the solution concept called initial rationalizability (see Section 5 for a formal definition). Under initial rationalizability, strategies which are never sequential best responses to any belief are removed iteratively. Moreover, along with every round of deletion, only beliefs at the beginning of the game are restricted. In the SR mechanism, we first deletes all strategies which misreports a player's own types in the arbitration stage. This ensures the Arbitration Stage Truth-Telling Condition. In the second round, we ensure the Inter-Stage Coordination Condition by deleting all strategies which misreport the other player's type at the first stage. Finally, the third round deletes all strategies which misreport a player's own type at the first stage. This ensures the Within-Stage Coordination Condition. In other words, implementation with truth-telling is achieved in three rounds of deletion.

2.2 Implementation

As discussed in detail in section 5, it is possible to satisfy all three conditions of the model for a price schedule $p(\hat{v}, \hat{c})$ that is monotonically increasing both in $\hat{v}$ and $\hat{c}$ by carefully choosing the dictator lotteries ($l(\cdot)$), the incentive transfers ($T_S(\hat{v}^S, \tilde{v})$ and $T_B(\hat{c}^B, \tilde{c})$), and the arbitration fees (F). Here, we show how this is done using the specific parameter values from the experiment.

2.2.1 Arbitration Stage Truth-Telling Condition

Table 2 shows the dictator lotteries that are used in our experiment which are designed explicitly to ensure that the Arbitration Stage Truth-Telling Condition holds.

As seen in panel (a), a buyer who enters into arbitration can make a second report that corresponds to one of the potential values of the object. The arbitrator takes a pre-specified action based on the second report of the buyer and the roll of a fair six-sided die. Likewise, a seller who enters into arbitration can make a second report that corresponds to one of the potential costs. The arbitrator again takes an action based on this report and the roll of a fair die.

Each potential buyer type has strict preferences over the potential lottery outcomes:

- If the buyer's value is 200, then the buyer's preference order is "No Trade" $\succ$ "Trade at 205" $\succ$ "Trade at 255";

- If the buyer's value is 250, then the buyer's preference order is "Trade at 205" $\succ$ "No Trade" $\succ$ "Trade at 255";
- If the buyer's value is 320, then the buyer's preference order is "Trade at 205" $\succ$ "Trade at 255" $\succ$ "No Trade."

Based on these preferences and the available lotteries, the buyer has pecuniary incentives to always make a truthful second report to receive his preferred lottery. A similar logic implies that the seller will truthfully report his cost.

Table 2: Trade Prices in Buyer and Seller Arbitration Stages

Panel (a): Buyer Enters into Arbitration

Buyer's Secondary Report	Outcome if Dice Roll is a 1-3	Outcome if Dice Roll is a 4-6
200	No Trade	No Trade
250	Trade at 205	No Trade
320	Trade at 205	Trade at 255

Panel (b): Seller Enters into Arbitration:

Seller's Secondary Report	Outcome if Dice Roll is a 1-3	Outcome if Dice Roll is a 4-6
130	No Trade	No Trade
80	Trade at 125	No Trade
10	Trade at 125	Trade at 75

2.2.2 Inter-stage Coordination Condition

To ensure the Inter-stage Coordination Condition, we set

$$T_S(\hat{v}^S, \tilde{v}) = \begin{cases} 300, & \text{if } \hat{v}^S = \tilde{v}; \\ -300, & \text{if } \hat{v}^S \neq \tilde{v}, \end{cases}$$

$$T_B(\hat{c}^B, \tilde{c}) = \begin{cases} 300, & \text{if } \hat{c}^B = \tilde{c}; \\ -300, & \text{if } \hat{c}^B \neq \tilde{c}, \end{cases}$$

and $F = 300$.

Recall that under initial rationalizability, we start by allowing buyers and sellers to have arbitrary initial beliefs about the strategy profile of the other party and then iteratively deletes strategies that are not sequential best replies to at least one potential set of initial

beliefs. By the arbitration truth-telling condition, lies in the second stage are dominated by announcing truthfully. This implies that all strategies that satisfy initial rationalizability will have truthful reports in the second stage. As a consequence, for the inter-stage coordination condition to be satisfied for the buyer, the buyer must be willing to make a truthful cost report for arbitrary beliefs about the cost and value reports of the seller in the first stage knowing that all reports in the arbitration stage will be truthful. A sufficient test for this condition is that the expected value for a truthful cost report exceeds the largest possible expected value for lying given the set of beliefs that would maximize the value of lying.

Table 3 shows this comparison for the state where $v = 320$ and $c = 130$. The logic used to test the inter-stage coordination condition in all other states and for the seller's value report is identical.

As can be seen from the table, we fix the reports of the seller and the buyer's value report and compare the value of truthtelling and lying on a case-by-case basis. In the first row, we look at the case where the buyer and seller's value reports coincide but where the seller misreports costs. The largest possible value for lying in this state would occur if the buyer believes that the sellers reports are $(200, 10)$. By matching these reports, the trade price would be 45 and the buyer's value is $320 - 45 = 275$. By contrast, reporting truthfully will result in the seller entering arbitration. In this case, the seller's arbitration report will result in no trade, but the buyer will receive 300 since his first-stage report matches the arbitration stage report of the seller.

In the second case, we assume that the seller has reported truthfully in the first stage and that the value reports match. For any lie by the buyer, the seller will enter into arbitration and make a truthful cost report in the arbitration stage. This will result in no-trade and a fine of -300 for the buyer. If instead the buyer tells the truth, he will trade at a price equal to $p(\hat{v}^B, c) = -35 + \hat{v}^B$ yielding an expected value of $320 + 35 - \hat{v}^B \geq 0$.

The third and fourth cases in the table represent cases where the buyer and seller disagree on the value reports and where even if the cost reports match, the buyer will enter into arbitration. In the third case, the lie that generates the highest value is matching the seller's lie, entering into arbitration, and reporting truthfully. In this case, trade always occurs at an expected price of $\frac{1}{2} * 205 + \frac{1}{2} * 255 = 230$ yielding an expected value of $320 - 230 - 300 = -210$. If the buyer reports the true cost, the seller enters arbitration with probability $1/2$ yielding a value of $0 - 300 + 300 = 0$ and the buyer enters arbitration with probability

1/2 yielding an expected value of −210. The expected value of truth-telling is thus −105. In the forth case, arbitration always occurs after a lie, but the seller entering arbitration is particularly bad for the buyer who is fined once for the two reports not matching and again for his cost report not matching the sellers arbitration report. Truthfully reporting the cost causes the buyer to enter into arbitration but avoids the double fine.

As can be seen by comparing across rows, the most difficult state to ensure truth-telling is the one in which the buyer believes he can coordinate with the seller on a set of reports that minimizes prices. To induce truthful reporting even in this case, we need to set a fine that is larger than the maximal absolute payoff difference between any two transactions in any two states excluding incentive transfers and arbitration fees for either the buyer or the seller. We will call this difference D and use it in the next condition and in the generalization of our SP mechanism developed in section 5. As can be seen in the table, $D = 275$ for the analyzed state. It is less than or equal to 275 in all other states.

Table 3: Comparison of Buyer's Expected Payoff for Lies and Truthful Announcements of Buyer when $v = 320$ and $c = 130$

Case	Highest Possible Expected Value of Lying	Expected Value of Truthful Report
$\hat{c}^S \neq c$ & $\hat{v}^S = \hat{v}^B$	275	300
$\hat{c}^S = c$ & $\hat{v}^S = \hat{v}^B$	−300	$35 + 320 - \hat{v}^B$
$\hat{c}^S \neq c$ & $\hat{v}^S \neq \hat{v}^B$	−210	−105
$\hat{c}^S = c$ & $\hat{v}^S \neq \hat{v}^B$	−405	−210

2.2.3 Within-stage Coordination Condition

Since $D = 275$, it again suffices to choose the arbitration fee $F = 300$. Then, by the Inter-stage Coordination Condition, a player who misreports his/her own type will be penalized by 300. Since $D = 275 < 300$, his/her unique best response is to truthfully announce his/her own type in the first stage.

2.3 Discussion

As the above example shows, the SR mechanism is constructed by first choosing a set of dictator lotteries $l(\cdot)$, one for each type of each player, such that it is a dominant strategy for each type to make a truthful report in the arbitration stage. As the outcome in the second

stage is based solely on the second reports of the party in arbitration and the dictator lotteries that were constructed prior to play, updating about the other player's type plays no direct role in our mechanism. This feature ensures that the mechanism is not dependent on the way in which players update their beliefs and makes the mechanism more robust to relaxations of assumptions surrounding common knowledge of rationality, information, and preferences.

It is not possible to elicit truthful secondary reports in our mechanism without lotteries when there are more than two states. In this sense, lotteries are an important component of our mechanism that cannot be eliminated. However, like the stochastic version of the BDM mechanism, the second-stage mechanism requires probability sophistication and dominance but does not require subjects to have Expected Utility preferences.[8] We also note that from a theory standpoint, there is no reason to divide the outcome space of the lottery equally. Similar to the approach used in virtual implementation, it is possible to construct second-stage lotteries where inefficient states occur with very small probability.

Experimental readers might (rightly) be concerned that our lottery method may lead to very small expected differences between reports and that the best response function may be quite flat. This has the potential to create noise in the second stage and could potentially undermine the performance of the mechanism. We are sympathetic to this concern and consider it to be one of the reasons to study the performance of the mechanism empirically. We note, however, that while second-stage noise may make implementation more difficult, variations of our mechanism can still uniquely implement the first best as long as a majority of second-stage reports are truthful and risk preferences are not too extreme.

Finally, we have restricted attention to the case where $F = T$. This case maximizes the potential rewards for being truthful subject to the requirement that money does not need to be injected into the mechanism in any state. We discuss situations where we may wish to set $F > T$ below.

[8]See Karni (2009) for a theoretical analysis of the stochastic BDM mechanism. Note that while we use lotteries in the case where both reports are wrong, it is also possible to fix priorities for entering arbitration. In this case, only the second stage lotteries are stochastic. See Section 5.5.2 for more details.

3 The Experiment

3.1 Main Treatment

Each session of our experiment consists of two phases in which participants play a total of 20 periods. Both phases are computerized and vary only in the rules governing the mechanism's adoption.

Phase 1: In periods 1-10 of the experiment, a seller is perfect-stranger matched with a buyer at the beginning of each period and both parties have the opportunity to invest to improve the joint surplus generated by trade. As seen in Table 4 below, the buyer's investment reduces the seller's true production cost while the seller's investment increases the true value of the produced good for the buyer. Both investments are made simultaneously.

Table 4: Buyer and Seller Investments

Buyer		Seller	
Buyer's Investment	True Cost	Seller's Investment	True Value
0	130	0	200
25	80	25	250
75	10	75	320

After making investments, both the buyer and the seller are informed of the true value and the true cost of production. The buyer and seller next enter into the SR mechanism to set prices and determine whether trade occurs.

The rules and parameters used in our SR mechanism are identical to those described in Section 2 above. Subjects begin in a "Report Stage" where the buyer is asked to make a value report $\hat{v}_B \in \{200, 250, 320\}$ and a cost report $\hat{c}_B \in \{10, 80, 130\}$ to the computer. The seller is also asked to make a value report $\hat{v}_S \in \{200, 250, 320\}$ and a cost report $\hat{c}_S \in \{10, 80, 130\}$. All four reports are made simultaneously.

The reports of the buyer and the seller are compared by the computer in a "Verification Stage". If all reports coincide, the buyer and seller trade at the report-specific prices given in Table 1. Prices in this table were constructed using the function

$$P^{SIM}(\hat{v}, \hat{c}) = (\hat{v} - \underline{v}) - (\overline{c} - \hat{c}) + \frac{\underline{v} + \overline{c}}{2}, \tag{1}$$

where $\hat{c}$ is the jointly reported cost, $\hat{v}$ is the jointly reported value, $\overline{c}$ is the highest possible cost, and $\underline{v}$ is the lowest possible value. The trade prices are structured such that if both the buyer and seller report the truth, the buyer receives the marginal surplus created from his investment and the seller receives the marginal surplus created from her investment.[9] Payments are also structured such that both parties receive the same surplus along the truth-telling path when they make the same investment choice.

If there is a discrepancy in the reports, one of the parties enters into the arbitration stage and is asked to make a second report. As described above, if only the value reports differ, the buyer enters into arbitration and is fined 300; if only the cost reports differ, the seller enters into arbitration and is fined 300; and if both reports differ, each party has a 50 percent chance of entering arbitration and both parties are fined.

If the buyer enters into arbitration stage he is asked to make a second report regarding the value of the good. As shown in Panel (a) of Table 2, we use the report along with a fair six-sided dice to determine whether trade occurs and the price. If the second report of the buyer matches the first-stage report of the seller, the seller is rewarded a bonus of 300 in addition to her earnings for the round. In other cases, the seller is also fined 300.

If the seller enters into the arbitration she is asked to make a second report regarding the cost of production. As shown in Panel (b) of Table 2, we use the report along with a fair six-sided dice to determine whether trade occurs and the price. If the second report of the seller matches the first-stage report of the buyer, the buyer is rewarded a bonus of 300 in addition to his earnings for the round. In other cases, the buyer is also fined 300.

In cases where no trade occurs, the investments made by the participants were sunk. However, the seller did not have to produce the good and had an effective production cost of zero.

Phase 2: In periods 11-20 of each session, the buyer and seller are given the choice to opt in or opt out of the mechanism at the beginning of each period. We framed opting out of the mechanism as "dismissing the arbitrator," so that opting in is the status quo. If the buyer

[9]For example, if the buyer invests 75 and the seller invests 0, the marginal surplus generated by the buyer's investment is 45 (120 - 75 = 45) and the marginal surplus generated by the seller's investment is 0. Starting from a baseline profit of 35, the mechanism should thus give the buyer a profit of 80 and the seller a profit of 35. This is indeed the case: if both parties report the true value of 200 and the true cost of 10, the trade price is 45; the buyer's profit is 80 (200 - 45 - 75 = 80) and the seller's profit is 35 (45 - 10 - 0 = 35).

and seller opt in, they are informed that the arbitrator is available, and play continues as in the first ten periods. If either party opts out, both parties are informed that the arbitrator is dismissed. They then make investment decisions as normal but always trade at a fixed price of 165. Both parties are informed about whether the arbitrator is available but are not informed about the dismissal decision of the other party. This implies that if a subject opts out, he cannot determine whether his counter party opted in or out.

3.2 Alternative Mechanisms

In order to benchmark the performance of the mechanism, we also ran three comparison treatments. The first of these treatments was a **Fixed Price** treatment where subjects chose investments but where the trade price was fixed at 165. As with the Main treatment, the Fixed Price treatment involved 20 periods. To maintain the same structure as the Main treatment we had subjects play 10 periods, read a short set of instructions that reminded subjects of the matching protocol, and then had them play the remaining 10 periods.

The other two treatments followed the exact protocol of the main treatment with subjects being forced to use the mechanism in Phase 1 and having the option of opting out of the mechanism in Phase 2. The mechanisms used in these treatments are as follows:

The KTH treatment Kartik et al. (2014) (KTH) show that if subjects have a preference for honesty, it may be possible to induce efficient trade in a one-stage mechanism if subjects use these preferences to break ties between indifferent reports. We test this mechanism in our **KTH** treatments.

In sessions using the KTH Mechanism, the buyer is asked to make a value report $\hat{v}_B \in \{200, 250, 320\}$ and a cost report $\hat{c}_B \in \{10, 80, 130\}$ to the computer. The seller is also asked to make a value report $\hat{v}_S \in \{200, 250, 320\}$ and a cost report $\hat{c}_S \in \{10, 80, 130\}$. All four reports are made simultaneously. Trade always occurs and price is set equal to:

$$P^H = (\hat{v}_S - \underline{v}) - (\overline{c} - \hat{c}_B) + \frac{\underline{v} + \overline{c}}{2}. \tag{2}$$

As before, prices are constructed so that if all reports are truthful, the buyer receives the marginal value of his investment and the seller receives the marginal value of her investment.[10]

[10]In principle we could have used any of the cost reports and value reports to set prices. We chose to

While trade always occurred in the KTH Mechanism, buyer's and seller's may incur fines if there was disagreement in the reports made by the buyer and seller. For the buyer, we assessed a fine equal to:

$$F_B^H = \max\{0, \hat{c}_S - \hat{c}_B\}, \tag{3}$$

where $\hat{c}_S$ and $\hat{c}_B$ are the cost reports of the seller and buyer. For the seller, we assessed a fine equal to:

$$F_S^H = \max\{0, \hat{v}_S - \hat{v}_B\}, \tag{4}$$

where $\hat{v}_S$ and $\hat{v}_B$ are the value reports of the seller and buyer. The fines are set such that (i) if the seller makes a truthful cost report, the buyer is indifferent between announcing a lower cost and the true cost and (2) if the buyer makes a truthful value report, the seller is indifferent between announcing a higher value or the true value.[11]

As structured, the prices and fees are set such that both the buyer and seller are indifferent between making a truthful report or making a lie when the other party always tells the truth. As shown in KTH, if buyers and sellers always receive a small utility for telling the truth, the truth-telling equilibrium is the unique equilibrium under two rounds of iterated deletion of strictly dominated strategies.[12]

The SPI treatment While earlier papers have documented issues that may arise in the use of subgame-perfect implementation (SPI) mechanisms, it is nonetheless useful to benchmark efficiency of the mechanisms for the experimental environment. We do this by running a three-stage **SPI** treatment that uses a mechanism based on Moore & Repullo (1988).

use the buyer's cost report as this was directly tied to his investment and it was easy for participants to understand how the cost arose. We also ran 2 pilot experiments where we used the buyer's value report and the seller's cost report to set prices. Results in these pilots were similar to those used in the main experiment, except for slightly lower investment levels.

[11] As an example, suppose that the true value is 320, the true cost is 130, and the seller makes truthful reports of $\hat{v}_S = 320$ and $\hat{c}_S = 130$. If the buyer makes truthful reports of $\hat{v}_S = 320$ and $\hat{c}_S = 130$, the trade price is 285. The buyer surplus is $35 (= 320 - 285)$. If, instead, the buyer lies and makes reports of $\hat{v}_S = 320$ and $\hat{c}_S = 10$, the trade price is 165, but the buyer is fined $120 (= 130 - 10)$. The buyer's surplus thus remains $35 (= 320 - 165 - 120)$.

[12] Our variant of the KTH mechanism uses a fine that exactly offsets the marginal gain associated with an advantageous lie. This departs from the original KTH construction where the authors consider a fine where the punishment exceeds the total gain associated with an advantageous lie. An advantage of our design is that buyers and sellers who invest optimally never have an incentive to make a misreport for any belief about the action of their counterparty. However, for a non-investing buyer or seller, our approach induces truth telling only in the case where an individual has a preference for honesty and believes the other party always makes truthful reports. See the appendix for a broader discussion.

In sessions using the SPI Mechanism, the buyer makes a value report $\hat{v}_B$ and the seller makes a cost report of $\hat{c}_S$. The buyer and seller observe the report of their counterparty and have the option to "call the arbitrator" or "not call the arbitrator." If both parties do not call the arbitrator, trade occurs at a price equal to:

$$p^{SEQ} = (\hat{v}_B - \underline{v}) - (\overline{c} - \hat{c}_S) + \frac{\underline{v} + \overline{c}}{2} \tag{5}$$

where, as before $\hat{v}_B$ is the buyer's report, $\hat{c}_S$ is the seller's report, $\underline{v}$ is the lowest possible value, and $\overline{c}$ is the highest possible cost.

If only the buyer calls the arbitrator, the seller enters into arbitration and is immediately fined 300. The seller is then given a counteroffer to sell the good at a counteroffer price of

$$\hat{p}_S^{SEQ} = \hat{c}_S - 5. \tag{6}$$

If the seller accepts the counteroffer, trade occurs at the counteroffer price. The buyer is given a bonus of 300 in this case. Otherwise the parties do not trade but still must pay their investment costs. In addition, the buyer is fined 300.

If only the seller calls the arbitrator, the buyer enters into arbitration and is immediately fined 300. The buyer is then given a counteroffer to buy the good at a counteroffer price of

$$\hat{p}_B^{SEQ} = \hat{v}_B + 5. \tag{7}$$

As in the other case, if the counteroffer is accepted, trade occurs at the counteroffer price and the seller is given a bonus of 300. If the counteroffer is rejected, the two parties do not trade and the seller is fined 300.

If both the buyer and seller call in the arbitrator, a virtual coin is flipped and either the buyer or the seller enters into the arbitration stage.

3.3 Protocol

Our experimental design utilizes a between-subjects design in which each subject is exposed to 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 phase of the experiment.

All of the experiments were run in the Experimental Economics Laboratory at the University of Melbourne in May and June of 2016. The experiments were conducted using the programming language z-Tree (Fischbacher, 2007). A total of 20 sessions were run: 8 sessions using the SR mechanism, 4 sessions using the No-Mechanism Baseline, 4 sessions using the KTH Mechanism, and 4 sessions using the SPI mechanism. All of the 400 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. 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.[13] Once all participants successfully completed the quiz, a verbal summary was read aloud that summarized the trading mechanism and emphasized the perfect-stranger matching. The purpose of the summary was to ensure that the main features of the experiment were common knowledge amongst the participants.

Subjects next played 6 periods where the computer played the role of their matched partner. In each period, the computer made maximal investments and truthful announcements.[14] In the event that the computer went into arbitration, the computer maximized its expected value by reporting the true cost or value. The first three periods against the computer were unpaid while the last three periods were paid. The rounds against the computer were used to allow participants to experiment with the mechanism, experiment with potential strategies, and to increase their initial surplus to reduce the potential for bankruptcies.

After completing the six rounds against the computer, we read additional oral instructions that reiterated the bankruptcy procedures (described below) and detailed the additional

[13]One potential criticism of implementation experiments is that in applied settings, individuals who enter into a contract will have time to discuss with each other how the game should be played and will naturally be able to to come to a general understanding of how the mechanism works. Keeping this criticism in mind but also being cognizant of introducing potential experimenter demand effects, our instructions are explicit about the incentives that exist in the mechanism but never state what a subject should do. Subjects are told that if all buyers and sellers "report the true value and the true cost" the prices will adjust so that each party receives the benefits from their investment. Subjects are also explicitly told that they cannot increase their material payoff by "misreporting" if the other party reports the true cost and the true value and that the other party cannot increase their material payoff by "misreporting" if they report their true cost and true value. We never use the words "lie" or "truthful reports" to mitigate demand effects.

[14]To be as close as possible to the other treatments, we also had the computer choose maximal investments in the Fixed Price treatment.

phase that would exist in Phase 2 of the experiment. Subjects were informed that their decisions in Phase 1 would not influence their position, matching, or available actions in Phase 2.

Subjects then entered and played Phase 1 and Phase 2 of the experiment. Payments were made in cash based on the earnings subjects had accumulated throughout the experiment with an exchange rate of 35 ECU to \$1 AUD. In addition subjects received a show-up fee of \$22. The average payment at the end of the experiment was \$51.14 AUD. At the time of the 2016 experiments. \$1 AUD = \$0.74 USD.

While we gave subjects a large show-up fee to offset losses, the fines that exist in all mechanisms created the potential for negative earnings and bankruptcies. Subjects were informed in the instructions and in the oral instructions that if they ever had negative earnings at the end of any period of the main experiment they would be removed from the experiment without payment. Subjects were also informed that if a subject was removed from the experiment, a computer agent would play the role of that particular buyer and seller and would play exactly like the computer player they traded with in the instruction phase of the experiment. There were no bankruptcies in sessions involving the KTH Mechanism and 6 out of 160 (2.5%) bankruptcies in the SR mechanism. Thus the bankruptcy protocols appeared to play a limited role in these sessions. In the SPI mechanism, however, 16 out of 80 (20.0%) subjects went bankrupt. We highlight the forces contributing to this large number of bankruptcies in the appendix.[15]

3.4 Hypotheses

The SR mechanism used in our experiment is designed to implement truthful announcements and to allow buyers and sellers to capture all surplus associated with their investment. Given the incentives induced by the mechanism we would predict the following pattern of behavior:

[15]In designing this experiment we also considered an alternative pay-one-period protocol to avoid the empirical difficulties that arise when dealing with bankruptcies in the data. We chose against this alternative protocol for two reasons. First, in order for payments to be credible, the show-up fee in a pay-one-period protocol must be set so that a buyer or seller never receive a negative payoff in any realization of any period. In our setting, this would have required us to either introduce an extremely large show-up fee or make the variable component of payment extremely small. Both of these policies are likely to have reduced the saliency of the incentive payments. Second, in AFHW (2017) sessions were run using a mechanism similar to the three-stage mechanism studied here under both the pay-one-period protocol and the pay-all-period protocol. AFHW finds similar behavior across the two treatments.

Hypothesis 1 *The path of play under the Simultaneous-Report mechanism involves both the buyer and seller making efficient investments and truthful reports by both parties. If either party enters into arbitration, they make a truthful secondary report.*

We refer to the behavior described in Hypothesis 1 as **efficient truth-telling behavior** and the resulting outcome as the **efficient outcome**. Note that in this equilibrium the buyer earns 80 and the seller earns 80. If either party opts out of the mechanism in the second phase, we would predict no investment by either party and earnings of 35. We thus would predict the following pattern of behavior in periods $11 - 20$:

Hypothesis 2 *Buyers and Sellers are predicted to opt into the Simultaneous Report Mechanism.*

Under different equilibrium refinements and preference assumptions, the Simultaneous Report Mechanism, the SPI Mechanism, and the KTH Mechanisms are predicted to induce truth-telling behavior and efficient investment while the Fixed Price mechanism is predicted to lead to no investment. We thus predict:

Hypothesis 3 *Efficiency in the SR treatment will be equal to efficiency in the KTH treatment and the SPI treatment. All three mechanisms will have higher efficiency than the Fixed Price treatment.*

4 Results

We describe the results of the main experiment in this section. Section 4.1 uses data from the eight sessions that use the Simultaneous Report Mechanism to study Hypothesis 1 and 2. Section 4.2 uses data from all sessions to make comparisons between the Simultaneous Report Mechanism and the other three treatments.

4.1 Behavior in the Simultaneous Report Mechanism

Result 1 *In Phase 1 of the experiment, the Simultaneous Report Mechanism induces truth telling behavior in over 93 percent of cases. Buyers and Seller make efficient investments in over 80 percent of cases. The efficient outcome occurs in* 73.8 *percent of cases.*

Figure 1 displays the patterns of play we observed in the first ten periods of the experiment. The left hand panels show the behavior of the buyers while the right hand panels show the behavior of the sellers. Panel (a) summarizes the investment decisions of both parties, Panel (b) summarizes decisions in the report, and Panel (c) summarizes reports in the secondary reports stage. The error bars in panel (b) are 95 percent confidence intervals of each proportion with errors clustered at the individual level.

Panel (a) shows that in the majority of observations, both the buyer and the seller chose the optimal level of investment. Aggregating over all 10 periods, buyers chose the optimal level of investment in 89.6 percent of cases while sellers chose the optimal level of investment in 84.8 percent of cases.

Panel (b) shows that in almost all periods, buyers and sellers make truthful cost and value reports. Looking at the left hand side, buyers made truthful value reports in 98.1 percent of cases and truthful cost reports in 94.0 percent of cases. Sellers made truthful value reports in 93.1 percent of cases and truthful cost reports in 97.8 percent of cases.

Finally, Panel (c) shows the types of secondary reports that were made by buyers and sellers. We divide these reports into four categories: truthful secondary reports, reports that are not truthful but match the report made by the counter party in the report stage, reports that are not truthful when the other party reported truthfully, and all other combinations. As can be seen by looking at the left hand side, buyers report truthfully in the second stage in 28 of 57 cases. However, they match the report of the seller who has lied in the first stage in 12 of 57 cases. This suggests that some buyer's may actively be trying to prevent pairwise losses by ensuring that the fines are transferred to their counter party. Similarly, seller's report truthfully in the second stage in 25 out of 46 cases and match the buyer's lie in 10 out of 46 cases.[16]

While the results in Figure 1 are presented as the aggregate of all 10 periods, there are only very small changes in investment and reporting decisions over time. Panel (a) and (b) of Figure 2 shows how investments and truthful reports evolve over the first 10 periods. As seen in panel (a), the proportion of buyers who chose high investment starts above 80 percent

[16]Panel (c) shows secondary reports in both the case where a buyer or seller enters arbitration due to their own lie or due to the lie of their counter party. Looking only at cases where a buyer enters into arbitration due to a seller lie, buyers make a truthful report in 25 of 42 cases and match the seller in 11 of 42 cases. In cases where a seller enters into arbitration due to a buyer lie, sellers make a truthful report in 21 of 35 cases and match the buyer in 9 of 35 cases. There is no combination of investments and reports where a buyer or seller has a positive return for lying.

in period 1 and increases to an average of 92.6 percent in periods $6-10$. The proportion of sellers who invest optimally also starts above 80 percent and increases to an average of 88.1 percent in periods $6-10$. As seen in Panel (b), the proportion of buyers and sellers who report truthfully is also stable with buyers and sellers making truthful cost and value reports at least 90 percent of the time in all periods.

Finally, panel (c) shows the aggregate number of lies that different buyers and sellers take over the first ten periods. The dark grey steps represent the two buyers and three seller who went bankrupt in the first 10 periods and whose lie frequency are truncated.[17] As can be seen, 81.3 percent of buyers and 77.5 percent of sellers make one lie or less suggesting that the mechanism is highly effective at inducing truth-telling.

As one might expect from the structure of fees, there is a strong connection between being rewarded for a lie in one period and making such a lie in a future period. Buyers and sellers who lie and are fined for such a lie have only a 28.4 percent chance of lying in the next period. By contrast, a buyer or seller who lies in a period and who is rewarded by having their counter-party match their misreport has a 69.6 percent chance of lying in the next period. Given that buyers and sellers who tell the truth in one period lie in the next only 5.1 percent of the time, the switching data suggests that a large proportion of lies are due to the poor learning dynamics that are generated by non-truthful secondary reports.[18]

Despite the potential learning issues noted above, our data suggest that the Simultaneous-Report Mechanism is highly effective in inducing truthful reports and in inducing efficient investment. In aggregate 86.2 percent of dyads improved their performance relative to the theoretical no mechanism benchmark and 73.8 percent of dyads exhibited efficient truth-telling behavior and achieved the efficient outcome. Truth-telling behavior also appears to be stable across the first 10 periods and there is high levels of efficiency even in period 1.[19]

We now turn to our second hypothesis and analyze opt-in behavior in periods 11-20:

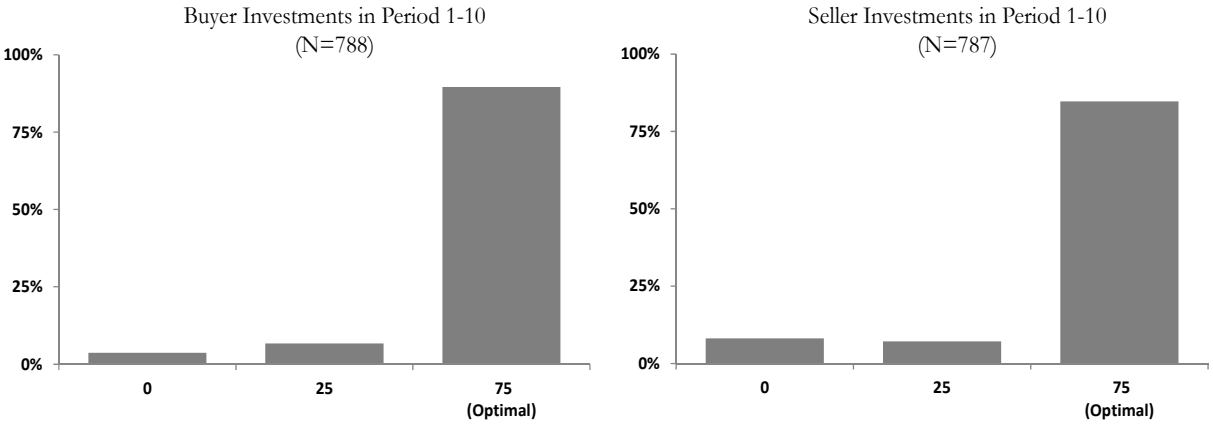
Result 2 *Buyers opt into the mechanism* 77.1 *percent of the time while sellers opt into the mechanism* 76.2 *percent of the time. Opt-in rates are increasing for both buyers and sellers.*

[17] One additional buyer went bankrupt in the second phase of the experiment

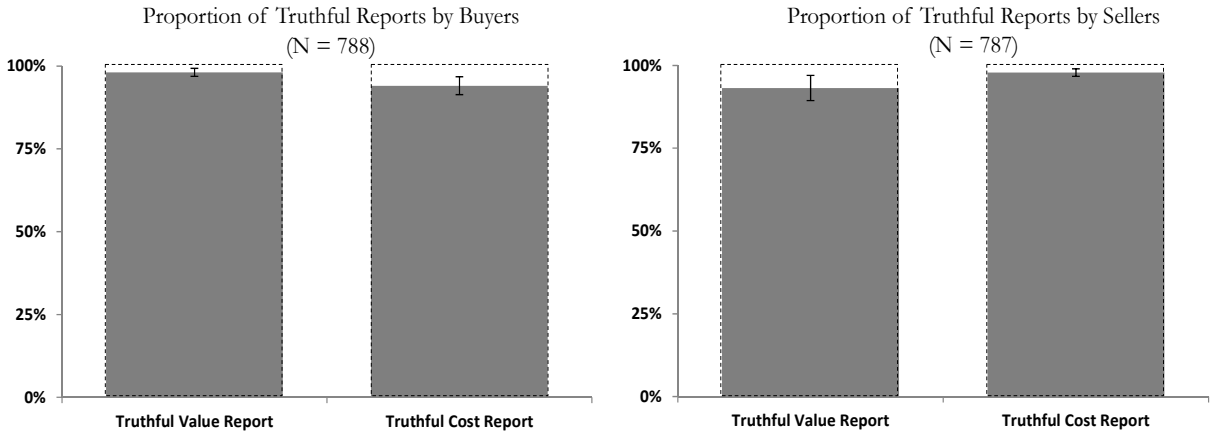
[18] The difference in switch rates is significant in a simple probit regression that restricts the sample to the 111 report decisions in a period following a lie and uses a dummy variable for cases where a buyer or seller lied in the last round and was rewarded (p-value $< .01$). The difference in learning dynamics is also apparent at the aggregate level.

[19] In period 1, 81.3 percent of dyad pairs improved their performance relative to the theoretical no mechanism benchmark and 67.5 percent of dyad pairs achieved the first best.

(a) Distribution of Investment Choices



(b) Proportion of Truthful Reports



(c) Secondary Reports in Arbitration Stage

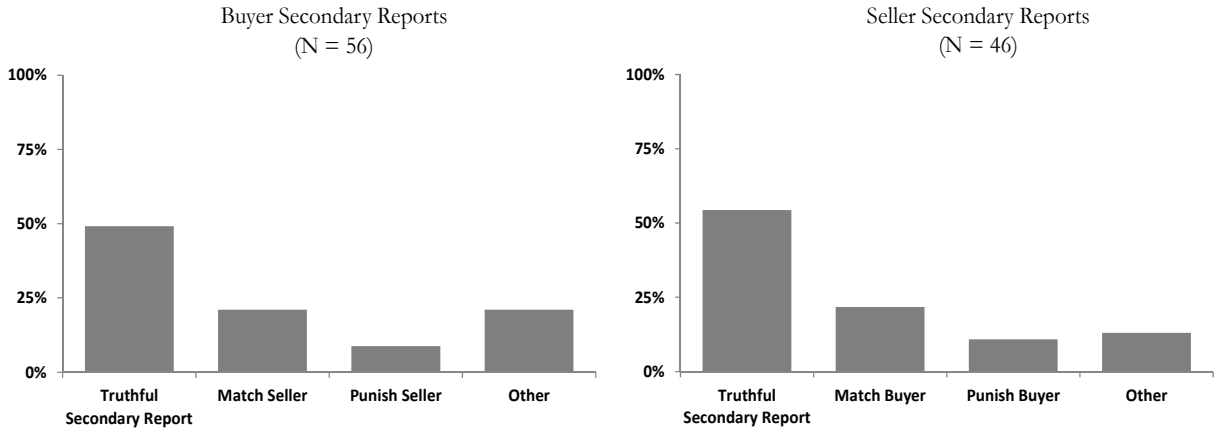
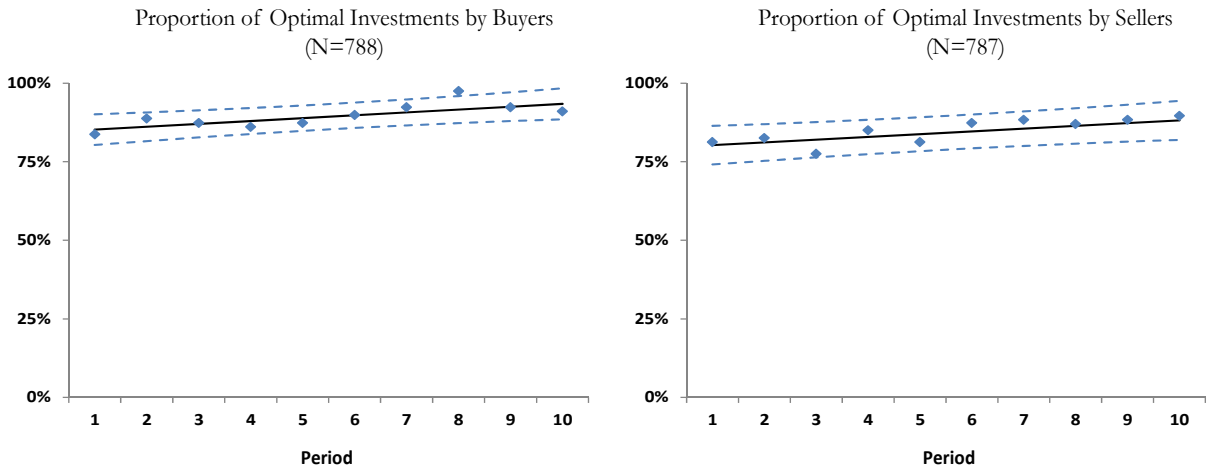


Figure 1: Pattern of Play in First 10 Periods of Simultaneous Report Mechanism

(a) Proportion of Optimal Investment Decisions over Time



(b) Proportion of Truthful Reports over Time

(c) Aggregate Number of Lies in Periods 1-10

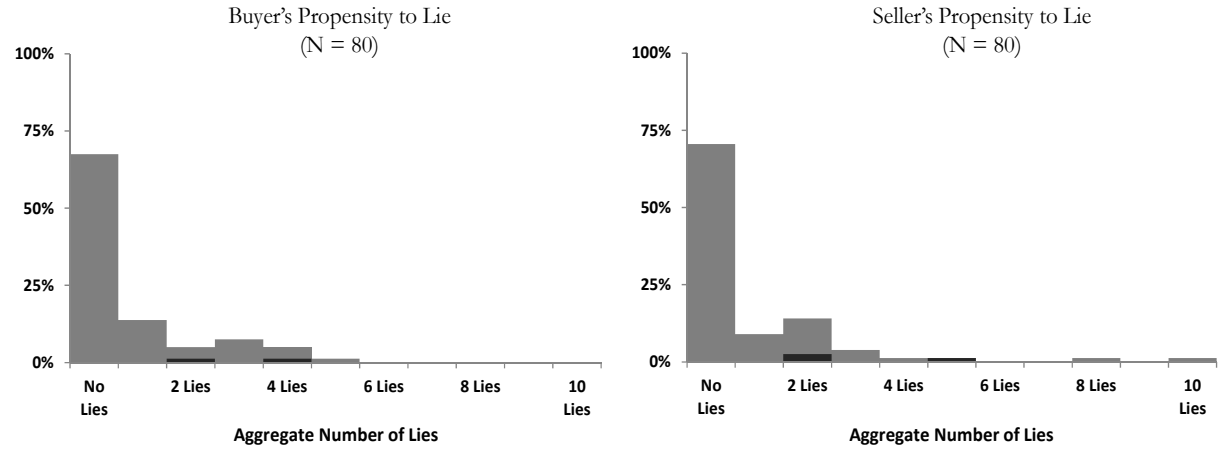


Figure 2: Evolution of Play in First 10 Periods of Simultaneous Report Mechanism

90.5 *percent of dyad pairs who opt into the mechanism exhibit efficient truth-telling behavior and achieve the efficient outcome.*

Panel (a) of Figure 3 shows opt-in rates of buyers and sellers in Phase 2 of the Simultaneous Report mechanism. As can be seen, opt-in rates for both buyers and sellers begin near 60 percent and increase to roughly 85 percent by periods $16 - 20$. On average, Buyer's opt into the mechanism 77.1 percent of the time while sellers opt into the mechanism 76.2 percent of the time. Given these opt-in rates, 59.4 percent of the groups had the SR mechanism available.[20]

Panel (b) shows the proportion of buyers and sellers who made optimal investments in groups where the mechanism was kept and where it was removed. Groups with the mechanism are represented by the blue diamonds while groups without the mechanism are represented with the red circle. As can be seen, optimal investment occurs in almost all periods and is stable over time in groups with the mechanism. By contrast, investment is decreasing in groups who opt out of the mechanism.

Panel (c) shows the proportion of truthful announcements by buyers and sellers in dyad pairs where buyers and sellers opt into the mechanism. Buyers are truthful in almost all periods while all but one seller is truthful in all periods.

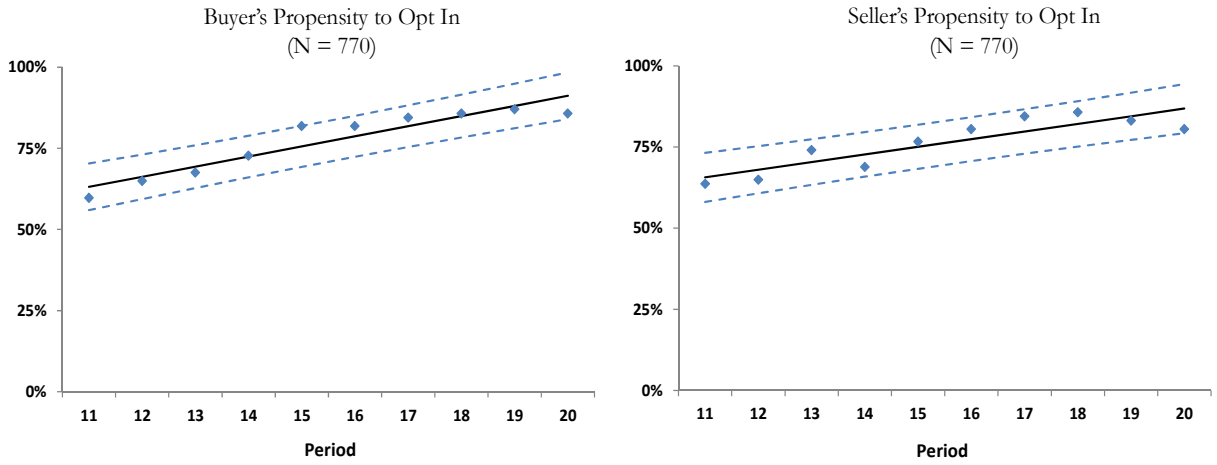
In aggregate, 90.5 percent of groups who opted into the mechanism exhibited efficient truth-telling behavior and achieve the efficient outcome. An additional 3.9 percent of groups made suboptimal investments but reported truthfully in the report stage. Buyers made truthful secondary reports in 9 of the 14 cases where the buyer entered in arbitration while sellers made truthful secondary reports in 4 of 5 cases. Buyers and Sellers who entered arbitration never matched their counter party's first-stage misreport.

4.2 The Relative Performance of the Simultaneous Report Mechanism

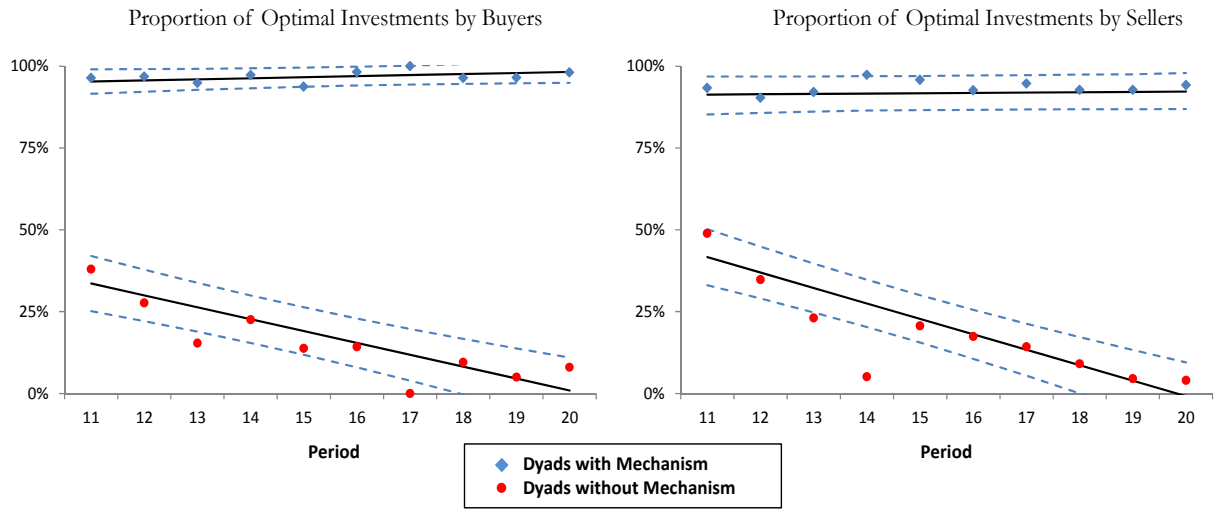
Thus far we have shown that the SR Mechanism is effective at inducing truthful reports and leads to the efficient outcome in the majority of cases. We have also shown that buyers and

[20]Looking at the aggregate number of opt-in decisions of buyers and sellers, 37.0 percent of buyers and sellers always opted in, while an additional 40.3 percent opted in between 7 and 9 times. 5.2 percent of buyers and sellers never opted in and the remaining 17.5 percent of buyers and sellers opted in between 1 and 6 times.

(a) Opt-in Rates in Periods 11-20



(b) Optimal Investment Rates in Periods 11-20



(c) Proportion of Truthful Reports in Periods 11-20 with Mechanism

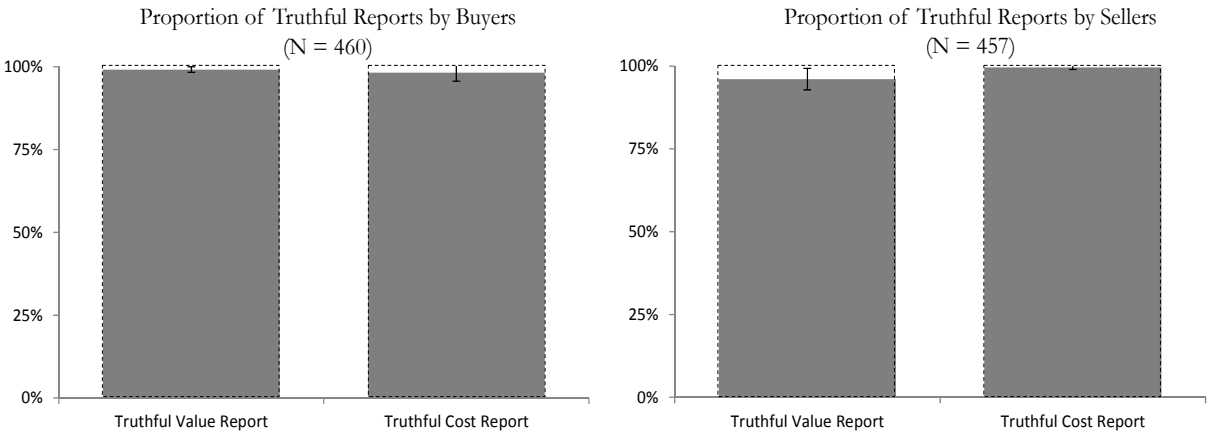


Figure 3: Pattern of Play in Periods 11-20 of Simultaneous Report Mechanism

sellers opt into the mechanism at a high frequency and that the efficient outcome occurs in over 90 percent of dyads where the parties have opted into the mechanism. We now compare the performance of the mechanism to the three other comparison mechanisms that were run in our main experiments.

We begin with the predictions in Hypothesis 3 that efficiency in the SR Mechanism should be equal to the efficiency found in the KTH Mechanism and the SPI Mechanism.

Result 3 *In contrast to Hypothesis 3, efficiency in the SR treatment is significantly higher than efficiency in each of the other three treatments. Efficiency in the SPI treatment is not significantly different than efficiency in the Fixed Price treatment. Efficiency in the KTH treatment is significantly lower than efficiency in each of the other three treatments.*

Support for Results 3 is provided in Panel (a) of Figure 4, which shows the average per-period earnings of each treatment using data from all 20 periods. An observation is a subject's earnings across the experiment divided by 20. The earnings of a subject who went bankrupt is equal to -38.5, which when multiplied by 20 is equal to the amount that could be lost before a subject was dismissed from the experiment.[21]

Average per-period efficiency in the SR treatment is 47.9. While below the theoretical benchmark of 80, efficiency in the SR treatment is 19.8 percent higher than the efficiency in the Fixed Price treatment, 35 percent higher than efficiency in the SPI treatment, and 62 percent higher than efficiency in the KTH treatment. All three differences are significant in a simple regression where average per-period earnings is regressed against the treatment dummies (SR vs Fixed Price: p-value $= .04$; SR vs SPI: p-value $< .01$; SR vs KTH: p-value $< .01$).[22]

[21] In the appendix, we also consider two alternative methods for calculating efficiency in cases where there were bankruptcies. In one method, we predict future behavior of bankrupt subjects using the behavior of other subjects who also made early lies. This is done by estimating switch rates between lying strategies and truthful strategies and constructing a Markov transition matrix using this switch data. The second method is to assume that bankrupt subjects lie in every period. The estimated per-period efficiencies of the SR mechanism using these alternative methods are 51.9 and 43.6 and similar to the efficiencies shown here. For the SPI mechanism, efficiencies are 28.6 and 4.5. The comparison of the efficiency of the SR mechanism to the other treatments is thus robust to the way we handle bankruptcies. The SPI mechanism is more sensitive to the way we handle bankruptcies but never has an estimated efficiency above the SR mechanism.

[22] We also compared treatments non-parametrically. The Kruskal–Wallis test that the four treatments are drawn from the same distribution is rejected at a p-value $< .001$ ($\chi^2(3) = 48.48$). As a follow-up post hoc test, we use Dunn's test of stochastic dominance using the Benjamini–Hochberg procedure to adjust for multiple hypotheses. The SR treatment has significantly higher efficiency than all three other treatments using a false discovery rate of .05. Both the Fixed Price treatment and the SPI treatment have a higher

The average per-period efficiency of the SPI treatment is 35.5. This level of efficiency is not significantly different from (or than) efficiency found in the Fixed Price treatment but is significantly greater than the efficiency found in the KTH treatment. As was noted in the earlier section, 20 percent of participants in the SPI treatment went bankrupt in the treatment. We show in the appendix that most bankruptcies occur early in the experiment and that many subjects lose money even in periods where they played against the computer. It thus appears that a significant proportion of individuals have a difficult time understanding this mechanism and that losses are driven in part by confusion. We also show that subjects who lie and are challenged reject the counter offer in the majority of cases and that subjects do not have pecuniary incentives to challenge. Thus, while efficiency is reasonably high in the SPI mechanism, the mechanism does not function as intended. This is consistent with results in FPW where the mechanism is not robust to negative reciprocity.

The efficiency of the KTH treatment is only 29.5 and significantly less than efficiency in all three other treatments. As shown in the appendix, the preference for honesty mechanism fails to induce truthful reporting for both buyers and sellers and truthful reports are decreasing over time. Buyer and seller investments are also decreasing over time and efficiency in this treatment is falling. Looking at the data, it appears that the inefficiency in this mechanism is driven by buyers and sellers who try to take advantage of potential mistakes by their counter party.[23]

Finally, efficiency in the Fixed Price treatment was 40.0. This efficiency is slightly higher than the theoretical benchmark of 35, but below the efficiency of the SR mechanism. The additional efficiency is due to a small subset of buyers and sellers who invest 25 in early periods. These positive investments decrease rapidly over time and an investment of 0 is observed in 95.6 percent of cases in periods 11-20.

Panel (b) of Figure 4 provides information on the number of dyads where the efficient outcome occurs. To maintain a similar comparison across treatments, we exclude pairs in which a buyer or seller was played by the computer. The error bars are 95 percent confidence intervals of each proportion with errors clustered at the individual seller level.[24] As can be

efficiency than the KTH treatment. There is no significant difference between the Fixed Price treatment and the SPI treatment.

[23]The original KTH mechanism uses a larger fine for disagreement that is likely to prevent buyers and sellers from trying to take advantage of potential mistakes. See the appendix for a discussion of the two variants of the mechanism.

[24]We use the seller data to avoid double counting. The confidence intervals are similar if only the buyer

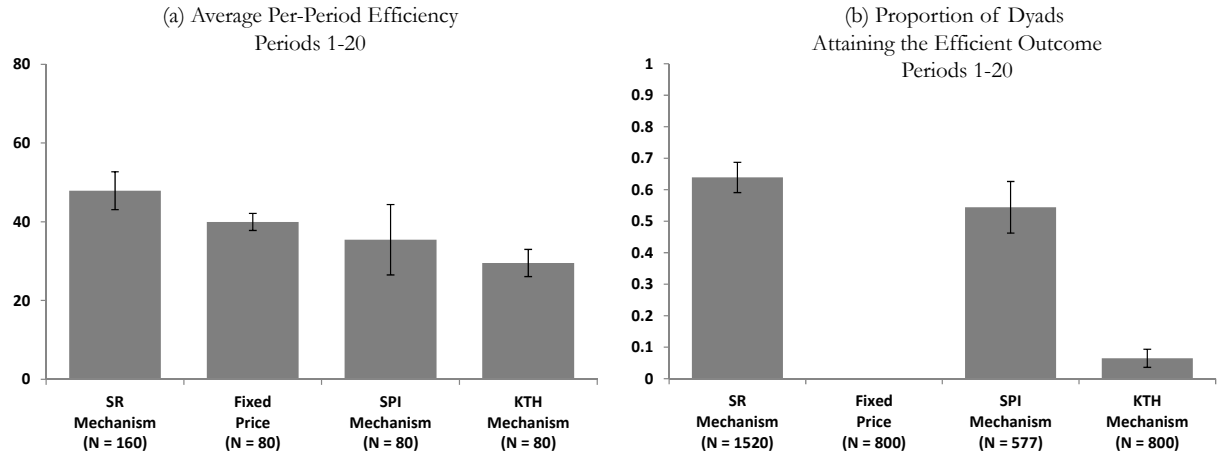


Figure 4: Average Per-Period Efficiency and the Proportion of Dyad Pairs Attaining the Efficient Outcome Across Treatments

seen, 63.9 percent of dyad pairs in the SR treatment achieve efficiency. This proportion is significantly higher than the proportion in any of the other treatments in a simple regression where a binary variable that is one if a dyad reaches the first best is regressed against the treatment dummies (SR vs Fixed Price: p-value $< .01$; SR vs SPI: p-value $= .04$; SR vs KTH: p-value $< .01$).

5 The Theory

We briefly review what we have done so far. In Section 2, we describe a two-sided hold-up environment with pure cooperative investments; propose the Simultaneous Report (SR) mechanism; and show that the SR mechanism achieves the first best under complete information. In Section 3, we adapt this environment to our experimental setup and in Section 4, we provide ample evidence that our SR mechanism works well experimentally. The main purpose of this section is to generalize the SR mechanism and to highlight the very permissive implementation results that exist when using initial rationalizability as a solution concept. We then show how the generalized SR mechanism is robust to the relaxation of the complete information assumption and to a wide variety of reasoning processes and behavioral

data is used.

assumptions.

The rest of this section is organized as follows. Section 5.1 describes the setup. In Section 5.2, we introduce initial rationalizability as our solution concept, define implementation under initial rationalizability, and state our Theorem 1: "any" social choice function is implementable in initial rationalizability. Section 5.3 defines the generalized SR mechanism that is used to prove our Theorem 1. In Section 5.4, we show that the our Theorem 1 holds even under information perturbations. In Section 5.5, we discuss the robustness of our Theorem 1 from a variety of aspects.

5.1 The Environment

Consider a finite set of players $\mathcal{I} = \{1, \ldots, I\}$ with $I \geq 2$ located around 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 and the predecessor of player 1 is player I. The set of pure social alternatives is denoted by A, and $\Delta(A)$ denotes the set of all probability distributions over A with countable supports. In this context, $a \in A$ denotes a pure social alternative and $l \in \Delta(A)$ denotes a lottery on 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 is identified with a 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 also assume $u_i(\cdot, \theta_i)$ has the expected utility representation and extend our result to non-expected utility preferences in Section 5.5.2. Finally, assume that any two distinct types θ_i and θ_i' induce different preference orders over $\Delta(A) \times \mathbb{R}$.

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.[25] The private-value assumption entails no loss of generality when information is complete. In Section 5.4, we study 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 we will prove that every social choice function can be implemented by the suitably adapted SR mechanism which still has only two stages. In Stage 1 each player i

[25]For our result to hold, it suffices to assume that each player's type is known by at least two players.

chooses one message m_i^1 from a finite set M_i^1. Denote by $M^1 \equiv \times_{i\in\mathcal{I}} M_i^1$ be the set of Stage 1 message profiles. In Stage 2, after observing some message profile $m^1 \in M^1$ chosen in Stage 1, each player i chooses one message m_i^2 from a another finite set $M_i^2(m^1)$. Again, write $M^2(m^1) \equiv \times_{i\in\mathcal{I}} M_i^2(m^1)$ as the set of Stage 2 message profiles following the Stage 1 message profile 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 the set of terminal histories into lotteries in $\Delta(A)$; and (5) τ_i is the transfer rule that maps each terminal history to the transfer to agent i.

Let $\Gamma(\theta)$ denote the two-stage game associated with Γ at state θ. A *message* m_i 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., $z(m) = (m^1, m^2(m^1))$.

5.2 Solution Concept and Implementation

We now define the solution concept of *initial rationalizability*. Given a two-stage game $\Gamma(\theta)$ and conditional on history $h \in \mathcal{H}$, player i's payoff from a message profile m is given by

$$v_i(m, \theta_i | h) = u_i(g(z(m); h), \theta_i) + \tau_i(z(m)),$$

where $g(z(m); h)$ stands for the lottery resulted from the message profile m conditional upon the history h being realized. In particular, for each $m^1 \in M^1$,

$$v_i\left(m, \theta_i | m^1\right) = u_i\left(g(z\left(m^1, m^2\left(m^1\right)\right)), \theta_i\right) + \tau_i\left(z\left(m^1, m^2\left(m^1\right)\right)\right).$$

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. A *conditional probability system* (CPS) μ_i specifies for each nonempty subset of M_{-i} a probability distribution over M_{-i} such that Bayes' rule applies whenever possible (see Appendix 7.4 for a precise formulation). Let $M_{-i}(h) \subset M_{-i}$ be the set of message profiles of player i's opponent that are consistent with history h. That is, $M_{-i}(\varnothing) = M_{-i}$ and for each $m^1 \in M^1$, we have $M_{-i}(m^1) = \left\{m_{-i} \in \Sigma_{-i} : \left(m_i^1, m_{-i}^1\right) = m^1\right\}$.

Conditional on history h, using message m_i, and holding CPS μ_i, player i receives the expected payoff of the game $\Gamma(\theta)$ as follows:

$$V_i(m_i, \theta_i, \mu_i | h) = \sum_{m_{-i}} \{u_i\left(g(z\left(m_i, m_{-i}\right); h), \theta_i\right) + \tau_i\left(z\left(m_i, m_{-i}\right)\right)\} \mu_i\left[m_{-i} | M_{-i}\left(h\right)\right].$$

Fix a player $i \in I$, a CPS μ_i, and a message $m_i \in M_i$. Then, m_i is a **sequential best response** to μ_i iff, for all $m_i' \in M_i$, and for all $h \in \mathcal{H}$,

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

We now define initial rationalizability and our notion of *full implementation* in initial rationalizable messages:

Definition 1 (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 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 i is $R_i^{\Gamma(\theta)} = \bigcap_{k=1}^{\infty} R_{i,k}^{\Gamma(\theta)}$.*

The solution concept is arguably the weakest among the standard equilibrium or non-equilibrium solution concepts which impose sequential rationality (see Dekel and Siniscalchi (2013) for more discussion). In particular, only beliefs at the beginning of the game (i.e., $\mu_i\left(\cdot \middle| M_{-i}\right)$) are restricted. The following is the definition of implementability that we use.

Definition 2 *A social choice function f is **implementable in initial rationalizable messages** if there exists a mechanism Γ such that, for all $\theta \in \Theta$ and $m \in M$, $R^{\Gamma(\theta)} \neq \varnothing$ and $m \in R^{\Gamma(\theta)}$, we have $g\left(z\left(m\right)\right) = f\left(\theta\right)$ and $\tau_i\left(z\left(m\right)\right) = 0$.*

We then prove the following permissive result for implementation in initial rationalizable messages. To prove Theorem 1, we construct what we call the SR (simultaneous-report) mechanism, which works in the more general environment than the setup we discussed in Section 2.1.

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

5.3 The SR Mechanism

The SR mechanism remains a finite two-stage mechanism which proceeds as follows. In the first stage, each player i announces simultaneously his own type as well as the type of his predecessor (i.e., player $(i-1)$). If player i's announcement about his own type is the same as his 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 under the consistent profile. Otherwise, each player who makes an inconsistent announcement in the first stage makes an additional announcement about his own type. We pick with equal probability one player i from those who make an inconsistent announcement in the first stage and implement a lottery based on player i's second stage announcement. Finally, any player i who makes an inconsistent announcement is imposed a large penalty; moreover, player $i+1$ is imposed a large reward if player $i+1$'s announcement of player i's type coincides with player i's second announcement; otherwise, player $i+1$ is imposed a large penalty.

We formally define the SR mechanism as follows.

5.3.1 Message Space

We specify the message space for each player i.

Stage 1: Each player i is asked to report his own type and player $(i-1)$'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 $\mathcal{I}^*(m^1) \equiv \left\{ i \in \mathcal{I} |\ \hat{\theta}_i^i \neq \hat{\theta}_i^{i+1} \right\}$ be the set of players who make an inconsistent announcement at m^1. For $m^1 = \left(\hat{\theta}_i^i, \hat{\theta}_{i-1}^i \right)_{i \in \mathcal{I}}$, each player $i \in \mathcal{I}^*(m^1)$ is asked to report his 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 $m_i^2 = \tilde{\theta}_i$.

5.3.2 Outcome Function

First, we construct the dictator lotteries by invoking the following result due to Abreu and Matsushima (1992a).

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

$$u_i\left(l_i\left(\theta_i\right), \theta_i\right) > u_i\left(l_i\left(\theta_i'\right), \theta_i\right), \text{ for any } \theta_i, \theta_i' \in \Theta_i \text{ with } \theta_i \neq \theta_i'. \tag{8}$$

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 chooses each element in $\left\{l_i(\tilde{\theta}_i) : i \in \mathcal{I}^*\left(m^1\right)\right\}$ with equal probability.

5.3.3 Transfers

We define the following transfer rule:

- Each player $i \in \mathcal{I}^*\left(m^1\right)$ pays a penalty F;
- If player $i \in \mathcal{I}^*\left(m^1\right)$ is the player who is selected by the planner, player $i+1$ gets the incentive transfer:
$$T_{i+1}\left(\hat{\theta}_i^{i+1}, \tilde{\theta}_i\right) = \begin{cases} T, & \text{if } \hat{\theta}_i^{i+1} = \tilde{\theta}_i; \\ -T, & \text{if } \hat{\theta}_i^{i+1} \neq \tilde{\theta}_i. \end{cases}$$
- We choose F and T large enough so that $\min\{F, T\} > D$ where[26]
$$D = \sup_{i,a,a',\theta_i} \left|u_i(a, \theta_i) - u_i(a', \theta_i)\right|.$$

In words, each player $i \in \mathcal{I}^*(m^1)$ is penalized by F for making an inconsistent announcement of his own type. Moreover, in case that player $i \in \mathcal{I}^*$ is selected by the planner, player $(i+1)$ is rewarded by T, if his Stage 1 announcement of i's type is confirmed by player i's second stage announcement; otherwise, player $(i+1)$ is penalized by T.

[26]Recall that we assume that u_i is finite and hence D is bounded.

5.3.4 Sketch of Proof of Theorem 1

We prove Theorem 1 in three steps. First, sequential rationality implies that every player $i \in \mathcal{I}^*(m^1)$ will truthfully announce his own type in the second stage (the Truth-Telling Condition in Section 2). Second, if every player $i \in \mathcal{I}^*(m^1)$ announces his own type truthfully, it is a strictly dominant message for each player to announce the type of his successor truthfully in the first stage (the Inter-stage Coordination Condition in Section 2). Third, if every player announces his successor's type truthfully, then it becomes a strictly dominant message for each player i to announce his own type truthfully (the Within-stage Coordination Condition in Section 2). In other words, implementation is achieved in the SR mechanism after the first three rounds of iterated deletion of never best sequential replies under initial rationalizability. We provide the complete proof in Appendix 7.5.

5.4 Almost Complete Information

We now formulate the robustness property of the SR mechanism. Suppose that players do not observe the state directly but are informed of the state via signals. Following Aghion et al. (2012), we set the signal space to be $S_i = \Theta$. 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, that is, $\text{marg}_S \pi(s) > 0$ for every $s \in S$ so that Bayes's rule is always well defined. For each π, we write $\pi(\cdot|s_i)$ (resp. $\pi(\cdot|s)$) for the probability measure over $\Theta \times S$ (resp. Θ) conditional on s_i (resp. s).

Let $\mathcal{P}$ denote the set of priors over $\Theta \times S$ endowed with the following metric $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$: for any $\pi, \pi' \in \mathcal{P}$,

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

We consider the following class of information perturbations.

Definition 3 *We say a sequence of priors* $\left\{\pi^k\right\}$ *is a* ***private-value perturbation*** *to* π^{CI} *(denoted by* $\pi^k \to \pi^{CI}$*) if* $d\left(\pi^k, \pi^{CI}\right) \to 0$ *and for all* $i \in \mathcal{I},\ \theta \in \Theta$*, and* $s_{-i} \in S_{-i}$*,*

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

Observe that the perturbations invoked in proving Theorems 1 and 2 of Aghion et al. (2012) are indeed private value perturbations. Both theorems of Aghion et al. (2012) and our model are concerned with the private-values environment. 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. For instance, in our bilateral trade example in Section 2, it is conceivable that aside from his opponent's investment, there is also some random/idiosyncratic factor that determines player i's value/cost. Hence, while player i is perfectly informed of his own preference after a state is realized, the other player $j \neq i$ may only have an approximate and imperfect idea about the preference of player i. Moreover, private-value perturbation is vacuously satisfied when each player's signal perfectly identifies his own payoff types as in Bergemann and Morris (2005).

Let $\Gamma(\pi)$ be the incomplete information game induced by a two-stage mechanism Γ under prior π. A pure strategy of player i now maps each $s_i \in S_i$ to a pair (σ_i^1, σ_i^2) such that $\sigma_i^1 \in M_i^1$ and $\sigma_i^2 \in \times_{m^1 \in M^1} M_i^2(m^1)$. Again, let Σ_i denote the set of pure strategies of player i, $\Sigma = \times_{i \in I}\Sigma_i$ be the set of pure strategy profiles, and $z(\sigma(s))$ be the unique terminal history induced by strategy profile σ under signal profile s.

Here, a CPS μ_i specifies for each nonempty subset E of $\Theta \times S_{-i} \times \Sigma_{-i}$ a distribution $\mu_i[\cdot|E]$ over $\Theta \times S_{-i} \times \Sigma_{-i}$ with the property that Bayes' rule applies whenever possible (see Appendix 7.4 for a precise formulation). Recall that $\Sigma_{-i}(h) \subset \Sigma_{-i}$ is the set of strategy profiles of player i's opponent that are consistent with history h. Now conditional on history h, using strategy σ_i and holding CPS μ_i, player i's expected payoff is computed as follows:

$$V_i(\sigma_i, s_i, \mu_i | h) = \sum_{\theta, s_{-i}, \sigma_{-i}} \{u_i(g(z(\sigma_i(s_i), \sigma_{-i}(s_{-i})); h), \theta_i) + \tau_i(z(\sigma_i(s_i), \sigma_{-i}(s_{-i}))\} \mu_i\left[(\theta, s_{-i}, \sigma_{-i}) | \Sigma_{-i}(h)\right].$$

We extend the notion of sequential rationality to incomplete information environments.

Definition 4 (Sequential Rationality under Incomplete Information) *Fix a player* $i \in I$*,* $s_i \in S_i$*, a CPS* μ_i*, and a strategy* $\sigma_i \in \Sigma_i$*. Say that* σ_i *is a* ***sequential best response***

to μ_i for s_i iff for every $h \in \mathcal{H}$, we have

$$V_i(\sigma_i, s_i, \mu_i|h) \geq V_i(\sigma_i', s_i, \mu_i|h), \forall \sigma_i' \in \Sigma_i.$$

We say a CPS μ_i is *consistent with* s_i if for every history h, $\text{marg}_{\Theta \times S_{-i}} \mu_i (\cdot|\Sigma_{-i}(h)) = \pi(\cdot|s_i)$. The following two definitions are the counterparts of Definitions 1 and 2 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) *Fix a two-stage game form $\Gamma(\pi)$. The set of initial rationalizable messages of player i with signal s_i is defined as $R_i(s_i|\Gamma(\pi)) = \bigcap_{k=1}^{\infty} R_{i,k}(s_i|\Gamma(\pi))$ where set $R_{i,0}(s_i|\Gamma(\pi)) = \Sigma_i$ and, inductively, for every integer $k \geq 1$,*

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

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 finite mechanism $\Gamma = (M, g)$ such that for any $\theta \in \Theta$, the following two conditions hold: (i) $R^{\Gamma(\theta)} \neq \varnothing$; (ii) for any $s^\theta \in S$, any private-value perturbation $\{\pi^k\}$ to π^{CI}, and any sequence of strategy profile $\{\sigma^k\}_{k=1}^{\infty}$ such that $\sigma^k \in R(s^\theta|\Gamma(\pi^k))$ for each k, we have $g\left(z\left(\sigma^k\right)\right) \to f(\theta)$ as $k \to \infty$ and $\tau_i\left(z\left(\sigma^k\right)\right) = 0$ for each k.*

Despite the restriction on private-value perturbation, the robustness notion is formulated with the permissive solution concept of initial rationalizability and we also allow each player's CPS to have any degree of correlations among player's strategies, other players' signals, and the payoff type profiles. We could also weaken the notion in requiring only $\text{marg}_{\Theta_i} \pi^k\left[\theta_i | s_i^\theta, s_{-i}\right] \to 1$ if we adopt a solution concept where each player's strategy only depends on his own signal but not on the payoff type profile such as the notion of sequential

equilibrium defined in the online appendix of Aghion et al. (2012). We are now ready to state the main result of this section.

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

In contrast to the negative result of Aghion et al. (2012) regarding the MR mechanism, here we obtain a permissive robust implementation result with respect to private-value perturbations. This is possible because we take advantage of the simultaneous move in the two-stage game and make full use of stochastic mechanisms. The proof of Theorem 2 is provided in Appendix 7.6.

5.5 Additional Robustness Results and Discussion

5.5.1 Renegotiation

Renegotiation-proofness is an important issue in subgame-perfect implementation and it is considered in detail in Maskin and Moore (1999) and Maskin and Tirole (1999). When there are only two agents and the efficiency frontier is linear, such as our experimental setting with risk-neutral agents, Maskin and Moore (1999) show that any social choice function that is partially implementable in Nash equilibrium with renegotiation must also be fully implementable in Nash equilibrium with renegotiation in a *direct* mechanism. In other words, it is unnecessary to appeal to indirect mechanisms for full implementation with renegotiation. In contrast, when the two agents are strictly risk-averse, Maskin and Tirole (1999) show that a modified version of the MR mechanism implements any social choice function in a renegotiation-proof fashion. The SR mechanism can also be made renegotiation-proof with strict risk aversion by invoking a similar modification. We provide the details in Appendix 7.7.

5.5.2 Expected Utility Hypothesis

Our result holds even if the agents are not expected utility maximizers. More precisely, suppose first that each agent i's utility function over lotteries $u_i(\cdot, \theta_i)$ is monotone in the sense that shifts in probability mass from less preferred to strictly preferred alternatives yield a

lottery which is strictly preferred. For instance, this is the case if the agents are probabilistically sophisticated in the sense of Machina and Schmeidler (1992). Second, instead of only assuming that different types have different preferences over lotteries $\Delta(A) \times \mathbb{R}$ suppose that different types have different preferences over pure allocations $A \times \mathbb{R}$. Then, by fixing priority for entering arbitration, we can modify the SR mechanism to deal with the case where the agents are only probabilistically sophisticated. We provide the details in Appendix 7.8.

5.5.3 Retaliation-Proof Equilibrium

Recently, FPW (2017) consider an implementation problem where agents care about both their material payoffs and retaliating against perceived unkind acts. FPW (2017) show that such retaliation behaviors undermine subgame-perfect implementation using the MR mechanism. We examine the possibility of retaliation behaviors in the SR mechanism. In the SR mechanism, the first-stage announcement is made simultaneously. Therefore, when the buyer moves to the second stage after the seller tells the truth, the buyer may still perceive the seller's behavior as kindness to him if the buyer believes that the seller believes the buyer told the truth in the first stage. We call such a belief as a *reasonable* belief. Given such a reasonable belief, the buyer will still follow the material incentive to tell the truth in the arbitration stage instead of choosing to revenge. In Appendix 7.9, we show that truth-telling remains the unique *retaliation-proof equilibrium* defined in FPW (2017), whenever agents hold such reasonable belief and care more about their own material payoff than their reciprocity payoff from retaliation. In this sense, the SR mechanism is robust to retaliation behaviors.

In contrast, the MR mechanism has no room for such reasonable belief to be sustained, as the previous history is solid evidence for unkindness in a perfect-information game. Specifically, consider a state (v, c) and a history where the buyer announces $v' \neq v$ followed by the seller calling the arbitrator. It is clear that when the seller calls the arbitrator, the seller knows that the buyer will suffer from the penalty $T = 300$. This may in turn make the buyer perceive the seller's previous move as being unkind. Then, in the arbitration stage, the buyer may misreport his type to retaliate against the seller rather than rewarding the seller by telling the truth. This observation turns out to be consistent with behavior in the MR mechanism reported in FPW (2017) and in the SPI treatment discussed in detail in the appendix.

5.5.4 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. To recap, we choose the dictator lotteries $l_{i^*}(\cdot)$, the incentive transfers T, and the arbitration fee F so that (1') sequential rationality ensures that player i^* will truthfully announce his type in the second stage (i.e., the Truth-Telling Condition holds); (2') the first-round deletion of strictly dominated strategies ensures that each player i wants to match his report on the type of player $(i-1)$ with the second stage report chosen by player $(i-1)$ (i.e., the Inter-stage Coordination Condition holds); (2") the second-round deletion of strictly dominated strategies ensures that each player i wants to match his report on his own type with the report chosen by player $(i+1)$ (i.e., the Within-Stage Coordination Condition holds). 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.

5.5.5 Perfect-Information Mechanism

A natural question is whether we can prove Theorem 1 using a simpler mechanism. Here, in particular, we argue that it is impossible to construct a two-stage, perfect-information, finite mechanism that fully implements any social choice function even in the stronger solution concept of subgame-perfect equilibrium. Consider our experimental setup. Recall that under the efficient SCF f, trade occurs with probability one at any state. Suppose instead that there exists a finite two-stage sequential mechanism (M, g) which fully implements f. For the sake of argument, we assume that S moves first. However, it is not difficult to see that a similar argument applies if B moves first.

Consider a pair of states (v, c) and (v, c') where S with cost c' strictly prefers $f(v, c)$ to $f(v, c')$.[27] Since the mechanism implements f, at state (v, c), there must be some equilibrium

[27]For example, S strictly prefers $f(200, 130)$ to $f(200, 80)$ where the social function is associated with trading prices of 165 and 115 at $(200, 130)$ and $(200, 80)$ respectively.

message m_S of S and B's best response m_B such that $g(m_S, m_B) = f(v, c)$. At state (v, c'), since B has the same value, m_B is still a best response after the history m_S occurs. Since the mechanism is finite, at state (v, c'), there is a subgame-perfect equilibrium where m_B is chosen after m_S is chosen. Since S can secure $f(v, c)$ by choosing m_S, in this equilibrium she must be able to get a strictly better outcome than $f(v, c')$. In other words, $f(v, c')$ cannot be the equilibrium outcome, which contradicts the hypothesis that f is implementable.

6 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".[28]

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. (2014) and Aghion et al. (2017) illustrate them as a practical matter in laboratory settings.

The key issue is that extensive-form mechanisms given 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 be-

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

liefs at the beginning of the game. Crucially, it makes no assumption about how beliefs evolve after zero probability events. This makes it the weakest rationalizability concept for extensive-form games.

As a theoretical matter, our mechanism is robust to moderate levels of reciprocity and to small amounts of incomplete information about the state of nature. In laboratory experiments, we show that the mechanism induces efficient investment in a two-sided hold-up problem with ex-ante investment and performs better than both the three-stage Moore-Reupllo mechanism and a one-stage mechanism introduced by Kartik, Tercieux and Holden (2014). We also show that the mechanism can be made renegotiation proof if the agents are strictly risk averse and we highlight the robustness of the mechanism to a wide variety of reasoning processes and behavioral assumptions.

Our mechanism performs very well experimentally. Buyers make truthful first-stage value and cost reports in 92.6 percent of cases. Likewise, sellers make truthful initial value and cost reports in 91.7 percent of cases. Buyers choose the optimal level of investment in 89.6 percent of cases while sellers choose the optimal level of investment in 83.3 percent of cases. In aggregate, 87.1 percent of dyads improve their performance relative to the theoretical no-mechanism benchmark and 72.9 percent of dyads exhibit first-best investments and truth-telling behavior.

Moreover, we consider a treatment where we add an "opt-in stage" to the mechanism where both parties have the option to eliminate the mechanism and trade at a fixed price. We find that both buyers and sellers are willing to use the mechanism and that opt-in rates are above 75 percent for both parties. Groups that opt into the mechanism behave very closely to theory with 90.5 percent of dyads achieving the first best.

This seems particularly relevant to economic environments where parties determine what governance structure to use to mediate their interactions. Indeed, such considerations underpin the literatures on the theory of the firm and the design of within-firm governance structures.

In general, one would expect that when mechanisms work well, economic and other 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 bilateral trading setting 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 (1992a): "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica*, 60, 993–1008.

Aghion, P., M. Dewatripont, and P. Rey (1994): "Renegotiation Design with Unverifiable Information," *Econometrica*, 257–282.

Aghion, P., E. Fehr, R. Holden, and T. Wilkening (2017): "Subgame Perfect Implementation under Approximate Common Knowledge: Evidence from a Laboratory Experiment," *Forthcoming in Journal of the Eurpoean Economic Association*.

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.

Becker, G. M., M. H. DeGroot, and J. Marschak (1964): "Measuring utility by a single-response sequential method," *Behavioral Science*, 9, 226—232.

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

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.

Che, Y.-K. and D. B. Hausch (1999): "Cooperative Investments," *American Economic Review*, 89, 125–147.

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.

Chung, T.-Y. (1991): "Incomplete Contracts, Specific Investment, and Risk Sharing," *Review of Economic Studies*, 58, 1031–1042.

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

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.

Dufwenberg, M. and G. Kirchsteiger (2004): "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, 47, 268–298.

Dufwenberg, M., A. Smith, and M. Van Essen (2011): "Hold-Up: With a Vengeance," *Economic Inquiry*, 51.

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 (2014): "Handing Out Guns at a Knife Fight: Behavioral Limitations of Subgame Perfect Implementation," CESIFO Working paper No. 4948, CESIFO Group Munich.

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.

Grossman, S. J. and O. Hart (1986): "A Theory of Vertical and Lateral Integration," *Journal of Political Economy*, 94.

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 (1990): "Property Rights and the Nature of the Firm," *Journal of Political Economy*, 98, 1119–1158.

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.

Karni, E. (2009): "A Mechanism for Eliciting Probabilities," *Econometrica*, 77, 603—606.

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

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. Moore (1999): "Implementation and Renegotiation," *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.

Masuda, T., Y. Okano, and T. Saijo (2014): "The minimum approval mechanism implements the efficient public good allocation theoretically and experimentally," *Games and Economic Behavior*, 83, 73–85.

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.

7 Appendix

7.1 Behavior in the SPI mechanism

Result 4 *In the SPI mechanism,* 35 *percent of subjects lose money in the three paid periods against the computer. Earnings in these periods are negative on average and significantly below the earnings in the SR and KTH treatments.*

Figure 5 shows the average earnings that are generated in the three paid periods against the computer in the SR, SPI, and KTH treatments. As can be seen, average earnings is negative in the SPI treatment and significantly below the earnings of the other two treatments in a simple regression where average earnings is regressed against the treatment dummies (SPI vs. SR: p-value $< .01$; SPI vs. KTH: p-value $< .01$; SR vs. KTH: p-value $= 0.07$). Looking across individuals, 35 percent of subjects lose money against the computer in the SPI treatment and only 37.5 percent achieve the theoretical first best. This is in sharp contrast with (i) the SR treatment where 11.8 percent lose money and 58.1 percent achieve the first best and (ii) the KTH treatment where no subject losses money and 46.3 percent of subjects achieve the first best.

In the instructions for all treatments, the strategy taken by the computer was fully explained in the oral instructions. Subjects were told that in the SPI mechanism, the computer would always make a maximal investment, report their true value or true cost, challenge any report below the true value and above the true cost, and make choices in the counter-offer stage that maximize the computer's profit. The large proportion of subjects who lose money against the computer suggest that not all subjects fully understand the strategic incentives generated by the SPI mechanism. This is supported by the fact that subjects who lose money against the computer lose money at the beginning of the main experiment: subjects

who lose money against the computer also lose money in the first period 67.1 percent of the

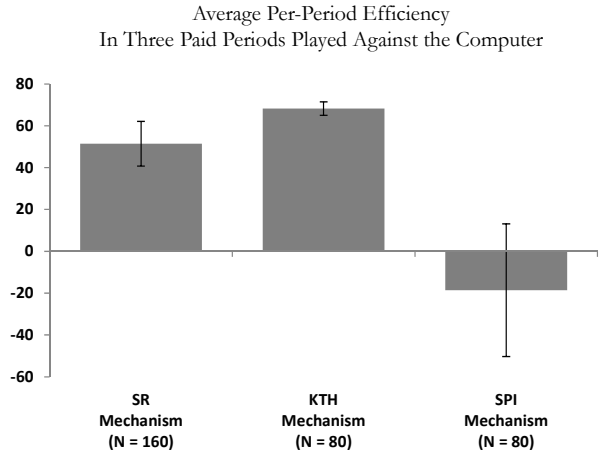


Figure 5: Average Per-Period Efficiency in the Three Paid Periods Played Against the Computer

We now describe aggregate behavior of subjects in the SPI mechanisms in Phase 1. We define a **advantageous lie** as a buyer announcement of value that is below the true value and a seller announcement of cost that is above the true cost. We will define a **false challenge** as a challenge of a truthful report, and a **legitimate challenge** as a challenge of an advantageous lie.

Result 5 *In periods 1-10, the SPI mechanism induces efficient investment in* 79.7 *percent of cases. Buyers make an advantageous lie in* 5.6 *percent of cases and make a false challenge in* 6.6 *percent of cases. Sellers make an advantageous lie in* 5.2 *percent of cases and false challenges in* 2.6 *percent of cases. However, subjects are reluctant to make legitimate challenges and such challenges are rejected in the majority of cases. The high proportion of disagreements coupled by losses in the periods against the computer lead to* 20 *percent of subjects going bankrupt.*

Figure 6 displays the pattern of behavior we observed in the first ten periods of the SPI treatment. The left hand panels shows the behavior of the buyers while the right hand

panels show the behavior of the sellers. Panel (a) summarizes the investment decision of both parties, Panel (b) shows the proportion of truthful reports, Panel (c) summarizes challenge behavior, and Panel (d) shows the proportion of challenges that are accepted after both a false and legitimate challenge. Finally, Panel (e) shows the aggregate number of lies and false challenges made over the 10 periods.

As can be seen in Panel (a), 76.3 percent of buyers and 82.9 percent of sellers exert an efficient level of investment. The proportion of buyers making an optimal investment is increasing over time, with 55.0 percent of buyers putting in optimal investment in the first period and 90.0 percent of buyers putting in optimal investment in period 10. Likewise, the proportion of seller making an optimal investment is increasing over time, with 72.5 percent of sellers making an optimal investment in the first period and 88.6 percent of sellers making an optimal investment in period 10.

Panel (b) and (c) show the proportion of buyers and sellers who make truthful reports and false challenges. As can be seen on the left hand side of these panels, buyers make a truthful announcement in 94.4 percent of cases and an advantageous lie in 5.6 percent of cases. Buyers also make a false challenge in 6.6 percent of cases. However, they make legitimate challenges in only 55.2 percent of cases. This suggests that some buyers are reluctant to make legitimate challenges.

Sellers make truthful announcements in 96.4 percent of cases and advantageous lies in 3.6 percent of cases. They make a false challenge in only 2.6 percent of cases. Sellers are also reluctant to make legitimate challenges and do so in only 55.6 percent of cases.

As can be seen in Panel (d), buyers and sellers are rightfully wary of making legitimate challenges. Buyers reject legitimate challenges in 77.8 percent of cases while sellers reject legitimate challenges in 62.5 percent of cases. Thus, it appears that buyers and sellers who enter into the arbitration stage are willing to forego their pecuniary incentives in order to reduce the payoff of their matched partner. Here, the rejection rates are high enough that if a buyer or seller was risk neutral and knew the empirical rejection rate of legitimate challenges, it would not be in their pecuniary interest to challenge.

Finally, Panel (e) shows the aggregate number of lies or false challenges that different buyers and sellers take over the first 10 periods of the experiment. The dark grey steps represent the ten buyers and five seller who went bankrupt in the first 10 periods and whose lie frequencies are truncated. Similar to the SR mechanism, over 75 percent of buyers and

sellers make one lie or less. However, buyers and sellers lies tend to be more persistent: buyers and sellers who make an advantageous lie have a 89 percent chance of making a lie in the next period if they are not challenged and have a 22 percent chance of lying if they are challenged. Further, a buyer or a seller who makes a false challenge in one period and does not go bankrupt has a 66 percent chance of making a false challenge in the next period if the counter-offer in the current period is accepted and a 55 percent chance of making a false challenge in the next period if the counter-offer in the current period is rejected.

In aggregate, the persistence of lies along with the losses that buyers and sellers incur in the pre-period stage leads 20 percent of our subjects to go bankrupt. This is roughly the same proportion of buyers and sellers who lie in each period of the SPI mechanism discussed in AFHW and is smaller than the proportion of buyers who lie in every period of the Main Treatment in FPW.

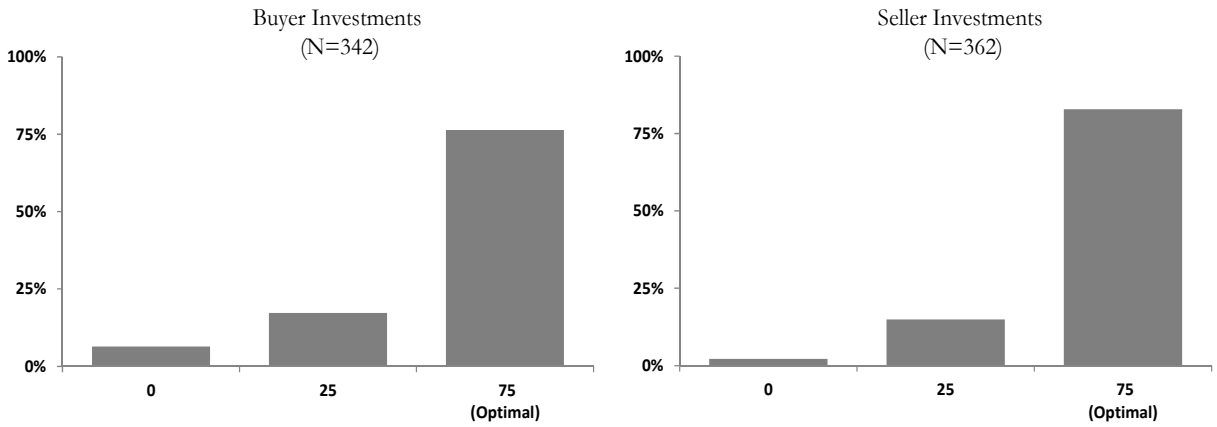
We now turn to behavior in periods 11-20, noting that the data here is a highly selected sample due to the high level of bankruptcies.

Result 6 *Buyers opt into the mechanism in* 77.5 *percent of cases while sellers opt into the mechanism in* 72.5 *percent of cases. Opt-in rates are increasing for both buyers and sellers.* 87.0 *percent of dyad pairs who opt into the mechanism exhibit efficient truth-telling behavior and achieve the efficient outcome.*

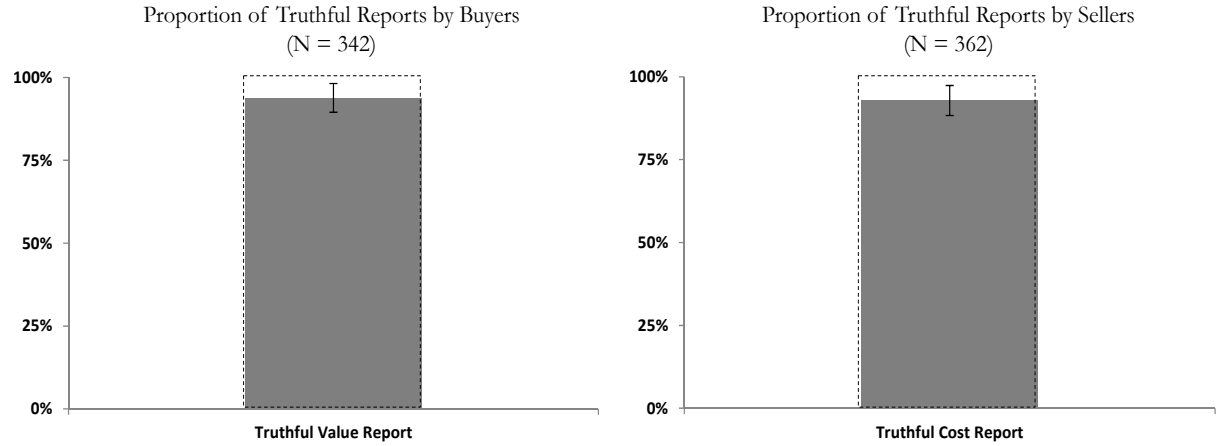
Figure 7 shows opt-in rates for buyers and sellers in Phase 2 of the SPI mechanism. As can be seen, opt-in rates for both buyers and sellers are increasing with opt-in rates near 50 percent early in the sample and near 75 at the end of the sample. Dyads who opt into the mechanism reach the efficient outcome over 90 percent of the time and buyers and sellers make truthful reports in all but 5 cases. All 5 lies are challenged and four of the five challenges end in rejections. Investments in groups that opt out of the mechanism decrease over time just as in the SR treatment.

Comparing the results here to the main text, it is clear that the SPI and SR mechanisms are similar in terms of efficiency in Phase 2 of the experiment but not Phase 1. At least in the current environment, the SR treatment's main advantage is that it is easier to understand by participants and subjects are less likely to incur early losses and end up going bankrupt. The SR mechanism also has the promising feature that truthful reporting in the first and second stage is a best response to the empirical distribution of counter-party be-

(a) Distribution of Investment Choices in Periods 1-10



(b) Proportion of Truthful Reports in Periods 1-10



(c) Challenges after Truthful Reports and Lies in Periods 1-10

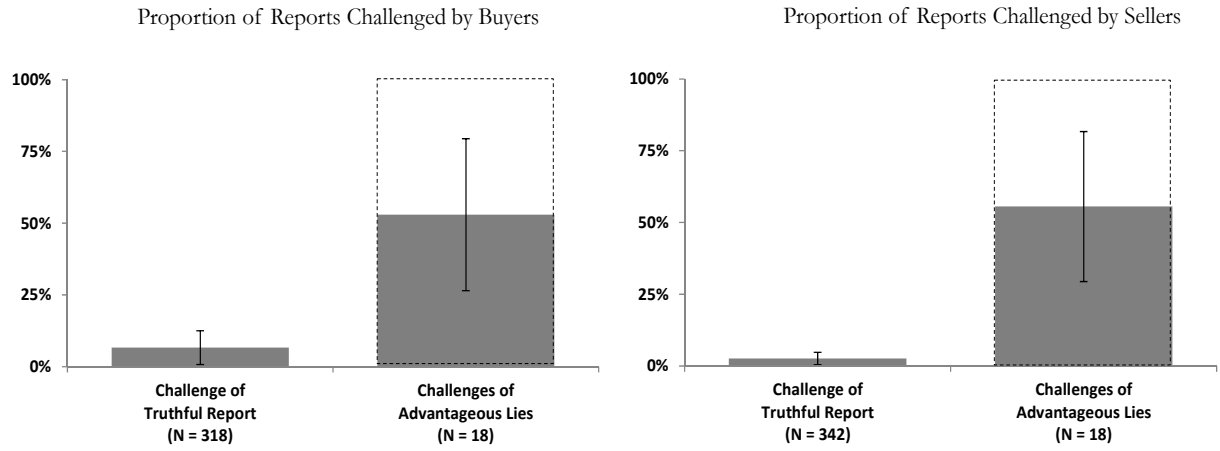
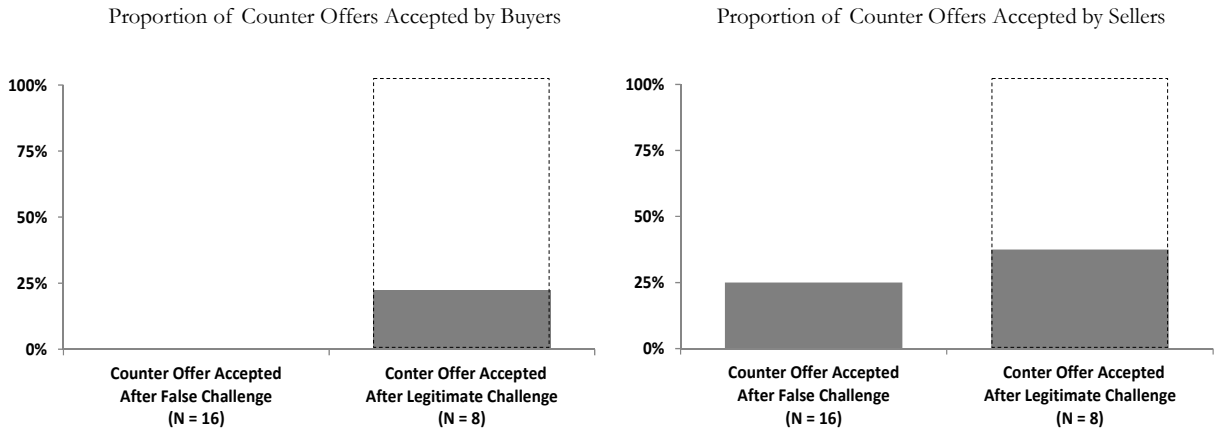
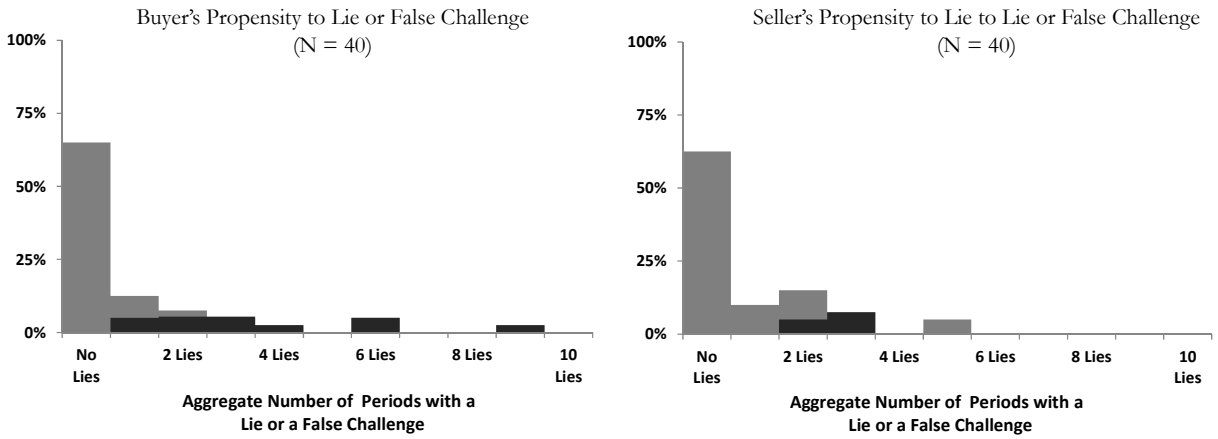


Figure 6: Pattern of Play in First 10 Periods of SPI Mechanism

(d) Acceptance of Counter Offers in Periods 1-10



(e) Aggregate Number of Lies or False Challenges in Periods 1-10



havior whereas in the SPI mechanism, buyers and sellers do not have a pecuniary incentive to make legitimate challenges. This finding is consistent with behavior in FPW where in a similar SPI mechanism buyers retaliate against legitimate challenges and sellers have a negative expected value for triggering arbitration.

7.2 Behavior in the KTH mechanism

Result 7 *In periods 1-10, the KTH mechanism induces efficient investments in only* 41.5 *percent of cases. The mechanism induces truthful reports in only* 50 *percent of cases. Both investments and truthful reports are decreasing over time.*

Figure 8 reports the pattern of behavior observed in periods 1-10 of the KTH mechanism. As with earlier figures, the behavior of buyers is shown in the left panels and the behavior of sellers is shown in the right panels. Panels (a) and (b) summarize the investment decisions of both parties, Panels (c) and (d) show the proportion of truthful reports, and Panel (e) shows the aggregate number of lies.

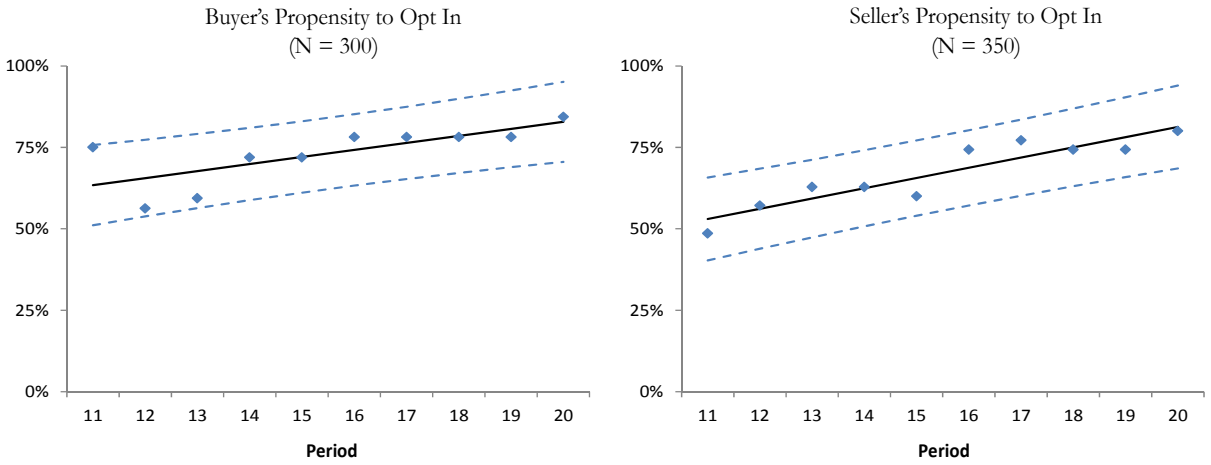
As can be seen in Panel (a), buyers make an optimal investment in 43.8 percent of cases and sellers make an optimal investment in 39.3 of cases. These proportions are much lower than those observed in the SR treatment and the SPI treatment. As seen in Panel (b), the proportion of subjects who make optimal investment is decreasing over time, with only 30 percent of buyers and 27.5 percent of sellers making optimal investments in period 10.

Panel (c) reveals that the mechanism fails to induce truthful reports for both buyers and sellers. Looking at the left side, buyers make truthful value reports in only 62.0 percent of cases and truthful cost reports in 80.3 percent of cases. Sellers make truthful value reports in 68.5 percent of cases and truthful cost reports in only 58.5 percent of cases. As seen in Panel (d) the frequency of truthful cost and value reports is decreasing for sellers and is not increasing for buyers.

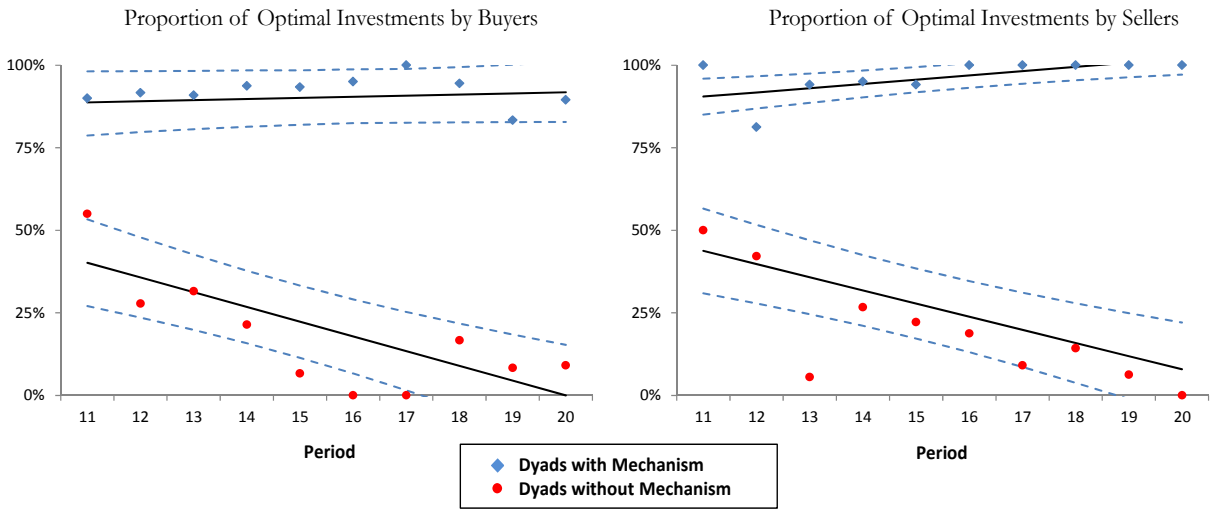
Finally, Panel (e) reveals strong heterogeneity in truth-telling behavior across the sample. Less than 10 percent of sample make truthful reports in all periods. Thus, the mechanism fails at inducing truth telling for almost all subjects.

To understand why lies are so prevalent in the data, it is useful to look at the action profiles of individual subjects. A feature of the data is that subjects who lie typically do so in a way that benefit them if there is a small chance that the other party makes a mistaken

(a) Opt-in Rates in Periods 11-20



(b) Optimal Investment Rates in Periods 11-20



(c) Proportion of Truthful Reports in Periods 11-20 with Mechanism

Truthful Reports by Buyers	159 of 159
Truthful Reports by Sellers	183 of 186
Challenges of Advantageous Lies by Buyers	3 of 3
Challenges of Advantageous Lies by Sellers	0 of 0
False Challenges by Buyers	1 of 159
False Challenges by Sellers	1 of 186
Counter Offer Accepted by Buyer After Lie	No Obs
Counter Offer Accepted by Buyer After Truth	0 of 1
Counter Offer Accepted by Seller After Lie	1 of 3
Counter Offer Accepted by Seller After Truth	0 of 1

Figure 7: Pattern of Play in Periods 11-20 of SPI Mechanism

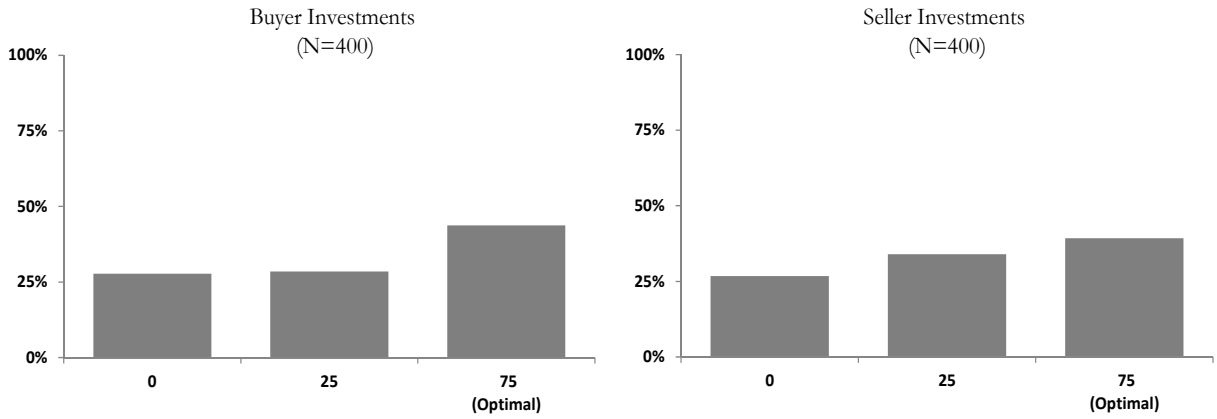
report. Buyers overstate their investment by reporting a cost below the true cost in 14.5 percent of cases and understate their investment in only 5.5 percent of cases. Likewise, sellers overstate their investment by reporting a value above the true value in 28 percent of cases and understate their investment in only 3.5 percent of cases. Overstating investment can increase the expected profit of a subject if (as in the data) there is a positive probability that their matched partner will match their misreport and cannot hurt a subject relative to telling the truth. However, they are extremely costly strategies for the counter party: whereas buyers who overstate their investment earn 22.4 ECU on average, their matched partners lose 59.0 ECU on average. Likewise, sellers who overstate their investment earn 38.9 ECU on average while their matched partners earn -67.6 ECU on average.

Buyers and sellers also tend to lie in the report that does not directly affect their payout in a way that hurts their matched partners. Buyers under report the value in 29.0 percent of cases and over-report the value in only 8.8 percent of cases. Sellers over report the cost in 32.8 percent of cases and under-report the cost in only 8.8 percent of cases. As it is only possible to under report values or over report costs when matched with a partner who has chosen to invest, lies in the buyer value report and the seller cost report reduces the expected value of investing. As a result, buyers who make an efficient investment and report truthfully earn 15.7 on average while sellers who make an efficient investments and report truthfully earn 24.2. These profits are strictly below the average profit from not investing and overstating one's investment.

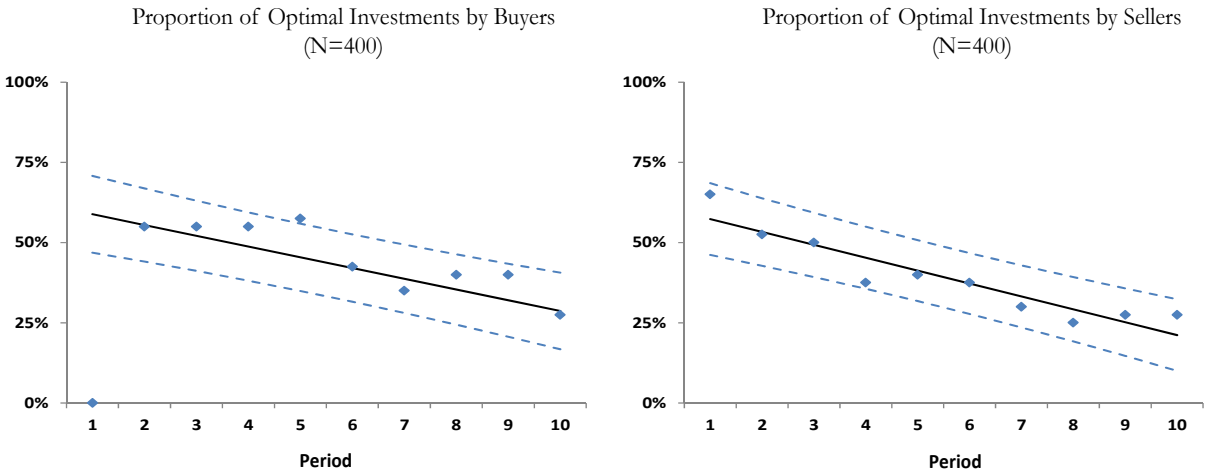
In fact, for a selfish buyer who does not have a preference for honesty, all strategies that are a best response to the empirical distribution involve an investment of zero and a cost report of 10 at all histories that occur with positive probability. For a selfish seller who does not have a preference for honesty, all strategies that are a best response to the empirical distribution involve an investment of zero and a value report of 320 at all histories that occur with positive probability.[29]

[29]It can also be shown that if one uses Agent Quantal Response Equilibrium (AQRE) as an equilibrium concept, the probability that buyers and sellers make zero investment and maximally overstate their investment goes to one as noise approaches zero. In contrast to the assumption made in KTH that subjects report honestly when indifferent, the AQRE assumes that buyers and sellers randomize uniformly over strategies where they are indifferent. This implies that buyers choosing the efficient truth-telling strategy will match with sellers who over-report their costs. Such matches lower the expected value of investment and truthful reporting. Non-investing buyers who lie may end up matched with sellers who under-report costs leading to an increase in the expected value of strategies involving lies and overstated investments. Models that combine AQRE with a preference for honesty rationalize the data well, though they cannot explain the asymmetry in the buyers' value reports and the sellers' cost reports without a force such as reciprocity that

(a) Distribution of Investment Choices



(b) Proportion of Optimal Investment Decisions over Time



(c) Proportion of Truthful Reports

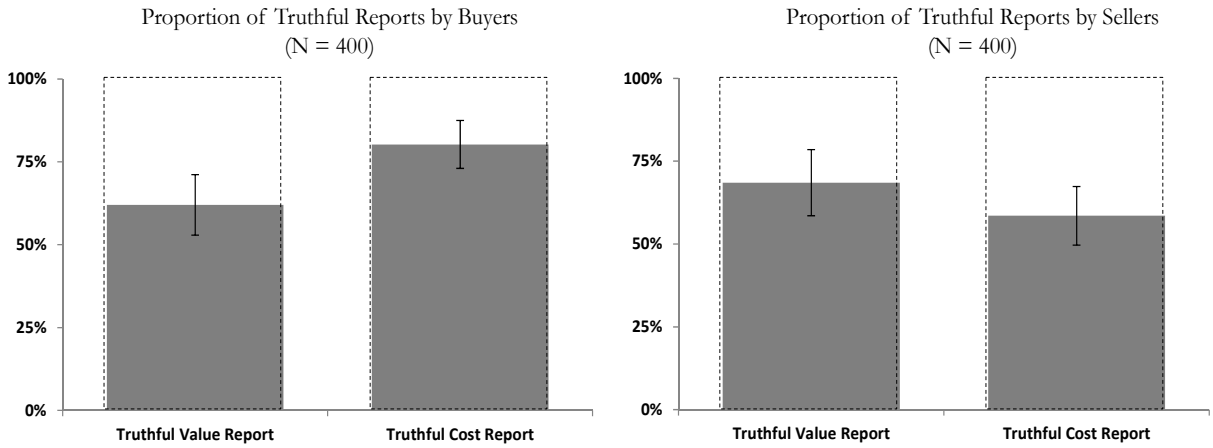
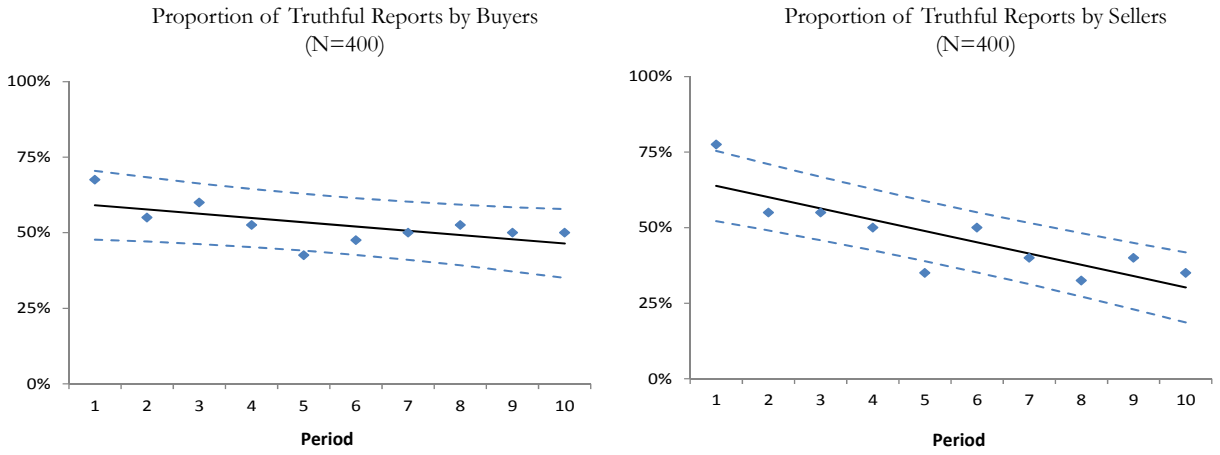
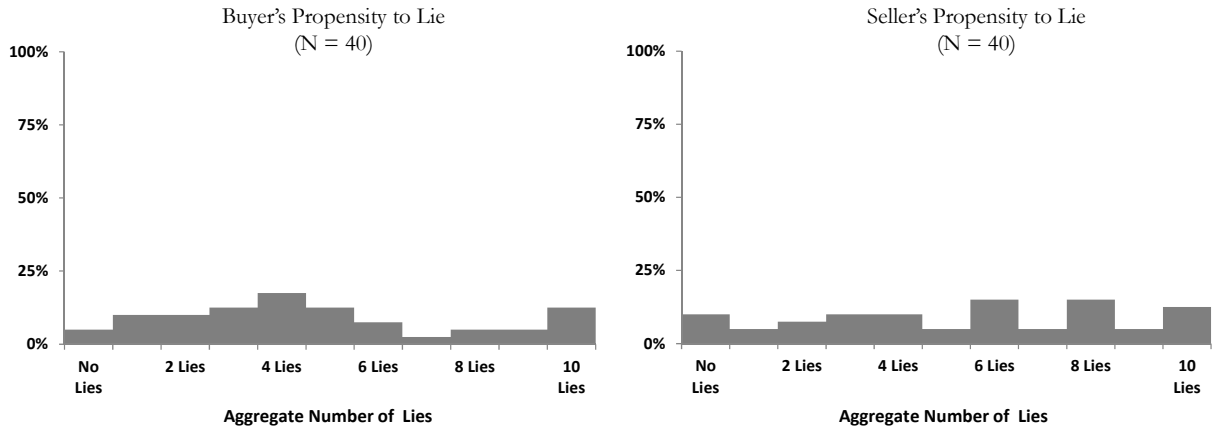


Figure 8: Pattern of Play in First 10 Periods of KTH Mechanism

(d) Proportion of Truthful Reports over Time



(e) Aggregate Number of Lies in Periods 1-10



The poor performance of the mechanism in periods 1-10 foreshadows the opt-in behavior in periods 11-20:

Result 8 *Buyers and sellers retain the mechanism in only* 20 *percent of cases. Groups that retain the mechanism have lower average profits than those who dismiss the mechanism.*

Buyers opt into the KTH mechanism in 35.5 percent of cases while sellers opt into the mechanism in 57.0 percent of cases. Opt in rates are increasing for both buyers and sellers but remain relatively low throughout the time series. Of the 400 observed dyads, only 80 of them retain the mechanism.

While groups that retain the mechanism in the SR and SPI mechanism tend to perform very well, groups in the KTH mechanism continue to perform worse than the no mechanism benchmark. Buyers choose efficient investment in 50 percent of cases, make truthful announcements in only 52.5 percent of cases, and earn $-.9$ ECU on average. Sellers choose efficient investment in only 35.0 percent of cases, make truthful announcements in only 41.3 percent of cases, and earn 37.4 ECU on average. On average, earnings of subjects in dyads that retain the mechanism are 21.9 ECU lower than individuals in groups without the mechanism, a difference that is significant in a simple regression where profit is regressed against a dummy that is one if the mechanism is retained and zero if the mechanism is dismissed (p-value $< .01$; errors clustered at the individual level).

In aggregate, the KTH mechanism is sensitive to systematic lies which attempt to take advantage of mistakes by the counter-party, but which are detrimental to aggregate welfare. Investments are falling over time and lies are increasing suggesting that the mechanism is unraveling over the course of the experiment. When given the chance, the majority of subjects choose to opt out of the mechanism and those who retain the mechanism lose money relative to groups where the mechanism is eliminated.

7.2.1 Discussion

In our variant of the KTH mechanism, we set the fine to be exactly equal to the marginal gain associated with an advantageous lie. Our fee structure implies that the buyer is strictly indifferent to all cost reports less than or equal to the seller's cost report while the seller is indifferent over all cost reports. By the construction of the fines, a buyer who makes an

generates disutility from taking actions that reward counterparties who lie.

efficient investment (i.e., the case where the true cost is 10) strictly prefers to report the true cost if he has a preference for honesty or believes there is a small probability that the seller reports the true cost. A drawback of our fine structure, however, is that a buyer who makes no investment and who has no preference for honesty is indifferent between all reports when the seller is truthful and may strictly prefer lies if he believes the seller is prone to mistakes.

In the original KTH construction, the authors consider a fine where the punishment exceeds the total gain associated with an advantageous lie. An advantage of the alternative approach is that buyers who make no investment have a strict preference to tell the truth if they believe that a large proportion of sellers have a preference for honesty and will report the true cost of 130. Thus, it is more likely to be robust to rent seekers who seek to exploit the mistakes of others. A disadvantage of the original approach, is that a buyer who makes an efficient investment may make a report above the true cost if the buyer is uncertain about the seller's cost reports.

Ex-ante, we felt that it was more important to design a mechanism that protected buyers and sellers who invested efficiently with the view that such mechanisms are more likely to be robust to communication between the investment stage and the report stage. This appears to be effective, as buyers and sellers who invest efficiently indeed almost always report truthfully in our KTH treatment. However, it is clear that our current implementation is not robust to attempts at rent seeking when the subjects do not invest efficiently. It is likely that a mechanism with fines that are slightly larger than the ones used in the current treatment could reduce attempts at rent seeking and have the potential to improve the mechanism relative to the variant described here.[30]

7.3 Efficiency Measures under Alternative Approaches for Dealing with Bankruptcy

In the main text, we used the average per-period earnings of each individual over the entire 20 period experiment as our main efficiency measure. For subjects who went bankrupt, we set their average per-period earnings equal to -38.5, which when multiplied by 20 is equal

[30] We should however note that more than 40% of the matched partners of the subjects who invest efficiently do not report the truth: sellers whose partner invests 75 report a cost of 10 in only 125 out of 198 cases (58.1%). The large number of counter-party misreports is at odds with preference for honesty and suggests that the alternative KTH mechanisms with higher fines may also have issue achieving the first-best investment due to the risky coordination.

to the amount that could be lost before the subject was dismissed from the experiment.

While we believe our "Original" method provides a simple measure of relative efficiency, readers may be concerned that it does not accurately reflect the impact that these subjects might have on future interactions if they remained in the sample. This section provides efficiency measures under two alternative methods for dealing with bankruptcy. In the first "Switch Rate" method, we estimate the probability that a subject switches between lying strategies and truth-telling strategies and use this estimate to construct a Markov transition matrix that can be used to predict the future behavior of subjects who go bankrupt. The underlying assumption of this approach is that bankrupt subjects are not fundamentally different from subjects who began the experiment by lying but eventually adopted a truth-telling strategy. Our second "Always Lie" method assumes that bankrupt subjects will lie in all periods following their bankruptcy. This is, in a sense, a worse-case scenario where bankrupt subjects generate losses for both themselves and their matched partner in every period.

Our switch rate method calculates efficiency as follows: for each period, we calculate the probability that an individual who is lying in period t will switch to telling the truth in period $t+1$ using the empirical switch rates of all subjects who lie in period t and do not go bankrupt. In periods where there are no observed lies, we interpolate the switch probability using the closest two periods for which there is data. We also calculate the (very small) switch rate that a subject will move from truth telling to lying. For both the SR treatment and the SPI treatment, switch rates are reasonably stable over time with the highest switch rates occurring in early periods and slightly lower switch rates occurring in later periods. Using the switch rates, we calculate the probability that a bankrupt subject will lie in each period. We then calculate the expected value of all dyads that involve a bankrupt subject using the empirical expected returns from lying as a proxy for a dyads profit after a lie. We assume that bankrupt subjects always opt into the mechanism in periods 11-20 as this maximizes the impact of these subjects on the final outcome.

For our worst-case method we assume that a bankrupt subject lies in every single period over the entire sample. As above, we use the empirical return from lying to calculate the outcome for the subject and their matched pair and assume that bankrupt subjects always opt into the mechanism.

For clarity, Figure 9 shows what the aggregate distribution of lies under each of our

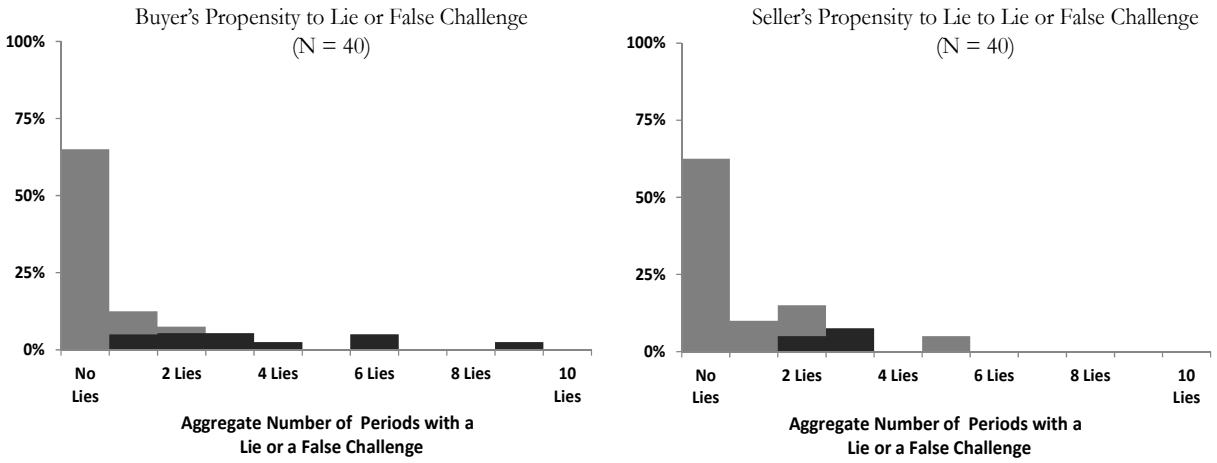
assumptions for the SPI treatment. In panel (a) we show the original aggregate distribution with bankrupt buyers and sellers highlighted. Panel (b) shows our switch rate method where, as can be seen, bankrupt subjects are distributed relatively evenly over each of the potential action profiles. Panel (c) shows our worst case scenario method. For the SPI treatment where bankruptcies are common, the resulting aggregate distribution is bimodal with buyers and sellers either lying infrequently or lying in almost all periods.

Table 5 shows the average per-period earnings using the original method, the switch rate method, and the worst-case scenario method. For the SR mechanism, the switch rate method generates a higher earnings estimate than the original fixed bankruptcy method. This is due to the fact that four of the six bankruptcies occur very early in the sample and these subjects are predicted to switch to truthful strategies relatively quickly. Given their low likelihood of lying, they impose only a small externality to their matched partners and increase their own earnings relative to the per-period loss of -38.5 assigned to them in the original method. For the SPI mechanism, the switch rate method predicts an average earning of 28.6. This estimate is below the earnings that we calculated in our original method because lies are more persistent in the SPI mechanism and expected losses after a lie are higher.

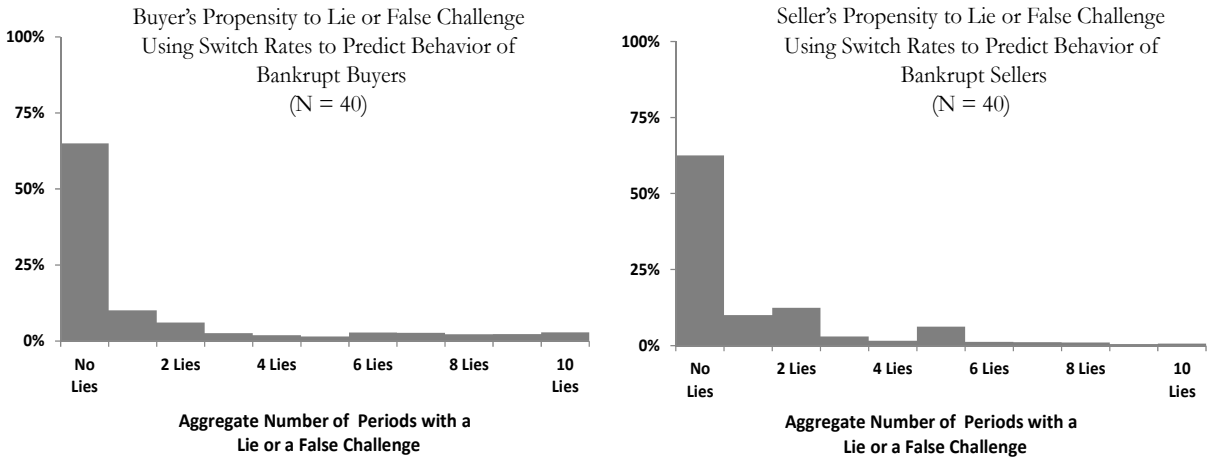
Using the worst-case scenario, subjects in the SR mechanism earn 43.6 ECU on average. This is not significantly different from earnings in the Fixed Price treatment (p-value $= .272$) but is significantly different from the theoretical benchmark prediction of 35 (p-value $< .01$) in a simple regression where the profits earned by a dyad pair are regressed against the treatment variables. By contrast, the earnings in the SPI treatment is only 4.5 ECU and significantly below the earnings in all other sessions (p-value $< .01$ in all comparisons).

Summarizing the results above, earnings in the SR treatment are robust to assumptions made about bankrupt subjects and the alternative methods of calculating efficiency do not change the ordering of this treatment relative to the other three treatments. We note, however, that the earnings estimate in the SPI treatment is more sensitive to the way in which bankruptcies are handled and overall efficiency of this treatment could potentially be quite low.

(a) Aggregate Number of Lies in SPI Mechanism in Periods 1-10



(b) Predicted Lie Distribution using Markov Switching Data to Predict Behavior of Bankrupt Subjects



(c) Predicted Lie Distribution Assuming Bankrupt Subjects Always Lie

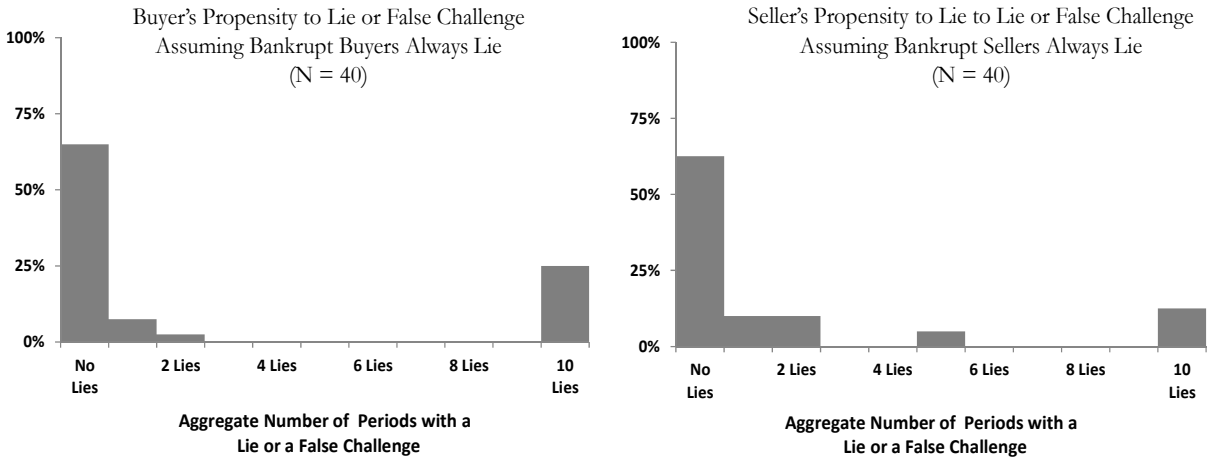


Figure 9: The Distribution of Aggregate Lies under Alternative Methods for Dealing with Bankruptcies

Table 5: Alternative Efficiency Measures

	SR Mechanism	SPI Mechanism
Original Method	47.9	35.5
Switch Rate Method	51.9	28.6
Always Lie Method	43.6	4.5

7.4 Conditional Probability System

Following Dekel and Siniscalchi (2015), we formulate the notion of CPS as follows.

Definition 7 *Fix a measurable space* $(\Omega, \mathcal{X})$ *and a countable collection* $\mathcal{B} \subset \mathcal{X}$. *A conditional probability system, or CPS, is a map* $\mu : \mathcal{X} \times \mathcal{B} \to [0,1]$ *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* $\mu(A|C) = \mu(A|B) \cdot \mu(B|C)$.

The set of CPSs on (Ω, Σ) *with conditioning events* $\mathcal{B}$ *is denoted* $\Delta^{\mathcal{B}}(\Omega)$.

In Section 5.2, we set $\Omega = \Sigma_{-i}$ and $\mathcal{B}$ to be the collection of all nonempty subsets of Σ_{-i}. In Section 5.4, we set $\Omega = \Theta \times S_{-i} \times \Sigma_{-i}$ and let $\mathcal{B}$ be the collection of all nonempty subsets of Ω.

7.5 Proof of Theorem 1

Let θ be the true state. We prove Theorem 1 in the following three steps as in Section 2.

7.5.1 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. ■

7.5.2 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.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 $\mu_i\left(R_{-i,1}^{\Gamma(\theta)} \middle| M_{-i}\right) = 1$. We fix such μ_i and $m^1 \in M^1$ as a message profile chosen at Stage 1. By Claim 1, it follows that for each $j \in \mathcal{I}$,

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

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 at Stage 2, always announces her true type. Given m_{-i} and $\hat{\theta}_i^i$ (i's announcement about her own type at Stage 1), let I^* be the number of player $(i-1)$'s opponent(s) who make an inconsistent announcement. 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 $D/(I^*+1)$.

We shall show that against any message profile $\sigma_{-i} \in R_{-i,1}^{\Gamma(\theta)}$ of player i's opponents, reporting $m_i^1 = (\hat{\theta}_i^i, \theta_{i-1})$ in the 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 in Stage 2 with probability $1/(I^*+1)$ (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 in Stage 2 with probability $1/(I^*+1)$. 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 $T/(I^*+1)$. Since $T > D$, it follows that $T/(I^*+1) > D/(I^*+1)$. This implies that reporting $\hat{\theta}_{i-1}^i = \theta_{i-1}$ in the first stage is strictly better for player i than reporting $\hat{\theta}_{i-1}^i \neq \theta_{i-1}$.

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

For player i, reporting $\hat{\theta}_{i-1}^i \neq \theta_{i-1}$ will result in the penalty T in Stage 2 with probability $1/(I^*+1)$, while reporting $\hat{\theta}_{i-1}^i = \theta_{i-1}$ will not induce any transfer. 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 $T/(I^*+1)$. Again, since $T > D$, we

obtain $D/(I^*+1) < T/(I^*+1)$ so that 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, reporting $(\hat{\theta}^i_i, \theta_{i-1})$ in the first stage is strictly better for player i than reporting any $(\hat{\theta}^i_i, \hat{\theta}^i_{i-1})$ with $\hat{\theta}^i_{i-1} \neq \theta_{i-1}$. We conclude that 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, $\sigma^1_i = (\hat{\theta}^i_i, \theta_{i-1})$ for some $\hat{\theta}^i_i \in \Theta$ ■

7.5.3 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 σ_i is a best reply to some CPS μ_i such that $\mu_i\left(R^{\Gamma(\theta)}_{-i,2} | M_{-i}\right) = 1$. We fix such μ_i. By Claim 2, μ_i has the following property:

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

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 her own type by announcing some $\hat{\theta}^i_i \neq \theta_i$, she will be penalized by F. Since $F > D$, player i's unique best response is to truthfully announce her own type in the first stage. Hence, every player i will truthfully report her type at the first stage, i.e., $\hat{\theta}^i_i = \theta_i$. Combining this with Claim 2, we conclude that $m^1_i = (\theta_i, \theta_{i-1})$. ■

7.6 Proof of Theorem 2

The proof of Theorem 2 consists of two claims below. Throughout the proof, we denote the true state as θ.

Claim 4 *Let* $\theta \in \Theta$, $m^1 \in M^1$, *and* $\{\pi^k\}$ *be a private-value perturbation to* π^{CI}. *For every* $i \in \mathcal{I}^*(m^1)$, *we have that* $\sigma^2_i(m^1) = \theta_i$ *for any* $\sigma_i \in R_{i,1}\left(s^\theta_i | \Gamma\left(\pi^k\right)\right)$ *and any* k *sufficiently large.*

Proof. Fix $k \geq 1$, $i \in \mathcal{I}$, and $\sigma_i \in R_{i,1}\left(s^\theta_i | \Gamma\left(\pi^k\right)\right)$. Observe that σ_i is a sequential best response against some CPS $\mu_{i,k}$ consistent with signal s^θ_i. We fix such μ_i. Consider any $m^1 \in M^1$ such that $i \in \mathcal{I}^*(m^1)$. Conditional on m^1, only m^2_i matters for player i's payoff; moreover, m^2_i matters only when $l_i(m^2_i)$ is chosen by the designer. Hence, by (8) in Lemma

1, it suffices to show that conditional on m^1, when k is chosen sufficiently large, $\mu_{i,k}$ assigns almost probability one to θ_i, namely,

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

This simply follows from Bayes' rule. Specifically, we compute the following:

$$\begin{aligned}
& \text{marg}_{\Theta_i}\mu_{i,k}\left(\theta_i|\Sigma_{-i}\left(m^1\right)\right) \\
= & \frac{\sum_{\theta_{-i},s_{-i}}\sum_{\sigma_{-i}\in\Sigma_{-i}(m^1)}\mu_{i,k}\left(\theta_i,\theta_{-i},s_{-i},\sigma_{-i}|s_i^\theta\right)}{\sum_{\theta_i',\theta_{-i},s_{-i}}\sum_{\sigma_{-i}\in\Sigma_{-i}(m^1)}\mu_{i,k}\left(\theta_i',\theta_{-i},s_{-i},\sigma_{-i}|s_i^\theta\right)} \\
= & \frac{\sum_{\theta_{-i},s_{-i}}\sum_{\sigma_{-i}\in\Sigma_{-i}(m^1)}\mu_{i,k}\left(\sigma_{-i}|\theta_i,\theta_{-i},s_{-i},s_i^\theta\right)\mu_{i,k}\left(\theta_i,\theta_{-i},s_{-i}|s_i^\theta\right)}{\sum_{\theta_i',\theta_{-i},s_{-i}}\sum_{\sigma_{-i}\in\Sigma_{-i}(m^1)}\mu_{i,k}\left(\sigma_{-i}|\theta_i',\theta_{-i},s_{-i}s_i^\theta\right)\mu_{i,k}\left(\theta_i',\theta_{-i},s_{-i}|s_i^\theta\right)} \\
= & \frac{\sum_{\theta_{-i},s_{-i}}\sum_{\sigma_{-i}\in\Sigma_{-i}(m^1)}\mu_{i,k}\left(\sigma_{-i}|\theta_i,\theta_{-i},s_{-i},s_i^\theta\right)\pi^k\left(\theta_i,\theta_{-i},s_{-i}|s_i^\theta\right)}{\sum_{\theta_i',\theta_{-i},s_{-i}}\sum_{\sigma_{-i}\in\Sigma_{-i}(m^1)}\mu_{i,k}\left(\sigma_{-i}|\theta_i',\theta_{-i},s_{-i}s_i^\theta\right)\pi^k\left(\theta_i',\theta_{-i},s_{-i}|s_i^\theta\right)} \\
= & \frac{\sum_{\theta_{-i},s_{-i}}\sum_{\sigma_{-i}\in\Sigma_{-i}(m^1)}\mu_{i,k}\left(\sigma_{-i}|\theta_i,\theta_{-i},s_{-i},s_i^\theta\right)\pi^k\left(\theta_i|s_i^\theta,s_{-i},\theta_{-i}\right)\pi^k\left(\theta_i,\theta_{-i},s_{-i}|s_i^\theta\right)}{\sum_{\theta_i',\theta_{-i},s_{-i}}\sum_{\sigma_{-i}\in\Sigma_{-i}(m^1)}\mu_{i,k}\left(\sigma_{-i}|\theta_i',\theta_{-i},s_{-i}s_i^\theta\right)\pi^k\left(\theta_i'|s_i^\theta,s_{-i},\theta_{-i}\right)\pi^k\left(\theta_i',\theta_{-i},s_{-i}|s_i^\theta\right)} \qquad (9)
\end{aligned}$$

where the third equality follows from the consistency of $\mu_{i,k}$ with s_i^θ. Finally, convergence in private values 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 (9) that $\text{marg}_{\Theta_i}\mu_{i,k}\left(\theta_i|\Sigma_{-i}\left(m^1\right)\right) \to 1$ as $k \to \infty$. ■

Claim 5 *Let $\theta \in \Theta$ and $\{\pi^k\}$ be a private-value perturbation to π^{CI}. Then, for any $i \in \mathcal{I}$, $k \geq 1$, and $\sigma_i \in R_{i,3}\left(s_i^\theta|\Gamma\left(\pi^k\right)\right)$, if k is sufficiently large, $\sigma_i^1 = (\theta_i, \theta_{i-1})$.*

Proof. First, note that $d\left(\pi^k,\pi^{\text{CI}}\right) \to 0$ as $k \to \infty$. This implies that $\pi^k\left(\theta, s_{-i}^\theta|s_i^\theta\right) \to 1$ as π^{CI} is a complete information prior. Thus, any player i with signal s_i^θ believes with probability close to 1 that all her opponents also receive signals s_{-i}^θ. 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 plays $\sigma_{i-1} \in R_{i-1,1}\left(s_{i-1}^\theta|\Gamma\left(\pi^k\right)\right)$ for k large enough.

Moreover, by Claim 2, player i finds it strictly better to report her predecessor's true type (i.e., $\hat{\theta}_{i-1}^i = \theta_{i-1}$) rather than tell a lie about it (i.e., $\hat{\theta}_{i-1}^i \neq \theta_{i-1}$). This strict better reply of telling the predecessor's true type as opposed to telling a lie about it remains the

same under the perturbed environment, so long as player $(i-1)$ reports type θ_{i-1} with probability close to one at 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 with type s_i^{θ} believes with probability close to one that all her opponents also receive s_{-i}^{θ} for any k large enough. Hence, $\sigma_i^1 = (\hat{\theta}_i^i, \theta_{i-1})$ for some $\hat{\theta}_i^i \in \Theta$ for any $\sigma_i \in R_{i,2}\left(s_i^{\theta}|\Gamma\left(\pi^k\right)\right)$ and any sufficiently large k.

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

7.7 Renegotiation

Suppose that both the buyer (B) and the seller (S) are risk averse. Assume also that when renegotiation occurs, B gets proportion α of the total surplus while S gets proportion $1-\alpha$ of the total surplus with some $\alpha \in (0,1)$. Consider the following modification of the SR mechanism. First, if B and S report different values in the first stage, we ask B to pay the arbitration fee F to S. Second, if B disagrees with the value report S made in the first stage, with equal probability we either ask B to make a payment of K_1 to S or S to make a payment of K_2 to B.

Utilizing risk aversion, we make this stochastic fine as harsh as a deterministic fine $T+F$; at the same time, the certainty equivalent of the payment B receives is adjusted to be zero. The random fine is realized as soon as the player makes his decision at the arbitration stage, so renegotiation will not be possible after that.

There remains the possibility of renegotiating after agents observe the outcome at Stage 1 and before B enters arbitration stage. In this case, it remains worse for S to report falsely as long as B's share of the surplus from renegotiation is positive and K_1 and K_2 are large enough. Thus, S will report B's true value. Therefore, it is not profitable for S to renegotiate when S tells the truth and disagrees with B on value at the first stage, since S will get a reward T.

7.8 Expected Utility Hypothesis

We shall show that the SR mechanism can be modified so that we can dispense with the expected utility hypothesis. In particular, we assume that preferences are only monotone with respect to first-order stochastic dominance. First, observe that the truth-telling condition holds with the dictator lottery in Table 2.[31] Second, we let B have the priority to enter the arbitration stage. That is, whenever they report different values in the first stage (and regardless of whether they report the same cost or not), we let B enter the arbitration stage. In contrast, we let S enter the arbitration stage only when they report the same value but different costs. Third, we double both the arbitration fee and the incentive transfer when B enters the arbitration stage, while we keep the same arbitration fee and incentive transfer as in our experiment when S enters the arbitration stage.

This modified SR mechanism implements the SCF. First, since we double the incentive transfer for seller, regardless of whether they agree on S's cost, S will prefer to report buyer's true value. Thus, the inter-stage coordination condition is satisfied. Similarly, since we also double the arbitration fee for the buyer, regardless of whether they agree on seller's cost, the buyer will prefer to report his own value to avoid paying the arbitration fee. Thus, within-stage coordination condition is satisfied. Finally, a similar argument shows that they will both report the true cost. Observe that the mechanism uses no randomization except for the dictator lotteries. In other words, players' preference over lotteries are irrelevant in establishing the inter-stage and within-stage conditions.

7.9 Retaliation-Proof Equilibrium

We will show that the SR mechanism is robust to retaliation behaviors. To see this, again, consider the setting in our experiment. We still adopt the notation introduced in Section 5.2. In particular, recall that in the SR mechanism, we have three types of histories, the initial history $\varnothing$, the set of first-stage histories M^1, and the set of terminal histories $\mathcal{Z}$. Also recall that $\mathcal{H} = \{\varnothing\} \cup M^1$ is the set of non-terminal histories and Σ_i is the set of pure strategies of $i \in \{B, S\}$. In this section, we use strategy and message interchangeability to be consistent

[31] In general, suppose that different types have different preferences over pure allocations $A \times \mathbb{R}$. Under monotonicity with respect to first-order stochastic dominance, the truth-telling condition also holds with the dictator lotteries employed in the proof of Theorem 1 (see the proof of Lemma in Abreu and Matsushima (1992a) for the details).

with FPW (2017) and our own section 5.

Following FPW (2017), we assume that the reciprocity payoff of player i depends on player i's belief about the other player's strategies and beliefs.[32] Denote by $\phi_{ij} \in \Sigma_j$ a generic belief of player i about the player j's strategy; similarly, denote by $\psi_{iji} \in \Sigma_i$ a generic belief of player i about the belief of player j's belief about player i's own strategy. For a given history h, the conditional expected utility of player i in state θ is given by:

$$\begin{aligned} U_i(\left(m_i, \phi_{ij}, \psi_{iji}\right), \theta|h) &= v_i\left(\left(m_i, \phi_{ij}\right), \theta_i|h\right) + \\ &\quad \rho_i \cdot \kappa_i\left(\left(m_i, \phi_{ij}\right), \theta|h\right) \cdot \lambda_i\left(\left(\phi_{ij}, \psi_{iji}\right), \theta_i|h\right) \end{aligned}$$

where the first term $v_i\left(\left(m_i, \phi_{ij}\right), \theta_i|h\right)$ is agent i's "material" payoff and the second term reflects his "psychological" utility from retaliation. The retaliation term is made up of three components: $\rho_i \geq 0$ is a sensitivity parameter, which reflects the relative size of monetary utility and psychological utility; $\kappa_i\left(\left(m_i, \phi_{ij}\right), \theta_i|h\right)$ measures player i's degree of unkindness for taking strategy m_i against the belief ϕ_{ij}. This measure compares the material payoff of player j under m_i and ϕ_{ij} with a reference material payoff level $v_j^{ref}\left(\phi_{ij}, \theta_j|h\right)$ of player j, i.e.,

$$\kappa_i\left(\left(m_i, \phi_{ij}\right), \theta|h\right) = v_j\left(\left(m_i, \phi_{ij}\right), \theta_j|h\right) - v_j^{ref}\left(\phi_{ij}, \theta_j|h\right).$$

Similarly, $\lambda_i\left(\left(\phi_{ij}, \psi_{iji}\right), \theta_i|h\right)$ measures player i's perceived unkindness of player j given player i's belief $\left(\phi_{ij}, \psi_{iji}\right)$ relative to a reference payoff level $v_i^{ref}\left(\psi_{iji}|^h, \theta_i\right)$:

$$\lambda_i\left(\left(\phi_{ij}, \psi_{iji}\right), \theta_i|h\right) = \min\left\{v_i\left(\phi_{ij}|_h, \psi_{iji}|^h, \theta_i|\varnothing\right) - v_i^{ref}\left(\psi_{iji}|^h, \theta_i\right), 0\right\},$$

where $\phi_{ij}|_h$ is i's updated belief about j's strategy after history h occurs and the minimum operation reflects FPW's specification that unlike the unkind acts, the kind acts do not result in any psychological utility of reciprocity.[33] Since ψ_{iji} is player i's belief about player j's belief (about player i's strategy) and in the SR mechanism player j never observes history h where he moves, $\psi_{iji}|^h$ need not be consistent with h. We thus use the notation $\psi_{iji}|^h$ to indicate the fact that $\psi_{iji}|^h$ is "not" updated after history h occurs, although it can still vary

[32]The specification used in FPW is based on the model of Dufwenberg et al. (2011), which is a simplified version of Dufwenberg and Kirchsteiger (2004) that allows for negative reciprocity but rules out positive reciprocity.

[33]The experimental evidence suggests that preferences for positive reciprocity is not as strong as those for negative reciprocity (see FPW (2017) for more discussion).

across different histories.

Finally, we specify the reference payoff level. First, following FPW (2017), we set $v_j^{ref}\left(\phi_{ij},\theta_j|h\right) > \min_{m_i}\left\{v_j\left(\left(m_i,\phi_{ij}\right),\theta_j|h\right)\right\}$. Second, we set $v_i^{ref}\left(\psi_{iji}|^h,\theta_i\right)$ as i's payoff that is obtained in the situation where no one enters the arbitration stage given $\psi_{iji}|^h$, i.e., $v_i^{ref}\left(\psi_{iji}|^h,\theta_i\right) \equiv v_i\left(\left(\psi_{iji}|^h,m_j\right),\theta_i|h\right)$ for some m_j such that m_j and $\psi_{iji}|^h$ make the same first-stage announcement.

Definition 8 *A pure strategy-belief profile* $\left(m_i^*,\left(\phi_{ij}|_h,\psi_{iji}|^h\right)_{h\in\mathcal{H}}\right)_{i\in\{B,S\}}$ *is a* ***retaliation-proof equilibrium*** *if, for each* $i\in\{B,S\}$ *and history* h*, the following two conditions hold:*

1. $m_i^* \in \arg\max_{m_i} U_i(\left(m_i,\phi_{ij},\psi_{iji}\right),\theta_i|h)$*;*
2. $\phi_{ij}=m_j^*$ *and* $\psi_{iji}|^h = m_i^*$ *for any history* h *on the path of* m^**.*

In Theorem 3 below, we first show that there is always a retaliation-proof equilibrium in which both players tell the truth at the first stage. Second, under two additional assumptions, we show that in any retaliation-proof equilibrium, both players tell the truth at the first stage. We now proceed by introducing these two assumptions:

Definition 9 *We say that player* i*'s belief is* ***reasonable*** *in a retaliation-proof equilibrium if the following condition holds: when* θ *is the true state, if* $\phi_{ij}|_\varnothing = m_j$ *in which* $m_{j,k}^1=\theta_k$ *for some player* k*, then* $\psi_{iji}^1=m_j^1$*.*

That is, player i's belief is said to be reasonable if, at the initial history, whenever player i believes that player j tells the truth on either i's type or j's type, player i also believes that player j believes that both of them make the same first-stage announcement on j's type.

Definition 10 *We say that player* i *cares more about material payoff than reciprocity payoff in a retaliation-proof equilibrium if* $\rho_i\cdot\lambda_i\leq 1$.[34]

[34]More precisely, we only need $\rho_i\cdot\lambda_i\leq(2T-D)/D$ in Case 2 of Step 3 in the proof of Theorem 3. Recall that $D\leq 200$ and hence $(2T-D)/D\geq 2$ in our experiment. In other words, the SR mechanism works (in the sense that Theorem 3 holds) so long as players do not want to burn two dollars in exchange of burning one dollar of their opponent for retaliation. In contrast, we need $\rho_i\cdot\lambda_i\leq D/(2T-D)$ to eliminate the retaliation motive in the MR mechanism. In particular, at the arbitration stage, a player may misreport to retaliate if he is willing to sacrifice one dollar in order to burn two dollars of his opponent.

Note that since $v_i^{ref}\left(\psi_{iji}|^h, \theta_i\right)$ involves neither penalty nor reward, we must have $\lambda_i \leq T + D$ in any retaliation-proof equilibrium in the SR mechanism. Hence, $\rho_i \cdot \lambda_i \leq 1$ is guaranteed if $\rho_i \leq 1/(T + D)$. Even if $\rho_i \cdot \lambda_i > 1$, both agents' telling the truth at the first stage can be sustained as part of a retaliation-proof equilibrium. This is because we can reward player i who at the first stage matches the second-stage report of his opponent j by $\rho_j D\,(T + D)$ while keeping the penalty for mismatch by T.[35]

We have the following result.

Theorem 3 *There exists a retaliation-proof equilibrium in which both players tell the truth at the first stage. Furthermore, in any retaliation-proof equilibrium where both players' beliefs are reasonable and they care more about material payoff than reciprocity payoff, they tell the truth at the first stage.*

Proof. Throughout the proof, we denote by θ the true state and for each agent $i \in N$, we let θ_i denote the true type of agent i. First, we show that there exists a retaliation-proof equilibrium where both players tell the truth at the first stage. We start our proof from establishing the following preliminary step.

Step 0: *If player j reports $m_{j,i}^1 \neq \theta_i$ at the first stage and player i enters the arbitration stage, then player i must report $m_i^2 \neq m_{j,i}^1$ at the arbitration stage.*

Proof of Step 0: Without loss of generality, consider $i = B$. By the argument in Section 2.2.1, telling the truth maximizes B's material payoff. Since $\phi_{ij}|_h$ is player i's updated belief about player j's strategy after history h occurs, it specifies the observed misreport $m_{j,i}^1$. If $\lambda_i((\phi_{ij}, \psi_{iji}), \theta_i|h) = 0$, then only the material payoff matters, which implies that it is optimal for B to tell the truth, i.e., $m_i^2 = \theta_i$; if $\lambda_i((\phi_{ij}, \psi_{iji}), \theta_i|h) < 0$, telling the truth, i.e., $m_i^2 = \theta_i \neq m_{j,i}^1$ results in a smaller $\kappa_i((m_i, \phi_{ij}), \theta|h)$ than reporting $m_i^2 = m_{j,i}^1$ does. Thus, B will not report $m_i^2 = m_{j,i}^1$. ■

Consider a strategy profile m where $m_i^1 = \theta$ for each $i \in \{B, S\}$, i.e., both players tell the truth at the first stage. Thanks to the finiteness of the SR mechanism, we can always

[35] If $\rho_i > 1/(T+D)$ and we still keep the reward as T, there will exist a "bad" equilibrium where retaliation is invoked. For example, both B and S misreport both value and cost and they match neither of their lie in the first stage, and then B reports the truth in the arbitration stage. If ρ_S is large, S would rather forgo the material gain of $2T$ from truth-telling to retaliate against B's mismatch by sticking to the lie in the first stage. More explicitly, S may forgo the gain of $2T = 600$ dollars in order to punish B by 2 dollars if $\rho_S \lambda_S > 300$.

specify a sequential best response against any possible arbitration stage. Conditional on initial history $\varnothing$, ϕ_{ij} and ψ_{iji} are associated with m so that $\lambda_i((\phi_{ij}, \psi_{iji}), \theta_i|\varnothing) = 0$. Thus, it suffices to show that conditional on the initial history, no player will deviate to any lie at the first stage. For any player i, given the initial history, we have ϕ_{ij} and ψ_{iji} such that $\lambda_i((\phi_{ij}, \psi_{iji}), \theta_i|\varnothing) = 0$. Thus, by Step 0, any deviation only results in either player i's arbitration fee or penalty from mismatching the opponent's announcement at the arbitration stage. This establishes the existence of retaliation-proof equilibrium.

Second, we show that if each player cares more about material payoff than reciprocity payoff and each player's belief is reasonable, both players tell the truth in the first stage in "any" retaliation-proof equilibrium. Fix a true state $\theta = (v, c)$ and m as a retaliation-proof equilibrium strategy profile. In the sequel, we write m_i^1 as $\left(m_{i,i}^1, m_{i,j}^1\right)$, i.e., $m_{i,i}^1$ (resp. $m_{i,j}^1$) denotes player i's report of player i's (resp. j's) type at the first stage. We shall prove that m is truth-telling at the first stage (i.e., $m_i^1 = \theta$ for every i) in the following three steps.

Step 1: *If player j reports $m_{j,i}^1 = \theta_i$ at the first stage and player i enters the arbitration stage, then player i must also truthfully report $m_i^2 = \theta_i$ at the arbitration stage.*

Proof of Step 1: Again, without loss of generality, we consider the case where $i = B$. By the argument in Section 2.2.1, telling the truth maximizes B's material payoff. Fix h as a history on the path of m. Moreover, by the second requirement of retaliation-proof equilibrium, $\phi_{ij}|_h = \phi_{ij}|_\varnothing = m_j$. Since B's belief is reasonable, $\psi_{iji}|^\varnothing = \psi_{iji}|^h = m_j$. Hence,

$$v_i\left(\phi_{ij}|_h, \psi_{iji}|^h, v|\varnothing\right) - v_i^{ref}\left(\psi_{iji}|^h, v\right) = 0$$

Thus, $\lambda_i\left(\left(\phi_{ij}, \psi_{iji}\right), v|h\right) = 0$. Thus, it remains a best response for B to tell the truth after history h occurs. ■

Step 2: *No one enters the arbitration stage.*

Suppose by way of contradiction that on the path of m, some player i enters the arbitration stage. Then, $m_{i,i}^1 \neq m_{j,i}^1$. We obtain a contradiction by considering each of the following two cases.

Case 1: $m_{i,i}^1 = \theta_i$.

Thus, $m_{j,i}^1 \neq \theta_i$. Let $\bar{m}_j$ be j's deviation strategy such that $\bar{m}_{j,i}^1 = \theta_i$ while we keep the other components of $\bar{m}_j$ the same as m_j. For the material payoff, by Step 1, player j can get rewarded by T which is larger than any payoff difference caused by an allocation

change. For the reciprocity payoff, since $\phi_{ji}|_{\varnothing} = m_i$ and player j's belief is reasonable, $\lambda_j((\phi_{ij}, \psi_{iji}), \theta_i|h) = 0$ for any history h on the path of m. Thus, $\bar{m}_j$ is a profitable deviation from m, which contradicts the hypothesis that m is the equilibrium strategy profile.

Case 2: $m^1_{i,i} \neq \theta_i$.

First, we show that $m^1_{j,i} = \theta_i$. By Steps 0 and 1, taking $m^1_{j,i} = \theta_i$ instead of any other report different from $m^1_{i,i}$, player j can obtain the material reward T rather than penalty T. Meanwhile, player j's reciprocity payoff results from making player i suffer from the arbitration fee T and the potential payoff change via allocations which is bounded by $D < T$. Since player j cares more about material payoff than reciprocity payoff, it is optimal for player j to choose $m^1_{j,i} = \theta_i$.

Second, let $\bar{m}_i$ be player i's deviation such that $\bar{m}^1_{i,i} = m^1_{j,i}$ while we keep the other components of $\bar{m}_i$ the same as m_i. For the material payoff, player i avoids penalty T which is larger than any difference caused by allocation change. Let h be a history on the path of m. For the reciprocity payoff, as $\phi_{ij}|_{\varnothing}$ and $\psi_{iji}|^{\varnothing}$ remain the same, the only change is that $\kappa_i((\bar{m}_i, \phi_{ij}), \theta|h) < \kappa_i((m_i, \phi_{ij}), \theta|h)$. Since λ_i always takes a nonpositive value, $\bar{m}_i$ is a profitable deviation from m. This contradicts the hypothesis that m is the equilibrium strategy profile. ■

Step 3: *Both B and S tell the truth at the first stage, i.e., $m^1_i = m^1_j = \theta$*

By Step 2, we know that $m^1_i = m^1_j$. Suppose on the contrary that $m^1_i = m^1_j = (\hat{v}, \hat{c}) \neq (v, c)$. Then, either $\hat{v} \neq v$ or $\hat{c} \neq c$. Suppose that $\hat{v} \neq v$. Then, let $\bar{m}_S$ be S's deviation strategy such that $m^1_S = v$ and we keep the rest of $\bar{m}_S$ the same as m_S. We claim that $\bar{m}_S$ is a profitable deviation from m because S can get the material reward T, which is guaranteed by Step 1 where B tells the truth at the arbitration stage. Let h be a history on the path of m. For the reciprocity payoff, as $\phi_{ij}|_{\varnothing}$ and $\psi_{iji}|^{\varnothing}$ remain the same, the only change is that $\kappa_S((\bar{m}_S, \phi_{SB}), \theta|h) < \kappa_S((m_S, \phi_{SB}), \theta|h)$. Since λ_S always takes a nonpositive value, S's reciprocity payoff will not be lower by changing from m_S into $\bar{m}_S$. This shows that $\bar{m}_S$ is a profitable deviation from m. Similarly, if $\hat{c} \neq c$, then it is profitable for B to deviate to a strategy $\bar{m}_B$ where B announces the true cost while keeping the rest the same as m_B. ■

Steps 1 through 3 together establish that both players tell the truth at the first stage at any retaliation-proof equilibrium. This completes the proof of the theorem. ■