2-2023

Getting dynamic implementation to work

Yi-Chun CHEN

Richard HOLDEN

Takashi KUNIMOTO
Singapore Management University, tkunimoto@smu.edu.sg

Yifei SUN

Tom WILKENING

Citation

CHEN, Yi-Chun; HOLDEN, Richard; KUNIMOTO, Takashi; SUN, Yifei; and WILKENING, Tom. Getting dynamic implementation to work. (2023). *Journal of Political Economy*. 131, (2), 285-387.
Available at: https://ink.library.smu.edu.sg/soe_research/2673

Getting Dynamic Implementation to Work

Yi-Chun Chen

National University of Singapore

Richard Holden

University of New South Wales Business School

Takashi Kunimoto

Singapore Management University

Yifei Sun

University of International Business and Economics

Tom Wilkening

University of Melbourne

We develop a new class of two-stage mechanisms, which fully implement any social choice function under initial rationalizability in complete information environments. We show theoretically that our simultaneous report (SR) mechanisms are robust to small amounts of incomplete information about the state of nature. We also highlight the robustness of the mechanisms to a wide variety of reasoning processes and behavioral assumptions. We show experimentally that an SR mechanism performs

We thank Sandro Brusco, Yeon-Koo Che, Navin Kartik, Siqi Pan, Patrick Rey, Steve Williams, and Jun Xiao for helpful comments and discussions as well as seminar participants at

Electronically published January 9, 2023

Journal of Political Economy, volume 131, number 2, February 2023.

https://doi.org/10.1086/721153

well in inducing truth telling in both complete and incomplete information environments and that it can induce efficient investment in a two-sided holdup problem with ex ante investment.

I. Introduction

In a classic paper, Maskin (1977, 1999) asked what social objectives can be fully implemented in a decentralized environment that respects the individual incentives of participants. Maskin showed that with a suitably constructed game form, one can fully implement a class of social choice functions—so-called monotonic social choice functions—in Nash equilibrium.[1] Monotonicity is, however, somewhat restrictive. In particular, it does not allow for social choice functions with distributional considerations.[2] Since then, there has been substantial interest in using extensive form mechanisms for implementation, as they hold the prospect of using refinements of Nash equilibrium (such as subgame perfection) to implement nonmonotonic social choice functions.

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 approach to common knowledge assumptions regarding rationality, payoffs, or preferences. For instance, Aghion et al. (2012, 2018) show that extensive form mechanisms are not robust to small deviations

Academia Sinica, Osaka University, Cardiff Business School, Hong Kong University of Science and Technology, Monash, New York University Shanghai, Shanghai University of Finance and Economics, University of Technology Sydney, University of Western Australia, the 10th Annual Organizational Economics Workshop at Sydney, the 18th Annual Society for the Advancement of Economic Theory Conference, and the 6th Xiamen University International Workshop on Experimental Economics. We are particularly grateful to three anonymous referees for their careful reading of our work and helpful feedback. We gratefully acknowledge the financial support of the Australian Research Council (ARC), including ARC Future Fellowship FT130101159 (Holden), ARC Discovery Early Career Research Award DE140101014 (Wilkening), the National Natural Science Foundation of China (NSFC71703020 and NSFC72273029; Sun), and Singapore Ministry of Education Academic Research Fund Tier 1 (R-122-000-296-115; Chen). Replication files are available in a zip file. This paper was edited by Emir Kamenica.

[1] A mechanism fully implements a social choice function if (1) the mechanism induces a unique equilibrium outcome in every possible state of the world and (2) this equilibrium outcome corresponds to the outcome of the social choice function.

[2] For instance, Maskin monotonicity rules out bilateral trading situations where the state is the value of the good, and we wish to implement a pricing rule where (1) trade always occurs and (2) the trade price between a buyer and seller is increasing in the value of the good. In such cases, the implementation mechanism affects only the distribution of surplus. We concentrate on these bilateral trading environments in the experimental parts of the paper.

from common knowledge about the state of nature,[3] while Fehr, Powell, and Wilkening (2021) 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's law) when the assumptions of common knowledge of rationality, payoffs, or preferences are relaxed. In such situations, belief updating occurs along the dimension of uncertainty, leading to equilibria that may be far away from the intended equilibrium even when uncertainty is small.

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 (2015), we use the notion of *initial rationalizability* as our solution concept. Like rationalizability in normal form games, this solution concept iteratively deletes strategies that are not best replies. However, unlike backward induction, it requires that there be sequential rationality and common belief of sequential rationality only at the beginning of the game, and it makes no assumption about how beliefs evolve after zero probability events occur. If we accommodate any belief revision assumption at any subsequent stages of the game when a zero probability event occurs, initial rationalizability is among the weakest rationalizability concepts for extensive form games. Hence, implementation under initial rationalizability is among the most robust notions of implementation that exist for dynamic mechanisms.

Part 1 of our paper provides very permissive implementation results when using initial rationalizability as a solution concept. Before getting into the details, we want to be clear from the outset about the domain of problems to 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 many applications where there are distributional concerns involve economies with money. Second, we employ stochastic mechanisms in which lotteries are explicitly used. Therefore, we assume that preferences over lotteries have an expected utility representation. Third, we focus on private value environments. That is, each player's utility depends on only his/her own payoff type as well as the lottery chosen and his/her monetary payment.

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

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

We use one of the second reports to select a lottery from a set of prespecified lotteries and use that lottery to determine the outcome. The set of lotteries are constructed so that it is a dominant strategy for a expected utility maximizer to make a truthful report. We can therefore use the second report as a part of a test to determine whether the successor was lying in the previous stage. We do this by comparing each second report with the initial report of the successor. We reward the successor with a bonus if the two reports match and punish her with a fine if they differ. The bonuses and fines can always be set to induce truthful reporting by the successor in the first stage without requiring money from an outside source. This in turn induces the self-reporting individual to make truthful first-stage reports.

In contrast to the canonical three-stage subgame perfect implementation (SPI) mechanism, the SR mechanism that we develop is robust to departures from the common knowledge assumption. Our notion of robustness—which we call robustness to private value perturbations—demands that a mechanism implement the desired social choice function under complete information and almost implement it in nearby environments where there is a small amount of incomplete information about the state of nature. Specifically, in such nearby environments, even conditional on the opponents' signals and types, each player's signal remains almost accurate in identifying her own type.[4] We prove that in the SR mechanism, any sequence of initially rationalizable message (e.g., sequential equilibrium) profiles under incomplete information converges to the truth-telling profile as the amount of incomplete information goes to zero. That is, any social choice function is robustly implementable under private value perturbations.[5]

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

[5] This result contrasts the impossibility result of robust SPI due to Aghion et al. (2012, theorem 3), which is proved by making use of non–private value perturbations.

In appendix B, we also show that the initial rationalizability correspondence in any finite mechanism is upper hemicontinuous with respect to private value perturbations. This result implies that if a finite mechanism can implement a social choice function in initial rationalizable messages under complete information, it can also implement it under a private value perturbation. It also suggests that implementation under initial rationalizability is likely to be a useful desideratum for achieving robust implementation; the SR mechanism we develop is one of the simplest dynamic mechanisms with this desired property.

Having developed a mechanism with promising robustness features, part 2 of our paper uses laboratory experiments to assess the performance of the mechanism in three different settings. In section III, we explore how the SR mechanism performs both in an environment with complete information and in an environment with noise. The setting we consider is identical to the one studied in Aghion et al. (2018) but with a private value perturbation. Specifically, a buyer is to receive a buyer-specific good of either high or low quality. Before learning the value of the good, the buyer and seller would like to write a contract where the buyer pays a high price if the good is of high quality and a low price if the good is of low quality. However, the quality of the good is not verifiable by a third party, such as a court, and thus a state-dependent contract cannot be directly enforced. Contracting parties must instead rely on some form of implementation mechanism.

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

Our first experiment explores whether the SR mechanism induces truthful first-stage revelation under complete information and under a private value perturbation and compares performance against a benchmark canonical SPI mechanism that uses a nearly identical set of prices, fines, and rewards.[6]

[6] The SPI mechanism is based on the work of Moore and Repullo (1988) and consists of three stages. In the first stage, the buyer reports the value of the good. The seller is informed about the buyer's report and has the option of calling or not calling the arbitrator in stage 2. If the arbitrator is not called, we use the buyer's report to set report-specific prices. If the arbitrator is called, the buyer is fined and is given a take-it-or-leave-it counteroffer. The counteroffer price is set so that it is in the buyer's material interest to accept if their first report was below their true value and reject otherwise. We use the counteroffer decision to either fine or reward the seller.

Each session consists of a single mechanism and two information treatments: a no-noise treatment, with complete information about the quality of the good, and a noise treatment, where buyers receive correct information about the quality of the good 97.5% of the time, while sellers receive correct information about the quality of the good 87.5% of the time. The SR mechanism we develop is predicted to induce truthful reports in both treatments. By contrast, the SPI mechanism has a unique subgame perfect equilibrium under complete information but has multiple initial rationalizable strategy profiles. Further, it is not predicted to be robust to the private value perturbation, and misreports by buyers in the high-quality scenario are predicted to increase when noise is introduced.

We find experimental evidence that is largely consistent with the behavior predicted by theory. In the no-noise treatment of the SR mechanism, buyers and sellers report truthfully in the vast majority of cases: in the low-quality scenario, buyers report truthfully in 97.7% of cases, while sellers report truthfully in 86.2% of cases. In the high-quality scenario, buyers report truthfully in 94.0% of cases and sellers report truthfully in 93.0% of cases. When noise is introduced, there is no significant change in the behavior of buyers and sellers. In particular, buyers report truthfully in the high-signal scenario in 90.0% of cases.

By contrast, buyer truth-telling rates in the high-quality scenario are lower in the canonical SPI mechanism, and the mechanism is not robust to noise. Buyers in the no-noise treatment are truthful in the high-quality scenario in only 77.5% of cases. This truth-telling rate falls to only 60.0% when noise is introduced. These truth-telling rates are significantly lower than the rates seen in the SR mechanism, and the difference in misreports between the no-noise and noise treatments of the SPI mechanism are significant. Thus, in terms of truth-telling rates, the SR mechanism strongly outperforms the SPI mechanism in complete information environments and appears robust to private value perturbations.

As an application of our mechanism, we also explore how the SR mechanism performs in a two-sided holdup environment with pure cooperative investments in section IV.[7] We choose this environment because it is the most important application of implementation with common knowledge in the literature. Grossman and Hart (1986) and Hart and Moore (1990) explore the implications of incomplete contracts by developing models that assume that key payoff-relevant information is observable but not

[7] As discussed in Che and Hausch (1999), the pure cooperative case is one where the buyer's investment reduces the cost of production for the seller and where the seller's investment increases the value for the buyer but investments offer no (or negative) direct benefits to the investor. See Chung (1991), Aghion, Dewatripont, and Rey (1994), and Nöldeke and Schmidt (1995) for option contracts that can solve the holdup problem under the alternative selfish investment case, where investment yields direct benefits.

verifiable by a third party, such as a court. These assumptions are intended to make formal contracting ineffective but allow parties to bargain ex post, thus creating a role for property rights and firm boundaries when ex ante investments are required. However, Maskin and Tirole (1999) show that if parties commonly observe payoff-relevant information, there often exists an auxiliary extensive form mechanism that can credibly make this information verifiable. Such mechanisms allow for the first best to be implemented and thus raise questions about the underlying foundations of the incomplete contracting literature.

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 privately 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 SR mechanism used in this experiment is similar to the one used in our first experiment except that the mechanism requires information on both costs and values. To elicit this information, both parties simultaneously report both the cost and the value of the good in the first stage of the mechanism. 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 to make a second report. We again use a lottery to make it a dominant strategy for an expected utility maximizer to make a truthful second report and use the second report as a part of a test to determine who was lying in the first 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 counterparty with a bonus if the two reports match and punish him or her with a fine if they differ.

In our experiments, subjects first choose investment levels and then enter 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.

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% of cases. Likewise, sellers make truthful first-stage value and cost reports in 91.7% of cases. Buyers choose the optimal level of investment in 89.6% of cases, while sellers choose the optimal level of investment in 84.8% of cases. In aggregate, 89.3% of dyads improve their performance relative to the theoretical no-mechanism benchmark of 70 experimental currency units (ECU), and 74.2% of dyads exhibit first-best investments and truth-telling behavior.[8]

Previous experiments have found that players are concerned about the strategic sophistication of their matched partners and avoid SPI mechanisms if given a choice (Fehr, Powell, and Wilkening 2021). 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% for both parties. Groups that opt in to the mechanism behave very closely to theory, with 90.5% of dyads reporting truthfully and achieving the first best.

In order to benchmark the performance of the mechanism, we also compare efficiency of the mechanism with 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 SPI mechanism based on Moore and Repullo (1988) and a one-stage mechanism proposed by Kartik, Tercieux, and Holden (2014). Using the average earnings of participants as a measure of efficiency, we find that our SR mechanism is 19.8% more efficient than the fixed price mechanism, 35.0% more efficient than the mechanism based on Moore and Repullo (1988), and 62.2% more efficient than the mechanism proposed by Kartik, Tercieux, and Holden (2014). However, relative to its theoretical benchmark, there is some efficiency loss due to fines.

In the holdup environment, participants must sign a contract prior to the investment stage but participate in the intended implementation mechanism after investments have taken place. Given that investment is not instantaneous in the real world, it is likely that participants will have the opportunity to communicate with one another between the investment decision and the beginning of the mechanism. Thus, it is important to also explore the extent to which implementation mechanisms are robust to communication.

[8] The only stage that does not confirm strongly to the theoretical prediction is the second-report stage, where in early periods, some 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 in all periods.

In appendix D, we report on a third set of experiments where we compare the SR mechanism to a coordination mechanism in a one-sided holdup setting with and without directed communication. The coordination mechanism is a simple one-stage mechanism that is common in the literature and that uses SRs and disagreement fines to weakly implement the first best. Given the similarities between the two mechanisms, we are interested in whether the SR mechanism provides any additional empirical advantage over this simpler mechanism.

We find that without communication, the true value of the good acts as a focal point in the coordination mechanism and the two mechanisms have similar investment and truth-telling properties. However, when directed communication is allowed, the multiple equilibria that exist in the coordination mechanism become problematic. In particular, when the seller makes an investment, buyers in the coordination mechanism frequently send messages to the seller that they plan to make a nontruthful report in the report stage. Both the buyers and the sellers use these messages to coordinate their reports on lies that are advantageous to the buyers but disadvantageous to the sellers. The coordination of reports on lies removes the incentives to invest in the coordination mechanism, leading to strong differences between the two mechanisms when directed communication is allowed. Thus, our experiments suggest that the SR mechanism provides additional robustness to communication relative to simpler mechanisms that only weakly implement the first best.

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, e.g., Chung and Ely 2003; Aghion et al. 2012).

In designing our SR mechanism, we took into consideration a number of findings from the experimental literature on implementation. Sefton and Yavas (1996) and Katok, Sefton, and Yavas (2002) study various versions of the Abreu-Matsushima mechanisms (Abreu and Matsushima 1992) and highlight issues that arise in mechanisms that use multiple iterations of backward induction.[9] Discussing the search for good mechanisms for the selection of arbitrators, de Clippel, Eliaz, and Knight (2014, 3436) argue that one desiderata in the search for good mechanisms is that a mechanism has “as few stages as possible so that backwards induction is relatively ‘simple’ to execute.” By concentrating on two-stage mechanisms and using a

[9] Note that these papers relate to a technique often used in virtual implementation to reduce the size of potential fines. They do not exhaust the set of potential virtual implementation mechanisms. In particular, it is an open question as to whether simpler virtual implementation mechanisms, such as those developed in Arya, Glover, and Young (1995), can be operationalized for the two-sided holdup problem in situations where the designer does not require exactness.

weaker solution concept, our paper directly addresses the issues raised in these papers.

Finding auxiliary mechanisms that have good empirical properties has proven difficult even in simple environments with complete information.[10] Yet our mechanism is robust to a range of reasoning processes. In particular, it remains valid for any solution concept that is stronger than deletion-of-never-sequential best replies followed by two rounds of deletion of strictly dominated strategies. This requirement is satisfied for almost all standard solution concepts in extensive form games as well as some behavioral solution concepts, such as the agent quantal response equilibrium.

Our focus is on exact implementation and specifying a clear mechanism for practical implementation problems. Abreu and Matsushima (1992) show that if one is satisfied with approximate or virtual implementation, a wide class of social choice functions can be implemented by static, stochastic mechanisms. Indeed, this class extends beyond the private value perturbations we consider in this paper. That said, there may be good reasons for a mechanism designer to seek an exact implementation. These relate to well-known issues pertaining to renegotiation proofness and the expected utility representation of preferences (see Jackson 2001 for an excellent discussion).

The remainder of the paper proceeds as follows. Section II contains our theoretical analysis and proves our main implementation results. Section III reports on our first experiment, while section IV reports on our second experiment. Section V contains some brief concluding remarks. Appendix A contains our theoretical proofs and a discussion of additional robustness properties of the SR mechanism. Appendix B explores the robustness of general finite mechanisms that implement under initial rationalizability. Appendix C contains additional empirical analysis and figures for the experiments in the main text, while appendix D reports on additional experiments comparing the SR mechanism to a coordination mechanism. Appendix E contains sample experimental instructions.

[10] Much of the experimental literature on implementation has centered on the public goods problems, Solomon's dilemma problems (Ponti et al. 2003; Giannatale and Elbittar 2010), or holdup problems. In the context of public goods, Chen and Plott (1996), Chen and Tang (1998), and Healy (2006) study learning dynamics in public good provision mechanisms. Andreoni and Varian (1999), Falkinger et al. (2000), and Chen and Gazzale (2004) study two-stage compensation mechanisms that build on work from Moore and Repullo (1988), while Harstad and Marrese (1981, 1982), Attiyeh, Franciosi, and Isaac (2000), Arifovic and Ledyard (2004), and Bracht, Figuières, and Ratto (2008) study the voluntary contribution game, Groves-Ledyard, and Falkinger mechanisms, respectively. In relation to the holdup problem, Aghion et al. (2012) draws attention to the issue of information perturbations, while Bierbrauer et al. (2017) and Fehr, Powell, and Wilkening (2021) draw attention to the issues of other-regarding preferences and reciprocity. Hoppe and Schmitz (2011) study option contracts developed in Nöldeke and Schmidt (1995) in a one-sided setting that allows for renegotiation and highlight how attempts at renegotiation are not always successful.

II. The Theory

In this section, we first define the solution concept of initial rationalizability. We argue that initial rationalizability makes no assumption about how beliefs evolve after zero probability events occur and is substantially more permissive than subgame perfect equilibrium. We then formally construct a two-stage mechanism—the SR mechanism—in a quasilinear setting and show that it can implement any social choice function in initial rationalizable strategy profiles. As a result, implementation by the SR mechanism is not sensitive to belief updating regarding other players' preferences, payoffs, or rationality.

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

A. The Environment

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

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

Let $\Theta \equiv \times_{i \in \mathcal{I}} \Theta_i$ be the set of type profiles, or *states*. We consider a *planner* who aims to implement a *social choice function* $f : \Theta \rightarrow \Delta(A)$. We start with the complete information environment; that is, the true type profile $\theta \in \Theta$ is commonly known to the players but unknown to the planner. In section II.D, we will turn to study the robustness of our result in an incomplete information environment where this common knowledge assumption is perturbed.

We will consider only finite two-stage mechanisms throughout section II. In particular, the SR mechanism that we are about to define has only two stages. In stage 1, each player i chooses one action m_i^1 from a finite set M_i^1. Denote by $M^1 \equiv \times_{i \in \mathcal{I}} M_i^1$ the set of stage 1 action profiles. In stage 2, after observing the stage 1 action profile $m^1 \in M^1$, each player i chooses an action m_i^2 from another finite set $M_i^2(m^1)$. Again, write $M^2(m^1) \equiv \times_{i\in\mathcal{I}} M_i^2(m^1)$ for the set of stage 2 action profiles, following the action 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 nonterminal histories; (3) $\mathcal{Z} = \{(m^1, \tilde{m}^2) : m^1 \in M^1, \tilde{m}^2 \in M^2(m^1)\}$ is the set of terminal histories; (4) g is the outcome function that maps each terminal history to a lottery in $\Delta(A)$; and (5) τ_i is the transfer rule that maps each terminal history to a transfer to player i.

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

B. Solution Concept and Implementation

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

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

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

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

In order to analyze each player's reasoning about other players' messages during the entire course of play of the game, we model players' beliefs by means of *a conditional probability system* (CPS). Following Battigalli and Siniscalchi (1999), we formulate the notion of CPS as follows.

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

1. for each $B \in \mathcal{B}$, $\mu[\cdot|B] \in \Delta(\Omega)$ and $\mu[B|B] = 1$; and
2. if $A \in \Sigma$ and $B, C \in \mathcal{B}$ with $A \subset B \subset C$, then $\mu[A|C] = \mu[A|B] \cdot \mu[B|C]$.

Let $M_{-i}(h) \subset M_{-i}$ be the set of message profiles of player i's opponents that are consistent with history h. Hence, $M_{-i}(\varnothing) = M_{-i}$, and for each $m^1 \in M^1$, we have $M_{-i}(m^1) = \{\tilde{m}_{-i} \in M_{-i} : (m_i^1, \tilde{m}_{-i}^1) = m^1\}$. Similarly, we set $M_i(\varnothing) = M_i$ and $M_i(m^1) = \{\tilde{m}_i \in M_i : (\tilde{m}_i, m_{-i}) = m^1\}$. In the current complete information setting, we set $\Omega = M_{-i}$ and $\mathcal{B} = \{M_{-i}(h)\}_{h\in\mathcal{H}}$. By conditions 1 and 2 of definition 1, a CPS μ_i specifies, for each history h, a probability distribution over M_{-i} such that Bayes's rule applies whenever possible.[11] To simplify the notation, we hereafter write $\mu_i[\cdot|h]$ to denote $\mu_i[\cdot|M_{-i}(h)]$ for $h \neq \varnothing$, and we write $\mu_i[\cdot]$ for $\mu_i[\cdot|\,\varnothing]$. By reporting message m_i and holding CPS μ_i, player i receives the following expected payoff conditional on history $h \in \mathcal{H}$:

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

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

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

We now define initial rationalizability.

Definition 2 (Initial rationalizability). Let $\Gamma(\theta)$ be a two-stage game. For every player $i \in I$, let $R^{\Gamma(\theta)}_{i,0} = M_i$. Inductively, for every integer $k \geq 1$, let $R^{\Gamma(\theta)}_{i,k}$ be the set of messages that are sequential best replies to some CPS μ_i such that $\mu_i[R^{\Gamma(\theta)}_{-i,k-1}] = 1$. Finally, the set of *initially rationalizable* messages for player i is $R^{\Gamma(\theta)}_i = \cap_{k=1}^{\infty} R^{\Gamma(\theta)}_{i,k}$.

This solution concept is arguably the weakest among standard notions of equilibrium or rationalizability which impose sequential rationality (see Dekel and Siniscalchi 2015 for more discussion). In particular, only beliefs at the beginning of the game (i.e., $\mu_i[\cdot]$) are restricted. In other words, a player can hold an arbitrary updated belief about his/her opponents once being surprised. For instance, at a history that is precluded by his/her opponents' rational moves, the player can simply cease believing that his/her opponents are rational. The feature sharply contrasts with subgame perfect equilibrium, where the opponents' irrational moves are always regarded as one shot and never upset a player's faith in his opponents' rationality in their subsequent moves. In particular, while the SPI mechanism due to Moore and Repullo (1988) also implements any social choice function in subgame perfect equilibrium, it fails to implement it when a player gives up the belief that his/her opponents will behave rationally upon seeing a history precluded by his/her opponents' rational moves.[12]

Let $R^{\Gamma(\theta)} \equiv \times_{i\in\mathcal{I}} R^{\Gamma(\theta)}_i$ be the set of initially rationalizable messages for all players. We now define our notion of implementability to be used later.

Definition 3. A social choice function f is *implementable in initial rationalizable messages* if there exists a mechanism Γ such that, for any state

[11] For instance, suppose that $B = M_{-i}(h)$, $C = M_{-i}$, and $A = \{m_{-i}\} \subset M_{-i}(h)$. Then, if player i initially believed that history h was possible and $\mu[M_{-i}(h)|M_{-i}] > 0$, condition 2 of definition 1 requires that player i use Bayes's rule and hold belief $\mu[m_{-i}|M_{-i}(h)] = \mu[m_{-i}|M_{-i}]/\mu[M_{-i}(h)|M_{-i}]$.

[12] We will revisit this point in sec. III.C when we discuss our experimental design and hypotheses.

$\theta \in \Theta$, we have $g(z(m)) = f(\theta)$ and $\tau_i(z(m)) = 0$ for every player $i \in \mathcal{I}$ and for every message profile $m \in R^{\Gamma(\theta)}$.

We omit the existential requirement from definition 3, since throughout the paper we consider only finite mechanisms for which $R^{\Gamma(\theta)} \neq \varnothing$.

C. *The SR Mechanism*

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

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

1. Message Space

First, we specify the message space.

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

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

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

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

$$M_i^2(m^1) = \begin{cases} \Theta_i & \text{if } i \in \mathcal{I}^*(m^1), \\ \{\hat{\theta}_i^i\} & \text{if } i \notin \mathcal{I}^*(m^1). \end{cases}$$

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

2. Outcome Function

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

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

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

Equipped with lemma 1, we now specify the outcome function of the SR mechanism: if all players' announcements in the first stage are consistent, then the planner implements $f(\hat{\theta})$, where $\hat{\theta} \equiv (\hat{\theta}_i^i)_{i\in\mathcal{I}}$ is the consistent state announcement. Otherwise, the planner randomly selects a player i^* from the set $\{i \in \mathcal{I}^*(m^1)\}$ with equal probability. The planner then implements $l_{i^*}(\tilde{\theta}_{i^*})$, which corresponds to the second-stage announcement $\tilde{\theta}_{i^*}$ of player i^*.

3. Transfers

We define the transfer rule. Transfers are incurred only when some player's stage 1 announcement is inconsistent. We impose the following rules:

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

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

- We choose $T > D$, where

$$D \equiv \sup_{i\in\mathcal{I}, a, a' \in A, \theta_i \in \Theta_i} |u_i(a, \theta_i) - u_i(a', \theta_i)|,$$

  where the supremum above is well defined because Θ_i is finite and $u_i(\cdot, \theta_i) : A \rightarrow \mathbb{R}$ is bounded for each $\theta_i \in \Theta_i$.

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

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

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

Proof. See appendix section A1.

We further elaborate on two features of the SR mechanism. First, the second stage in the SR mechanism is constructed by first choosing a set of lotteries, one for each type of each player, according to which it is the unique optimal choice for each player to truthfully report his/her own type. Since the outcome in the second stage is solely based on the second reports of the party and the dictator lotteries were constructed prior to the play of the mechanism, a player's belief about the other player's type plays no role in his/her choice made at the second stage. This feature ensures that the mechanism is insensitive to the way in which players update their beliefs about other players. As a result, the SR mechanism is less susceptible to relaxations of common knowledge assumptions on rationality, information, and preferences.[13] This feature carries over to the situation where the original information structure is slightly perturbed, as long as each (active) player's own signal is more informative over his/her own payoff types than the other players' signals/payoff types (see sec. II.D).

Second, the truthful report in the second stage plays the same role as a behavioral anchor in the level k model (see, e.g., Crawford and Iriberri 2007; de Clippel, Saran, and Serrano 2019). At the first stage, once everyone knows that telling the truth is the unique optimal choice for any active player in the second stage, truth telling must also be the uniquely optimal action for the successor(s) of each active player and hence the optimal action of each player. By leveraging on these two features, we obtain the informational robustness of the SR mechanism in section II.D.

D. Robustness to Information Perturbations

We now formulate the second robustness property of the SR mechanism. Suppose that the players do not observe the state directly but are informed of the state via a noisy signal. Following Aghion et al. (2012), we set the space of signals as $S_i = \Theta$ for each $i \in \mathcal{I}$. A signal profile is an element $s = (s_1, \dots, s_I) \in S = \times_{i\in\mathcal{I}} S_i$. Let s_i^θ denote the signal in S_i that

[13] In app. sec. A3, we discuss how the structure of the SR mechanism leads to additional robustness features related to the common knowledge assumption of preferences and rationality. In app. sec. A3.1, we discuss how the insensitivity to belief updating makes the SR mechanism more robust to retaliatory preferences than the SPI mechanism and the retaliatory seller mechanism of Fehr, Powell, and Wilkening (2021). In app. sec. A3.2, we discuss how the simple two-stage structure of the mechanism makes the mechanism robust to small amounts of noise in the best-response function and to beliefs about the strategic sophistication of others. In app. sec. A3.3, we explain how the SR mechanism can be modified so that our implementation result works even with nonexpected utility preferences.

corresponds to θ and s^θ denote the signal profile such that $s_i = s_i^\theta$ for every player $i \in \mathcal{I}$.

Suppose that the state and signals are jointly distributed according to a prior distribution $\pi \in \Delta(\Theta \times S)$. We assume that for each i and θ, the marginal distribution on i's signals places strictly positive weight on each of i's signals in every state. That is, for each θ, $\text{marg}_{S_i}\pi[s_i^\theta] > 0$, so that Bayes's rule is well defined. Then, for each π, we write $\pi[\cdot|s_i]$ for the probability measure over $\Theta \times S_{-i}$ conditional on s_i and $\pi[\cdot|s, \theta_{-i}]$ for the probability measure over Θ_i conditional on s and θ_{-i}, when it is well defined. A prior π is said to be a *complete information* prior if $\pi[\theta, s] = 0$ whenever $s \neq s^\theta$. Henceforth, we denote a complete information model by π^{CI}.

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

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

We consider the following class of information perturbations.

Definition 4. A sequence of priors $\{\pi^k\}_{k=1}^\infty$ is a *private value perturbation* to π^{CI} (which we denote by $\pi^k \to \pi^{CI}$) if, for any $\varepsilon > 0$, there exists $K \in \mathbb{N}$ such that for all $k \geq K$, the following two properties hold: (i) $d(\pi^k, \pi^{CI}) < \varepsilon$ and (ii) for every $i \in \mathcal{I}$ and $\theta \in \Theta$,

$$\text{marg}_{\Theta_{-i}\times S_{-i}}\pi^k\left[\theta'_{-i}, s_{-i}|s_i^\theta\right] > 0 \Rightarrow \pi^k\left[\theta_i|s_i^\theta, s_{-i}, \theta'_{-i}\right] > 1 - \varepsilon. \quad (4)$$

That is, conditional on the opponents' signal and payoff type profile on the support of $\pi^k[\cdot|s_i^\theta]$, each player i's signal s_i^θ is asymptotically accurate in identifying his/her own type θ_i. Indeed, theorems 1 and 2 of Aghion et al. (2012) both invoke private value perturbations in proving the nonrobustness of the SPI mechanism. To wit, when players' values are private, it is natural to assume that a player's own signal is more informative over their own payoff type than others' signals/payoff types.

One special case of private value perturbations depicts a situation in which player i knows precisely his/her own type θ_i (e.g., Bergemann and Morris 2005) even under information perturbations and entertains only a small amount of uncertainty about his/her opponents' types. This amounts to assuming that $\text{marg}\Theta_i\pi^k[\theta_i|s_i^\theta] = 1$ for every k. This assumption also implies that the sequence of priors $\{\pi^k\}_{k=1}^\infty$ is a private value perturbation, as long as $d(\pi^k, \pi^{\text{CI}}) \to 0$ as $k \to \infty$. To see this, observe that since $\text{marg}\Theta_i\pi^k[\theta_i|s_i^\theta] = 1$, we have that $\pi^k[\theta'_i, \theta'_{-i}, s_{-i}|s_i^\theta] = 0$ for all $\theta'_i \neq \theta_i$. This implies that condition (4) is satisfied for every k.[14]

[14] Eccles and Wegner (2016) explore robust implementation in this special case where players know their own types with certainty. They show that a two-player perfect information mechanism is robustly implementable under subgame perfection if it can be implemented by a two-stage sequential move mechanism. The class of sequential move mechanisms they consider is more restrictive than ours and cannot fully implement all social

We now adapt our definitions of mechanisms and solution concepts to the incomplete information setup. We denote by $\Gamma(\pi)$ the incomplete information game induced by a two-stage mechanism Γ under prior π. Here, for each history h, a CPS μ_i specifies a distribution $\mu_i[\cdot|\Theta \times S_{-i} \times M_{-i}(h)]$ over $\Theta \times S_{-i} \times M_{-i}$ with the property that Bayes's rule applies whenever possible. Using the formal notation introduced for CPS's in definition 1, we let $\Omega = \Theta \times S_{-i} \times M_{-i}$ and $\mathcal{B} = \{\Theta \times S_{-i} \times M_{-i}(h)\}_{h\in\mathcal{H}}$. Again, to simplify the notation, we write $\mu_i[\cdot|h]$ to denote $\mu_i[\cdot|\Theta \times S_{-i} \times M_{-i}(h)]$ for $h \neq \varnothing$, and we write $\mu_i[\cdot]$ instead of $\mu_i[\cdot|\varnothing]$.

By sending message m_i and holding CPS μ_i, player i receives the following expected payoff conditional on history $h \in \mathcal{H}$:

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

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

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

As a CPS represents a player's belief, it should also be based on the player's signal. The connection is formalized via the following definition.

Definition 5. A CPS μ_i is said to be consistent with s_i under prior π if there exists a sequence of totally mixed probability distributions $\{\mu_i^q\}_{q=1}^{\infty}$ over $\Theta \times S_{-i} \times M_{-i}$. such that (i) for every $(\theta, s_{-i}, m_{-i}) \in \Theta \times S_{-i} \times M_{-i}$ and $h \in \mathcal{H}$, we have

$$\mu_i[\theta, s_{-i}, m_{-i}|h] = \lim_{q\to\infty} \mu_i^q[\theta, s_{-i}, m_{-i}|h];$$

and (ii) for every $q \geq 1$, there exists $\sigma_{-i}^q : \Theta_{-i} \times S_{-i} \to \Delta(M_{-i})$ such that

$$\mu_i^q[\theta, s_{-i}, m_{-i}] = \sigma_{-i}^q[m_{-i}|\theta_{-i}, s_{-i}]\pi[\theta, s_{-i}|s_i]. \tag{5}$$

Myerson (1986) shows that any CPS can be approximated by a sequence of totally mixed probability distributions. In addition, definition 5 requires that in each of the totally mixed probability distributions in the sequence, agent i's belief over the other agents' strategy profiles depends on only the other agents' signal and payoff type profiles. In other words, agent i believes that the other agents' strategies do not signal the information about agent i's payoff type (which the other agents do not know anyway).[15] Both requirements i and ii in definition 5 are satisfied in the standard formulation of sequential equilibrium adopted in Aghion

choice functions. In particular, sequential move mechanisms cannot robustly implement the first best in the two-sided holdup environment considered in experiment 2.

[15] We thank a referee for suggesting this formulation of the consistency requirement and its interpretation to us.

et al. (2012); see definition B.1 and definition σ in the online appendix of Aghion et al. (2012).

The following two definitions are the counterparts of definitions 2 and 3 in the incomplete information environments that we study here. We first define the solution concept of initial rationalizability under incomplete information.

Definition 6 (Initial rationalizability under incomplete information). Let $\Gamma(\pi)$ be the game induced by a two-stage mechanism Γ under prior π. The set of initial rationalizable messages of player i with signal s_i is defined as $R_i(s_i|\Gamma(\pi)) = \cap_{k=1}^{\infty} R_{i,k}(s_i|\Gamma(\pi))$, where $R_{i,0}(s_i|\Gamma(\pi)) = M_i$ and, inductively, for every integer $k \geq 1$,

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

As we did under complete information, we consider only finite mechanisms for which $R(s^\theta|\Gamma(\pi)) \equiv \times_{i\in\mathcal{I}} R_i(s_i^\theta|\Gamma(\pi)) \neq \emptyset$. The following is the definition of robust implementation that we adopt.

Definition 7. A social choice function f is *robustly implementable in initial rationalizable strategies* if there exists a mechanism $\Gamma = (M, g)$ such that for any state $\theta \in \Theta$, any signal profile $s^\theta \in S$, any private value perturbation $\{\pi^k\}_{k=1}^{\infty}$ to π^{CI}, and any sequence of message profiles $\{m^k\}_{k=1}^{\infty}$ with $m^k \in R(s^\theta|\Gamma(\pi^k))$ for each k, we have $g(z(m^k)) = f(\theta)$ and $\tau_i(z(m^k)) = 0$ for every player i and for all sufficiently large k.

Within the class of private value perturbations, our robustness notion is based on the permissive solution concept of initial rationalizability. Specifically, we allow each player's CPS to have any degree of correlations among other players' strategies, signals, and payoff types, subject to the consistency requirement in definition 5. Hence, our robust implementation result holds even when the stronger yet more standard solution concept of sequential equilibrium (formally defined in the online appendix of Aghion et al. 2012) is used.[16] We are now ready to state our robust implementation result using the SR mechanism defined in section II.C.

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

Proof. See appendix section A2.

Note that theorem 2 does not contradict the impossibility result of robust SPI proved in theorem 3 of Aghion et al. (2012). Indeed, theorem 3

[16] If we adopt a solution concept under which each player's strategy depends on only his own signal but not on the payoff type profile (such as sequential equilibrium), we can modify definition 4 in requiring that $\text{marg}_{S_{-i}} \pi^k[s_{-i}|s_i^\theta] > 0 \Rightarrow \text{marg}_{\Theta_i} \pi^k[\theta_i|s_i^\theta, s_{-i}] > 1 - \varepsilon$ for all $k \geq K$.

of Aghion et al. (2012) invokes a perturbation that is not a private value perturbation. Our theorem 2 shows that it is necessary for Aghion et al. (2012) to invoke a perturbation outside the class of private value perturbations that is covered by our theorem 2.

One might wonder whether other mechanisms that implement under initial rationalizability have the same robustness properties as the SR mechanism. In appendix B, we show that the initial rationalizability correspondence in any finite mechanism is upper hemicontinuous with respect to private value perturbations. This result implies that if a social choice function f is implementable in initial rationalizable strategies (under complete information) by a multistage mechanism with observable actions, then f is also robustly implementable in initial rationalizable strategies by the same mechanism. This result suggests that implementation in initial rationalizable strategies is a useful desideratum for achieving robust implementation, and the SR mechanism we develop is one of the simplest dynamic mechanisms that have this desired property.

III. Experiment 1: The SR Mechanism in Complete and Incomplete Information Environments

Part 1 of this paper suggests that the SR mechanism is robust to private value perturbations and may be robust to a variety of alternative reasoning processes and behavioral assumptions. In this section, we study the empirical properties of the mechanism using laboratory experiments in a two-person environment where buyers and sellers seek to implement a state-dependent contract with observable but nonverifiable information. To concentrate directly on the robustness properties of the SR mechanism, we consider behavior in both a complete information environment and an environment with a private value perturbation. We further benchmark behavior against a canonical SPI mechanism that is not predicted to be robust to private value perturbations.

A. Environment and Mechanisms

Our experimental environment is based on Aghion et al. (2018), which borrows its setup from Hart and Moore (2003). In each of 20 periods, a buyer and a seller are matched and the seller is randomly assigned one of two sealed containers with equal probability. One container is worth 70 ECU to the buyer, and the other container is worth 20 ECU.[17] The assigned container is always worth 0 ECU to the seller if no trade occurs.

[17] The exchange rate of ECU to Australian dollars was a rate of 2 ECU = A$1. As discussed in the text, we randomly paid two periods: one from periods 1–10 and one from periods 11–20.

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

1. In the no-noise treatment, both the buyer's compartment and the seller's compartment of the container worth 70 ECU is filled with 40 red balls and 0 blue balls. Likewise, both the buyer's compartment and the seller's compartment of the container worth 20 ECU is filled with 40 blue balls and 0 red balls.
2. In the noise treatment, the buyer's compartment of the container worth 70 ECU is filled with 39 red balls and 1 blue ball. The seller's compartment of the container worth 70 ECU is filled with 35 red balls and 5 blue balls. Similarly, the buyer's compartment of the container worth 20 ECU is filled with 39 blue balls and 1 red ball. The seller's compartment of the container worth 20 ECU is filled with 35 blue balls and 5 red balls.

The two parties in a group do not initially know which container has been allocated to the seller. However, over the course of a period, the buyer privately observes a ball drawn from the buyer's compartment of the container, and the seller privately observes a ball drawn from the seller's compartment. These signals provide complete information about the container being traded and the signal observed by their matched partner in the no-noise treatment. In the noise treatment, the buyer's signal is more accurate in identifying the value of the container than the seller's signal: the buyer will receive the correct signal 97.5% of the time, while the seller will receive the correct signal 87.5% of the time.[18] The signals will coincide 85.3% of the time. Throughout the rest of the paper, we refer to the red signal as the *high signal* and the blue signal as the *low signal*.

In each period, the buyer and seller have the task of trading the container using either an SR mechanism or an SPI mechanism. Both mechanisms use near identical price schedules, bonuses, and fines to implement a state-contingent trading scheme that (under complete information) trades containers worth 20 ECU at a price of 10 ECU and containers worth 70 ECU at a price of 35 ECU. The mechanisms are implemented as follows:

The SR mechanism.—In treatments using the SR mechanism, each period is comprised of four stages: a report stage, a signal stage, a verification stage, and an arbitration stage. In the report stage, both the buyer and the seller are asked to privately report the value of the container under two

[18] In the theory section of the paper, private value perturbations represented signals over every player's payoff type. In our experiment, sellers have only a single payoff type, and thus we only need to generate signals about the buyer's payoff type. The value of the container fully describes this payoff type.

scenarios: the scenario where he or she observes the high signal and the scenario where he or she observes the low signal. The buyer and seller may report a high value of 70 or a low value of 20 in each scenario.

In the signal stage, the buyer privately draws a ball from the buyer's compartment of the container. After observing the signal, the computer makes a formal report that corresponds to the buyer's decision in the report stage for that signal. Likewise, the seller privately draws a ball from the seller's compartment of the container. After observing the signal, the computer makes a formal report that corresponds to the seller's decision in the report stage for that signal.

Following the signal stage, each party is made aware of the formal report made by their matched party but not their matched partner's signal or their strategy. Thus, the strategy method that we employ generates a compete set of reports in each period but does not affect the information observed in the mechanism itself. Obtaining a complete panel of first-stage reports improves our ability to control for heterogeneity across individuals. It also reduces variation across periods that is driven by the random assignment of containers and signals to different buyers and sellers.[19]

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

In the arbitration stage, the buyer is asked to make a second report. As shown in table 1, the buyer may report a value of 0, 20, or 70.[20] We use the second report along with a fair six-sided die to determine whether trade occurs and the price.[21] If the second report of the buyer matches the first-stage report of the seller, the seller is rewarded a bonus of 40 ECU in addition to her earnings from trade. In other cases, the seller also pays a fine of 40 ECU.

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

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

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

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

TABLE 1
Trade Prices If Buyer Enters Arbitration

Buyer's Secondary Report	Outcome If Roll Is 1–3	Outcome If Roll Is 4–6
0	No trade	No trade
20	Trade at 10	No trade
70	Trade at 10	Trade at 35

$$\pi_B = \text{Value} - \text{Price} - \text{Buyer Arbitration Fee},$$

$$\pi_S = \text{Price} + \text{Seller Bonus} - \text{Seller Arbitration Fee}.$$

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

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

The seller in the mechanism next draws a signal from the seller's compartment and is informed of the formal report of the buyer. The seller next has the option to call or not call the arbitrator.[22] If the arbitrator is not called, the parties trade at a price equal to one-half of the reported value. If the seller calls the arbitrator, the buyer is fined 40 ECU and enters into the arbitration stage.

In the arbitration stage, the buyer is given a counteroffer equal to 35 if he reported a value of 20 and 85 if he reported a value of 70.[23] The buyer may accept or reject the counteroffer. If the counteroffer is accepted, trade occurs and the seller receives a bonus of 40 ECU. If the counteroffer is rejected, no trade occurs and the seller is fined 40 ECU.

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

[23] Note that the counteroffer price of 35 in the SPI mechanism is the same price that a buyer in the SR mechanism will trade at if he or she makes a second report of 70 and the dice roll is a 4–6. Thus, in both mechanisms, the price used to induce a buyer to reveal that he or she has a high signal in the arbitration stage after reporting a value of 20 in the report stage is the same.

B. Protocol

Our experiment utilized a 2 × 2 design in which we generated within-subject variation in noise and between-subject variation in mechanism. Within a session, subjects played 10 periods of the no-noise treatment followed by 10 periods of the noise treatment, using a single mechanism. All sessions consisted of exactly 20 participants who were evenly divided between buyers and sellers at the beginning of the experiments. Buyers and sellers were matched with each other at most once in each of the two information treatments.

All of the experiments were run in the Experimental Economics Laboratory at the University of Melbourne in March 2019. The experiments were conducted using the programming language z-Tree (Fischbacher 2007). A total of 12 sessions were run: six sessions using the SR mechanism and six sessions using the SPI mechanism. All of the 240 participants were undergraduate students at the university and were invited from a pool of more than 6,000 volunteers using the Online Recruitment System for Economic Experiments (ORSEE; Greiner 2015).

Upon arrival at the laboratory, participants were randomly assigned buyer and seller roles and asked to read the instructions for their assigned mechanism in the no-noise treatment. Consistent with previous implementation experiments, the instructions described the game in detail, walked through a series of examples that calculated the payoffs of both parties along the equilibrium path and along the off-equilibrium paths, and culminated in a quiz. In the quiz, the subjects were required to calculate the payoffs that each buyer and seller would receive for both potential values of the container and both on- and off-equilibrium actions.[24]

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

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

[24] While the instructions for both mechanisms were complete in describing the outcome of each set of actions in the mechanism, we did not explicitly state that the mechanism is designed to induce truthful reports for both buyers and sellers because the SPI mechanism does not have this property in the noise treatment. This precluded us from running training periods against a Nash best-responding computer, like we do in experiment 2.

We randomly selected one period from the no-noise treatment and one period from the noise treatment for payment at an exchange rate of 2 ECU to A$1. To avoid bankruptcies, participants received a show-up fee of A$35. The average payment at the end of the experiment was A$51.84. At the time of the 2019 experiments, A$1 ≈ US$0.71.

The experimental design, instructions, and analysis plan were preregistered at Open Science (https://osf.io/p6ukx). Prior to preregistration, we ran one pilot session of the SR mechanism and one session of the SPI mechanism to obtain a better estimate of the distribution of buyer misreports in order to perform power calculations. The pilots were identical to the main experiment. We do not include these pilots in the results, as they were done before finalizing the analysis plan.

C. Hypotheses

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

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

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

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

Truth telling is also the path of play in the unique subgame perfect equilibrium of the SPI mechanism in the no-noise treatment and is one of potentially many initial rationalizable strategy profiles. However, when noise is introduced, the original truth-telling equilibrium of the SPI mechanism is no longer supported by both solution concepts. Instead, two equilibria emerge: (1) a unique mixed strategy equilibrium, in which the buyer with a high signal reports a low value with a positive probability, and (2) a pure strategy equilibrium, in which the buyer with the low signal reports a high value.[25] Using subgame perfection as the basis for the null hypothesis, we predict the following:

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

As discussed earlier, truth telling corresponds to the unique initial rationalizable strategy profile of the SR mechanism, while it corresponds to one of potentially many initial rationalizable strategy profiles in the SPI mechanism.[26] The SR mechanism is also more robust to noise in the best response functions and less sensitive to retaliatory preferences. Thus, under initial rationalizability and for a number of alternative theoretical assumptions, we would predict that the SR treatment will have a greater level of truth telling than the SPI mechanism in the no-noise treatment. In testing how noise influences the two mechanisms, we provide both direct proportion tests of buyer misreports using data from only the noise treatments and a difference-in-difference estimator that uses data from both information treatments to control for potential differences in the no-noise treatment. These estimates provide both an absolute difference between the two noise treatments and the relative change in buyer misreport rates that occur when noise is introduced.

D. *Results*

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

[25] Previous experiments in Aghion et al. (2018) suggest that the introduction of noise tends to increase the proportion of buyers with a high signal who misreport but has a limited impact on the behavior of buyers with a low signal. Thus, for the purpose of conducting power calculations, we concentrated on the mixed strategy equilibrium, where buyers with the high signal mix between misreporting and telling the truth and sellers with the high signal challenge a report that does not match their signal with a positive probability. In the mixed strategy equilibrium, buyers misreport on average 13% of the time. The sample size was thus selected to detect an effect size of 0.13 at a significance level of .05 and a power of 0.80.

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

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

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

1. The SR Mechanism under Complete Information

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

Result 1. In the no-noise treatment of the SR mechanism, buyers report truthfully in 97.7% of cases in the low-signal scenario and in 94.0% of cases in the high-signal scenario. Sellers report truthfully in 86.2% of cases in the low-signal scenario and in 93.0% of cases in the high-signal scenario. Buyers who enter arbitration report the true value in 77.2% of cases.

Figure 1 shows the pattern of play in the no-noise treatment in sessions using the SR mechanism. Panel A shows the proportion of reports that were truthful in the report stage for both the buyers and the sellers, while panel B shows how buyers responded in the arbitration stage. The left-hand side of panel C shows a histogram of the aggregate number of buyer misreports over the 10 periods of the no-noise treatment. The right-hand side of panel C shows the aggregate number of periods where each seller made a misreport in the low-signal scenario.[27]

As seen in panel A, misreports were rare for buyers and uncommon for sellers. Buyers told the truth 97.7% of the time in the low-signal scenario and 94.0% of the time in the high-signal scenario. Sellers told the truth 86.2% of the time in the low-signal scenario and 93.0% of the time in the high-signal scenario.[28]

Aggregating the behavior of the buyer and seller, the SR mechanism induces truth-telling first-stage strategies in 74.7% of groups. The buyer's

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

[28] As shown in app. sec. C2, there are no apparent time trends in the behavior of buyers in the SR mechanism. However, sellers appear to be learning to tell the truth in the low-signal scenario through experience: a seller who lies in the low-signal scenario in one period lies in the same scenario only 26.1% of the time in the next period if this scenario arises and they are

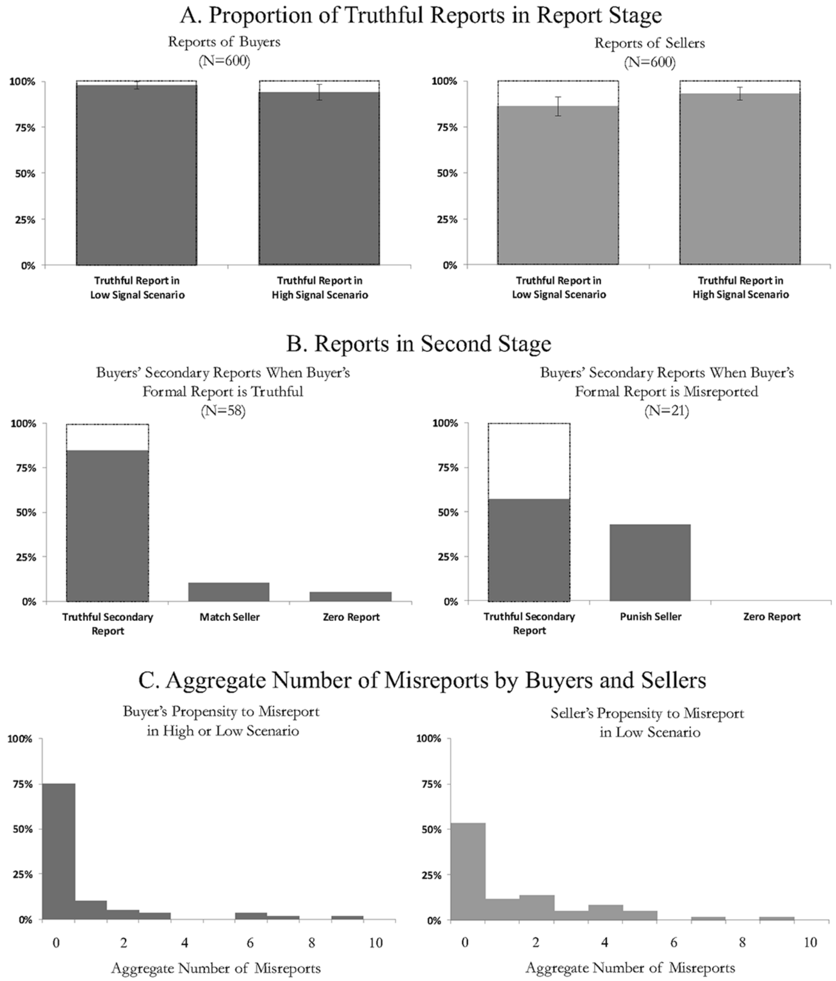


FIG. 1.—Pattern of play in no-noise treatment of SR mechanism.

formal report was truthful in 96.0% of cases, while the seller's formal report was truthful in 89.8% of cases. This led to 86.8% of groups displaying truth-telling behavior. As such, we observe only 58 cases where the buyer enters into the second stage after following his signal and 21 cases where the buyer enters the second stage after lying. As seen in panel B, buyers report truthfully in 49 out of 58 cases (84.5%) where arbitration is due to a seller misreport and in 12 out of 21 cases (57.1%) where arbitration

able to observe the buyers response. Since sellers rarely switch from a truthful strategy to a lying strategy, the aggregate truth-telling rate for this scenario increases from 82.3% in periods 1–5 to 90.0% in periods 6–10.

is due to the buyers own misreport.[29] On the basis of the distribution of buyer secondary reports, sellers who tell the truth earn an average of 21.6 ECU more than sellers who misreport in the low-signal scenario and 17.2 ECU more than sellers who misreport in the high-signal scenario. Buyers earn 13.6 ECU more for telling the truth in the low-signal scenario and 33.3 ECU more in the high-signal scenario. Thus, the mechanism generates strong incentives for truthful reporting for both parties.

Finally, as seen in panel C, the majority of buyers and sellers are truthful in all 10 periods, and there are no individuals who misreport in every period. This implies that (1) the mechanism induces truth-telling strategies for the majority of individuals from the very start of the experiment and (2) there is little heterogeneity in strategy. On the buyer side, 45 out of 60 buyers are truthful in every period, and an additional nine buyers make one or two misreports. On the seller side, 32 sellers never make a misreport with a low signal, and an additional 15 make one or two misreports.

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

2. The SR Mechanism under a Private Value Perturbation

Our second hypothesis predicts that in the noise treatment, both buyers and sellers will continue to make truthful reports in the first stage in both the scenario where they receive a high signal and the scenario where they receive a low signal. We also predict that buyers will continue to report truthfully in the second stage. The data are largely consistent with these predictions:

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

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

[29] While a truth-telling rate of only 57.1% is low, the sample here is highly selected and based on only a few buyers who misreport in multiple periods. Thus, it is unlikely to be indicative of how most participants are likely to play in the second stage.

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

Figure 2 shows the pattern of play in the noise treatment of the SR mechanism and is directly comparable to figure 1. As can be seen in the left-hand side of panel A, the frequency of truthful reports by buyers remains high, with the buyer reporting truthfully in 96.5% of cases with

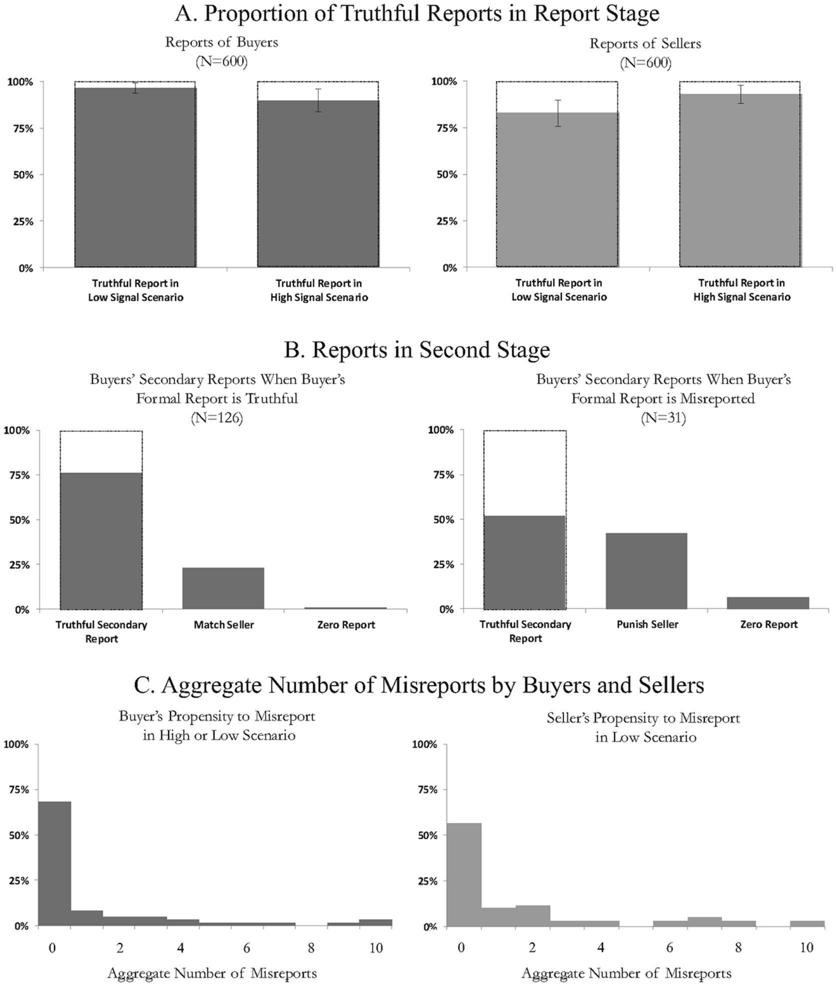


Fig. 2.—Pattern of play in noise treatment of SR mechanism.

the low signal and in 90.0% of cases after the high signal. The proportion of misreports made by buyers in the noise treatment is not significantly different from the proportion of misreports made in the no-noise treatment in a simple regression where buyer misreports is regressed on the noise treatment dummy ($p = .12$).[30]

Sellers report truthfully in 82.8% of cases after the high signal and 92.8% of cases after the low signal. The proportion of seller misreports in the low-signal scenario is not significantly different from the no-noise treatment using a simple regression where seller misreports in the low-signal scenario are regressed on the noise treatment dummy ($p = .354$).

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

Finally, as seen in panel C, the majority of buyers make truthful reports in both the high- and the low-signal scenario, and the majority of sellers make truthful reports in the low-signal scenario. The distribution of buyer misreports in the noise treatment is not significantly different in the distribution of misreports in the no-noise treatment using a Wilcoxon sign-ranked test ($p = .14$). Likewise, there is no significant difference in the distribution of seller challenges with the low signal ($p = .65$).

At the aggregate level, 68.5% of groups exhibit first-stage truth-telling strategies in the noise treatment. This is relatively similar to the 74.7% of groups that exhibit first-stage truth-telling strategies in the no-noise treatment, and there is no significant difference in these proportions when a variable that is 1 if a group exhibits first-stage truth-telling strategies and zero otherwise is regressed on the treatment variable ($p = .15$).

We note that the proportion of groups exhibiting first-stage truth-telling strategies is jointly determined by the proportion of buyers who report the truth and the proportion of sellers who report the truth. As such, the difference of 6.2% in the proportion of groups exhibiting first-stage truth-telling strategies reflects a decrease in buyer truth-telling rates of 4.8% and a decrease in seller truth-telling rates of 2.5%. As such, the changes in economically relevant behavior when introducing noise is relatively small.[31]

[30] Unless otherwise specified, regressions on buyer's behavior is clustered by buyer, regressions on seller's behavior is clustered by seller, and regressions on group outcomes are clustered by both buyers and sellers, using the procedure by Cameron, Gelbach, and Miller (2008). We use clustering rather than random effects to allow for more complex error structures that might arise between the two treatments. Alternative session-level specifications (including a random effects regression with session-level clustering) are provided in app. sec. C2.

[31] Our original design was powered to detect a 16% decrease in the truth-telling rate (81% to 65%) when noise was introduced, with a significance level of .05 and a power

E. *The Relative Performance of the SR Mechanism*

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

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

Figure 3 compares the proportion of periods in which the buyer misreports his signal in the no-noise treatments (*top*) and noise treatments (*bottom*). The error bars are 95% confidence intervals. As can be seen in the top panel, buyers misreport in 7.7% of cases in the no-noise treatment with the SR mechanism and in 25.5% of cases in the no-noise treatment with the SPI mechanism.[32] This difference is significant in a simple regression where buyer lies are regressed on the SPI treatment dummy with data restricted to the no-noise treatments ($p < .01$).[33]

As can be seen in the bottom panel, buyers misreport in 12.5% of cases in the noise treatment with the SR mechanism and in 43.3% of cases with the SPI mechanism.[34] Consistent with the theoretical predictions, there are significantly more lies in the SPI mechanism in the noise treatment than the SR mechanism ($p < .01$).

Finally, using the no-noise treatment as a baseline, there is a 4.8% increase in buyer misreports when noise is introduced in the SR mechanism

of 0.8. The effect size was chosen because it represented a 10% decrease in the truth-telling rate of the buyer and the seller, starting from a baseline in the no-noise treatment of 90%.

[32] In app. sec. C1, we show that buyers in the high-signal scenario of the no-noise treatment are truthful in 77.5% of cases and misreport in 22.5% of cases. Thus, most of these misreports are cases where a buyer misreports in the high-signal scenario. We also show that sellers do not always challenge lies in the high-signal scenario and that buyers who are legitimately challenged often reject welfare-improving counteroffers. Thus, the data are consistent with the alternative initial rationalizable strategy profile outlined in n. 26, where a seller (correctly) believes that the buyer may not be rational after observing the buyer report a low value with the high signal. The pattern of play is also consistent with a model where some buyers have retaliatory preferences.

[33] An extended analysis of the SPI mechanism and its performance relative to the SR mechanism is provided in app. sec. C2. As seen there, we also perform Mann-Whitney-Wilcoxon tests on the distribution of buyer reports across the treatments at the buyer level and the session level. The null hypothesis of these tests are rejected at the .01 level for all comparisons between the SR and SPI treatments and between the noise and no-noise treatments of the SPI mechanism.

[34] In app. sec. C1, we show that buyers in the high-signal scenario of the noise treatment are truthful in only 60.0% of cases and misreport in 40.0% of cases. Thus, consistent with theory, the high aggregate misreport rate is again a result of buyer lies in the high-signal scenario.

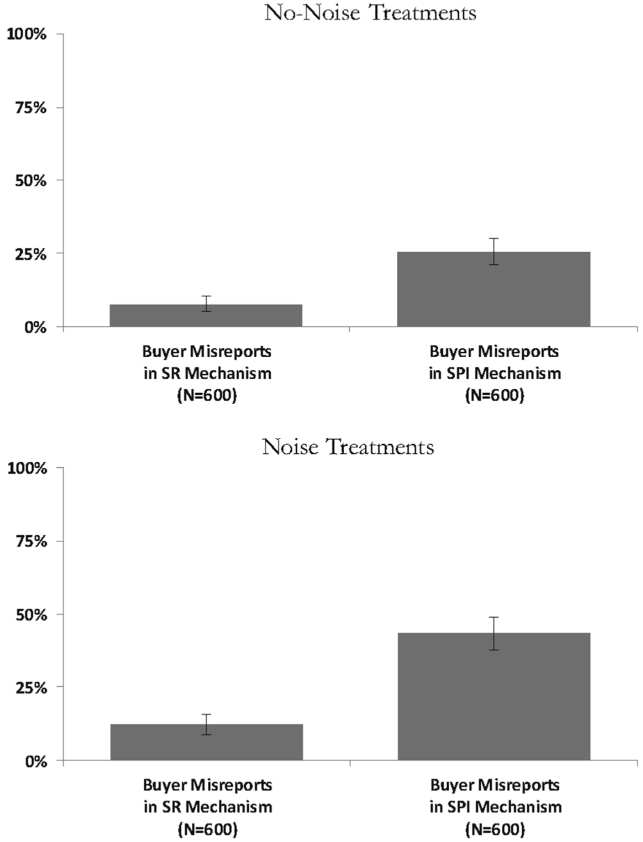


FIG. 3.—Proportion of buyer misreports in SR and SPI mechanisms in no-noise and noise treatments.

and an 18.8% increase in buyer lies when noise is introduced in the SPI mechanism. Using a simple difference-in-difference estimator where buyer lies is regressed on the SPI treatment variable, the noise treatment variable, and the interaction of noise and the SPI treatment variable, we find that the interaction term is significantly different from zero ($p = .03$). Thus, the introduction of noise increases the number of lies both in absolute terms and when considering only the relative change in misreport rates that occur when noise is introduced.

At the aggregate level, groups display truth-telling behavior in 78.3% of cases in the no-noise treatment with the SPI mechanism. This is significantly different from the 86.3% of groups that display truth-telling behavior in the no-noise treatment of the SR mechanism when a variable that is 1 if a group displays truth-telling behavior and zero otherwise is regressed on the SPI mechanism treatment variable ($p = .02$). Groups display truth-telling behavior in 72.3% of cases in the noise treatment with the SPI mechanism. This is not significantly lower than in the no-noise treatment with the SPI mechanism ($p = .06$). However, it is significantly lower than the 82.5% of groups that display truth-telling behavior in the noise treatment of the SR mechanism ($p = .01$).

We note that although the truth-telling rates of the SR mechanism are higher than the SPI mechanism, our first experiment does not have any explicit benefit for inducing truthful reports and differences in welfare between the two mechanisms are due to periods where both parties were fined and where trade may not have occurred. In appendix section C1, we show that while buyers enter into the arbitration stage more often in the SPI mechanism, the buyer frequently accepts the counteroffer after a legitimate challenge reducing the overall costs of arbitration in this mechanism. As a result, the two mechanisms actually have very similar levels of efficiency. The overall per-period earnings of 15.8 in the SR mechanism are not significantly different from the per-period efficiency of 15.6 in the SPI mechanism ($p = .85$). In the next experiment, we explore an environment where a high truth-telling rate is predicted to lead to increased investments and where there are both benefits and costs associated with introducing an implementation mechanism.

IV. Experiment 2: The Two-Sided Holdup Problem

In this section, we report on a second set of experiments that are designed to test whether the SR mechanism can be used to induce efficient investments in a bilateral trade environment where values and costs are observable but nonverifiable. To keep the environment as simple as possible for participants, we explore behavior in an environment where both costs and values are commonly observed and where there is no noise.

As discussed in the introduction, it has traditionally been assumed that state-contingent contracts are infeasible in settings with observable but nonverifiable information and that formal contracting is ineffective. Thus, identifying implementation mechanisms that have good empirical properties has the potential of expanding the scope of environments for which first-best outcomes can be achieved.

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

Prior to bargaining over the production and exchange of the widget, both the buyer and the 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 costs the seller e_S 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 $v(75) = 320$. On the basis of 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 nonverifiable by a court. These assumptions imply that while the true cost and true value are 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 for holdup 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.

Finally, the set of prices $p = \theta v(e_S) + (1 - \theta)c(e_B)$ represent the set of expected prices that can arise in the second stage if bargaining power is randomly assigned to the seller with probability θ and the buyer with probability $1 - \theta$. These prices also correspond to the set of prices that are predicted by the weighted Nash bargaining solution when the seller is assigned a weight of θ. There is no price in this set that provides incentives for both the buyer and the seller to choose the efficient level of effort. Consequently, both parties can be made better off if they can commit to a trade price that is sensitive to both v and c if the resulting price schedule can induce efficient effort for both parties.

Our experiment explores whether it is possible to construct a contract that is based solely on publicly observable reports that can generate the price schedule given in table 2, using our SR mechanism. This price

TABLE 2
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

schedule makes first-best investment for both parties the unique Nash equilibrium, provided that truth telling is realized in the mechanism.

A. *Main Treatment*

Each session of our experiment consists of 20 periods split into two phases, each consisting of 10 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 3, 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.

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 SR mechanism consists of a report stage, a verification stage, and an arbitration stage. In the report stage, 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 the verification stage. If all reports coincide, the buyer and seller trade at the report-specific prices given in table 2. Prices in this table were constructed using the function

$$P^{SR}(\hat{v}, \hat{c}) = (\hat{v} - \underline{v}) - (\bar{c} - \hat{c}) + \frac{\underline{v} + \bar{c}}{2}, \tag{6}$$

where $\hat{c}$ is the jointly reported cost, $\hat{v}$ is the jointly reported value, $\bar{c}$ 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 the seller report the truth, the buyer receives the marginal surplus created from his investment and

TABLE 3
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

the seller receives the marginal surplus created from her investment.[35] 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. 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% chance of entering arbitration and both parties are fined 300.[36]

If the buyer enters into the arbitration stage, he is asked to make a second report regarding the value of the good. As shown in panel A of table 4, 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 4, 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 are sunk. However, the seller does not have to produce the good and has 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 and seller opt in, they are informed that the arbitrator is available, and play continues as in the first 10 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

[35] 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. If we start 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 ($45 - 10 - 0 = 35$).

[36] Note that there is a slight difference between this implementation of the SR mechanism and the one described in sec. II.C. In sec. II.C, the buyer and seller both make a second report in situations where the value report and cost report differed and one of these would then randomly be used to determine the actual allocation. The incentive properties of the two variants of the SR mechanism are actually the same.

TABLE 4
Trade Prices in Buyer and Seller Arbitration Stages

	Outcome If Roll Is 1–3	Outcome If Roll Is 4–6
	A. Buyer Enters into Arbitration	
Buyer's secondary report:		
200	No trade	No trade
250	Trade at 205	No trade
320	Trade at 205	Trade at 255
	B. Seller Enters into Arbitration	
Seller's secondary report:		
130	No trade	No trade
80	Trade at 125	No trade
10	Trade at 125	Trade at 75

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/she cannot determine whether his/her counterparty opted in or out.

B. 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.

1. The KTH Treatment

Kartik, Tercieux, and Holden (2014) 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

$$P^H = (\hat{v}_S - \underline{v}) - (\bar{c} - \hat{c}_B) + \frac{\underline{v} + \bar{c}}{2}. \tag{7}$$

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.[37]

While trade always occurred in the KTH mechanism, buyers and sellers 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

$$F_B^H = \max\{0, \hat{c}_S - \hat{c}_B\}, \tag{8}$$

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

$$F_S^H = \max\{0, \hat{v}_S - \hat{v}_B\}, \tag{9}$$

where $\hat{v}_S$ and $\hat{v}_B$ are the value reports of the seller and buyer. The fines are set such that (1) 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.[38]

As structured, the prices and fees are set such that both the buyer and the seller are indifferent between making a truthful report or making a lie when the other party always tells the truth. As shown in Kartik, Tercieux, and Holden (2014), 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.[39]

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

[38] 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 $\hat{v}_B = 320$ and $\hat{c}_B = 130$, the trade price is 285. The buyer surplus is 35 ($= 320 - 285$). If, instead, the buyer lies and makes reports of $\hat{v}_B = 320$ and $\hat{c}_B = 10$, the trade price is 165, but the buyer is fined 120 ($= 130 - 10$). The buyer's surplus thus remains 35 ($= 320 - 165 - 120$).

[39] 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 noninvesting 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 app. sec. C6 for a broader discussion.

2. The SPI Treatment

While our initial experiment has documented issues that arise in the use of 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 and Repullo (1988).

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

$$p^{SPI} = (\hat{v}_B - \underline{v}) - (\bar{c} - \hat{c}_S) + \frac{\underline{v} + \bar{c}}{2}, \tag{10}$$

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 $\bar{c}$ 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

$$\hat{p}_S^{SPI} = \hat{c}_S - 5. \tag{11}$$

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

$$\hat{p}_B^{SPI} = \hat{v}_B + 5. \tag{12}$$

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

C. Protocol

Our experimental design utilizes a between-subject 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 2016. The experiments were conducted using the programming language z-Tree (Fischbacher 2007). A total of 20 sessions were run: eight sessions using the SR mechanism, four sessions using the no-mechanism baseline, four sessions using the KTH mechanism, and four 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 6,000 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.[40] 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 among the participants.

Subjects next played six periods where the computer played the role of their matched partner.[41] In each period, the computer made maximal investments and truthful announcements.[42] 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 increase their initial surplus to reduce the potential for bankruptcies.

[40] 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 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.

[41] In the third set of experiments discussed in app. D, we ran sessions of the SR mechanism without the initial training periods. We find that behaviors in the two treatments are similar but that the training periods tend to reduce noise in early periods that appears to be a result of experimentation and confusion. This is consistent with the results in Fehr, Powell, and Wilkening (2021), where initial training periods tend to decrease lies in a SPI mechanism.

[42] To be as close as possible to the other treatments, we also had the computer choose maximal investments in the fixed price treatment.

After completing the six rounds against the computer, we read additional oral instructions that detailed the additional 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 on the basis of the earnings subjects had accumulated throughout the experiment, with an exchange rate of 35 ECU to A$1. In addition, subjects received a show-up fee of A$22. The average payment at the end of the experiment was A$51.14. At the time of the 2016 experiments, A$1 = US$0.74.

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 player 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 six bankruptcies (out of 160 sessions; 2.5%) 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 appendix section C5.[43]

D. *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:

[43] 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 to ensure that incentive payments were salient: 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 receives 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. In app. D, we report on a third set of experiments with the SR mechanism in a one-sided holdup setting where a pay-one-period protocol was feasible. The behavior in the two sets of experiments is similar, suggesting that bankruptcy rules are not a major factor in behavior. Aghion et al. (2018) also conducts sessions using both a pay-one-period protocol and a pay-all-period protocol with the SPI mechanism and finds similar behavior across the two treatments.

Hypothesis 4. The path of play under the SR mechanism involves both the buyer and the 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 5. Buyers and sellers are predicted to opt in to the SR mechanism.

Under the assumptions of subgame perfection and a weak preference for honesty, the SR 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 use this set of assumptions as the basis for the following null hypothesis:

Hypothesis 6. 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.

The SR treatment is predicted to be robust to noise in the best response function, moderate levels of negative reciprocity, heterogeneity in honesty preferences, and small deviations from common knowledge. Further, it relies on a less stringent assumption on how beliefs evolve. Thus, under a number of alternative assumptions, we would predict that the SR treatment will be more efficient than the other mechanisms. Our alternative H_1 hypothesis is thus that the SR treatment has higher efficiency than the other three treatments, with no explicit ordering between the fixed treatment, the KTH treatment, or the SPI treatment.

E. Results

We describe the results of the main experiment in this section. Section IV.E.1 uses data from the eight sessions that use the SR mechanism to study hypotheses 4 and 5. Section IV.E.2 uses data from all sessions to make comparisons between the SR mechanism and the other three treatments.

1. Behavior in the SR Mechanism

Result 5. In phase 1 of the experiment, the SR mechanism induces truth-telling behavior in over 93% of cases. Buyers and sellers make efficient investments in over 80% of cases. The efficient outcome occurs in 74.2% of cases.

Figure 4 displays the patterns of play we observed in the first 10 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% 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% of cases, while sellers chose the optimal level of investment in 84.8% 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, we find that buyers made truthful value reports in 98.1% of cases and truthful cost reports in 94.0% of cases. Sellers made truthful value reports in 93.1% of cases and truthful cost reports in 97.8% 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 counterparty 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 counterparty. 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.[44]

While the results in figure 4 are presented as the aggregate of all 10 periods, there are only very small changes in investment and reporting decisions over time. In appendix section C4, we report on the investment and report decisions over time. We show that the proportion of buyers and sellers who chose high investment starts above 80% in period 1 and increases to about 90% in periods 6–10 for both parties. We also show that the proportion of buyers and sellers who report truthfully is stable over time and that 81.3% of buyers and 77.5% of sellers make one lie or less.

[44] Panel C shows secondary reports in both the case where a buyer or seller enters arbitration because of their own lie or because of the lie of their counterparty. Looking only at cases where a buyer enters into arbitration because of a seller lie, we find that buyers make a truthful report in 21 of 42 cases and match the seller in 11 of 42 cases. Looking only at cases where a seller enters into arbitration because of a buyer lie, we find that sellers make a truthful report in 21 of 35 cases and match the buyer in nine of 35 cases. There is no combination of investments and reports where a buyer or seller has a positive return for lying.

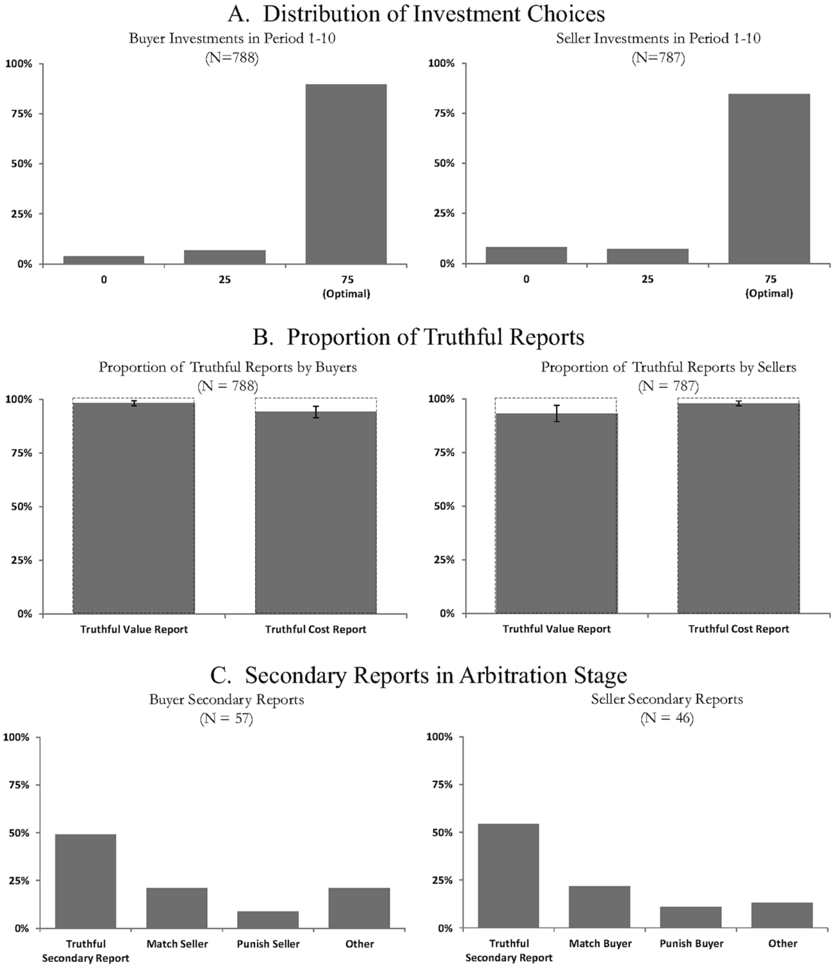


FIG. 4.—Pattern of play in first 10 periods of SR mechanism.

Overall, our data suggest that the SR mechanism is highly effective in inducing truthful reports and in inducing efficient investment. In aggregate, 89.3% of dyads improved their performance relative to the theoretical no-mechanism benchmark of 70, and 74.2% 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 are high levels of efficiency even in period 1.[45]

[45] In period 1, 83.8% of dyad pairs improved their performance relative to the theoretical no-mechanism benchmark of 70, and 67.5% of dyad pairs achieved the first best.

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

Result 6. Buyers opt in to the mechanism in 77.1% of cases, while sellers opt in to the mechanism in 76.2% of cases. The proportion of buyers and the proportion of sellers who opt in to the mechanism are increasing over time, and 90.5% of dyad pairs who opt in to the mechanism exhibit efficient truth-telling behavior and achieve the efficient outcome.

Panel A of figure 5 shows opt-in rates of buyers and sellers in phase 2 of the SR mechanism. As can be seen, opt-in rates for both buyers and sellers begin near 60% and increase to roughly 85% by periods 16–20. On average, buyers opt in to the mechanism 77.1% of the time, while sellers opt in to the mechanism 76.2% of the time. Given these opt-in rates, 59.4% of the groups had the SR mechanism available.[46]

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. Diamonds indicate groups with the mechanism, and circles indicate groups without the mechanism. 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.

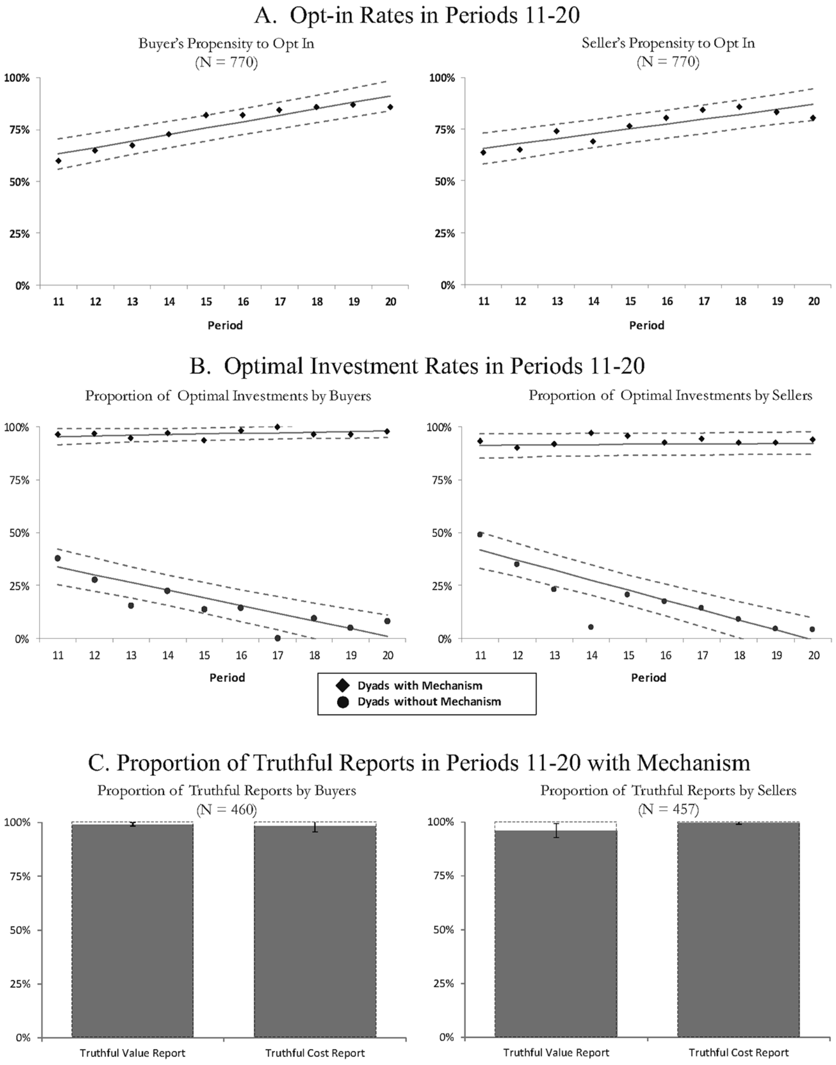
Panel C shows the proportion of truthful announcements by buyers and sellers in dyad pairs where buyers and sellers opt in to the mechanism. Buyers are truthful in almost all periods, while all but one seller is truthful in all periods.

In aggregate, 90.5% of groups who opted in to the mechanism exhibited efficient truth-telling behavior and achieve the efficient outcome. An additional 3.9% of groups made suboptimal investments but reported truthfully in the report stage. Buyers made truthful secondary reports in nine of the 14 cases where the buyer entered in arbitration, while sellers made truthful secondary reports in four of five cases. Buyers match their counterparty's first-stage report in two out of 14 cases, while sellers match their counterparty's first-stage misreport in one out of five cases.

2. The Relative Performance of the SR 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 sellers opt in to the mechanism at a high frequency and that the efficient outcome occurs in over 90% of

[46] Looking at the aggregate number of opt-in decisions of buyers and sellers, we find that 37.0% of buyers and sellers always opted in, while an additional 40.3% opted in between seven and nine times; 5.2% of buyers and sellers never opted in, and the remaining 17.5% of buyers and sellers opted in between one and six times.



A. Opt-in Rates in Periods 11-20

Buyer's Propensity to Opt In (N = 770)

Seller's Propensity to Opt In (N = 770)

B. Optimal Investment Rates in Periods 11-20

Proportion of Optimal Investments by Buyers

Proportion of Optimal Investments by Sellers

C. Proportion of Truthful Reports in Periods 11-20 with Mechanism

Proportion of Truthful Reports by Buyers (N = 460)

Proportion of Truthful Reports by Sellers (N = 457)

Fig. 5.—Pattern of play in periods 11–20 of SR mechanism.

dyads where the parties have opted in to the mechanism. We now compare the performance of the mechanism with 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 7. 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 from efficiency in the fixed price treatment or the KTH treatment. Efficiency in the KTH treatment is significantly lower than efficiency in the SR and fixed price treatments.

Support for result 3 is provided in panel A of figure 6, 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.[47] The error bars are 95% confidence intervals.

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% higher than efficiency in the fixed price treatment, 35% higher than efficiency in the SPI treatment, and 62% 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: $p = .03$; SR vs. SPI: $p < .01$; SR vs. KTH: $p < .01$).[48]

The average per-period efficiency of the SPI treatment is 35.5. This level of efficiency is not significantly different from efficiency found in the fixed price treatment ($p = .33$) or the KTH treatment ($p = .22$). As was noted in section III.C, 20% of participants in the SPI treatment went bankrupt in the treatment. We show in appendix section C5 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.

[47] In app. sec. C7, 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 49.3 and 43.6 and similar to the efficiencies shown here. For the SPI mechanism, efficiencies are 44.3 and 10.6. 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.

[48] We also compared treatments nonparametrically. The Kruskal-Wallis test of whether the four treatments are drawn from the same distribution is rejected at $p < .001$ ($\chi^2(3) = 48.48$). 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 0.05. Both the fixed price treatment and the SPI treatment have a higher efficiency than the KTH treatment. There is no significant difference between the fixed price treatment and the SPI treatment.

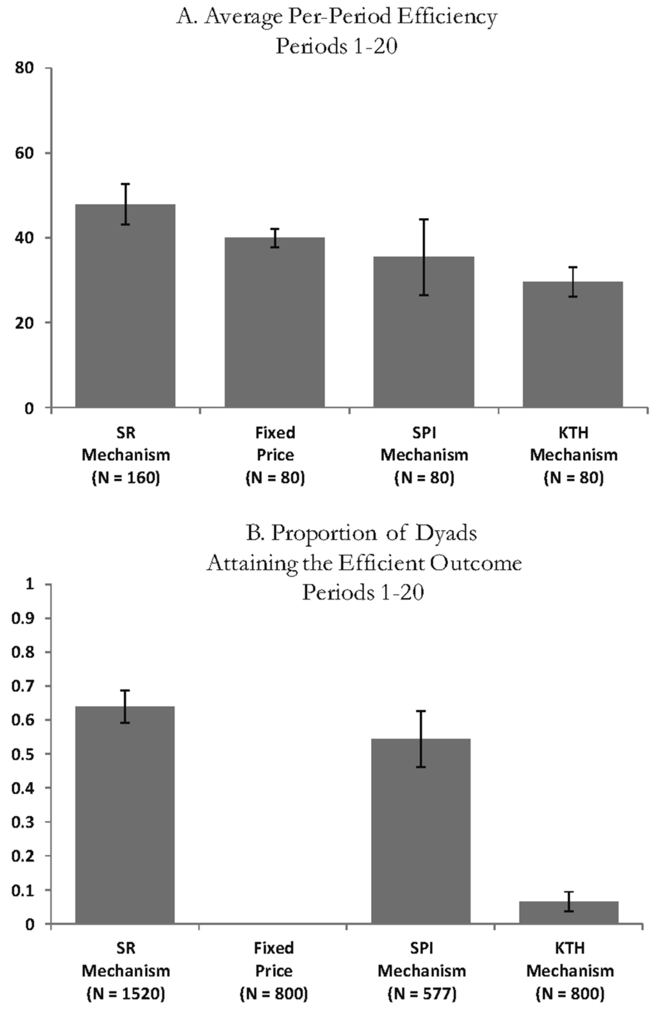


Fig. 6.—Average per-period efficiency and proportion of dyad pairs attaining efficient outcome across treatments.

We also show that subjects who lie and are challenged reject the counteroffer 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 Fehr, Powell, and Wilkening (2021), 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 the SR treatment ($p < .01$) and fixed price treatment

($p < .01$). As shown in appendix section C6, 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. When we look 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 counterparty.

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 zero is observed in 86.3% of cases in periods 11–20.

Panel B of figure 6 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% confidence intervals of each proportion with errors clustered at the individual seller level.[49] As can be seen, 63.9% 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 1 if a dyad reaches the first best is regressed against the treatment dummies (SR vs. fixed price: $p < .01$; SR vs. SPI: $p = .046$; SR vs. KTH: $p < .01$).

V. 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 by Moore and Repullo (1988), who show that any social choice function can be implemented as the unique subgame perfect equilibrium of a suitably constructed multistage mechanism in economic environments.[50]

However, there is also a long tradition in game theory (see, e.g., Fudenberg, Kreps, and Levine 1988; Monderer and Samet 1989; Dekel and Fudenberg 1990; Kajii and Morris 1997) of skepticism about the robustness of refinements of Nash equilibrium in extensive form games to

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

[50] that is, with transferable utility or with at least one divisible private good.

small perturbations of the environment. Aghion et al. (2012) raise these types of concerns in the context of implementation theory, and Aghion et al. (2018) and Fehr, Powell, and Wilkening (2021) illustrate them as a practical matter in laboratory settings.

The key issue is that extensive form mechanisms give rise to consideration of how beliefs evolve when unexpected play occurs. These considerations drive the nonrobustness 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 SR mechanism fully implements any social choice function under initial rationalizability in complete information environments. This solution concept iteratively deletes strategies that are not best replies but only mandates rationality and common beliefs at the beginning of the game. Crucially, it makes no assumption about how beliefs evolve after zero probability events occur.

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

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

Relative to the virtual implementation approach of Abreu and Matsushima (1992), we have concentrated on exact implementation. The question of what can be accomplished with virtual implementation under initial rationalizability remains an open but interesting one. In particular, it may be interesting to explore whether virtual implementation can be combined with the SR mechanism to accommodate a larger class of information perturbations and/or to decrease the size of off-equilibrium transfers.

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 simple revelation game studied in our experiments. Understanding the efficacy of our SR mechanism—or a suitably adapted variant—in these richer environments may be a fruitful direction for further work.

Appendix A

Theory

A1. Proof of Theorem 1

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

A1.1. 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 an action 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. QED

A1.2. Interstage 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 $\hat{\theta}_i^i \in \Theta_i$; that is, player i must report the type of player $i - 1$ truthfully at stage 1.

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

$$\mathrm{marg}_{M_j}\mu_i\left[\left\{m_j \in M_j : m_j^2\left(m^1\right) = \theta_j\right\}\right] = 1 \text{ if } j \in \mathcal{I}^*\left(m^1\right).$$

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

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

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

For player i, reporting $\hat{\theta}_{i-1}^i \neq \theta_{i-1}$ will result in either the penalty T (if $\hat{\theta}_{i-1}^i \neq \hat{\theta}_{i-1}^{i-1}$) or no transfer (if $\hat{\theta}_{i-1}^i = \hat{\theta}_{i-1}^{i-1}$), while reporting $\hat{\theta}_{i-1}^i = \theta_{i-1}$ will result in the reward T, which is due to the incentive transfer triggered at stage 2. Thus, the transfer gain from reporting $\hat{\theta}_{i-1}^i = \theta_{i-1}$ relative to $\hat{\theta}_{i-1}^i \neq \theta_{i-1}$ is at least T. Since $T > D$, reporting $\hat{\theta}_{i-1}^i = \theta_{i-1}$ 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, 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. Again, since $T > D$, reporting $\hat{\theta}_{i-1}^i = \theta_{i-1}$ in stage 1 is strictly better for player i than reporting $\hat{\theta}_{i-1}^i \neq \theta_{i-1}$.

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

A1.3. Within-Stage Coordination Condition

Claim 3. If $m_i \in R_{i,3}^{\Gamma(\theta)}$, then $m_i^1 = (\theta_i, \theta_{i-1})$.

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

$$\mu_i[m_{-i}^1] > 0 \Rightarrow m_{i+1}^1 = (\hat{\theta}_{i+1}^{i+1}, \theta_i) \text{ for some } \hat{\theta}_{i+1}^{i+1} \in \Theta_{i+1}.$$

That is, player $i + 1$ makes a truthful announcement about player i's type in the first stage. Hence, if player i misreports his/her own type by announcing some $\hat{\theta}_i^i \neq \theta_i$, he/she will be penalized by T. Since $T > D$, player i's unique best response is to truthfully announce his/her own type in the first stage. Hence, every player i will truthfully report his/her type at the first stage, that is, $\hat{\theta}_i^i = \theta_i$. Combining this with claim 2, we conclude that $m_i^1 = (\theta_i, \theta_{i-1})$. It follows that if $m_i \in R_{i,3}^{\Gamma(\theta)}$ for all i, then $g(z(m)) = f(\theta)$ and $\tau_i(z(m)) = 0$. QED

A2. Proof of Theorem 2

To prove theorem 2, we continue to use the same SR mechanism that we define in section II.C. Like the proof of theorem 1, we prove theorem 2 in three steps. In the proof, let $\{\pi^k\}_{k=1}^{\infty}$ denote a private value perturbation to π^{CI}.

A2.1. Truth-Telling Condition

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

Proof. Let $m^1 \in M^1$ be a stage 1 action profile such that $i \in \mathcal{I}^*(m^1)$. For any $\tilde{m}_i \in R_{i,1}(s_i^\theta | \Gamma(\pi^k))$, $\tilde{m}_i$ is a sequential best response to some CPS $\mu_{i,k}$ over $\Theta \times S_{-i} \times M_{-i}$ such that $\mu_{i,k}$ is consistent with s_i^θ under π^k. By lemma 1, there exists $\varepsilon \in (0, 1)$ small enough such that if $\text{marg}_{\Theta_i} \mu_{i,k}[\theta_i | m^1] > 1 - \varepsilon$, then $\tilde{m}_i^2(m^1) = \theta_i$ for any sequential best reply $\tilde{m}_i$ against $\mu_{i,k}$. We now fix one such ε and show that for any k large enough, we have

$$\text{marg}_{\Theta_i} \mu_{i,k}[\theta_i | m^1] > 1 - \varepsilon.$$

Since $\mu_{i,k}$ is consistent with signal s_i^θ under π^k, by definition 5, there is a sequence of totally mixed probability distributions $\{\mu_{i,k}^q\}_{q=1}^{\infty}$ such that (i) for every $(\theta', s_{-i}, m'_{-i}) \in \Theta \times S_{-i} \times M_{-i}$, we have $\mu_{i,k}[\theta', s_{-i}, m'_{-i} | h] = \lim_{q \to \infty} \mu_{i,k}^q[\theta', s_{-i}, m'_{-i} | h]$ for any $h \in \mathcal{H}$; and (ii) for every $q \geq 1$, there exists $\sigma_{-i.k}^q : \Theta_{-i} \times S_{-i} \to \Delta(M_{-i})$ such that

$$\mu_{i,k}^q[\theta', s_{-i}, m'_{-i}] = \sigma_{-i,k}^q[m'_{-i} | \theta'_{-i}, s_{-i}] \pi^k[\theta', s_{-i} | s_i^\theta]. \tag{A1}$$

First, since $\mu_{i,k}$ is consistent with signal s_i^θ under π^k, we have

$$\begin{aligned}\mu^q_{i,k}[\theta'_i, \theta'_{-i}, s_{-i}, m'_{-i}] &= \sigma^q_{-i\,.\,k}[m'_{-i}|\theta'_{-i}, s_{-i}]\pi^k[\theta'_i, \theta'_{-i}, s_{-i}|s^\theta_i] \\ &= \sigma^q_{-i\,.\,k}[m'_{-i}|\theta'_{-i}, s_{-i}]\pi^k[\theta'_i|s^\theta_i, s_{-i}, \theta'_{-i}]\pi^k[\theta'_{-i}, s_{-i}|s^\theta_i] \qquad \text{(A2)} \\ &= \mathrm{marg}_{\Theta_{-i}\times S_{-i}\times M_{-i}}\mu^q_{i,k}[m'_{-i}, \theta'_{-i}, s_{-i}]\pi^k[\theta'_i|s^\theta_i, s_{-i}, \theta'_{-i}].\end{aligned}$$

Second, since $\{\pi^k\}^\infty_{k=1}$ is a private value perturbation to π^{CI}, we have, for any k sufficiently large,

$$\mathrm{marg}_{\Theta_{-i}\times S_{-i}}\pi^k\left[\theta'_{-i}, s_{-i}|s^\theta_i\right] > 0 \Rightarrow \pi^k\left[\theta_i|s^\theta_i, s_{-i}, \theta'_{-i}\right] > 1 - \frac{\varepsilon}{|\Theta_i|}, \qquad \text{(A3)}$$

where $|\Theta_i|$ denotes the cardinality of Θ_i. Denote by Ω^k_{-i} the support of $\mathrm{marg}_{\Theta_{-i}\times S_{-i}}\pi^k[\cdot|s^\theta_i]$. Then, it follows from (A3) that

$$\frac{\min_{(s_{-i},\theta'_{-i})\in\Omega^k_{-i}}\pi^k[\theta_i|s^\theta_i, s_{-i}, \theta'_{-i}]}{\sum_{\theta'_i}\max_{(s_{-i},\theta'_{-i})\in\Omega^k_{-i}}\pi^k[\theta'_i|s^\theta_i, s_{-i}, \theta'_{-i}]} > 1 - \varepsilon. \qquad \text{(A4)}$$

Now given $m^1 \in M^1$, we compute player i's conditional belief on θ_i under $\mu^q_{i,k}$ as follows:

$$\begin{aligned}&\mathrm{marg}_{\Theta_i}\mu^q_{i,k}[\theta_i|m^1] \\ &= \sum_{s_{-i}}\sum_{m'_{-i}\in M_{-i}(m^1)}\mu^q_{i,k}[\theta_i, s_{-i}, m'_{-i}|m^1] \\ &= \frac{\sum_{\theta'_{-i},s_{-i}}\sum_{m'_{-i}\in M_{-i}(m^1)}\mu^q_{i,k}[\theta_i, \theta'_{-i}, s_{-i}, m'_{-i}]}{\sum_{\theta'_i}\sum_{\theta'_{-i},s_{-i}}\sum_{m'_{-i}\in M_{-i}(m^1)}\mu^q_{i,k}[\theta'_i, \theta'_{-i}, s_{-i}, m'_{-i}]} \\ &= \frac{\sum_{\theta'_{-i},s_{-i},m'_{-i}}\mathrm{marg}_{\Theta_{-i}\times S_{-i}\times M_{-i}}\mu^q_{i,k}[\theta'_{-i}, s_{-i}, m'_{-i}]\pi^k[\theta_i|s^\theta_i, s_{-i}, \theta'_{-i}]}{\sum_{\theta'_i}\sum_{\theta'_{-i},s_{-i},m'_{-i}}\mathrm{marg}_{\Theta_{-i}\times S_{-i}\times M_{-i}}\mu^q_{i,k}[\theta'_{-i}, s_{-i}, m'_{-i}]\pi^k[\theta'_i|s^\theta_i, s_{-i}, \theta'_{-i}]} \qquad \text{(A5)} \\ &\geq \frac{\min_{s_{-i},\theta'_{-i}}\pi^k[\theta_i|s^\theta_i, s_{-i}, \theta'_{-i}]\sum_{\theta'_{-i},s_{-i},m'_{-i}}\mathrm{marg}_{\Theta_{-i}\times S_{-i}\times M_{-i}}\mu^q_{i,k}[\theta'_{-i}, s_{-i}, m'_{-i}]}{\sum_{\theta'_i}\max_{s_{-i},\theta'_{-i}}\pi^k[\theta'_i|s^\theta_i, s_{-i}, \theta'_{-i}]\sum_{\theta'_{-i},s_{-i},m'_{-i}}\mathrm{marg}_{\Theta_{-i}\times S_{-i}\times M_{-i}}\mu^q_{i,k}[\theta'_{-i}, s_{-i}, m'_{-i}]} \\ &= \frac{\min_{s_{-i},\theta'_{-i}}\pi^k[\theta_i|s^\theta_i, s_{-i}, \theta'_{-i}]}{\sum_{\theta'_i}\max_{s_{-i},\theta'_{-i}}\pi^k[\theta'_i|s^\theta_i, s_{-i}, \theta'_{-i}]} \\ &> 1 - \varepsilon,\end{aligned}$$

where the second equality follows from Bayes's rule, as $\mu^q_{i,k}$ is a totally mixed probability distribution; the third equality follows because of (A2); and the last inequality is obtained from (A4). Then, since $\mu_{i,k}[\theta', s_{-i}, m'_{-i}|m^1] = \lim_{q\to\infty}\mu^q_{i,k}[\theta', s_{-i}, m'_{-i}|m^1]$ and inequality (A5) holds for every q, we conclude that $\mathrm{marg}_{\Theta_i}\mu_{i,k}[\theta_i|m^1] > 1 - \varepsilon$ for every sufficiently large k. Since $\tilde{m}_i$ is a sequential best response to $\mu_{i,k}$, it follows that $\tilde{m}^2_i(m^1) = \theta_i$. QED

A2.2. Interstage Coordination Condition

Claim 5. For any $i \in \mathcal{I}$, k sufficiently large, and $m_i \in R_{i,2}(s^\theta_i|\Gamma(\pi^k))$, we have $m^1_i = (\hat{\theta}^i_i, \theta_{i-1})$ for some $\hat{\theta}^i_i \in \Theta$; that is, player i must report the type of player $i - 1$ truthfully at stage 1.

Proof. First, by claim 4, for any $m^1 \in M^1$, then player $i - 1 \in \mathcal{I}^*(m^1)$ will report θ_{i-1} truthfully, as long as he/she plays $\tilde{m}_{i-1} \in R_{i-1,1}(s^\theta_{i-1}|\Gamma(\pi^k))$ for any k large enough. Moreover, since claim 2 holds under complete information, player i is strictly better off by reporting his/her predecessor's true type (i.e., $\hat{\theta}^i_{i-1} = \theta_{i-1}$) than telling a lie. This strict truth-telling incentive remains the same under π^k with sufficiently large k, so long as player $i - 1$ reports type θ_{i-1} with probability close to 1 in stage 2. Since $d(\pi^k, \pi^{\mathrm{CI}}) \to 0$, which implies $\pi^k[\theta, s^\theta_{-i}|s^\theta_i] \to 1$ as $k \to \infty$, player i of type s^θ_i believes with probability close to 1 that player $i - 1$ also receives s^θ_{i-1} for any k large enough. Hence, player i reports the type of player $i - 1$ truthfully in the first stage, that is, $m^1_i = (\hat{\theta}^i_i, \theta_{i-1})$ for any $m_i \in R_{i,2}(s^\theta_i|\Gamma(\pi^k))$ and any k sufficiently large. QED

A2.3. Within-Stage Coordination Condition

Claim 6. For any $i \in \mathcal{I}$, $k \in \mathbb{N}$ sufficiently large, and $m_i \in R_{i,3}(s^\theta_i|\Gamma(\pi^k))$, we have $m^1_i = (\theta_i, \theta_{i-1})$.

Proof. Since claim 3 holds under complete information, player i finds it strictly better to report his/her own type at stage 1 rather than to tell a lie about it. This strict better reply of telling his/her true type as opposed to misreporting his/her type remains the same under π^k (with k large), so long as player $i + 1$ reports player i's type θ_i truthfully with probability close to 1 at stage 1. Therefore, it follows that $m^1_i = (\theta_i, \theta_{i-1})$ for any $m_i \in R_{i,3}(s^\theta_i|\Gamma(\pi^k))$ and any sufficiently large k. Since $m^1_i = (\theta_i, \theta_{i-1})$ for every i, it follows that $g(z(m)) = f(\theta)$ and $\tau_i(z(m)) = 0$. QED

A3. Additional Robustness Results

In this appendix, we discuss how the structure of the SR mechanism leads to additional robustness features related to the common knowledge assumption of preferences and rationality. We also discuss an alternative construction of the SR mechanism that requires preferences over lotteries that are monotone with respect to first-order stochastic dominance rather than requiring that preferences over lotteries have an expected utility representation.

A3.1. Retaliatory Preferences

Fehr, Powell, and Wilkening (2021) consider an implementation problem where players care about not only material payoffs but also psychological payoffs obtained from retaliating against perceived unkind acts. They show that for any canonical SPI mechanism that implements a pricing rule that increases with the value of the good, there exists a distribution of reciprocal preferences where truth telling is not a sequential reciprocity equilibrium at least one-fourth of the time. These results suggest that negative reciprocity has the potential to impact canonical SPI mechanisms.

Using the (modified) sequential reciprocity equilibrium of Fehr, Powell, and Wilkening (2021) as the solution concept, we find that if the SR mechanism (fully) implements the social choice function with all selfish players, it will partially implement the social choice function for any distribution of reciprocity parameters.[51] Thus, for any pricing rule that increases with the value of the good, there always exists a distribution of reciprocity parameters where the SR mechanism has a truth-telling equilibrium while the SPI mechanism does not. In this sense, it is more robust to reciprocity.[52]

A3.2. Agent Quantal Response Equilibrium

The SR mechanism is robust to alternative reasoning processes and behavioral assumptions. In particular, the SR mechanism implements the social choice function after (1) deleting strategies that violate sequential rationality and (2) deleting strictly dominated strategies for two rounds. As discussed in greater detail in the proof of theorem 1, we choose the dictator lotteries $l_i^*(\cdot)$, the incentive transfers T, and the arbitration fee T so that (1′) sequential rationality ensures that player i^* will truthfully announce his/her type in the second stage; (2′) the first-round deletion of strictly dominated strategies ensures that each player i wants to match his/her first-stage report on the type of player $i - 1$ with the second-stage report chosen by player $i - 1$; and (2″) the second-round deletion of strictly dominated strategies ensures that each player i wants to match his/her first-stage report on his/her own type with the first-stage report chosen by player $i + 1$. Consequently, our result remains valid for any solution concept that 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 for almost all standard solution concepts in extensive form games as well as some behavioral solution concepts, such as the agent quantal response equilibrium proposed by McKelvey and Palfrey (1998), provided that the noise parameter is sufficiently small.

The SR mechanism relies on strict inequalities at all stages of the game and is thus remarkably robust to level k reasoning. If play is anchored by truth telling of level 0 types and all other level k types play sequential best replies to level $k - 1$ types even at probability zero information sets, all types find it optimal to tell the truth. If we instead allow for any level 0 play but retain the sequential best reply assumption, then the SR mechanism is level 3 implementable.[53] These results are consistent with de Clippel, Saran, and Serrano (2019), who show that a slight weakening of standard strict incentive constraints is necessary for level k implementation.

[51] The solution concept in Fehr, Powell, and Wilkening (2021) uses the pricing rule being implemented as the reference point and considers only negative reciprocity. Thus, it differs slightly from the original one proposed by Dufwenberg and Kirchsteiger (2004).

[52] Fehr, Powell, and Wilkening (2021) also introduces a retaliatory seller (RS) mechanism that is more robust to reciprocity than the SPI mechanism. This mechanism was inspired by the SR mechanism in this paper and uses simultaneous reports in the first stage. An advantage of the RS mechanism is that it is deterministic and does not require lotteries. However, since truth telling is always an equilibrium of the SR mechanism but is not always an equilibrium in the RS mechanism, the SR mechanism is more robust to reciprocity than the RS mechanism. Further, the RS mechanism is not robust to private value perturbations.

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

A3.3. Expected Utility

In the main text, we constructed a version of the SR mechanism that required that preferences over lotteries have an expected utility representation. In this appendix, we demonstrate that it is always possible to construct an alternative version of the SR mechanism that relaxes this assumption. This alternative SR mechanism requires only that preferences over lotteries are monotone with respect to first-order stochastic dominance.[54] We also note that in some cases the SR mechanism can be made entirely deterministic so that the implementation requires no specification of agents' preferences over lotteries.

To illustrate how the alternative SR mechanism is constructed, we build on the two-sided holdup problem we explored in experiment 2, where the buyer and seller make a cost and value report in the first stage. Recall that our original mechanism was symmetric and used a coin flip to randomly assign the buyer or seller to arbitration in the case where both the cost and the value report did not coincide. The expected utility representation was necessary to ensure that this lottery generated the correct incentives to satisfy the truth-telling, interstage, and within-stage conditions that need to be satisfied if we want the SR mechanism to implement any social choice function in initial rationalizability (see sec. A1).

Rather than use a lottery, it is possible to use fixed priorities instead. Consider the following two modifications to the mechanism described in section IV. First, assign the buyer priority for entering the arbitration stage: whenever the value reports of the buyer and seller differ (and regardless of the cost reports), the buyer enters the arbitration stage. In this construction, the seller enters the arbitration stage only in the case where the value reports are the same but the cost reports differ. Second, double both the arbitration fee and the incentive transfer when the buyer enters the arbitration stage, but keep the arbitration fee and incentive transfer as in our experiment 2 when the seller enters the arbitration stage.

The modified SR mechanism implements the social choice function depicted in table 2 in our experiment 2. To see this, first note that the buyer will always report truthfully in the arbitration stage. Since we have doubled the incentive transfer for the seller, the seller will always prefer to report the buyer's true value regardless of the cost reports. Thus, the interstage coordination condition (see sec. A1.2) is satisfied for the value report. Similarly, since we also double the arbitration fee to the buyer, the buyer will prefer to report his own value in the report stage to avoid paying the arbitration fee. Thus, the within-stage coordination condition (see sec. A1.3) is satisfied for the value report. Next, noting that the value reports will now coincide, the seller will enter arbitration if the cost reports differ. As such, the buyer will always prefer to report the true cost, and as a result, the seller will also prefer to make a truthful cost report.

Observe that the modified mechanism uses no randomization except for the dictator lotteries. In other words, players' preferences over lotteries are irrelevant in establishing the interstage and within-stage conditions. Further, the logic

[54] Monotonicity with respect to first-order stochastic dominance means that any shift of probability weight from a less preferred to a more preferred pure alternative yields a lottery that is preferred.

above can be repeated in cases where there are more than two players, and thus priorities can be used for any finite number of players.

Although we still use dictator lotteries as part of the SR mechanism to satisfy the truth-telling condition, there are environments where the elicitation of agents' preferences can be done via a menu of deterministic allocation-transfer pairs (in the sense that [3] in lemma 1 still holds),[55] which will render the modified SR mechanism deterministic. In such cases, the modified SR mechanism works robustly in the sense that it requires no specification of the agents' preferences over lotteries. In contrast, the use of random allocations and thereby the specification of the agents' preferences over lotteries are both essential for virtual implementation (e.g., Abreu and Matsushima 1992; Arya, Glover, and Young 1995); see section 4 of Jackson (2001) for more discussion.

Appendix B

Robustness of General Mechanisms That Implement under Initial Rationalizability

In this appendix, we show that any multistage mechanism that can implement a social choice function in initial rationalizable messages under complete information can also robustly implement the social choice function under a private value perturbation. We document this result as proposition 2 at the end of this appendix.

B1. Preliminary

We consider multistage mechanisms with observable actions. A multistage mechanism with observable actions is defined as a tuple $\Gamma = (\mathcal{I}, \mathcal{H}, \mathcal{Z}, (A_i)_{i\in\mathcal{I}}, g, (\tau_i)_{i\in\mathcal{I}})$, where $\mathcal{I}$ is the set of players and, for each player i, A_i is the finite set of player i's possible actions, $\mathcal{H}$ is the finite set of partial histories, and $\mathcal{Z}$ is the set of terminal histories. For each $h \in \mathcal{H}$ and $i \in \mathcal{I}$, let $A_i(h)$ denote the finite set of actions available to player i at history h, and let $A(h) = \times_{i\in\mathcal{I}} A_i(h)$ and $A_{-i}(h) = \times_{j\neq i} A_j(h)$. Without loss of generality, $A_i(h)$ is assumed to be nonempty for each h. In particular, player i is inactive if $|A_i(h)| = 1$ and he is active otherwise. In the following, we adapt the notation of section II.A to the more general class of multistage mechanisms with observable actions.

A message/pure strategy, which we denote by m_i, specifies an action in $A_i(h)$ for each history h. Let M_i denote the set of messages of player i. Each message profile m induces a unique terminal history $z(m) \in \mathcal{Z}$. Let $M_i(h)$ be the set of strategies m_i that allow h to be reached (i.e., there exists m_{-i} such that h is on the path to $z(m_i, m_{-i})$) and, likewise, $M_{-i}(h)$ be the set of strategy profiles of player i's opponents that allow h to be reached. Let $\mathcal{H}(m_i) = \{h \in \mathcal{H} : m_i \in M_i(h)\}$ denote the set of partial histories that are not precluded by m_i. As in section II.A, g is the outcome function that maps the set of terminal histories into outcomes in

[55] In the Hart-Moore example in Aghion et al. (2012), e.g., we may set the menu for the buyer as choosing between no trade and trade at a price equal to the average of the high value and the lower value of the buyer.

$\Delta(A)$, and $(\tau_i)_{i\in\mathcal{I}}$ is the transfer rule where each τ_i maps the set of terminal histories into a transfer to player *i*.

B2. Implementation under Complete Information

Let $\Gamma(\theta)$ denote the multistage game with observable actions induced by Γ at state θ. Player *i*'s payoff from a message profile *m* is given by

$$v_i(m_i, m_{-i}, \theta_i) \equiv u_i(g(z(m)), \theta_i) + \tau_i(z(m)). \tag{B1}$$

We define a CPS according to definition 1 in section II.B, where we also set $\Omega = M_{-i}$ and $\mathcal{B} = \{M_{-i}(h)\}_{h\in\mathcal{H}}$. That is, a CPS μ_i specifies, for each history *h* (identified with $M_{-i}(h)$), a probability distribution over M_{-i} such that Bayes's rule (i.e., condition 2 of definition 1) applies whenever possible. To simplify the notation, we write $\mu_i[\cdot|h]$ as opposed to $\mu_i[\cdot|M_{-i}(h)]$ for $h \neq \varnothing$, and we write $\mu_i[\cdot]$ instead of $\mu_i[\cdot|\varnothing]$.

By reporting message m_i and holding CPS μ_i, player *i* receives the expected payoff conditional on history $h \in \mathcal{H}$:

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

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

$$V_i(m_i, \theta_i, \mu_i|h) \geq V_i(m_i', \theta_i, \mu_i|h), \forall\ m_i' \in M_i(h). \tag{B2}$$

We define initial rationalizability under complete information:

Definition 8 (Initial rationalizability). Let $\Gamma(\theta)$ be the multistage game with observable actions induced by mechanism Γ with respect to state θ. 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 that are sequential best replies to some CPS μ_i such that $\mu_i[R_{-i,k-1}^{\Gamma(\theta)}] = 1$. Finally, the set of *initially rationalizable* messages for player *i* is $R_i^{\Gamma(\theta)} = \cap_{k=1}^{\infty} R_{i,k}^{\Gamma(\theta)}$.

Note that this definition respects the players' common knowledge that the true state is θ. We now define the notion of implementability for general mechanisms:

Definition 9. A social choice function *f* is *implementable in initial rationalizable messages* if there exists a multistage mechanism with observable actions Γ such that, for any state $\theta \in \Theta$, we have $g(z(m)) = f(\theta)$ and $\tau_i(z(m)) = 0$ for every player $i \in \mathcal{I}$ and for every message profile $m \in R^{\Gamma(\theta)}$.

B3. Implementation under Incomplete Information

We define a prior $\pi \in \Delta(\Theta \times S)$ as in section II.D. Let $\Gamma(\pi)$ be a multistage game with observable actions induced by Γ with respect to prior π. A conjecture for agent *i* is a conditional probability system over $\Theta \times S_{-i} \times M_{-i}$. Using the formal notation introduced for CPS's in definition 1, we set $\Omega = \Theta \times S_{-i} \times M_{-i}$ and let $\mathcal{B} = \{\Theta \times S_{-i} \times M_{-i}(h)\}_{h\in\mathcal{H}}$. Denote the set of CPSs over $\Theta \times S_{-i} \times M_{-i}$ by $\Delta^{\mathcal{H}}(\Theta \times S_{-i} \times M_{-i})$. Again, to simplify the notation, we write $\mu_i[\cdot|h]$ for $\mu_i[\cdot|\Theta \times S_{-i} \times$

$M_{-i}(h)]$ for $h \neq \varnothing$, and we write $\mu_i[\cdot]$ for $\mu_i[\cdot|\varnothing]$. As a CPS represents a player's beliefs, it should also be based on the player's signal. The connection is formalized via the following definition:

Definition 10. A CPS μ_i is said to be consistent with s_i under prior π if there exists a sequence of totally mixed probability distributions $\{\mu_i^q\}_{q=1}^{\infty}$ over $\Delta^{\mathcal{H}}(\Theta \times S_{-i} \times M_{-i})$ such that (i) for every $(\theta, s_{-i}, m_{-i}) \in \Theta \times S_{-i} \times M_{-i}$ and $h \in \mathcal{H}$, we have

$$\mu_i[\theta, s_{-i}, m_{-i}|h] = \lim_{q \to \infty} \mu_i^q[\theta, s_{-i}, m_{-i}|h];$$

and (ii) for every $q \geq 1$, there exists $\sigma_{-i}^q : \Theta_{-i} \times S_{-i} \to \Delta(M_{-i})$ such that

$$\mu_i^q[\theta, s_{-i}, m_{-i}] = \sigma_{-i}^q[m_{-i}|\theta_{-i}, s_{-i}]\pi[\theta, s_{-i}|s_i]. \tag{B3}$$

Definition 11. A strategy m_i is said to be a sequential best response to CPS μ_i if for every h and every m_i',

$$\sum_{\theta_i, m_{-i}} v_i(m_i, m_{-i}, \theta_i) marg_{\Theta_i \times M_{-i}} \mu_i(\theta_i, m_{-i}|h) \geq \sum_{\theta_i, m_{-i}} v_i(m_i', m_{-i}, \theta_i) marg_{\Theta_i \times M_{-i}} \mu_i(\theta_i, m_{-i}|h).$$

Denote by $sr_i : \Delta^{\mathcal{H}}(\Theta \times S_{-i} \times M_{-i}) \rightrightarrows M_i$ player i's sequential best response correspondence. Equipped with the previous two definitions, we are now ready to define the solution concept of initial rationalizability for $\Gamma(\pi)$.

Definition 12. Let $\Gamma(\pi)$ be the multistage game with observable actions induced by Γ with respect to prior π. The set of initial rationalizable messages of player i with signal s_i is defined as $R_i(s_i|\Gamma(\pi)) = \cap_{k=1}^{\infty} R_{i,k}(s_i|\Gamma(\pi))$, where $R_{i,0}(s_i|\Gamma(\pi)) = M_i$ and, inductively, for every integer $k \geq 1$,

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

As the mechanism is finite, we have $R(s^{\theta}|\Gamma(\pi)) \neq \varnothing$. The following is the definition of robust implementation that we adopt.

Definition 13. A social choice function f is *robustly implementable in initial rationalizable strategies* if there exists a multistage mechanism with observable actions Γ such that for any state $\theta \in \Theta$, any signal profile $s^{\theta} \in S$, any private value perturbation $\{\pi^k\}_{k=1}^{\infty}$ to π^{CI}, and any sequence of message profiles $\{m^k\}_{k=1}^{\infty}$ with $m^k \in R(s^{\theta}|\Gamma(\pi^k))$ for each k, we have $g(z(m^k)) = f(\theta)$ and $\tau_i(z(m^k)) = 0$ for every player i and for all sufficiently large k.

B4. Upper Hemicontinuity of the Initial Rationalizability Correspondence

We first establish the following preliminary result.

Claim 7. Let $\{\pi^k\}_{k=1}^{\infty}$ be a private value perturbation to π^{CI}, and for each $k \in \mathbb{N}$, let $\mu_{i,k}$ be a CPS that is consistent with signal s_i^{θ} under π^k. Then, for any $\varepsilon > 0$ and $h \in \mathcal{H}$, $\text{marg}_{\Theta_i} \mu_{i,k}[\theta_i|h] > 1 - \varepsilon$ for any sufficiently large k.

Proof. Fix $k \geq 1$, player $i \in \mathcal{I}$, and a CPS $\mu_{i,k}$ that is consistent with signal s_i^{θ} under π^k. We show that for any $\varepsilon > 0$, we have

$$\text{marg}_{\Theta_i}\mu_{i,k}[\theta_i|h] > 1 - \varepsilon, \forall\ h \in \mathcal{H}$$

for all sufficiently large k. Since $\mu_{i,k}$ is consistent with signal s_i^θ, we first have that, for any $(\theta'_i, \theta'_{-i}, s_{-i}) \in \Theta \times S_{-i}$ such that $\pi^k[\theta'_i, \theta'_{-i}, s_{-i}|s_i^\theta] > 0$,

$$\mu^q_{i,k}[\theta'_i, \theta'_{-i}, s_{-i}, m'_{-i}] = \text{marg}_{\Theta_{-i}\times S_{-i}\times M_{-i}}\mu^q_{i,k}[m'_{-i}, \theta'_{-i}, s_{-i}]\pi^k[\theta'_i|s_i^\theta, s_{-i}, \theta'_{-i}]. \tag{B4}$$

In addition, since $\{\pi^k\}_{k=1}^\infty$ is a private value perturbation to π^{CI}, for all sufficiently large k, we have

$$\frac{\min_{(\theta'_{-i}, s_{-i})\in\Omega^k_{-i}}\pi^k[\theta_i|s_i^\theta, s_{-i}, \theta'_{-i}]}{\sum_{\theta'_i}\max_{(\theta'_{-i}, s_{-i})\in\Omega^k_{-i}}\pi^k[\theta'_i|s_i^\theta, s_{-i}, \theta'_{-i}]} > 1 - \varepsilon, \tag{B5}$$

where Ω^k_{-i} is the support of $\text{marg}_{\Theta_{-i}\times S_{-i}}\pi^k[\cdot|s_i^\theta]$. The reader is referred to the proof of claim 4 in appendix A for a detailed argument for (B4) and (B5).

Now consider $\mu^q_{i,k}$ such that (B4) and (B5) hold and an arbitrary history $h \in \mathcal{H}$. To see $\text{marg}_{\Theta_i}\mu^q_{i,k}[\theta_i|h] > 1 - \varepsilon$, we can similarly compute player i's conditional belief $\text{marg}_{\Theta_i}\mu^q_{i,k}[\theta_i|h]$ following the proof of claim 4. QED

The following upper hemicontinuity result is the main result of this section.

Proposition 1. For any private value perturbation $\{\pi^k\}_{k=1}^\infty$ to π^{CI} and $s_i^\theta \in S_i$, if $m_{i,k} \in R_i(s_i^\theta|\Gamma(\pi^k))$ for all k and $m_{i,k} \to m_i$, then we have $m_{i,k} \in R_i^{\Gamma(\theta)}$ for all sufficiently large k.

Proof. For each nonnegative integer n, we define $H(n)$ as the statement that if $m_{i,k} \in R_{i,n}(s_i^\theta|\Gamma(\pi^k))$ for all k and $m_{i,k} \to m_i$, then there exists $K \in \mathbb{N}$ such that $m_{i,k} \in R_{i,n}^{\Gamma(\theta)}$ for all $k \geq K$. We prove this by induction on n. $H(0)$ is trivially true. Now we fix $n \geq 1$ and suppose that $H(n - 1)$ is true. Fix $k \in \mathbb{N}$. Since $m_{i,k} \in R_{i,n}(s_i^\theta|\Gamma(\pi^k))$, there exists $\mu_{i,k} \in \Delta^{\mathcal{H}}(\Theta \times S_{-i} \times M_{-i})$ such that

(1) $\mu_{i,k}[\theta', s_{-i}, m_{-i}] > 0 \Rightarrow m_{-i} \in R_{-i,n-1}(s_{-i}|\Gamma(\pi^k))$;

(2) $m_{i,k} \in sr_i(\mu_{i,k})$; and

(3) $\mu_{i,k}$ is consistent with $s_{i,k}$ under π^k.

Since Θ, S_{-i}, and M_{-i} all are finite sets, $\Delta^{\mathcal{H}}(\Theta \times S_{-i} \times M_{-i})$ is compact. Therefore, there exists a convergent subsequence $\{\mu_{i,k}\}_{k=1}^\infty$ in $\Delta^{\mathcal{H}}(\Theta \times S_{-i} \times M_{-i})$, and we denote its limit by $\bar{\mu}_i \in \Delta^{\mathcal{H}}(\Theta \times S_{-i} \times M_{-i})$. Define $\mu_i(\cdot|h) \equiv \text{marg}_{M_{-i}}\bar{\mu}_i(\cdot|h)$ for every $h \in \mathcal{H}$. Observe that $\mu_i \in \Delta^{\mathcal{H}}(M_{-i})$ is a CPS in the complete information game $\Gamma(\theta)$. We now claim that there exists $K \in \mathbb{N}$ such that $m_{i,k} \in R_{i,n}^{\Gamma(\theta)}$ for any $k \geq K$. To prove this claim, we shall show that there exists $K \in \mathbb{N}$ such that for any $k \geq K$, (1′) $\mu_i[m_{-i}] > 0 \Rightarrow m_{-i} \in R_{-i,n-1}^{\Gamma(\theta)}$ and (2′) $m_{i,k}$ is a sequential best response to μ_i in $\Gamma(\theta)$.

To see why (1′) holds, suppose that $\mu_i[m_{-i}] > 0$ for some m_{-i}. Since $\bar{\mu}_i$ is a limit point of $\{\mu_{i,k}\}_{k=1}^\infty$, each $\mu_{i,k}$ is consistent with $s_{i,k}$ under π^k, and $d(\pi^k, \pi^{\text{CI}}) \to 0$, we have $\text{marg}_{\Theta\times S_{-i}}\bar{\mu}_i(\theta, s_{-i}^\theta) = 1$. Then, since $\bar{\mu}_i$ is a limit point of $\{\mu_{i,k}\}$ and each $\mu_{i,k}$ satisfies (1), $\mu_i[m_{-i}] = \text{marg}_{M_{-i}}\bar{\mu}_i[m_{-i}] > 0$ implies that there exists $K_1 \in \mathbb{N}$ such that $m_{-i} \in R_{-i,n}(s_{-i}^\theta|\Gamma(\pi^k))$ for all $k \geq K_1$. Hence, by the induction hypothesis that $H(n - 1)$ is true, we have $m_{-i} \in R_{-i,n-1}^{\Gamma(\theta)}$.

To establish (2′), observe that by theorem of the maximum, $sr_i(\cdot)$ is upper hemicontinuous. Since $m_{i,k} \in sr_i(\mu_{i,k})$ for every k and $m_{i,k} \to m_i$, it follows from the upper hemicontinuity of $sr_i(\cdot)$ that there exists $K_2 \in \mathbb{N}$ such that for every $h \in \mathcal{H}$, every $k \geq K_2$, and every m_i',

$$m_{i,k} = m_i,$$

$$\sum_{\theta_i', m_{-i}} v_i(m_{i,k}, m_{-i}, \theta_i')\text{marg}_{\Theta_i \times M_{-i}}\bar{\mu}_i(\theta_i', m_{-i}|h) \geq \sum_{\theta_i', m_{-i}} v_i(m_i', m_{-i}, \theta_i')\text{marg}_{\Theta_i \times M_{-i}}\bar{\mu}_i(\theta_i', m_{-i}|h).$$

It follows from claim 7 that $\text{marg}_{\Theta_i}\bar{\mu}_i[\theta_i|h] = 1$. Furthermore, since $\mu_i(\cdot|h) = \text{marg}_{M_{-i}}\bar{\mu}_i(\cdot|h)$ for every $h \in \mathcal{H}$, we have that for every $k \geq K_2$ and every m_i',

$$\sum_{\theta_i', m_{-i}} v_i(m_{i,k}, m_{-i}, \theta_i)\mu_i(m_{-i}|h) \geq \sum_{\theta_i', m_{-i}} v_i(m_i', m_{-i}, \theta_i)\mu_i(m_{-i}|h).$$

That is, $m_{i,k}$ is a sequential best response to μ_i in $\Gamma(\theta)$ for any $k \geq K_2$. It then follows from (1′) and (2′) that $m_{i,k} \in R_{i,n}^{\Gamma(\theta)}$ for any $k \geq K_2$. Setting $K = \max\{K_1, K_2\}$, we conclude that $H(n)$ is true. This completes the proof. QED

The result below follows immediately from proposition 1.

Proposition 2. If the social choice function f is implementable in initial rationalizable strategies by a multistage finite mechanism with observable actions Γ, then f is also robustly implementable in initial rationalizable strategies by the same mechanism Γ.

Proof. Suppose that f is implementable in initial rationalizable strategies by a multistage finite mechanism with observable actions Γ. Then, f is robustly implementable in initial rationalizable strategies by Γ if for any state $\theta \in \Theta$, any signal profile $s^\theta \in S$, and any private value perturbation $\{\pi^k\}_{k=1}^\infty$ to π^{CI}, we have $R(s^\theta|\Gamma(\pi^k)) \subset R^{\Gamma(\theta)}$ for all sufficiently large k. Suppose to the contrary that for some state $\theta \in \Theta$, some signal profile $s^\theta \in S$, and some private value perturbation $\{\pi^k\}_{k=1}^\infty$ to π^{CI}, we have some subsequence $\{\pi^{k_q}\}_{q=1}^\infty$ of $\{\pi^k\}_{k=1}^\infty$ such that $m^q \in R(s^\theta|\Gamma(\pi^{k_q}))$ and $m^q \notin R^{\Gamma(\theta)}$ for all q. By finiteness of M, we may take a further subsequence so that $m^q = m$ for all q. Since $m^q = m \in R(s^\theta|\Gamma(\pi^{k_q}))$ for every q, by proposition 1 we must have $m \in R^{\Gamma(\theta)}$. This is a contradiction. QED

Appendix C

Additional Analyses and Treatments

C1. Experiment 1: The SPI Mechanism

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

Result C1. In the no-noise treatment with the SPI mechanism, buyers misreport the value of their good with the high signal in 22.5% of cases. Sellers challenge buyers who misreport a high signal in 78% of cases, and the buyer rejects a legitimate challenge in 37.7% of cases.

Figure C1 displays the patterns of play we observed in the first 10 periods of the experiment. The left-hand column examines play in the low-signal scenario, and the right-hand column examines play in the high-signal scenario. Panel A summarizes the buyers' announcement decisions, panel B summarizes the sellers' challenge decisions for different announcements, and panel C summarizes the buyers' decisions to accept or reject counteroffers.

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

The left-hand side of panel B shows that sellers mistakenly challenge a low report with a low signal in 11.0% of cases. This rate of false challenges is not significantly different from the proportion of sellers who misreport in the low-signal scenario in the SR mechanism in a simple regression that regresses misreporting behavior on the SPI treatment ($p = .47$).[56] Data for this test include all observations from the low-signal scenario of the SR mechanism but use only the observations in the SPI mechanism where the low-signal is observed and the buyer has reported a low value because the seller's challenge behavior is not observed in the other cases.

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

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

Result C2. The introduction of noise leads to a significant increase of misreports by buyers with the high signal and a significant decrease in the proportion of challenges made by sellers who have high signals and observe a low report.

Figure C2 shows the path of play in the noise treatment with the SPI mechanism and is directly comparable with figure C1. As seen in panel A, buyers lie in 40% of observations in the high-signal scenario but in only 6.3% of observations in the low-signal scenario. The misreport rate in the high-signal scenario is significantly higher in the noise treatment than the no-noise treatment in a simple regression that regresses buyer misreports in the high-signal scenario on the noise treatment dummy ($p < .01$).

Panel B shows that sellers challenge in only 62.4% of cases when they have the high signal and the buyer has made a low announcement. This challenge rate is significantly lower than in the no-noise treatment in a simple regression that regresses seller challenges on the noise treatment dummy using data from observations where the seller has the high signal and receives a low report ($p < .034$).[57] Finally, panel C shows that buyers accept a counteroffer in 78% of cases when they misreport their value, have the high signal, and are challenged. This is slightly higher than in the no-noise baseline treatment, but the difference is not significant ($p = .059$).

In both the no-noise treatment and the noise treatment, buyers have a higher expected value for telling the truth in the high-signal scenario (30.9 in the no-noise

[56] This regression was not in the preanalysis plan and has been added on the basis of the relatively high misreport rate of sellers observed in the SR mechanism.

[57] This regression was not in the preanalysis plan, but the result is consistent with theory.

treatment and 31.5 in the noise treatment) than they are expected to receive by lying and accepting all challenges (9.3 in the no-noise treatment and 19.4 in the noise treatment). Thus, we might expect to see truth-telling rates increase over time. Figure C3*A* tracks the proportion of truthful announcements in the high-signal scenario over time. These data are overlaid with the predictions and 95% confidence intervals from a simple linear random effects regression that regresses the reporting decision on the period. While there appears to be a small decrease in misreports over time, the time series is not significant in either random effects regression at the .05 level (no-noise treatment: $p = .059$; noise treatment: $p =.121$).

Finally, panel C of figure C3 shows the distribution of buyer misreports in the no-noise and noise treatments of the SPI mechanism. While truth telling is the modal action in the no-noise treatment, behavior here is heterogeneous, with a small number of buyers misreporting in every period. When noise is introduced, the distribution of lies shifts to the right and the distribution becomes bimodal, with some buyers lying in every period.[58]

[58] See also fig. C5, which shows buyer reports in both treatments simultaneously.

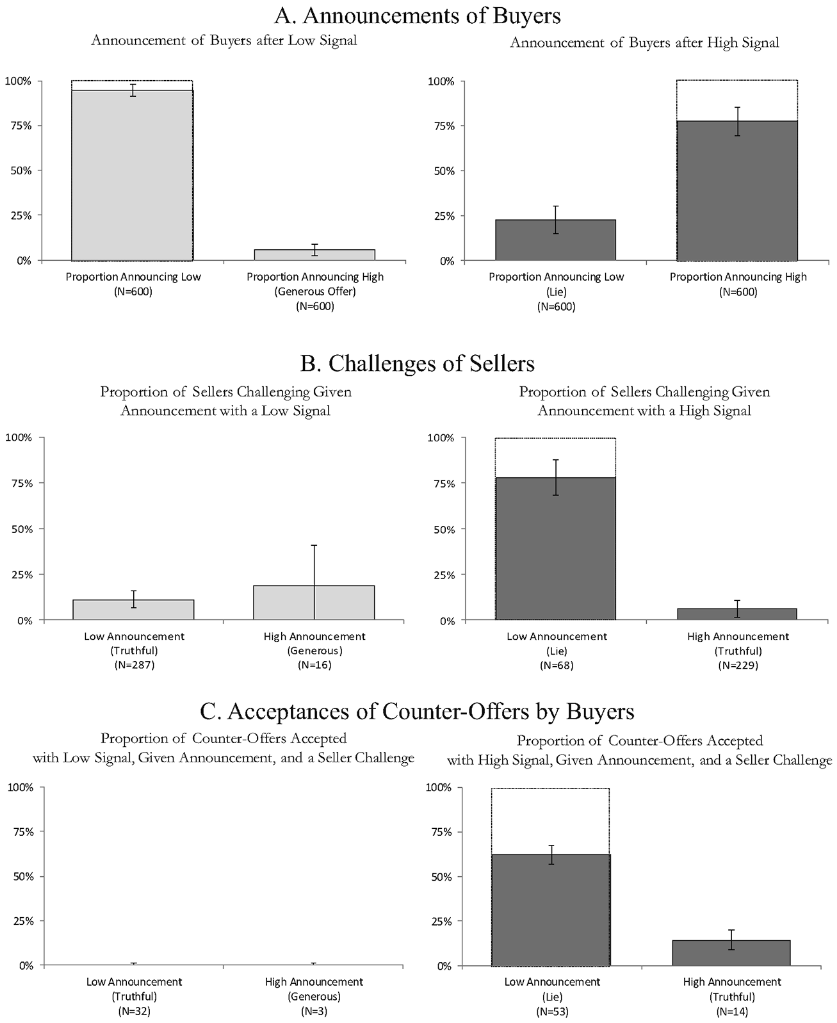


Fig. C1.—Path of play in no-noise treatment with SPI mechanism.

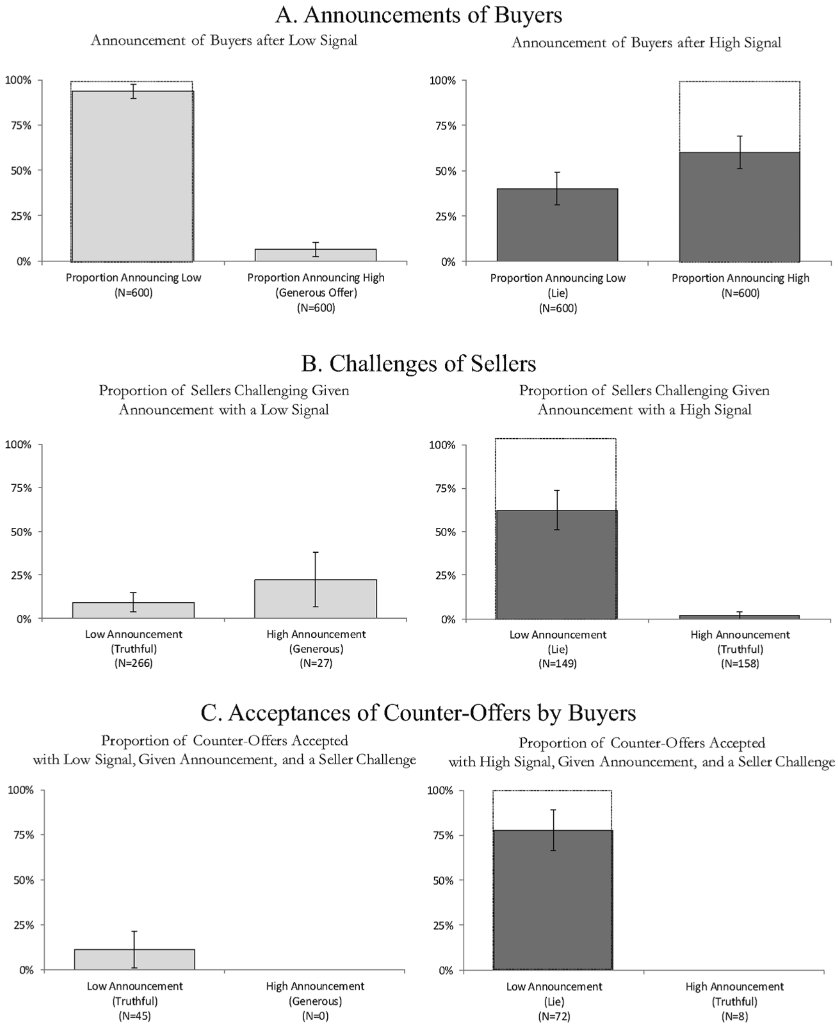
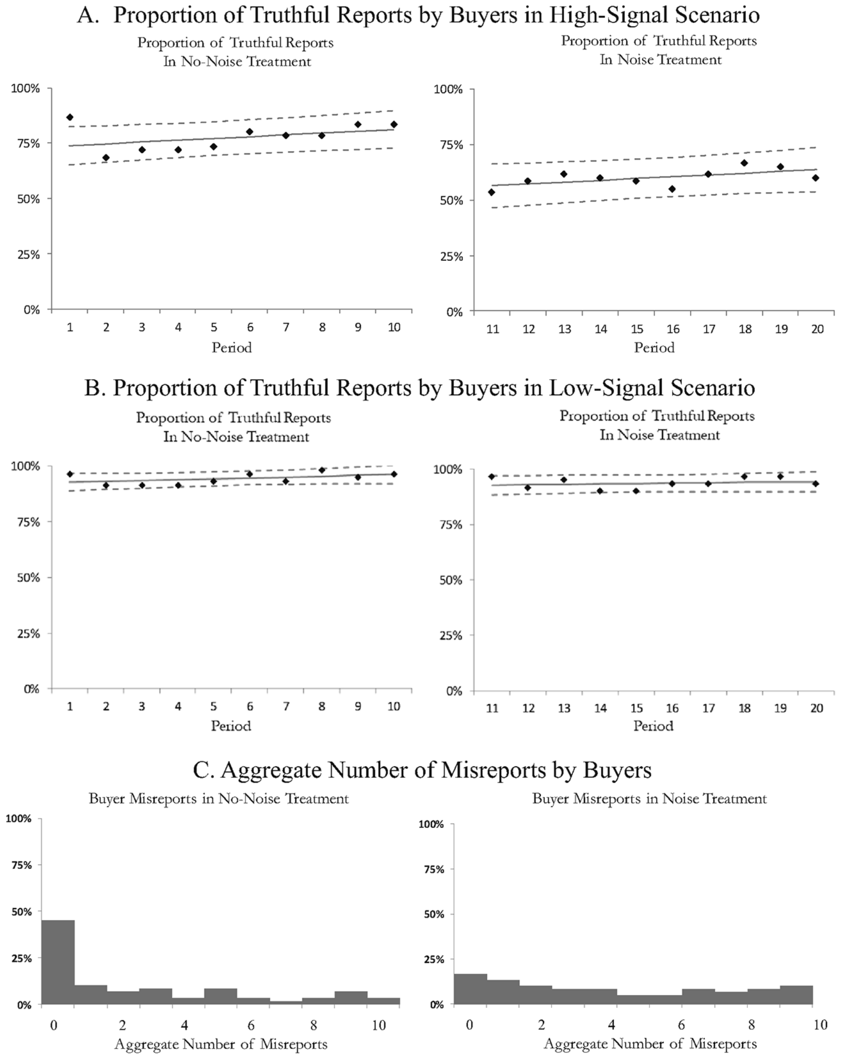


FIG. C2.—Path of play in noise treatment with SPI mechanism.



A. Proportion of Truthful Reports by Buyers in High-Signal Scenario

B. Proportion of Truthful Reports by Buyers in Low-Signal Scenario

C. Aggregate Number of Misreports by Buyers

Fig. C3.—Evolution and distribution of buyer misreports with SPI mechanism.

C2. Experiment 1: Additional Figures Comparing Misreports in the SR and SPI Mechanisms

This appendix provides additional analysis that compares the behavior of buyers in each of the four treatments. Table C1 reports coefficients from ordinary least squares regressions, with buyer lies on the left-hand side and treatments on the right-hand side. Column 1 includes data from only the noise treatments, while column 2 includes all four treatments. Column 3 is a random effects regression and clusters data at the session level. Note that while asterisks represent two-sided significance levels, the predicted difference between the SR and the SPI mechanisms in the noise treatments is one-sided in the preanalysis plan.

Table C2 provides nonparametric comparisons of the treatments with data aggregated at either the individual buyer level or the session level, while figure C4 shows buyer misreports over time in all four treatments. The prediction and confidence intervals that have been overlaid on the data in figure C4 are from simple random effects regressions with a treatment-specific linear time trend.

Figure C5 shows the aggregate number of misreports made by buyers in both the no-noise and the noise treatments of both mechanisms. As seen in panel A, 34 out of 60 buyers reported truthfully in all 20 periods of sessions using the SR mechanism. Lies are more prevalent in the SPI mechanism, and 47% of buyers lie more frequently in the noise treatment than the no-noise treatment, while only 25% of buyers lie less.

Finally, figures C6 and C7 show the proportion of truthful reports for buyers and sellers in the SR mechanism over time. The prediction and confidence intervals that have been overlaid on the data are from simple random effects regressions with a treatment-specific linear time trend. The only time trend in these graphs that is significant is the time series for sellers in the no-noise treatment in the low-signal scenario. Recall that by the construction of the mechanism, buyers are always punished if they enter the arbitration stage of the mechanism, while sellers may be rewarded or punished on the basis of the actions of the buyer. If a seller is uncertain about the incentives generated in the mechanism, they may experiment with lies until they are able to observe how the buyers behave and experience the loss associated with their lie. Such experimentation is apparent in an ex post analysis of the data: a seller who lies in the low-signal scenario in period t lies in the next period 76.5% of the time if the high-signal scenario occurred and the repercussions of the lie are not observable. By contrast, if a seller lies in the low-signal scenario and the low-signal scenario occurs, sellers lie only 26.1% of the time in the next period. Thus, the time trend appears to be based on sellers who initially lie in the low scenario and switch to truthful reporting after experiencing losses when the low-signal scenario occurs.

TABLE C1
Buyer Misreports in SR and SPI Mechanisms

	(1)	(2)	(3)
SPI treatment	.308***	.178***	.178***
	(.055)	(.048)	(.044)
Noise treatment		.048	.048
		(.030)	(.043)
SPI × noise treatment		.130**	.130*
		(.058)	(.079)
Constant	.125***	.077***	.077***
	(.032)	(.024)	(.017)
R^2	.118	.392	.110
Observations	1,200	2,400	2,400

Note.—Dependent variable is 1 if the buyer lies by announcing low with a high signal and zero otherwise. Regression 1 is a linear probability model that includes data from only the noise treatment. Regressions 2 and 3 are linear probability models that include data from all four treatments. Regressions 1 and 2 are clustered at the buyer level. Regression 3 uses individual-level random effects and is clustered at the session level.

* Two-tailed significance at the 10% level.

** Two-tailed significance at the 5% level.

*** Two-tailed significance at the 1% level.

TABLE C2
Nonparametric Tests Comparing Proportion of Misreports Made by Buyers

Nonparametric Tests of Buyer Lies and Unit of Observation	p
No-noise treatment vs. noise treatment in SPI mechanism (Wilcoxon signed-rank test):	
Individual level	.0030
Session level	.0464
No-noise treatment vs. noise treatment in SR mechanism (Wilcoxon signed-rank test):	
Individual level	.1421
Session level	.4630
SR vs. SPI in no-noise treatments (Mann-Whitney-Wilcoxon test):	
Individual level	.0002
Session level	.0039
SR vs. SPI in noise treatments (Mann-Whitney-Wilcoxon test):	
Individual level	.0000
Session level	.0039

Note.—The table shows the nonparametric comparison of treatments, with p-values for the two-sided version of each test.

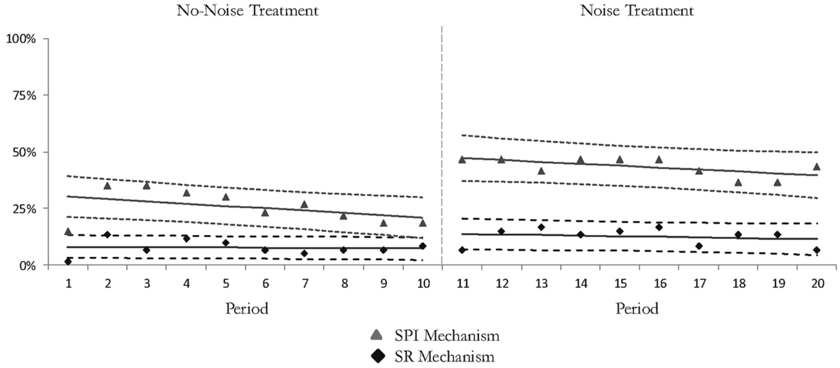
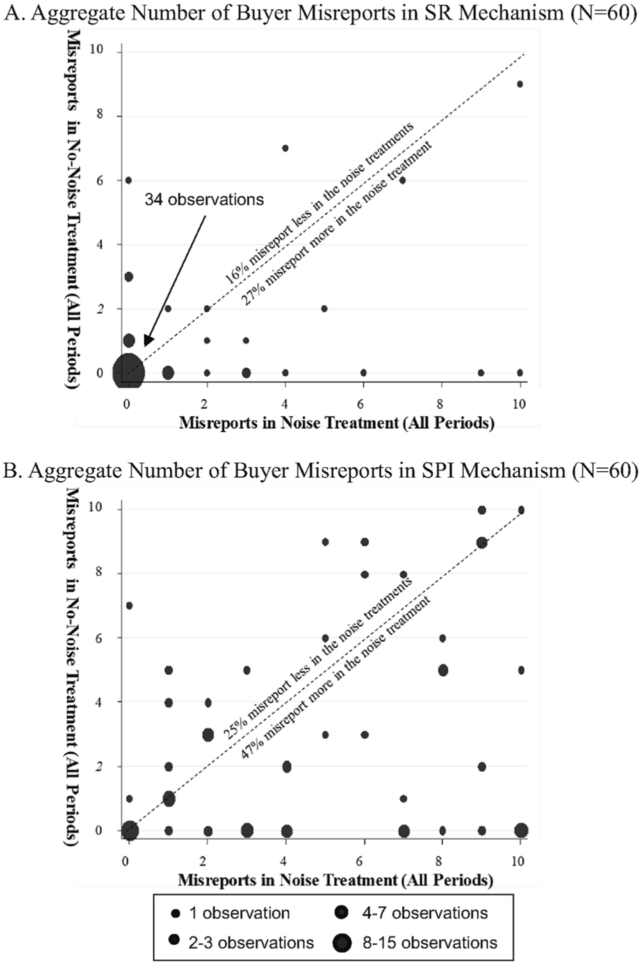


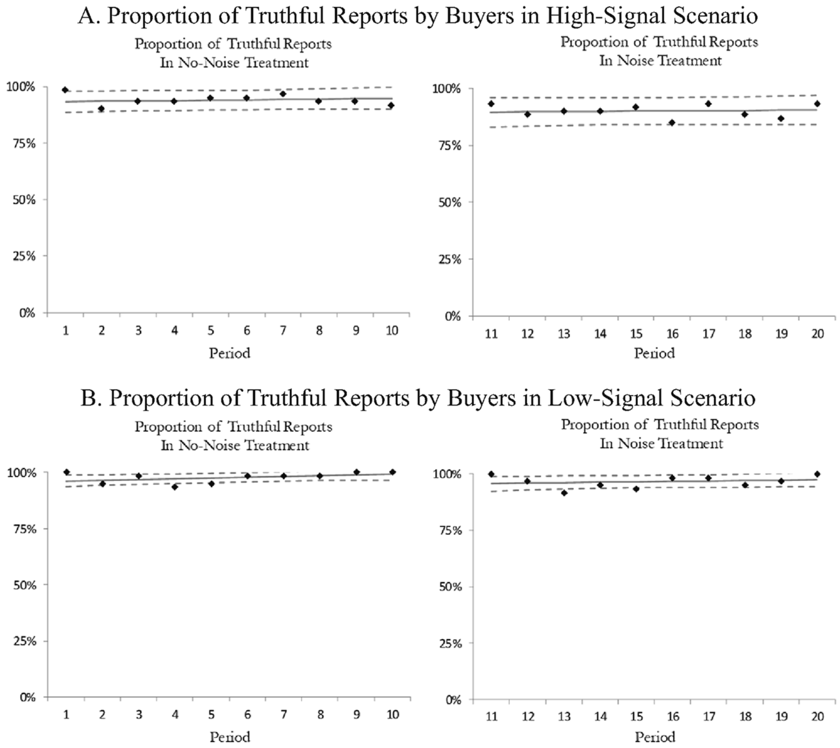
FIG. C4.—Buyer misreports over time. The figure shows 95% confidence intervals constructed from a random effects regression with a treatment-specific linear time trend.



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

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

Fig. C5.—Aggregate number of misreports made by buyers in both no-noise and noise treatments of SR and SPI mechanisms.



A. Proportion of Truthful Reports by Buyers in High-Signal Scenario

B. Proportion of Truthful Reports by Buyers in Low-Signal Scenario

FIG. C6.—Evolution of buyer misreports in no-noise and noise treatments of SR mechanism.

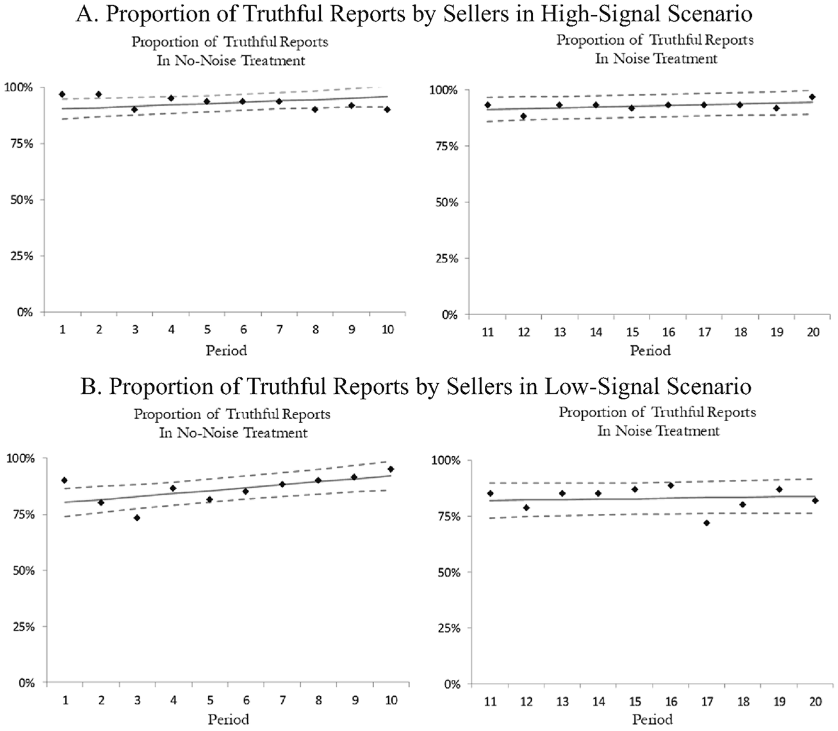


FIG. C7.—Evolution of seller misreports in no-noise and noise treatments of SR mechanism.

C3. *Experiment 1: Preanalysis Plan*

Both the experimental design and the analysis plan were preregistered at Open Science prior to the initial experiment (https://osf.io/p6ukx).

We preregistered the design, experimental hypotheses, and analysis plan. All statistics and figures in the preanalysis plan have been included in the main text or the appendix. On the basis of the initial analysis, we also included the following in the appendix: (1) a short analysis of seller misreports in the SPI mechanism, (2) session-level clustered analysis of the main treatment effects, and (3) time series graphs of all treatments. The comparison of efficiency in the main treatment is also not part of our preanalysis plan and was introduced as part of the referee process.

C4. *Experiment 2: Additional Figures from the SR Treatment*

In the main text, we presented the results of phase 1 of the SR treatment in the aggregate over all 10 periods. In this section, we show that there are only very small changes in investment and reporting decisions over time. Panels A and B of figure C8 show 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% in period 1 and increases to an average of 92.6% in periods 6–10. The proportion of sellers who invest optimally also starts above 80% and increases to an average of 88.1% 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% of the time in all periods.

Finally, panel C shows the aggregate number of lies that different buyers and sellers take over the first 10 periods. The dark gray bars represent the two buyers and three sellers who went bankrupt in the first 10 periods and whose lie frequencies are truncated.[59] As can be seen, 81.3% of buyers and 77.5% 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% 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 counterparty match their misreport has a 69.6% 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% of the time, the switching data suggest that a large proportion of lies are due to the poor learning dynamics that are generated by nontruthful secondary reports.[60]

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

[60] 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 ($p < .01$). The difference in learning dynamics is also apparent at the aggregate level.

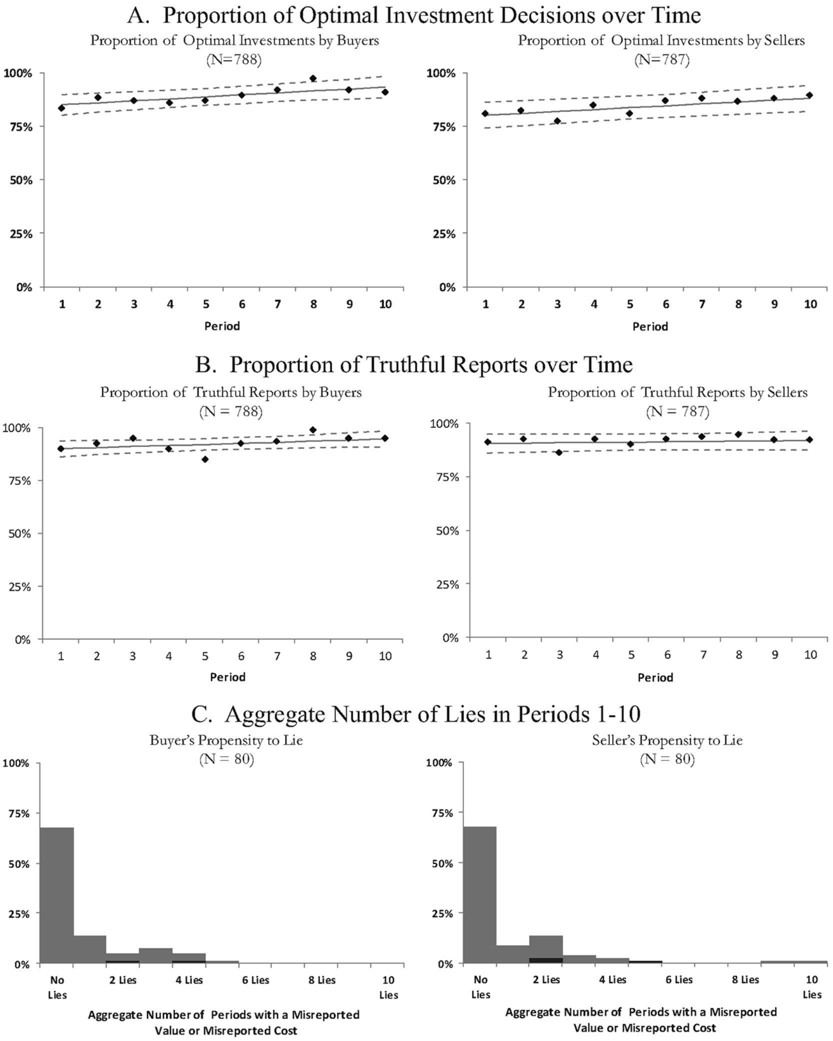


FIG. C8.—Evolution of play in first 10 periods of SR mechanism.

C5. Experiment 2: Behavior in the SPI Mechanism

In this section, we describe the behavior observed in the SPI mechanism in experiment 2. We begin with a discussion of behavior in the periods played against the computer.

RESULT C3. In the SPI mechanism, 35% 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 C9 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 are negative in the SPI treatment and significantly below the earnings of the other two treatments in a simple regression where average earnings are regressed against the treatment dummies (SPI vs. SR: $p < .01$; SPI vs. KTH: $p < .01$; SR vs. KTH: $p = .03$). Looking across individuals, we find that 35% of subjects lose money against the computer in the SPI treatment and only 37.5% achieve the theoretical first best. This is in sharp contrast with (1) the SR treatment, where 11.8% lose money and 58.1% achieve the first best, and (2) the KTH treatment, where no subject loses money and 46.3% 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 counteroffer stage that maximize the computer's profit. The large proportion of subjects who lose money against the computer suggests 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 57.1% of the time, while subjects who earn money against the computer lose money in the first period 11.5% of the time.

We now describe aggregate behavior of subjects in the SPI mechanisms in phase 1. We define an *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 C4. In periods 1–10, the SPI mechanism induces efficient investment in 79.7% of cases. Buyers make an advantageous lie in 5.6% of cases and make a false challenge in 6.6% of cases. Sellers make an advantageous lie in 5.2% of cases and false challenges in 2.6% 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 with losses in the periods against the computer lead to 20% of subjects going bankrupt.

Figure C10 displays the pattern of behavior we observed in the first 10 periods of the SPI treatment. 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 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 a 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% of buyers and 82.9% of sellers exert an efficient level of investment. The proportion of buyers making an optimal investment is increasing over time, with 55.0% of buyers putting in optimal investment in the first period and 90.0% of buyers putting in optimal investment in period 10.

Likewise, the proportion of sellers making an optimal investment is increasing over time, with 72.5% of sellers making an optimal investment in the first period and 88.6% of sellers making an optimal investment in period 10.

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

Sellers make truthful announcements in 96.4% of cases and advantageous lies in 3.6% of cases. They make a false challenge in only 2.6% of cases. Sellers are also reluctant to make legitimate challenges and do so in only 55.6% 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% of cases, while sellers reject legitimate challenges in 62.5% 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 gray steps represent the 10 buyers and five sellers who went bankrupt in the first 10 periods and whose lie frequencies are truncated. Similar to the SR mechanism, over 75% 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% chance of making a lie in the next period if they are not challenged and have a 22% 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% chance of making a false challenge in the next period if the counteroffer in the current period is accepted and a 55% chance of making a false challenge in the next period if the counteroffer in the current period is rejected.

In aggregate, the persistence of lies along with the losses that buyers and sellers incur in the preperiod stage leads 20% 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 Aghion et al. (2018) and is smaller than the proportion of buyers who lie in every period of the main treatment in Fehr, Powell, and Wilkening (2021).

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

Result C5. Buyers opt in to the mechanism in 77.5% of cases, while sellers opt in to the mechanism in 72.5% of cases. Opt-in rates are increasing for both buyers and sellers, and 87.0% of dyad pairs who opt in to the mechanism exhibit efficient truth-telling behavior and achieve the efficient outcome.

Figure C11 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% early in the sample and near 75% at the end of the sample.

Dyads who opt in to the mechanism reach the efficient outcome over 90% of the time, and buyers and sellers make truthful reports in all but five cases. All but one lie or false challenge end in a rejected counteroffer.[61] Investments in groups that opt out of the mechanism decrease over time just as in the SR treatment.

If we compare the results here with those of 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 counterparty behavior, 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 Fehr, Powell, and Wilkening (2021), where—in a similar SPI mechanism—buyers retaliate against legitimate challenges and sellers have a negative expected value for triggering arbitration.

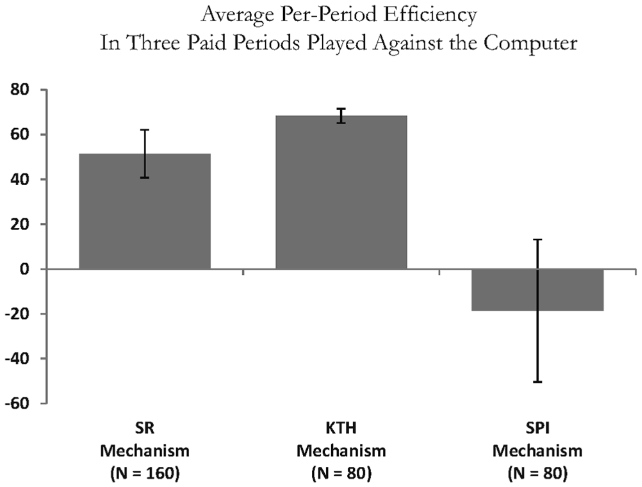
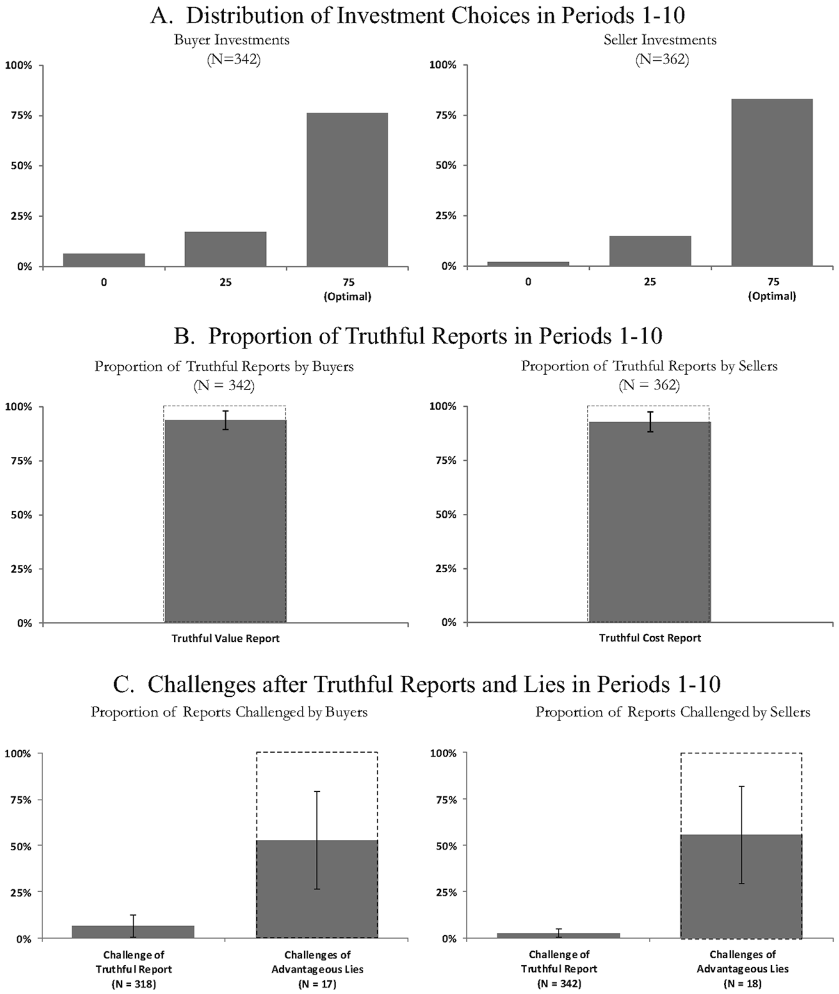


FIG. C9.—Average per-period efficiency in three paid periods played against computer.

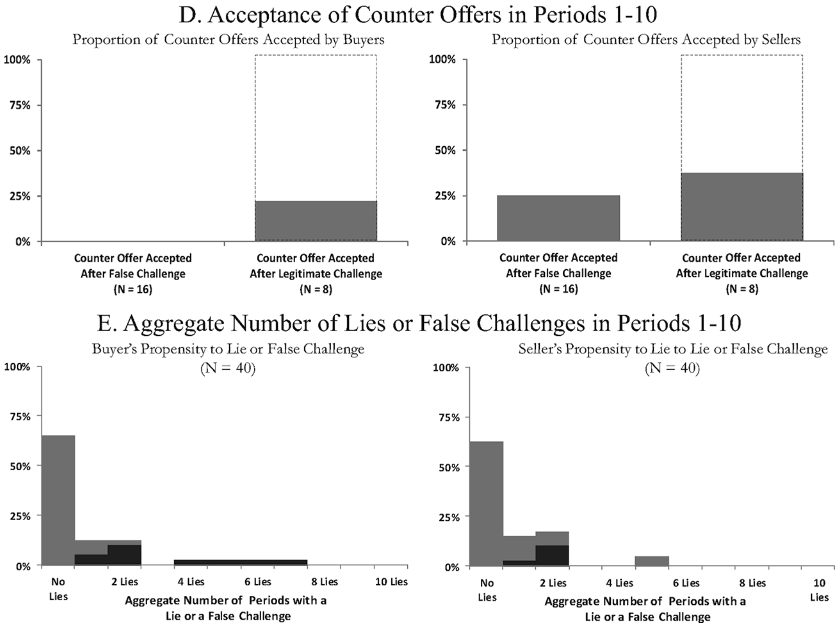
[61] One seller lies and makes a false challenge; thus, there are only four observations in the counteroffer stage. One of the buyer's counteroffer decisions is also missing because of an error in one of the matching groups in one period.



A. Distribution of Investment Choices in Periods 1-10

B. Proportion of Truthful Reports in Periods 1-10

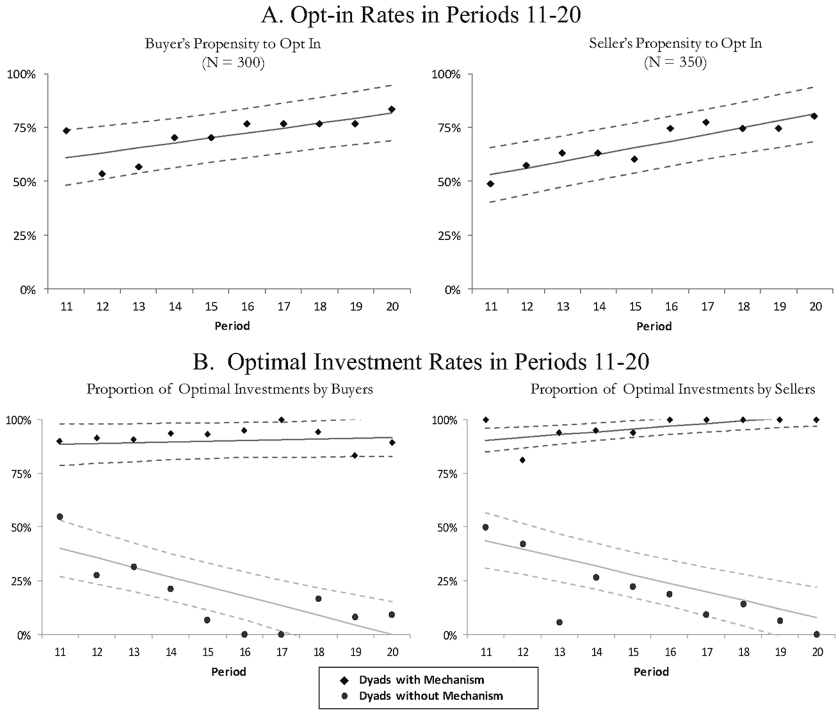
C. Challenges after Truthful Reports and Lies in Periods 1-10



D. Acceptance of Counter Offers in Periods 1-10

E. Aggregate Number of Lies or False Challenges in Periods 1-10

FIG. C10.—Pattern of play in first 10 periods of SPI mechanism.



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 155
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 2
Counter Offer Accepted by Seller After Truth	0 of 1

Fig. C11.—Pattern of play in periods 11–20 of SPI mechanism.

C6. Experiment 2: Behavior in the KTH Mechanism

In this section, we describe the behavior observed in the KTH mechanism in experiment 2.

Result C6. In periods 1–10, the KTH mechanism induces efficient investments in only 41.5% of cases. The mechanism induces truthful reports in only 50% of cases. Both investments and truthful reports are decreasing over time.

Figure C12 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-hand panels, and the behavior of sellers is shown in the right-hand 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% 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 an optimal investment is decreasing over time, with only 27.5% of buyers and 27.5% 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-hand side, buyers make truthful value reports in only 62.0% of cases and truthful cost reports in 80.3% of cases. Sellers make truthful value reports in 68.5% of cases and truthful cost reports in only 58.5% 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% of the 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 benefits them if there is a small chance that the other party makes a mistaken report. Buyers overstate their investment by reporting a cost below the true cost in 14.5% of cases and understate their investment in only 5.5% of cases. Likewise, sellers overstate their investment by reporting a value above the true value in 28% of cases and understate their investment in only 3.5% 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 counterparty: 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 underreport the value in 29.0% of cases and overreport the value in only 9.0% of cases. Sellers overreport the cost in 32.8% of cases and underreport the cost in only 8.8% of cases. As it is only possible to underreport values or overreport costs when matched with a partner who has chosen to invest, lies in the buyer value report and the seller cost report reduce the expected value of investing. As a result, buyers who make an efficient investment and report truthfully earn 15.8 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. 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.[62]

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

Result C7. Buyers and sellers retain the mechanism in only 20% of cases. Groups that retain the mechanism have lower average profits than those who dismiss the mechanism.

Buyers opt in to the KTH mechanism in 35.5% of cases, while sellers opt in to the mechanism in 57.0% 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% of cases, make truthful announcements in only 52.5% of cases, and earn −0.7 ECU on average. Sellers choose efficient investment in only 35.0% of cases, make truthful announcements in only 41.3% 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 1 if the mechanism is retained and zero if the mechanism is dismissed ($p < .01$).

In aggregate, the KTH mechanism is sensitive to systematic lies that attempt to take advantage of mistakes by the counterparty but that 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.

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 efficient investment (i.e., the case where the true cost is 10) strictly prefers to report the true cost if he has

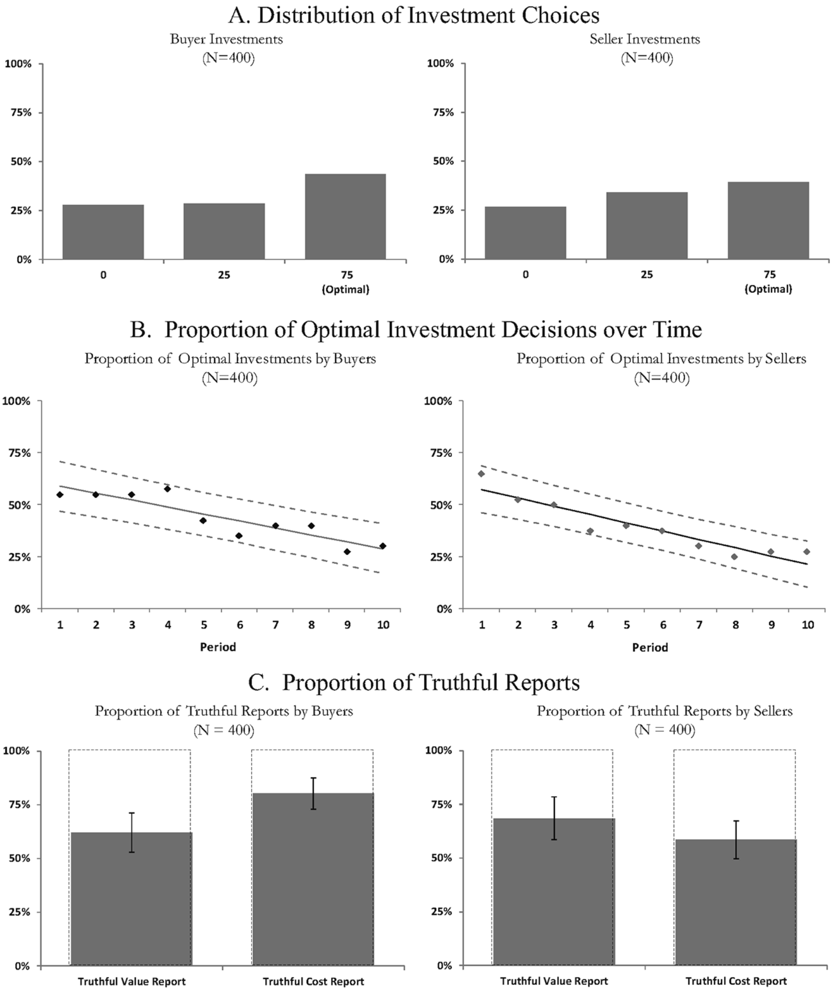
[62] It can also be shown that if one uses the agent quantal response equilibrium as an equilibrium concept, the probability that buyers and sellers make zero investment and maximally overstate their investment goes to 1 as noise approaches zero. In contrast to the assumption made in Kartik, Tercieux, and Holden (2014) that subjects report honestly when indifferent, the agent quantal response equilibrium 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 overreport their costs. Such matches lower the expected value of investment and truthful reporting. Noninvesting buyers who lie may end up matched with sellers who underreport costs, leading to an increase in the expected value of strategies involving lies and overstated investments. Models that combine the agent quantal response equilibrium with a preference for honesty rationalize the data reasonably well, though they cannot explain why sellers overstate their investment more frequently than buyers.

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 also consider a fine where the punishment exceeds the total gain associated with an advantageous lie. An advantage of the original 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 this approach is that there are multiple equilibria in the report stage in cases where efficient investments are made. As seen in experiment 3 below, the existence of multiple equilibrium may be problematic when communication between the two parties is allowed.

Ex ante, we chose the design where the incentives for truth telling for investing buyers and sellers were independent of the choice made by the other party. This design appears to have generated strong incentives for investing buyers and sellers to report truthfully. 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.[63]

[63] However, we should 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%), whereas buyers whose partner invests 75 report a value of 320 in only 108 out of 185 cases (58.4%). The large number of counterparty misreports is at odds with a preference for honesty and suggests that the alternative KTH mechanisms with higher fines may also have issue achieving the first-best investment because of the risk of miscoordination.



A. Distribution of Investment Choices

B. Proportion of Optimal Investment Decisions over Time

C. Proportion of Truthful Reports

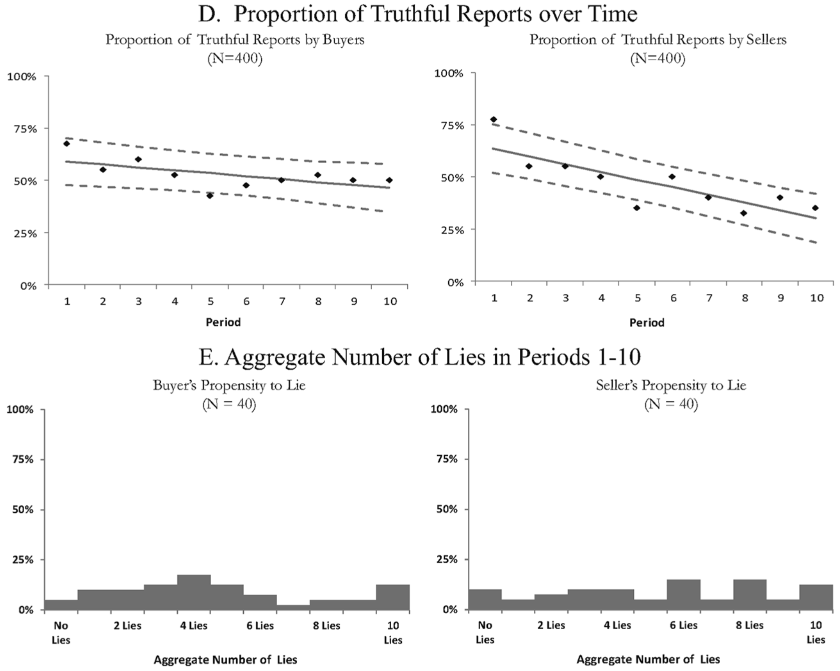


FIG. C12.—Pattern of play in first 10 periods of KTH mechanism.

C7. Experiment 2: 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 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 worst-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 are 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 in to the mechanism in periods 11–20, as this maximizes the impact of these subjects on the final outcome.

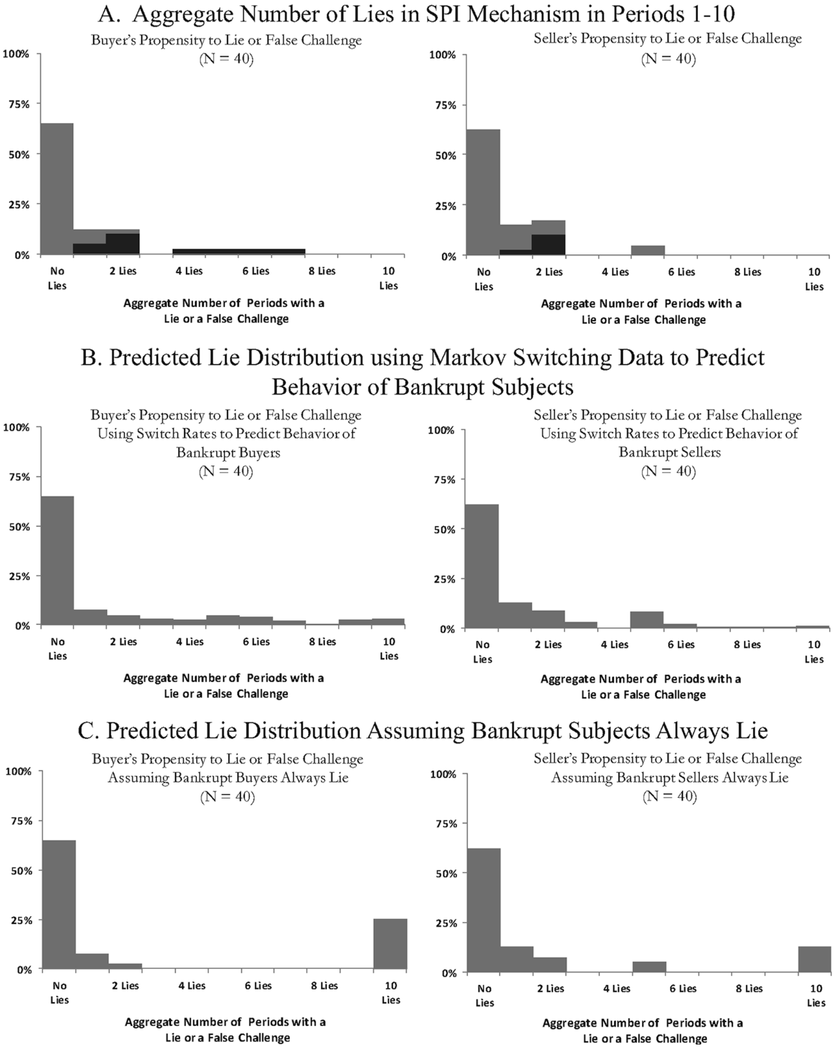
For our worst-case scenario method, we assume that a bankrupt subject lies in every single period over the entire sample. As above, we use the empirical return from lying to calculate the outcome for the subject and their matched pair and assume that bankrupt subjects always opt in to the mechanism.

For clarity, figure C13 shows 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 C3 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 44.3. This estimate is again above the earnings that we calculated in our original method because most bankruptcies occur early in the sample and most individuals are predicted to switch to truth-telling strategies before the end of the first 10 periods.

Using the worst-case scenario, we find that subjects in the SR mechanism earn 43.6 ECU on average. This is not significantly different from earnings in the fixed price treatment ($p = .363$) but is significantly different from the theoretical benchmark prediction of 35 ($p = .03$) 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 10.6 ECU and significantly below the earnings in the SR and fixed price treatments ($p < .01$ in both 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 that 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

FIG. C13.—Distribution of aggregate lies under alternative methods for dealing with bankruptcies.

TABLE C3
Alternative Efficiency Measures

Method	SR Mechanism	SPI Mechanism
Original	47.9	35.5
Switch rate	49.3	44.3
Always lie	43.6	10.6

Appendix D

Experiment 3: Comparison of the SR Mechanism and a Coordination Mechanism

In this section, we report on a third set of experiments that were designed to compare performance of the SR mechanism with a coordination mechanism that is common in the literature. The coordination mechanism is a simple one-stage mechanism that can weakly implement a pricing rule using only an SR stage and disagreement fines. We conjecture that the coordination mechanism is easier to understand than the SR mechanism and that it is likely to perform well in ideal circumstances where truthful reporting is likely to be focal. However, because of the existence of multiple equilibria, we predict that this type of mechanism is sensitive to directed communication, which can be used by participants to coordinate on lies.

Our design considers a simplified one-sided holdup problem where the seller can produce a nondivisible widget for a buyer. Prior to exchanging the good, the seller can choose to make a relationship-specific investment of $e_S \in \{0, 25, 75\}$, which increases the value of the final good for the buyer. Investment costs the seller e_S but results in a value of $v(e_S$ to the buyer, where $v(0) = 100$, $v(25) = 150$, and $v(75) = 220$. On the basis of the values and investment costs, a seller investment of 75 is efficient.

The object has no outside option value to the seller, and there are no additional production costs beyond the relationship-specific investments. We further assume that the buyer's value is *observable* to both parties but *nonverifiable* by a court. The parties would like to write a state-dependent contract using the price schedule in table D1. Under this contract, the buyer always receives a payment of 55 and the seller is the residual claimant of the surplus generated by his or her investment. Thus, if the social choice function is implemented, the seller always has an incentive to choose the investment that maximizes the joint surplus.

The parameters chosen in this experiment are similar to those of experiment 2, with two exceptions. First, since we are dealing with a one-sided problem, we removed the seller's production costs and adjusted the buyer values to simplify the instructions. Second, we have concentrated on an asymmetric equilibrium where in the efficient outcome the buyer receives 55 and the seller receives 90. Our interest in the asymmetric environment is to explore whether the mechanisms studied have reasonable truth-telling properties even in cases where payoffs are asymmetric and where the efficient outcome might be deemed unfair.[64]

D1. The Experiment

Our experiment utilizes a 2 × 2 design in which we generate between-subject variation in mechanism and within-subject variation in communication. Within a session, subjects participate in eight periods where we do not allow for communication between the participants followed by 16 periods where we introduce an additional stage where the buyer can send directed messages to the seller. These messages occur after the seller has made his or her investment choice but before the parties participate in the revelation mechanism.

We first describe the no-communication and communication treatments in sessions where the SR mechanism is used. As seen below, the coordination mechanisms is identical to the SR mechanism except that the arbitration stage is removed and disagreements in the report stage always lead to fines and no exchange.

D1.1. No-Communication Treatment

In periods 1–8 of the experiment, a seller is matched with a buyer and given the opportunity to invest to increase the value of the good to the buyer. The seller can choose an investment of $e_S \in \{0, 25, 75\}$. Investment costs the seller e_S but results in a value of $v(e_S)$ to the buyer, where $v(0) = 100$, $v(25) = 150$, and $v(75) = 220$.

After making investments, the buyer is informed of the true value of the good and the seller is reminded of this value. The buyer and seller next enter into the SR mechanism to set prices and determine whether trade occurs.

The SR mechanism consists of a report stage, a verification stage, and an arbitration stage. In the report stage, the buyer is asked to make a value report $\hat{v}_B \in \{100, 150, 220\}$ to the computer. The seller is also asked to make a value report $\hat{v}_S \in \{100, 150, 220\}$ to the computer. The two reports are made simultaneously.

The reports of the buyer and the seller are compared by the computer in the verification stage. If all reports coincide, the buyer and seller trade at the report-specific prices given in table D2. Prices in this table were constructed using the function

[64] Social preference robust mechanism design is explored theoretically and experimentally in Bierbrauer et al. (2017). The paper shows that when individuals have social preferences, truth-telling rates can be improved by making mechanisms externality free by ensuring that the reports of one party do not influence the payoffs of the other party. In the holdup setting that we consider, Fehr, Powell, and Wilkening (2021) show that only fixed price contracts are externality free. Thus, it is not possible to design mechanisms that are fully robust to social preferences that can also implement the first best when the holdup problem is two sided. Our focus here is to assess whether inequality aversion—one of the more common forms of social preferences observed in bilateral settings—leads to large deviations in truth telling when compared with our original experiments where payoffs were equal along the equilibrium path. We note that a full treatment of social preferences in the incomplete contracting setting would require the additional exploration of the contracting stage because it is possible for parties to make initial transfers at the point of signing the contract to achieve distributional objectives. Such transfers could potentially be used to relax outcome-based social preference constraints.

$$P^{SR}(\hat{v}) = (\hat{v} - \underline{v}) + 45,$$

where $\hat{v}$ is the jointly reported value and $\underline{v}$ is the minimum value of 100. The trade prices are structured so that the seller receives the marginal surplus created from her investment.

If there is a discrepancy in reports, the buyer enters into the arbitration stage and must pay an arbitration fee of 150. The buyer is then asked to make a second report regarding the value of the good. This report can be {0, 100, 150, 220}. As shown in table D2, we use the report along with a fair six-sided dice to determine whether trade occurs and the price.[65] If the second report of the buyer matches the first-stage report of the seller, the seller is rewarded an arbitration bonus of 150 in addition to his or her earnings for the round. In other cases, the seller must also pay an arbitration fee of 150.

If trade occurs, the profits of the buyer and seller are given by

$$\pi_B = \text{Value} - \text{Price} - \text{Buyer's Arbitration Fee},$$

$$\pi_S = \text{Price} - \text{Investment Costs} \pm \text{Seller's Arbitration Fee or Bonus}.$$

If trade does not occur, the object is destroyed. However, both parties must sill pay their arbitration fees, and the seller must still pay his or her investment costs.

D1.2. Communication Treatment

Periods 9–24 are identical to the first eight periods except that we introduce a communication stage between the investment stage of the seller and the SR mechanism. At the start of the communication stage, the buyer is informed of the investment decision of the seller and the true value of the good. The buyer may then send a message to the seller indicating which of the value reports he or she is planning to make. The buyer may send any of the following nonbinding messages:

- "I plan on reporting a value of 100."
- "I plan on reporting a value of 150."
- "I plan on reporting a value of 220."
- "I have chosen not to send a message."

The seller is informed of this message in the report stage and is reminded of the true value and the message when making their decision. For instance, if the true value is 220 and the buyer's message is "I plan on reporting a value of 100," the seller's screen will display "The true value is 220 and the buyer plans on reporting a value of 100. What is your report?" next to the decision box.

Our choice of restricting communication only to a prespecified set of messages for the buyer is based on Cooper et al. (1989), which explored different communication structures in the battle of the sexes game. It finds that unidirectional

[65] In experiment 2, we did not allow participants to report zero in the arbitration stage. However, we observed cases where the buyer reported a value of 100 and matched a seller's lie on this value. It was not possible to identify whether this was a buyer trying to avoid mutual fines or a buyer who was trying not to trade and accidently matching the seller. We added the zero report to help distinguish between these two cases.

communication is effective at coordinating outcomes on the sender's preferred outcome. Thus, although the communication space is restricted, the message space allows for the types of messages that have been found in previous work to be effective at influencing equilibrium selection.

D1.3. The Coordination Mechanism

The coordination mechanism uses the same report stage and verification stage as the SR mechanism. However, if the reports disagree in the verification stage, both the buyer and the seller are fined 150 and trade does not occur. The instructions for the two mechanisms were similar except that we discuss a verification system in the instructions for the coordination mechanism and an arbitration and verification system in the instructions for the SR mechanism.

D2. Protocol

Each session of experiment 3 consisted of either 14 or 16 participants who were evenly divided between buyers and sellers at the beginning of the experiments. Each session was divided into three eight-period phases. In sessions consisting of 16 participants, buyers and sellers were matched with each other at most once in each eight-period phase. In sessions consisting of 14 participants, we implemented perfect stranger matching in the first seven periods of a phase and randomly matched participants to a partner they had in periods 1–6 in the last period. This ensured that they never played against the same person in two consecutive periods.

All experiments were run in the Experimental Economics Laboratory at the University of Melbourne in April and May 2021. The experiments were conducted using the programming language z-Tree (Fischbacher 2007). A total of eight sessions were run: four sessions using the SR mechanism and four sessions using the coordination mechanism. All of the 124 participants were undergraduate students at the university and were invited from a pool of more than 4,000 volunteers using ORSEE (Greiner 2015).

Experiment 3 took place between under tight social distancing restrictions and a restricted maximum lab capacity of 16. 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. Once all participants successfully completed the quiz, a verbal summary was read aloud that summarized the trading mechanism and emphasized the matching used. In the initial instructions, participants were told that there would be three phases in the experiment and that phases 2 and 3 would be similar to the first phase except that there would be an additional stage where some of the participants could send messages. Subjects were also informed that their decisions in phase 1 would not influence their position, matching, or available actions in phase 2 or phase 3.

Subjects then entered and played phases 1–3 of the experiment. We handed out new instructions after phase 1 that explained the additional communication stage and reiterated the matching protocols. We also displayed additional instructions after phase 2 that reiterated the matching protocol. The division of the communication treatment into two phases was used to keep the matching protocols the same in each eight-period phase but to allow for additional time for behavior to evolve in the no-communication treatments.

We randomly selected one period from each eight-period phase for payment with an exchange rate of 10 ECU = A$1. To ensure that participants did not go bankrupt, we also gave participants a A$35 one-off payment for completing the experiment. Any losses that a participant incurred in the experiment were subtracted from this initial payment. The average payment at the end of the experiment was A$49.22. At the time of the 2021 experiments, A$1 ≈ US$0.77.

D3. Hypotheses

The SR mechanism is designed to implement truthful announcements and to allow buyers and sellers to capture all surplus associated with their investment. The mechanism is also predicted to be robust to communication. Thus, we would predict the following:

HYPOTHESIS 7. The path of play under the SR mechanism involves the seller making efficient investments and both parties making truthful reports in both the no-communication treatment and the communication treatment.

If truth telling acts as a focal report, we would also expect truthful reports and efficient investments in the no-communication treatment with the coordination mechanism. However, we would predict that in the communication treatment, the buyers will send messages that are used to coordinate reports on values that are below the true value in cases where the seller has exerted costly effort. The seller is worse off in these equilibria than they would be if they chose no effort, and thus we would predict that the mechanism cannot support efficient effort and that play will converge to an equilibrium with no investments.

HYPOTHESIS 8. In the no-communication treatment, the path of play under the coordination mechanism involves the seller making an efficient investment and both parties making truthful reports. In the communication treatment, the path of play under the coordination mechanism involves the sellers making no investment.

D4. Results

D4.1. Results and Support

RESULT D1. The introduction of communication does not significantly change the distribution of investment choices made by sellers in sessions using the SR mechanism. By contrast, the introduction of communication leads to a significant decrease in the proportion of sellers choosing efficient effort in sessions using the coordination mechanism.

Support for result D1 is provided in panels A–C of figure D1. Panel A shows the distribution of investment choices with and without communication in sessions using the SR mechanism. As seen in the left-hand panel, sellers choose the efficient investment in 68.6% of cases and an investment of 25 in a further 21.5% of cases. As seen in the right-hand panel, investments are similar in the communication treatment, with efficient investments selected in 63.9% of cases and an investment of 25 selected in a further 23.0% of cases. There is no significant difference in the observed distribution of investment choices across the two treatments in a random effects ordered probit model where investment choice is regressed on a dummy that is 1 when communication is allowed ($p = .383$).[66]

Panel B shows the distribution of investments with and without communication in the coordination mechanism. Without communication, 80.8% of investment choices are efficient. However, when communication is introduced, only 28.5% of investment choices are efficient. The observed distribution of investment choices are significantly different across the two treatments using the random effects ordered probit specification described above ($p < .01$).

Panel C shows the proportion of efficient investment choices made in the two treatments over time. The data are overlaid with the predictions and 95% confidence intervals from a simple linear random effects regression that regressed a dummy variable that is 1 if efficient effort is chosen and zero otherwise on the period. As seen in the left-hand panel, the investments made in the SR mechanism and coordination mechanism are similar in the no-communication treatment, and there is no significant difference in the proportion of optimal investments using a random effects regression where efficient investment is regressed on the mechanism ($p = .115$). However, as seen in the right-hand panel, efficient investment is decreasing over time in the coordination mechanism when communication is introduced, and the difference in mechanisms is significant using the same specification ($p = .031$).

The rapid decrease in efficient investment choices observed in the coordination mechanism when communication is introduced suggests that sellers are not able to recover their investments in this treatment. The following result shows that this is due to buyers using their messages to coordinate on lies.

Result D2. In the treatment with the coordination mechanism and communication, buyers often send messages indicating that they plan to lie when sellers make a costly investment. These messages lead buyers and sellers to coordinate on a lie. Buyers are less likely to send messages indicating that they plan to lie in the SR mechanism, and sellers typically ignore these messages.

Support for result D2 is given in panel A of figure D2, which concentrates exclusively on the treatments with communication. As seen on the left-hand side of this panel, buyers in sessions with the coordination mechanism frequently send messages indicating that they plan to lie when the seller has made a costly investment and the true value of the good is 150 or 220. As seen on the right-hand side of the panel, these messages are frequently used by both parties to coordinate on the message sent by the buyer.

Coordinating on a lie increases the buyer's profit but reduces the profit of the investing seller. Empirically, sellers in the communication treatment who invest receive less profits than they receive by investing zero. Thus, the reduction in investments observed in the coordination mechanism when communication is introduced is a rational response to the buyer's message choices.

[66] As with the main paper, we cluster all errors at the individual level for all tests reported in app. D.

In the SR mechanism, by contrast, sellers typically ignore nontruthful messages by the buyer and report the true value in the report stage. As a result, nontruthful messages are not profitable and decrease in frequency from 18.4% of potential cases in periods 9–16 to 11.7% of potential cases in periods 17–24. Given the distribution of buyer messages, the seller's expected earnings for optimally investing are higher than not investing, and the high level of investments observed in this treatment is again a rational response to the buyer's message choices.

Taken together, results D1 and D2 suggest that the coordination mechanism is simpler than the SR mechanism and performs well without communication. However, when communication is possible, it may be possible to use messages to coordinate on lies. We view such strategic communication to be particularly problematic in the bilateral investment setting, where contracts must be signed before investments take place but the revelation mechanism occurs after investments have been realized. In such settings, it seems unlikely that communication can be prevented over the length of the contract.

D4.2. Comparison of the SR Mechanism in Experiments 2 and 3

Although results D1 and D2 are consistent with hypotheses 7 and 8, the SR mechanism is not as effective at inducing optimal investment and truthful reports when compared with the behavior of sellers in the two-sided mechanism discussed in experiment 2. In particular, optimal investments are chosen in only 68.6% of cases in the no-communication treatment of experiment 3, but the optimal investment was chosen by sellers in 84.8% of cases in phase 1 of experiment 2. The truth-telling rates of buyers and sellers are also slightly lower in experiment 3, with buyers making a truthful report in 93% of cases and the seller making truthful reports in 89.5%. These truth-telling rates are below those found in experiment 2, where 98% of buyers and 93% of sellers made truthful value reports. Aggregate payoffs are also lower, with sellers receiving an average payoff of 62.25 and buyers receiving an average payoff of only 27.0 in the no-communication treatment.[67] These payoffs are slightly lower than what would occur if there was no mechanism and the seller always made no investment. They are also lower than the average earnings in the no-communication treatment with the coordination mechanism, where the average earnings of sellers was 69.4 and the average earnings of buyers was 42.2.

There are two potential reasons for the difference in the SR treatment across the two experiments. First, experiment 3 explores an environment where the buyer and seller receive different payoffs after optimal investment. This inequity may lead buyers to lie after optimal investments and reduce the overall efficiency of the SR mechanism. Second, experiment 2 uses a preplay treatment, where participants play against the computer. Thus, the difference may be due to the additional training provided in experiment 2, which may have influenced the extent to which participants experimented with different strategies.[68]

[67] Payoffs are similar in the communication treatment, with sellers earning an average of 51 and buyers receiving an average of 27.2.

[68] We did not have participants play against the computer in experiment 3 because we were interested in whether participants in the coordination mechanism naturally adopt the truth-telling equilibrium on their own in the no-communication treatment.

The data do not suggest that inequity is a main reason for the difference in truth-telling rates and investments across the two experiments. Panel A of figure D3 reports on the proportion of truthful reports of buyers and sellers in the no-communication and communication treatments of the SR mechanism, with the data separated by the value of the object. As seen in this figure, when the seller chooses the optimal investment, the seller makes a truthful report in 98.3% of cases and the buyer makes a truthful report in 93.8% of cases. Thus, truth-telling rates are highest after optimal investments where the payoff difference between the buyer and seller is the largest.

Instead, the data suggest that both suboptimal investments and seller misreports are driven by sellers who experiment with a strategy of choosing low effort and then reporting a value of 220. Recall that a seller can be fined or rewarded in the SR mechanism on the basis of the secondary reports made by the buyer. If the seller is uncertain about the incentives of the mechanism, they may experiment with a strategy where they place the buyer into arbitration and hope that the buyer takes an action that provides them both with a high price and a bonus. Such actions are not optimal for selfish buyers but may occur if the buyer is trying to minimize pairwise losses by ensuring that fines are transferred to their counterparty.

Experimentation with a strategy of choosing low effort and lying can be seen in panel A of figure D3, where sellers who choose an effort of zero in the no-communication treatment are truthful in only nine of 25 cases but lie and report a value of 220 in 12 of the remaining 16 cases. Sellers who choose an effort of 25 in the no-communication treatment are truthful in 47 out of 55 cases. However, the seller reports a value of 220 in seven of the remaining eight cases.

Restricting attention to the arbitration stage in the 19 cases where the seller makes a suboptimal investment and announces a value of 220, we find that the buyer makes a truthful report in nine cases and punishes the seller with a report below the true value in four cases but matches the seller's lie in six cases. Thus, although the expected value of putting the seller into arbitration is less than the optimal strategy, a subset of sellers are rewarded for using a suboptimal strategy. These rewards tend to reinforce the seller's investment and reporting strategy: a seller who reports a value of 220 after making a suboptimal investment in period *t* repeats this strategy in 80% of cases if the buyer matches their lie. By contrast, a seller who reports a value of 220 after making a suboptimal investment in period *t* lies in the next period in only 22.2% of cases if the buyer does not match their lies. Thus, a large portion of suboptimal investment choices and seller lies appear to be due to sellers who experiment with a strategy of low investment and lies and who are rewarded for this strategy.

The results here are similar to those seen in figure 4 from experiment 2, where a small subset of sellers also experimented with lies and some buyers chose not to punish them in early periods. However, experimentation is more frequent in experiment 3, suggesting that the early training periods against the computer were important for reducing seller experimentation in experiment 2. These results suggest that initial training is particularly important for the SR mechanism, which is likely to be unfamiliar to potential users.

TABLE D1

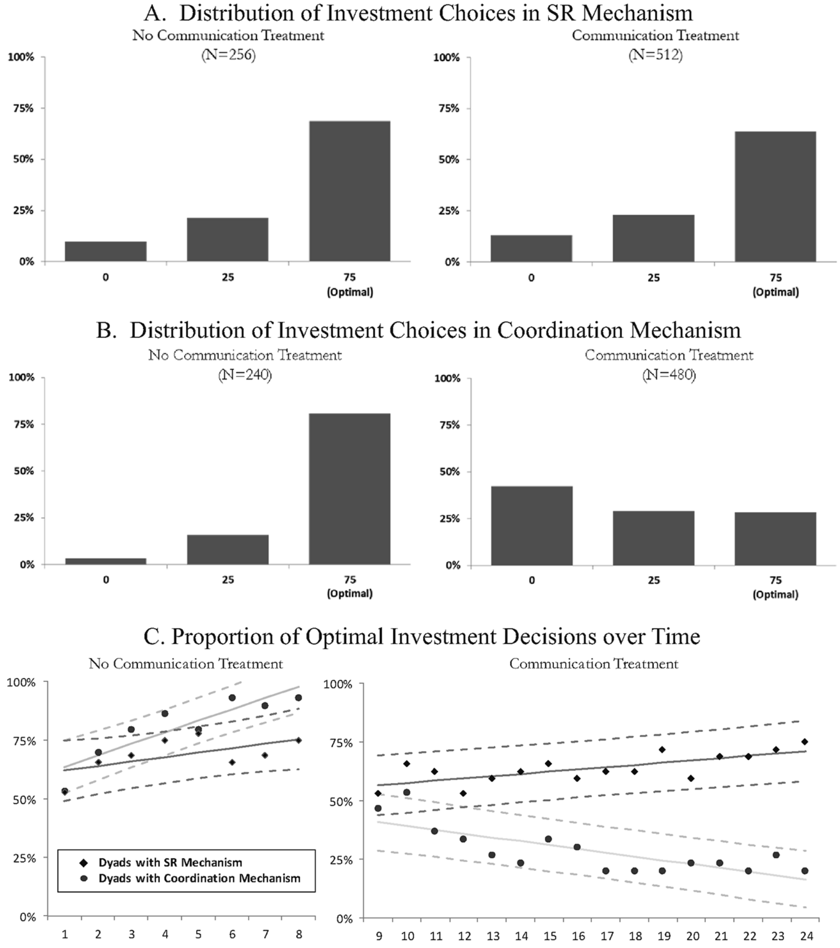
Price Schedule

v	$p(v)$
100	45
150	95
220	165

TABLE D2

Trade Prices in Buyer and Seller Arbitration Stages

Buyer's Secondary Report	Outcome If Roll Is 1 or 2	Outcome If Roll Is 3 or 4	Outcome If Roll Is 5 or 6
0	No trade	No trade	No trade
100	Trade at 55	No trade	No trade
150	Trade at 55	Trade at 105	No trade
220	Trade at 55	Trade at 105	Trade at 175

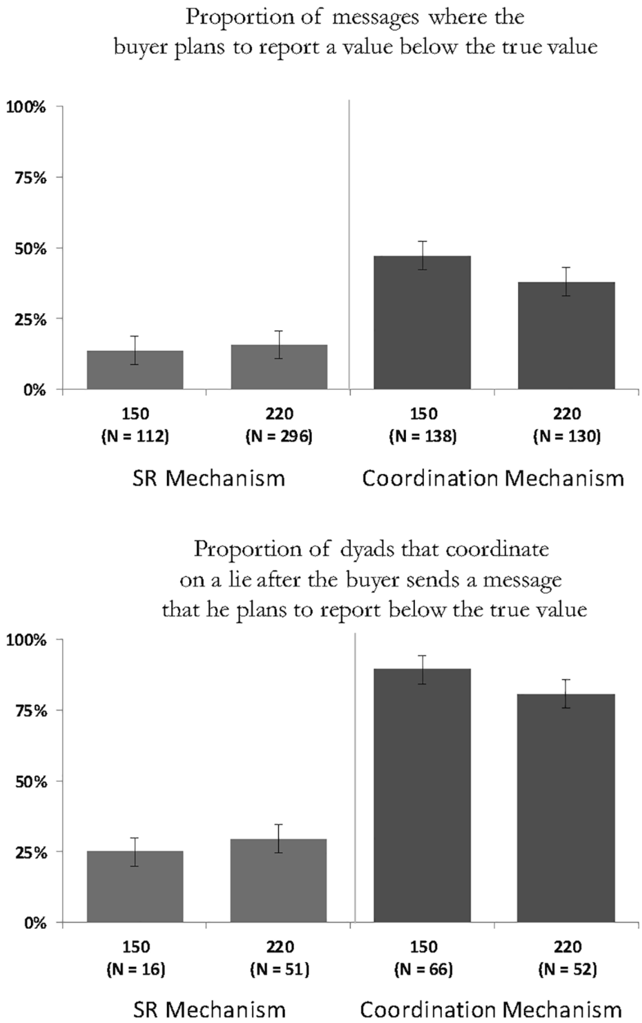


A. Distribution of Investment Choices in SR Mechanism

B. Distribution of Investment Choices in Coordination Mechanism

C. Proportion of Optimal Investment Decisions over Time

FIG. D1.—Investment choices in SR mechanism and coordination mechanism.


Proportion of messages where the buyer plans to report a value below the true value

Proportion of dyads that coordinate on a lie after the buyer sends a message that he plans to report below the true value

FIG. D2.—Buyer messages and subsequent coordination behavior in communication treatment for true values of 150 and 220.

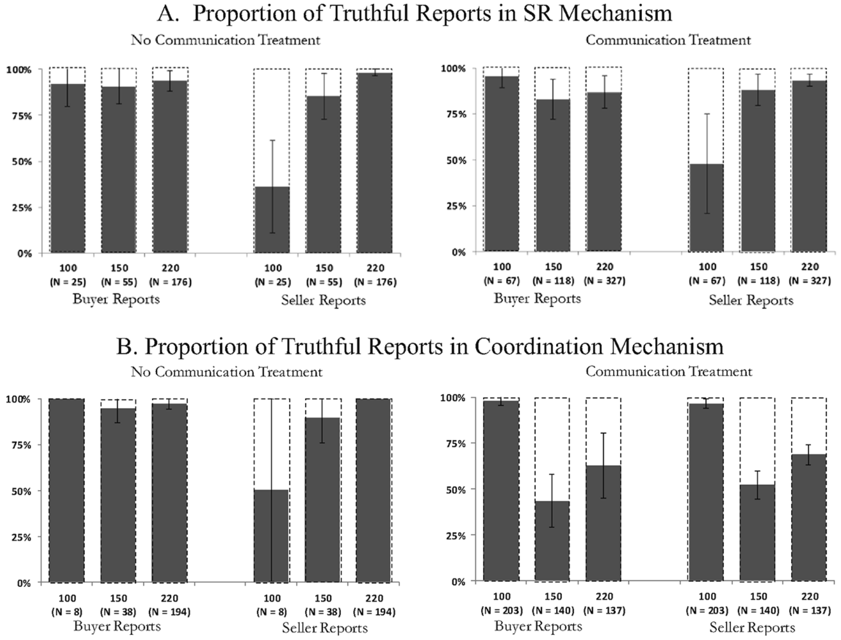


FIG. D3.—Frequency of truthful reports in SR and coordination mechanism by true value.

References

Abreu, Dilip, and Hitoshi Matsushima. 1992. “Virtual Implementation in Iteratively Undominated Strategies: Complete Information.” *Econometrica* 60:993–1008.

Aghion, Philippe, Mathias Dewatripont, and Patrick Rey. 1994. “Renegotiation Design with Unverifiable Information.” *Econometrica* 62 (2): 257–82.

Aghion, Philippe, Ernst Fehr, Richard Holden, and Tom Wilkening. 2018. “The Role of Bounded Rationality and Imperfect Information in Subgame Perfect Implementation—An Empirical Investigation.” *J. European Econ. Assoc.* 16 (1): 232–74.

Aghion, Philippe, Drew Fudenberg, Richard Holden, Takashi Kunimoto, and Olivier Tercieux. 2012. “Subgame-Perfect Implementation under Information Perturbations.” *Q.J.E.* 127 (4): 1843–81.

Andreoni, James, and Hal Varian. 1999. “Pre-Play Contracting in the Prisoners’ Dilemma.” *Proc. Nat. Acad. Sci. USA* 96:10933–38.

Arifovic, Jasmina, and John Ledyard. 2004. “Scaling Up Learning Models in Public Good Games.” *J. Public Econ. Theory* 6 (2): 203–38.

Arya, Anil, Jonathan Glover, and Richard Young. 1995. “Virtual Implementation in Separable Bayesian Environments Using Simple Mechanisms.” *Games and Econ. Behavior* 9 (5): 127–38.

Attiyeh, Greg, Robert Franciosi, and R. Mark Isaac. 2000. “Experiments with the Pivot Process for Providing Public Goods.” *Public Choice* 102 (1–2): 95–114.

Battigalli, Pierpaolo, and Marciano Siniscalchi. 1999. “Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games.” *J. Econ. Theory* 88 (1): 188–230.

Ben-Porath, Elchanan. 1997. “Rationality, Nash Equilibrium and Backwards Induction in Perfect-Information Games.” *Rev. Econ. Studies* 64 (1): 23–46.

Bergemann, Dirk, and Stephen Morris. 2005. “Robust Mechanism Design.” *Econometrica* 73 (6): 1771–813.

Bierbrauer, Felix, Axel Ockenfels, Andreas Pollak, and Désirée Rückert. 2017. “Robust Mechanism Design and Social Preferences.” *J. Public Econ.* 149:59–80.

Bracht, Juergen, Charles Figuières, and Marisa Ratto. 2008. “Relative Performance of Two Simple Incentive Mechanisms in a Public Goods Experiment.” *J. Public Econ.* 92 (1–2): 54–90.

Cameron, Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. “Bootstrap-Based Improvements for Inference with Clustered Errors.” *Rev. Econ. and Statis.* 90 (3): 414–27.

Che, Yeon-Koo, and Donald B. Hausch. 1999. “Cooperative Investments.” *A.E.R.* 89 (1): 125–47.

Chen, Yan, and Robert Gazzale. 2004. “When Does Learning in Games Generate Convergence to Nash Equilibria? The Role of Supermodularity in an Experimental Setting.” *A.E.R.* 94 (5): 1505–35.

Chen, Yan, and Charles Plott. 1996. “The Groves-Ledyard Mechanism: An Experimental Study of Institutional Design.” *J. Public Econ.* 59 (3): 335–64.

Chen, Yan, and Fang-Fang Tang. 1998. “Learning and Incentive-Compatible Mechanisms for Public Goods Provision: An Experimental Study.” *J.P.E.* 106 (3): 633–62.

Chung, Kim-Sau, and Jeffrey C. Ely. 2003. “Implementation with Near-Complete Information.” *Econometrica* 71 (3): 857–71.

Chung, Tai-Yeong. 1991. “Incomplete Contracts, Specific Investment, and Risk Sharing.” *Rev. Econ. Studies* 58 (5): 1031–42.

Cooper, Russell, Douglas V. DeJong, Robert Forsythe, and Thomas Ross. 1989. “Communication in the Battle of the Sexes Game: Some Experimental Results.” *RAND J. Econ.* 20 (4): 568–87.

Crawford, Vincent P., and Nagore Iriberri. 2007. “Level-*k* Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner’s Curse and Overbidding in Private-Value Auctions?” *Econometrica* 75 (6): 1721–70.

de Clippel, Geoffroy, Kfir Eliaz, and Brian Knight. 2014. “On the Selection of Arbitrators.” *A.E.R.* 104 (11): 3434–58.

de Clippel, Geoffroy, Rene Saran, and Roberto Serrano. 2019. “Level-*k* Mechanism Design.” *Rev. Econ. Studies* 86 (3): 1207–27.

Dekel, Eddie, and Drew Fudenberg. 1990. “Rational Behavior with Payoff Uncertainty.” *J. Econ. Theory* 52 (2): 243–67.

Dekel, Eddie, and Marciano Siniscalchi. 2015. “Epistemic Game Theory.” In *Handbook of Game Theory with Economic Applications*, vol. 4, edited by H. Peyton Young and Shmuel Zamir, 619–702. Amsterdam: Elsevier.

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. “A Theory of Sequential Reciprocity.” *Games and Econ. Behavior* 47 (2): 268–98.

Eccles, Peter, and Nora Wegner. 2016. “Robustness of Subgame Perfect Implementation.” Working Paper no. 601, Bank England, London.

Falkinger, Josef, Ernst Fehr, Simon Gächter, and Rudolf Winter-Ebrner. 2000. “A Simple Mechanism for the Efficient Provision of Public Goods: Experimental Evidence.” *A.E.R.* 90 (1): 247–64.

Fehr, Ernst, Michael Powell, and Tom Wilkening. 2021. “Behavioral Constraints in the Design of Subgame-Perfect Implementation Mechanisms.” *A.E.R.* 111 (4): 1–37.

Fischbacher, Urs. 2007. “z-Tree: Zurich Toolbox for Ready-Made Economic Experiments.” *Experimental Econ.* 10 (2): 171–78.

Fudenberg, Drew, David M. Kreps, and David K. Levine. 1988. “On the Robustness of Equilibrium Refinements.” *J. Econ. Theory* 44 (2): 354–80.

Giannatale, Sonia Di, and Alexander Elbittar. 2010. “King Solomon’s Dilemma: An Experimental Study on Implementation.” CIDE Working Paper 477, Centro de Investigación y Docencia Económicas, Mexico City.

Greiner, Ben. 2015. “Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE.” *J. Econ. Sci. Assoc.* 1 (1): 114–25.

Grossman, Sanford J., and Oliver Hart. 1986. “The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration.” *J.P.E.* 94:691–719.

Harstad, Ronald M., and Michael Marrese. 1981. “Implementation of Mechanism by Processes: Public Good Allocation Experiments.” *J. Econ. Behavior and Org.* 2 (2): 129–51.

———. 1982. “Behavioral Explanations of Efficient Public Good Allocations.” *J. Public Econ.* 19 (3): 367–83.

Hart, Oliver, and John Moore. 1990. “Property Rights and the Nature of the Firm.” *J.P.E.* 98 (6): 1119–58.

———. 2003. “Some (Crude) Foundations of Incomplete Contracts.” Working paper.

Healy, Paul J. 2006. “Learning Dynamics for Mechanism Design: An Experimental Comparison of Public Goods Mechanisms.” *J. Econ. Theory* 129 (1): 114–49.

Hoppe, Eva I., and Patrick W. Schmitz. 2011. “Can Contracts Solve the Hold-Up Problem? Experimental Evidence.” *Games and Econ. Behavior* 73 (1): 186–99.

Jackson, Matthew O. 2001. “A Crash Course in Implementation Theory.” *Social Choice and Welfare* 18 (4): 655–708.

Kajii, Atsushi, and Stephen Morris. 1997. “The Robustness of Equilibria to Incomplete Information.” *Econometrica* 65:1283–309.

Kartik, Navin, Olivier Tercieux, and Richard Holden. 2014. “Simple Mechanisms and Preferences for Honesty.” *Games and Econ. Behavior* 83:284–90.

Katok, Elena, Martin Sefton, and Abdullah Yavas. 2002. “Implementation by Iterative Dominance and Backward Induction: An Experimental Comparison.” *J. Econ. Theory* 104:89–103.

Maskin, Eric. 1977. “Nash Equilibrium and Welfare Optimality.” Working paper.

———. 1999. “Nash Equilibrium and Welfare Optimality.” *Rev. Econ. Studies* 66 (1): 23–38.

Maskin, Eric, and Jean Tirole. 1999. “Unforeseen Contingencies and Incomplete Contracts.” *Rev. Econ. Studies* 66 (1): 83–114.

McKelvey, Richard D., and Thomas R Palfrey. 1998. “Quantal Response Equilibria for Extensive Form Games.” *Experimental Econ.* 1 (1): 9–41.

Monderer, Dov, and Dov Samet. 1989. “Approximating Common Knowledge with Common Beliefs.” *Games and Econ. Behavior* 1 (2): 170–90.

Moore, John, and Rafael Repullo. 1988. “Subgame Perfect Implementation.” *Econometrica* 56:1191–220.

Myerson, Roger B. 1986. “Multistage Games with Communication.” *Econometrica* 54 (2): 323–58.

Nöldeke, Georg, and Klaus Schmidt. 1995. “Option Contracts and Renegotiation: A Solution to the Hold-Up Problem.” *RAND J. Econ.* 26 (2): 163–79.

Ponti, Giovanni, Anita Gantner, Dunia López-Pintado, and Robert Montgomery. 2003. “Solomon’s Dilemma: An Experimental Study on Dynamic Implementation.” *Rev. Econ. Design* 8 (2): 217–39.

Sefton, Martin, and Abdullah Yavas. 1996. “Abreu-Matsushima Mechanisms: Experimental Evidence.” *Games and Econ. Behavior* 16 (2): 280–302.