Fehr, Ernst; Powell, Michael; Wilkening, Tom

Article — Accepted Manuscript (Postprint)

Behavioral Constraints on the Design of Subgame-Perfect Implementation Mechanisms

American Economic Review

Suggested Citation: Fehr, Ernst; Powell, Michael; Wilkening, Tom (2021) : Behavioral Constraints on the Design of Subgame-Perfect Implementation Mechanisms, American Economic Review, ISSN 1944-7981, American Economic Association, Pittsburgh, PA, Vol. 111, Iss. 4, pp. 1055-1091, https://doi.org/10.1257/aer.20170297

Behavioral Constraints on the Design of Subgame-Perfect Implementation Mechanisms

By Ernst Fehr and Michael Powell and Tom Wilkening*

We study subgame-perfect implementation (SPI) mechanisms that have been proposed as a solution to incomplete contracting problems. We show that these mechanisms — which are based on off-equilibrium arbitration clauses that impose large fines for lying and the inappropriate use of arbitration — have severe behavioral constraints because the fines induce retaliation against legitimate uses of arbitration. Incorporating reciprocity preferences into the theory explains the observed behavioral patterns and helps us develop a new mechanism that is more robust and achieves high rates of truth-telling and efficiency. Our results highlight the importance of tailoring implementation mechanisms to the underlying behavioral environment.
JEL: D23, D71, D86, C92
Keywords: Implementation Theory, Incomplete Contracts, Experiments

Incomplete contracts pervade economic and political life. Politicians in executive positions as well as bureaucrats in ministries and agencies act on the basis of loose objectives, and the obligations of employees and managers in private organizations are often described in vague terms. Economists have explored the implications of incomplete contracts by developing models that assume that key payoff-relevant information is observable but not verifiable by a third-party enforcer.[1] Such observable but non-verifiable information implies that third-party enforcement of state-contingent contracts is infeasible and that formal contracting

* Fehr: Department of Economics, Zurich University. Blümlisalpstrasse 10, 8006 Zurich, Switzerland. E-mail: ernst.fehr@econ.uzh.ch. Powell: Strategy Department, Kellogg School of Management, Northwestern University. 2211 Campus Drive, Evanston, IL 60208. E-mail: mike-powell@kellogg.northwestern.edu. Wilkening: Department of Economics, The University of Melbourne. FBE Building, 111 Barry Street, Carlton VIC 3053, Australia. E-mail: Tom.Wilkening@unimelb.edu.au.
We thank James Bland, Sanket Patil, and Hans Zhu for excellent research assistance. We also thank Mathias Dewatripont, Martin Dufwenberg, Greg Fischer, Robert Gibbons, Lorenz Goette, Oliver Hart, Eric Maskin, Jean Tirole, and Christian Zehnder for helpful comments. We gratefully acknowledge the financial support of the Australian Research Council through the Discovery Early Career Research Award DE140101014 as well as the Faculty of Business and Economics at the University of Melbourne. Ernst Fehr acknowledges support by the Swiss National Science Foundation (Project on "Distribution and Determinants of Social Preferences"; Project no. `100018_140734\1` and the European Research Council (Advanced Grant on "Foundations of Economic Preferences")).

[1] The assumption has been used to understand property rights and firm boundaries (Grossman & Hart 1986; Hart & Moore 1990; Hart 1995), the optimal scope of governments (Hart, Shleifer, & Vishny 1997; Besley & Ghatak 2001), problems of privatization (Schmidt 1996a, 1996b), the control of insiders by outsiders through voting rights (Grossman & Hart 1988; Gromb 1993) or financial contracts (Aghion & Bolton 1992; Dewatripont & Tirole 1994; Hart & Moore 1998), and patterns of international trade and technology adoption (Antras 2003; Nunn 2007; Acemoglu, Antras, & Helpman 2007).

is ineffective.

The tractable nature of models using the assumption of observable but non-verifiable information has made them an essential tool for evaluating trade-offs in institutional design. However, despite its widespread influence, the assumption that payoff-relevant information is observable but non-verifiable stands on controversial theoretical foundations. Building on work by Moore & Repullo (1988), Maskin & Tirole (1999) show that if parties commonly observe payoff-relevant information, there often exists an auxiliary extensive-form mechanism that induces truthful revelation of the relevant information in the unique subgame perfect equilibrium of the game generated by the mechanism.

Maskin and Tirole's critique of the microfoundations of incomplete contracting models that use the observable-but-non-verifiable information assumption is troubling because it implies that the payoffs that are attainable with verifiable variables are also attainable with variables that are only commonly observable. Comparing the effectiveness of second-best institutional arrangements under incomplete contracts is moot when a mechanism exists that is capable of achieving the same payoffs as the best contract with verifiable information. However, the very limited use of implementation mechanisms leads to the question of whether they can indeed costlessly reveal this information and overcome contracting problems via indirect verification.

In this paper, we experimentally explore the performance and adoption of an SPI mechanism described in Maskin & Tirole (1999) that is designed to resolve the hold-up problem in bilateral exchange with observable but non-verifiable ex ante effort. In our experiment, a seller is selling a good to a buyer and may provide costly effort to increase the value of the good. Effort and the value of the good are commonly known to the trading parties, but they are not verifiable by a third-party court. This implies that the two parties cannot write a contract that conditions payments on effort or the value of the good and hence, any effort made by the seller is prone to hold up.

While effort is not verifiable by a third-party court, public announcements can be recorded and used in legal proceedings. Thus, the two parties can in principle write a contract that specifies trade prices as an increasing function of the buyer's announcement of the good's value. If the buyer always announces the true value of the good, then his announcements can be used to set prices that promote efficient effort. One way of doing this is to implement an arbitration mechanism that allows announcements to be challenged by the seller and to punish the buyer any time he is challenged. If the seller challenges only when the buyer has underreported the good's value, then the threat of punishment will ensure truth telling.

The crux of the implementation problem is to give the seller the incentive to challenge only those buyer announcements that are below the value of the good. A key property of the SPI mechanism is that it provides incentives for selfish buyers to tell the truth and for selfish sellers to challenge only in the case that the buyer lied. This is achieved by combining the seller's challenge with an immediate fine

for the buyer and a counteroffer that the buyer will accept only if he lied. If the buyer accepts the counteroffer and thus reveals that he was lying, the mechanism rewards the seller for appropriately challenging the buyer. If, however, the buyer rejects the challenge, the mechanism also fines the seller, and no trade occurs.

Since the value of the good is common knowledge between the buyer and the seller, the seller will only challenge if he knows the buyer will accept the counteroffer (i.e., fails the truth test), because otherwise the seller will be fined. The buyer understands that the seller has the incentive to only challenge lies, and thus he will make a truthful announcement. Truth-telling is therefore part of the unique subgame-perfect Nash equilibrium of the game, and truthful announcements can be used as part of a formal contract.

In our experiment, we constructed the SPI mechanism so that (i) the sellers have an incentive to choose high effort levels and (ii) truth-telling is the unique subgame-perfect equilibrium outcome. Instead, we find that the mechanism does not induce high effort, and buyer lies are prevalent. By construction, the mechanism uses off-equilibrium arbitration clauses that impose large fines for lying and the inappropriate use of arbitration. While arbitration is predicted never to occur in the subgame perfect equilibrium, buyers frequently lie under the mechanism and retaliate against sellers who legitimately use arbitration to challenge buyers' lies. These deviations from the predicted equilibrium lead to the imposition of sizeable fines on both parties. Due to the mechanism's negative effects on parties' pecuniary payoffs, the trading parties opt out of the mechanism in the majority of the cases when given the chance to do so. These results are not just observed in one parametrization of the mechanism. In two additional treatments that implemented different cost and benefit parameters, frequent lies and low efficiency prevail.

Why does the mechanism perform so badly relative to the theoretical predictions? It is often argued that SPI mechanisms are complicated and impose strong rationality requirements in the form of, for example, backward induction or sequential rationality. For this reason, it is thought that SPI mechanisms are likely to fail. Our subjects, however, do well in terms of backward induction: Sellers correctly forecast retaliation against the legitimate use of arbitration and, therefore, only infrequently invoke arbitration. Buyers forecast this reluctance and make lies that are unlikely to be challenged. Finally, sellers correctly forecast these lies when making their investment decisions. These behavioral patterns also prevail when we provide our subjects intense training opportunities, which include a direct description of the incentives the mechanism provides and the opportunity to play against a computer that acts in a payoff-maximizing way. Thus, it is not a lack of rationality that is fundamental to the failure of the mechanism.

Instead, our data suggests that negative reciprocity is the primary force inhibiting efficiency. The intuitive reason for the important role of negative reciprocity is that the mechanism imposes a large fine on a lying buyer if the seller triggers arbitration. Buyers motivated by negative reciprocity therefore retaliate against

sellers who trigger arbitration which — under the rules of the mechanism — imposes large costs on the seller. As a consequence, sellers who anticipate buyers' retaliation are reluctant to trigger arbitration, generating lying incentives for the buyers.

Many laboratory experiments have shown that a substantial share of people seem to be motivated by negative reciprocity (e.g., Blount 1995; Fehr, Gächter, & Kirchsteiger 1997; Offerman 2002; Falk, Fehr, & Fischbacher 2008) and field evidence also points towards the importance of this motive (e.g., Kube, Maréchal, & Puppe 2013; Cohn, Fehr, Hermann, & Schneider 2014). However, theories of social preferences and reciprocity (e.g., Fehr & Schmidt 1999; Falk & Fischbacher 2006; Dufwenberg, Smith, & Van Essen 2011) as well as experimental evidence (e.g., Roth, Prasnikar, Okuno-Fujiwara, & Zamir 1991; Fischbacher, Fong & Fehr 2009; Güth, Marchand, & Rulliére 1998) have shown that such preferences do not automatically become behaviorally relevant in all settings. For example, in some competitive markets, they play little role. Thus, whether negative reciprocity affects behavior depends on the institutional environment. Our empirical results suggest that these preferences play a key role in the Maskin-Tirole mechanism.

Because the empirical evidence strongly points towards the importance of negative reciprocity for SPI mechanisms, we apply (a slightly adapted version of) the Sequential Reciprocity Equilibrium (SRE) concept of Dufwenberg & Kirchsteiger (2004) to our context. We show that if buyers are motivated by reciprocity, they are willing to reject counter offers after small lies, even if they have only weak preferences for reciprocity. However, the rejection of counteroffers triggers a large fine for the seller and, thus, constitutes an unkind act. This raises the question why reciprocal sellers do not retaliate against the expected rejection of counteroffers by challenging buyers' lies. In other words, reciprocal sellers could, in principle, discourage buyers to lie by threatening to challenge lies, even if they know that buyers will reject the subsequent counteroffer. In this way, seller reciprocity could be the remedy for the problems generated by the buyers' negative reciprocity.

However, our theoretical analysis shows that a very large amount of seller reciprocity is required to induce them to challenge buyers' lies, while only a little bit of buyer reciprocity suffices to induce buyers to reject counteroffers. These asymmetric reciprocity requirements are a result of the inherent asymmetry in the timing of the fines in the canonical SPI mechanism that we study. When the seller decides whether to retaliate against the buyer's lie and the anticipated rejection of the counteroffer, she incurs a large fine in case of a challenge. She can avoid paying this fine by refraining from the challenge. In contrast, when the buyer decides whether to reject a counter offer, the fine has already been imposed on him and thus does not count as a part of the cost of rejecting the offer. Retaliation by challenging a lie is thus much more expensive than rejecting a counteroffer, implying that much stronger reciprocity motives are required to challenge a lie compared to rejecting a counteroffer.

We also show theoretically that the sequential structure of fines may lead to

deviations from truth-telling in any canonical SPI mechanism. In particular, we show that for any canonical SPI mechanism that implements, under selfish preferences, 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 1/4 of the time. Thus, negative reciprocity has the potential to impact all canonical SPI mechanisms.

Based on these insights, we developed an alternative mechanism, the Retaliatory Seller (RS) mechanism, that reduces the strong reciprocity requirement for the seller to challenge buyers' lies.[2] The key idea behind the RS mechanism is that at the announcement stage both the buyer and the seller announce the value of the good. If they announce the same value, the game stops and trade occurs at the announced value. If they disagree, the seller is fined and given the option to challenge the buyer. Thus, when the seller decides whether to challenge, the fine is sunk and only a moderate amount of reciprocity suffices to ensure that the seller will challenge a buyer's lie even when she believes with certainty that the buyer will retaliate.

We show that truth-telling is an equilibrium outcome of the RS mechanism for a wider range of reciprocity parameters when using the same pricing rules (the mapping of announcements into trade prices and counteroffers) as our original experiment. We also show generally that for any SPI mechanism and RS mechanism that use the same pricing rules and fines, there always exists a distribution of reciprocity parameters where the RS mechanism has a truth-telling equilibrium while the SPI mechanism does not. In this sense, the RS mechanism is more robust to negative reciprocity than the SPI mechanism.[3]

Finally, we test the new mechanism and find that the RS mechanism outperforms the SPI mechanism, and if we implement the same intense training protocol as for the SPI mechanism, it achieves truthful reports in over 90 percent of the cases, induces high effort in over 90 percent of the cases, and achieves very high levels of aggregate efficiency. However, despite these high performance scores, the RS mechanism does not appear to meet the participation constraint of both parties because it is only adopted in 20 to 60 percent of the cases. Buyers are particularly reluctant to opt into the mechanism. This might be due to the fact that in roughly 5 percent of the cases, the RS mechanism is still associated with disagreements and the payment of large fines. In addition, there is a subset of "trusting" sellers who initially exert high effort even in cases where the mechanism is dismissed. The buyers exploit these sellers, which boosts their average earning in the absence of the mechanism.

[2]We also considered the approach pursued by Bierbrauer & Netzer (2016) and Bierbrauer, Ockenfels, Pollak, & Rückert (2017) who developed a retaliation-robust class of mechanisms that eliminate players' desires or abilities to act on their retaliatory preferences. It turns out, however, that such mechanisms are tantamount to a fixed-price contract in the hold-up setting such that they cannot solve bilateral hold-up problems with cross investments.

[3]We also show that the converse of this statement is not true: there are no psychological environments where truth-telling is an equilibrium of the SPI mechanism but not an equilibrium in the RS mechanism.

Taken together, our findings suggest that reciprocity and other-regarding preferences may cripple proposed mechanisms in many settings and that real-world mechanisms need to be tailored to the underlying behavioral environment. Subgame-perfect implementation mechanisms designed under the assumption that participants are self-interested may perform very poorly and be abandoned by participants. Viable real-world mechanisms must take into consideration the retaliatory inclinations of the people involved and their beliefs about other players' retaliatory propensities.

Apart from speaking to the debate on the micro-foundation of incomplete contracts and the justifiability of the "observable but not verifiable information" assumption, our paper is also related to the theoretical literature on the role of reciprocity in contract design (Cabrales & Charness 2010; Englmaier & Leider 2012; Netzer & Volk 2014), mechanism design (Bierbrauer & Netzer 2016; Bartling & Netzer 2016; Bierbrauer, Ockenfels, Pollak, & Rückert 2017), and implementation (de Clippel, Eliaz, & Knight 2014), as well as to the experimental literature that examines how negative reciprocity affects behavior in settings with a hold up problem (e.g., Dufwenberg, Smith, & Van Essen 2011). The interesting study by de Clippel, Eliaz & Knight (2014), in particular, corroborates the conclusion that reciprocity preferences need to be taken into account in mechanism design. They examine a short-listing mechanism used to select arbitrators and show that the underperformance of this mechanism is consistent with intentions-based reciprocity. We contribute to the literature by showing that the functioning of an important class of SPI mechanisms — ones that have played a prominent role in the debate on the microfoundation of incomplete contracts — is undermined by retaliatory behaviors. We show that a model of reciprocity explains the major regularities of the SPI mechanism and we use the model to develop an alternative mechanism that is predicted to perform well under realistic assumption on the distribution of reciprocity preferences. The new mechanism in fact outperforms the original SPI mechanism and achieves very high levels of truth-telling and efficiency when intense training opportunities prevail.

Our paper also contributes more generally to the experimental literature on implementation.[4] Sefton & Yavas (1996) study extensive-form Abreu-Matsushima mechanisms that vary in the number of stages and find that incentive-compatible

[4]An extensive experimental literature also exists looking at efficiency of implementation mechanisms in the public goods provision problem. Chen & Plott (1996), Chen & Tang (1998), and Healy (2006) study learning dynamics in public good provision mechanisms. Andreoni & Varian (1999) and Falkinger, Fehr, Gächter, & Winter-Ebmer (2000) study two-stage compensation mechanisms that build on work from Moore-Repullo (1988), while Harstad & Marese (1981, 1982), Attiyeh, Franciosi, & Isaac (2000), Arifovic & Ledyard (2004), and Bracht, Figuieres, & Ratto (2008) study the voluntary contribution game, Groves–Ledyard, and Falkinger mechanisms respectively. Masuda, Okano & Saijo (2014) study approval mechanisms and emphasize the need for implementation mechanisms to be robust to multiple reasoning processes and behavioral assumptions. Cabrales, Charness, and Corchón (2003) study Nash implementation in an abstract setting with three-player groups and find that a preference for honesty may play a role. Ponti et al. (2003) study a two-stage mechanism that theoretically solves King Solomon's Dilemma, but this mechanism does not solve the hold-up problem studied here. In addition, none of the above-mentioned papers gives subjects the opportunity to voluntarily select into the mechanism.

mechanisms with 8 and 12 stages perform worse than a mechanism with 4 stages that is not incentive compatible. Katok, Sefton, & Yavas (2002) study both simultaneous and sequential versions of the Abreu-Matsushima mechanism and conclude that individuals use only a limited number of iterations of dominance and steps of backward induction. Based on these papers, we restricted our attention to mechanisms that required only two levels of backward induction. Our paper is also related to the recent experimental work of Aghion, Fehr, Holden, & Wilkening (2018), which tests the theoretical predictions of Aghion, Fudenberg, Holden, Kunimoto, & Tercieux (2012) in an environment where the impact of reciprocity is predicted to be small. The theory paper shows that the absence of common knowledge about the state of nature limits the performance of SPI mechanisms, and the experimental paper confirms this prediction.[5]

I. Subgame-Perfect Implementation

We begin with a description of a simplified version of the Maskin and Tirole argument and highlight how a subgame-perfect implementation mechanism can potentially solve the classic hold-up problem when effort is non-contractible. A seller and buyer bargain over the production and exchange of a good. The seller can choose an effort level e that determines the value of a good that he can costlessly produce and sell to the buyer. Effort costs e to the seller and determines a distribution over the buyer's valuation $v \in \mathcal{V}$, where $\mathcal{V}$ is a finite set of possible buyer valuations. Let e^{FB} be the first-best effort level, which maximizes $E[v|\, e] - e$. Given the buyer's valuation v and the seller's effort e, if trade occurs at price p, the seller receives a payoff of $p - e$, and the buyer receives a payoff of $v - p$.

The good's value to the buyer is *observable* to both parties but *non-verifiable* by a court. To highlight the hold-up problem, assume that after the seller's effort choice has been sunk, the buyer makes a take-it-or-leave-it offer to the seller, resulting in a trade price of $p = 0$. Since the trade price does not depend on the seller's effort choice, the seller has no incentives to choose a costly effort level even if doing so would be socially efficient. Consequently, both parties would prefer a **pricing rule**, $p(v)$, that is more sensitive to the actual value of the good, as such a pricing rule would provide incentives for the seller to choose high effort. Formal contracts written directly on this value cannot be used because the value is non-verifiable. However, Maskin & Tirole (1999) argue that a contract in which the trade price depends on a public message can achieve the first-best outcome if it is augmented with a verification system based on Moore & Repullo (1988). In particular, consider the following class of subgame-perfect-implementation (SPI) mechanisms that is designed to implement a non-decreasing pricing rule $p(v)$:

1) The buyer and seller sign a contract with a third party, whom we will call

[5] Chen, Holden, Kunimoto, Sun, & Wilkening (2020) explore how mechanisms can be made robust to small perturbations in common knowledge when initial rationalizability is used as a solution concept and lotteries are allowed.

the arbitrator. The contract specifies *(i)* an **initial-price schedule** $p(\hat{v})$ at which trade may occur, given an announcement $\hat{v}$ that the buyer makes in stage 3, and *(ii)* a **counter-offer schedule** $\hat{p}(\hat{v})$ and fines F_B and F_S, which may jointly be used to mediate disagreement and will be discussed below. Note that both $p(\cdot)$ and $\hat{p}(\cdot)$ are based only on the buyer's announcement, which can be made publicly observable (and therefore verifiable). The initial price schedule $p(\cdot)$ corresponds to the desired pricing rule if $\hat{v} = v$ for all v.

2) The seller chooses effort e, which determines a distribution over the buyer's valuations $v \in \mathcal{V}$. The realized value v is commonly observed by both the buyer and seller.

3) The buyer announces $\hat{v} \in \mathcal{V}$. The announcement $\hat{v}$ is observable to the seller and the arbitrator.

4) The seller may challenge the announcement. If he does not, trade occurs at price $p(\hat{v})$, and the game ends. If he does, the buyer pays a fine F_B to the arbitrator, and play proceeds.

5) The buyer is given a counter offer $\hat{p}(\hat{v})$. If the buyer accepts the counter offer and buys, he pays $\hat{p}(\hat{v})$ and receives the good, and the seller is paid $F_S \leq F_B$ by the arbitrator.

6) If the buyer rejects the counteroffer, the seller gives the good to the arbitrator, and it is destroyed. Additionally, the seller must also pay a fine F_S to the arbitrator.

A **SPI mechanism**, which we will denote by γ^{SPI}, is therefore a collection $(\hat{p}(\cdot), F_B, F_S)$ consisting of a counter-offer schedule, a buyer fine, and a seller fine, that is designed to implement pricing rule $p(.)$. The logic of this mechanism is that the counter-offer schedule and fines are constructed so that if the buyer and seller are commonly known to be sequentially rational, the buyer never has an incentive to announce a $\hat{v} \neq v$. We will say that SPI mechanism γ^{SPI} **subgame-perfect-equilibrium (SPE)-implements pricing rule** $p(v)$ if under every subgame-perfect equilibrium of the game, trade occurs at price $p(v)$ if v is the buyer's valuation. We will also say that SPI mechanism γ^{SPI} **achieves efficiency** if under every subgame-perfect equilibrium, the seller chooses e^{FB}, and trade always occurs.

In the appendix, we show that the following three conditions are sufficient to ensure that γ^{SPI} SPE-implements $p(\cdot)$:

(a) **Counter-Offer Condition.** The buyer prefers to accept any counter offer $\hat{p}(\hat{v})$ for which he has announced $\hat{v} < v$ and reject any counter offer for which he has announced $\hat{v} \geq v$.

(b) **Appropriate-Challenge Condition.** The seller prefers to challenge announcements $\hat{v} < v$ and not challenge announcements $\hat{v} \geq v$.

(c) **Truth-Telling Condition.** The buyer prefers to announce $\hat{v} = v$ rather than $\hat{v} \neq v$.

We also show that for any increasing and non-negative pricing rule $p(\cdot)$, there always exists a SPI mechanism γ^{SPI} that SPE-implements $p(\cdot)$. This result implies that for any increasing and non-negative pricing rule that motivates the seller to choose an optimal effort level, we can design an SPI mechanism that implements this rule, that is, the parties can trade as if contracts were complete.

II. Experimental Design: The SPI Treatment

In this section, we describe the SPI mechanism we implement experimentally in the **SPI Treatment** and highlight the predicted patterns of play when buyers and sellers have selfish preferences. The SPI treatment uses the SPI mechanism of the class described in Section I and is divided into two phases that vary only in the rules governing the mechanism's adoption.

Phase 1: Phase 1 of the experiment consists of 10 periods. In each period, a seller is perfect-stranger matched with a buyer and the two parties play the following four-stage game:

1) **Effort Stage**: In the effort stage the seller chooses either high or low effort. Low effort costs 30 and generates a good the buyer values at 120. High effort costs 120 and generates a good the buyer values at 260.

2) **The Announcement Stage**: The buyer is informed about the value of the good. The buyer then announces $\hat{v} \in \hat{V} = \{100, 120, \ldots, 260, 280, 300\}$. Note that $\hat{V}$ includes (i) the true value for each potential effort choice, (ii) small lies below each true value, and (iii) generous offers above each true value. We discuss this choice of announcement space in Section II.A.

3) **The Arbitration Stage**: The seller is informed about the buyer's announcement and reminded of the true value. The seller then has the option to "call the arbitrator" or to "not call the arbitrator." We will often refer to the act of "calling the arbitrator" as a challenge.

   a) If the seller chooses to call the arbitrator, the buyer is charged an arbitration fee of $F_B = 250$ and enters the Arbitration Response Stage.

   b) If the seller chooses to not call the arbitrator, the two parties trade at
   $$p(\hat{v}) = 70 + 0.75(\hat{v} - 100).$$
   Note that this price is based on the buyer's original announcement. This price function is shown in column 2 of Table 1.

4) **The Arbitration Response Stage**: If the buyer enters the arbitration stage, he is given a counter offer of $\hat{p}(\hat{v}) = \hat{v} + 5$. This price is again based on the buyer's announcement.

a) If the buyer accepts the counter offer, the seller is given an arbitration reward of $F_S = 250$ and trade occurs at $\hat{p}(\hat{v})$.
b) Otherwise trade does not occur and the seller is also fined $F_S = 250$. Note that the sellers initial production costs are sunk in the effort stage and thus the seller's losses are equal to -280 if the seller chose low effort and -370 if the seller chose high effort.

Phase 2: In periods $11 - 20$, the buyer and seller are again perfect-stranger matched at the start of each period. The buyer and seller are then given the choice to opt in or opt out of the mechanism prior to the seller's effort choice. 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 ten periods. If either party opts out, the game is identical to the game in the first phase, except that the seller may not challenge the buyer's announcement, and trade must occur at price $p(\hat{v})$. Both parties are informed about whether the arbitrator is available but are not informed about the dismissal decision of the other party. This implies that if a subject opts out, he cannot determine whether his counterparty opted in or out.

As seen in Table 1, the mechanism γ^{SPI} satisfies the Counter-Offer, Appropriate-Challenge, and Truth-Telling Conditions described in Section I, and there is a unique subgame-perfect equilibrium, which involves the following predictions:

SPI Hypothesis 1. Along the equilibrium path, the seller chooses high effort, the buyer makes a truthful announcement, and the seller does not challenge. If the seller challenges an announcement of $\hat{v}$, the buyer accepts the counter offer if and only if $\hat{v} < v$.

We refer to the equilibrium-path behavior described in SPI Hypothesis 1 as **efficient truth-telling behavior** and the resulting outcome as the **efficient outcome**. Note that under the efficient outcome, the buyer earns 70 and the seller earns 70. If either party opts out of the mechanism in the second phase, the arbitrator is not available, and the buyer will make the lowest possible announcement, $\hat{v} = 100$, regardless of the true value. The seller has no incentive to choose high effort in this case and will therefore choose low effort. Consequently, the SPNE payoffs if either party opts out are 50 for the buyer and 40 for the seller. As both parties have higher pecuniary payoffs with the mechanism than without it, we have the following prediction:

SPI Hypothesis 2. The buyer and seller opt into the mechanism in periods $11 - 20$.

A. Discussion of Design Features

As the goal of our experiment is to assess the plausibility of using SPI mechanisms in real-world contracting environments, we make a number of design choices

Table 1—Correspondence Between Announcement, Prices, and Outcomes in SPI Treatment

			Low Effort (Value = 120, Effort Cost = 30)			High Effort (Value = 260, Effort Cost = 120)		
Value Announced $\hat{v}$	**Price to Seller** $p(\hat{v})$	**Counter-Price** $\hat{p}(\hat{v})$	**Buyer's Surplus if No Challenge Occurs**	**Seller's Surplus if No Challenge Occurs**	**Buyer's Net Profit of Accepting Counter Offer**	**Buyer's Surplus if No Challenge Occurs**	**Seller's Surplus if No Challenge Occurs**	**Buyer's Net Profit of Accepting Counter Offer**
100	70	105	50	40	**15**	190	−50	**155**
120	85	125	**35**	**55**	−5	175	−35	**135**
140	100	145	20	70	−25	160	−20	**115**
160	115	165	5	85	−45	145	−5	**95**
180	130	185	−10	100	−65	130	10	**75**
200	145	205	−25	115	−85	115	25	**55**
220	160	225	−40	130	−105	100	40	**35**
240	175	245	−55	145	−125	85	55	**15**
260	190	265	−70	160	−145	**70**	**70**	−5
280	205	285	−85	175	−165	55	85	−20
300	220	305	−100	190	−185	40	100	−45

Note: Bolded numbers in the "Buyer's Net Profit of Accepting Counter Offer" Column show announcements for which a selfish buyer would accept the counter offer if challenged. A selfish buyer will make the lowest possible announcement that is not challenged. This will be an announcement of 260 after high effort and 120 after low effort. As these are the true values, this mechanism induces truth telling.

that can be divided into roughly two categories: features that make the mechanism easier to implement experimentally and features that broaden the applicability of the mechanism to richer settings.

To work toward this first objective, we focus on a subset of SPI mechanisms in which the counter-offer schedule is independent of the good's actual value. In more general environments, following the buyer's announcement, the seller chooses a particular counter offer that depends on the buyer's announcement as well as the good's actual value. For example, if the good is worth v, and the buyer announces any value other than v, the seller offers to sell the good to the buyer at a price strictly between the buyer's announcement and v. Additionally, to further reduce the cognitive complexity of the experiment, we assume there are only two effort choices and two possible values for the good.

Our choice of initial-price and counter-offer schedules is intended to encourage truth-telling behavior, under which both players receive an equal payoff of 70.[6] Our expectation is that preferences for equity, for which there is substantial evidence in laboratory experiments, makes such behavior more salient. We also transferred the entire fine F_B to the seller in the case of a successful challenge to maximize the seller's expected value to challenging.

Finally, to ensure that the buyer has strict incentives to adopt the mechanism in the second phase, we give the buyer some of the surplus generated from efficient

[6]The experimental literature on implementation (e.g., Cabrales, Charness, and Corchón 2003; Aghion et. al 2018) and contract theory (e.g., Sanchez-Pages & Vorsatz 2007; Ederer & Fehr 2007) suggest that some individuals have a preference for honesty. In our SPI mechanism, such preferences should reinforce the SPNE since buyers are expected to report truthfully along the equilibrium path.

effort. Absent the mechanism, under the unique SPNE, the seller chooses $e = 30$, and the buyer announces $\hat{v} = 100$, yielding payoffs of 50 to the buyer and 40 to the seller. If the mechanism induces efficient truth-telling behavior, the buyer's gain from adopting it is 20, and the seller's gain is 30.

Moore and Repullo show that in a broad class of environments, any social choice function can be implemented using a three-stage mechanism. In simpler environments, some social choice functions can be implemented using two-stage mechanisms. For example, in our environment, the efficient outcome can be implemented using a two-stage "option contract" (see, for example, Nöldeke & Schmidt (1995)). We deliberately explore the performance of a three-stage mechanism in our simple environment with one-sided hold-up and no uncertainty because if such mechanisms fail to work well in a simple environment, they are even more likely to fail in the more complex environments that necessitate their use.[7]

In the experiment, we restricted the set of possible values of the good to be a strict subset of the announcement space. This restriction simplifies the experiment substantially relative to an experiment with eleven possible values. We view this feature of the experiment as an approximation of a more realistic environment in which no potential values can be completely ruled out in advance. For example, it approximates an environment in which the probability of the value being 120 after low effort and 260 after high effort is equal to $1 - \epsilon$ and the probability of one of the other values is ϵ. It also approximates an environment in which the announcement space is the set of potential values at the time of signing the initial contract and that at some later date some of the values are no longer possible. Such contracts are in the spirit of Maskin & Tirole (1999), which discusses at length the possibility of using SPI mechanism to write contracts that are flexible and that can adapt when new physical contingencies arise that cannot be described ex ante.

Finally, a larger fine slackens the Appropriate-Challenge and Truth-Telling Conditions, and in our SPI Treatment both are satisfied for any fines $F_B > 85$ and $F_S > 85$. According to SPI Hypothesis 1, since a larger fine would also satisfy these conditions, our choice of $F_B = F_S = 250$ should not affect the performance of the mechanism. We deliberately chose a high fine, because one of the key steps in the constructive proofs of SPI mechanisms in the literature is showing that all incentive-compatibility constraints can be satisfied if arbitrarily large fines are allowed.

[7]Hoppe & Schmitz (2011) experimentally study simple single-price option contracts in a one-sided hold-up environment and find promising efficiency improvements even when renegotiation is allowed. Unfortunately, the mechanisms that they consider cannot implement the first-best solution in the environment most commonly used in the incomplete contracts literature where the buyer's investment reduces the seller's cost and the seller's investment increases the buyer's value.

B. *Experimental Protocol*

The experiments were run in the Experimental Economics Laboratory at the University of Melbourne between May and September of 2009 and between November of 2017 and February of 2018. Experiments were conducted using z-Tree (Fischbacher 2007). All 520 subjects participating in the SPI Treatment and follow-up treatments (described in Section IV and V) were undergraduate students at the university and were randomly invited from a pool of more than 5000 volunteers using ORSEE (Greiner 2015).[8] Session sizes varied from 20 to 26. We ran two additional control sessions without the mechanism in 2015 (N = 38). In these control experiments, subjects played 20 periods of our SPI Treatment without the possibility for buyer announcements to be challenged. We use these sessions to estimate average efficiency in the absence of the mechanism.

In sessions run in 2009, subjects participated in a Personal Norms of Reciprocity (PNR) survey developed by Perugini et al. (2003). This survey consisted of 27 questions related to a subject's inclination to punish hostile or reward kind acts. Using principal-components analysis, these questions were combined into orthogonal measures of positive and negative reciprocity for each subject. Subjects earned \$10 for the survey and a \$10 show-up fee, which were used to insulate individuals from bankruptcy. The survey was conducted two weeks prior to the experiment at the point of sign up in order to mitigate demand effects that might occur from running the SPI Treatment and survey during the same session.

Upon arrival to the laboratory, subjects began by playing a lottery game to elicit aversion to gambles that involve the risk of losses. Each subject was presented with the opportunity to participate in six different lotteries, each having the following form:

> Win \$12 with probability 1/2, lose X with probability 1/2. If subjects reject the lottery, they receive \$0.

The six lotteries varied in the amount X that could be lost, where $X \in \{4, 6, 8, 10, 12, 14\}$. One of the six gambles was randomly selected at the end of the experiment and paid.[9] These lotteries enable us to construct a measure of heterogeneity in the willingness to accept actuarially fair gambles. Discussion of the lottery task can be found in Fehr & Goette (2007).

Following the lottery task, subjects were assigned the role of a buyer or a seller, which was fixed for the duration of the experiment. Subjects were then asked to read the instructions and answer a series of practice questions that were checked by the experimenter. These instructions explained the first phase of the experiment (in which the arbitrator is exogenously available) as well as the rules

[8] All data and code for this paper is available in Fehr, Powell & Wilkening (2020).

[9] The lottery treatment was run prior to the experiment to prevent strategic choices by subjects with large losses from the main experiment who might have negative earnings under a subset of the lotteries. The lottery treatment was resolved after the experiment to prevent endowment effects from impacting decisions made in the experiment.

regarding random matching and payment. The instructions were accompanied by a detailed payment chart showing the price and counter offer for each announcement as well as the payment to the buyer and the seller for each potential outcome of the game. The instructions explicitly explained how to read this chart, and subjects were required to work through examples of play with announcements of 180 and 260 to ensure that everyone understood the pecuniary incentives of buyers and sellers after a truthful announcement and a lie. All subjects were required to answer all questions correctly before continuing.

Once the answers of all subjects were checked, the experimenter read aloud a summary of the instructions. The purpose of the summary was to ensure that the main features of the experiment were common knowledge amongst the participants. The oral instructions also explained that there would be a second phase of the experiment and that instructions would be handed out for this phase after the first phase was complete. Subjects were explicitly informed that the second phase would be similar to the first and that their actions in the first phase would have no influence on the rules and potential earnings of the second phase.

To better understand the rationale for subjects' choices, we also elicited buyers' and sellers' beliefs about the other parties' likely actions. For the buyers, we elicited the likelihood that the seller would challenge for each of the possible announcements given the effort level actually chosen by the the seller. These likelihoods were elicited using a 4-point Likert scale (Never/Unlikely/Likely/Always) in each period following the buyer's announcement. Similarly, we asked each seller the likelihood that their challenge would be rejected if they were to challenge the buyer's announcement. This belief was elicited directly after the decision to challenge or not challenge the buyer's announcement.

The choice of unpaid beliefs for our main experiment were based on three considerations. First, we wanted to have a full set of belief information including beliefs about counterfactual actions. In order to elicit these beliefs in an incentive-compatible way, we would have had to use the strategy method for eliciting the seller's challenges and the buyer's acceptance or rejection decision. Given that the solution concept of subgame perfection is such an important part of the implementation mechanism, we were averse to using the strategy method at interior nodes. Second, we felt explaining an additional belief elicitation mechanism would take attention away from the main experiment. Third, in games where both beliefs and action are compensated, risk averse individuals may find it optimal to hedge risk by stating beliefs which differ from their true estimates.[10]

The large fine size in the SPI Treatment opened up the possibility that subjects could go bankrupt. As such, the protocol for bankruptcy was made explicit to all subjects. Subjects began the experiment with a $10 show-up fee and the $10 from the online survey. If a subject accumulated $10 in losses, their money from the online survey payment was liquidated, and they received a warning. If they lost all $20 of their initial endowment, they were removed from the experiment.

[10]See Blanco, Engelmann, Koch, & Normann (2010) for a discussion of hedging.

There were no bankruptcies in the SPI treatment and a total of five bankruptcies in all other treatments. All these subjects were buyers. In these cases, the lab manager took over the terminal and played the SPNE equilibrium path actions. All tests reported in the paper are robust to dropping sessions where there was a bankruptcy.

III. Experimental Results of the SPI Treatment

We describe the results of the SPI Treatment in this section. For purposes of categorizing data, we define $\hat{v} < v(e)$ as a **lie**, $v(e) - 60 \leq \hat{v} < v(e)$ as a **small lie**, $\hat{v} = v(e)$ as a **truthful announcement**, and $\hat{v} > v(e)$ as a **generous announcement**. We define an **appropriate challenge** as a challenge of a lie and an **inappropriate challenge** as a challenge of a truthful announcement or a generous announcement. Note that the terms lying, challenge, and truthful announcement are never used in the experiment.

A. Behavior Under the Mechanism

Under SPI Hypothesis 1, our experimental design generates sharp predictions about the course of play: the seller will always choose high effort, the buyer will always announce the actual value of the good, the seller will challenge if and only if doing so is appropriate, and the buyer will accept counter offers if and only if they result from an appropriate challenge. The data from periods $1 - 10$ of our SPI Treatment provide strikingly little support for SPI Hypothesis 1.

RESULT 1: (*a*) *In a majority of cases buyers make small lies,* (*b*) *the large majority of these lies are not challenged by the sellers,* (*c*) *the buyers reject counter offers in most cases, and* (*d*)*, the mechanism does not induce high effort in many cases. On average,* (*e*) *the parties have higher pecuniary payoffs without the mechanism.*

Figure 1 displays the patterns of play we observed in the first ten periods of the experiment. The left column examines play following low effort ($N = 200$), and the right column examines play following high effort ($N = 260$). Panel (*a*) summarizes the buyers' announcement decisions, Panel (*b*) summarizes the sellers' challenge decisions for different announcements, and Panel (*c*) summarizes the buyers' decisions to accept or reject counter offers. An observation is a dyad-period.

Panel (*a*) shows that in the majority of observations, buyers lied: following high (low) effort, only 37 percent (31 percent) of buyers announce the true value of the good, while 54 percent (61 percent) make small lies. Downward lies are increasingly less frequent the larger they are.

Panel (*b*) shows the proportion of sellers who challenge each announcement $\hat{v}$. SPI Hypothesis 1 predicts that sellers challenge 100 percent of the time after a

lie and never challenge after a truthful or generous announcement. In the data, the challenge probability for small lies is less than 30 percent.

Further, SPI Hypothesis 1 predicts that buyers will accept all counter offers following appropriate challenges and reject all counter offers following inappropriate challenges. Panel (c) shows that in the case of low effort, 21 out of 27 appropriate challenges are rejected; in the case of high effort, 43 out of 52 appropriate challenges are rejected.

Finally, average surplus in periods $1-10$ of the experiment for a buyer and seller pair was only 7.2. To put this number into perspective, average total surplus in periods 1-10 of our control treatment without the mechanism was 97.1, total surplus in the unique SPNE when the mechanism is unavailable is 90, and the total surplus under the efficient outcome is 140. The introduction of the mechanism thus leads to a 93 percent reduction in efficiency relative to the control treatment. This difference is significant (p-value < 0.01) in a comparison of means.[11] Normalizing the actual gain generated by the mechanism by the predicted theoretical gain of the mechanism, the realized gain from the mechanism is $\frac{7.2-90}{140-90} = -166\%$.

While the results in Figure 1 are presented as the aggregate of all 10 periods, there is very little change in the pattern of play when looked at on a period by period basis. In Appendix C1, we show how effort, announcements, and challenges of small lies evolve over the first ten periods. As seen there, the proportion of sellers exerting high effort is relatively stable at roughly 55 percent, the proportion of small lies is stable at roughly 55 percent, and the likelihood of a seller challenging a small lie is decreasing over time. This implies that the mechanism is actually moving away from the truth-telling equilibrium since sellers are becoming more reluctant to challenge over time.

B. The Role of Beliefs

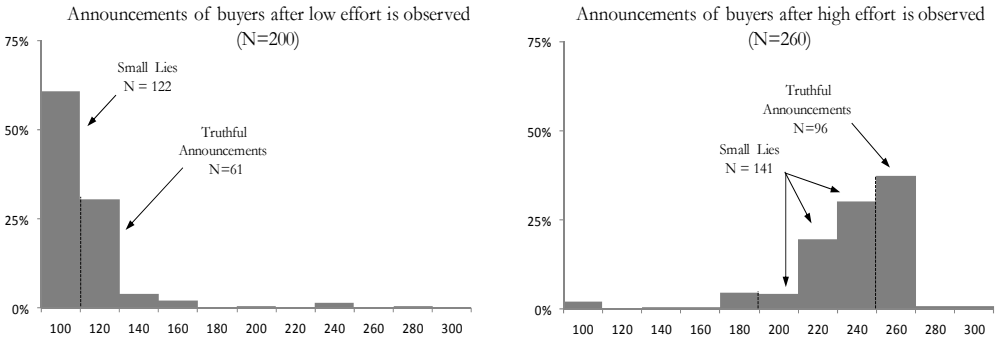
In Appendix B1, we explore the role of subject's beliefs in shaping his or her decision under the mechanism. As shown there, the majority of buyers correctly believe that small lies are unlikely to be challenged or that challenges of small lies will never occur. Similarly, the majority of sellers correctly believe that a challenge of a small lie is unlikely to be accepted or will never be accepted.

Subjects also respond to their beliefs in a consistent manner. Buyers who believe that a small lie is unlikely to be challenged or believe that a small lie will never be challenged are more likely to make a small lie than buyers with other beliefs. Likewise, sellers who believe that a challenge is unlikely to be accepted or will never be accepted are less likely to challenge than sellers with other beliefs.

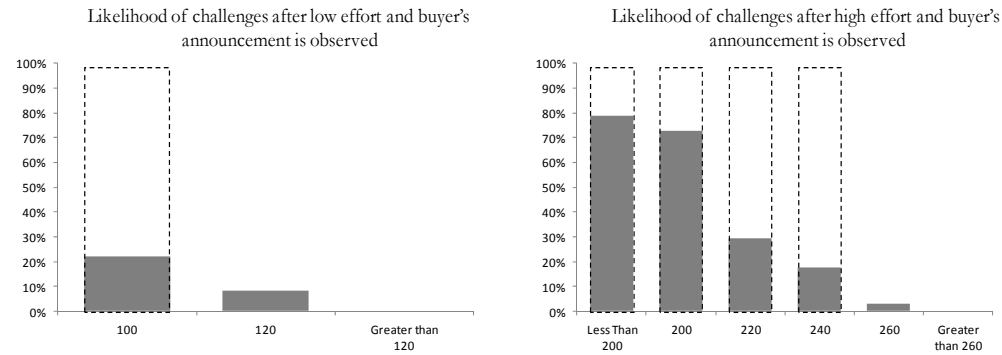
The belief data suggests that individuals are correctly predicting deviations from the SPI predictions in later stages of the game and are responding to these beliefs in a consistent manner. Persistent deviations from the SPI hypothesis and the fact that these deviations were expected by the players suggests that the

[11] All statistical tests in the paper are clustered at the individual level unless otherwise specified.

(a) Distribution of announcements after low and high effort



(b) Likelihood of a challenge after each announcement



(c) Number of challenges accepted and rejected

Number of challenges accepted and rejected
after low effort, given announcement and a seller challenge

Announcement	Challenge Accepted	Challenge Rejected
100	6	21
120	0	5

Grey boxes are predicted action by SPI hypothesis

Number of challenges accepted and rejected
after high effort, given announcement and a seller challenge

Announcement	Challenge Accepted	Challenge Rejected
Less than 200	7	8
200	2	6
220	0	15
240	0	14
260	0	3

Grey boxes are predicted action by SPI hypothesis

FIGURE 1. PATTERN OF PLAY IN FIRST 10 PERIODS OF SPI TREATMENT

model on which our predictions are based may be missing an important force which exerts a systematic influence on beliefs and behavior. We return to this issue after reporting the results from the second phase of the experiment.

C. Selection of the Mechanism

We now examine data from the second phase of the experiment, where subjects were given the option to opt out of the mechanism. SPI Hypothesis 2 predicts all buyers and sellers would opt into the mechanism, since absent the mechanism, sellers would always choose low effort. The results are largely inconsistent with this hypothesis.

RESULT 2: *A majority of dyads opt out of the mechanism. Although the proportion of sellers who choose high effort is greater when the mechanism exists, both buyers and sellers have higher pecuniary payoffs when the mechanism is unavailable than when it is available.*

Panel (a) of Figure 2 shows the opt-out behavior for buyers and sellers over the last 10 periods of the experiment. On average, 65 percent of groups have at least one subject choosing to opt out of the mechanism. While this opt-out rate is decreasing over periods 11-15, the opt-out rate continues to be high, with at least 50 percent of groups opting out of the mechanism in every period. Buyers are much more likely to opt out of the mechanism (as they did in 58 percent of the cases) than sellers are. The latter opt out of the mechanism in only 17 percent of the cases.

In the unique SPNE of the game without the mechanism available, the hold-up problem is predicted to be unresolved: sellers are predicted to choose low effort and buyers are predicted to make the smallest possible announcement. As can be seen on the right hand side of panel (b), these predictions hold true. When either party opts out of the mechanism, 273 out of 298 sellers exert low effort. In 262 of these cases, buyers announces $\hat{v} = 100$. Of the 25 observations where the seller put in high effort, the buyer was truthful in only 3 cases, made a small lie in 7 cases, and made the maximal lie of $\hat{v} = 100$ in 15.

For those periods in which both subjects opted in, we conjectured that the mechanism would perform better than it did in the first phase of the experiment, since opting into the mechanism ought to serve as a positive signal to the other subject in the dyad. From the perspective of effort, this conjecture appears to hold; 114 out of 162 sellers (70 percent) who had access to the mechanism exerted high effort in periods 11-20 whereas high effort was observed in only 260 out of 460 cases (57 percent) in the first 10 periods. This difference is significant (p-value < 0.01) in a probit regression.

However, when the mechanism is kept, buyers still make small lies in 32 out of 48 cases (66 percent) after low effort and in 66 out of 114 cases (57 percent) after high effort. These lying rates are similar to the first 10 periods where the rate of small lies was 61 percent after low effort and 54 percent after high effort. The

rate of small lies in the first 10 periods is not significantly different in either case using a probit regression (low-effort case: p-value = 0.52; high-effort case: p-value = 0.59). Across both effort levels, small lies were challenged in only 13 out of 98 cases (13 percent), a rate that is not significantly different to the challenge rate in periods 8-10 (probit regression: p-value = 0.72).

Empirically, both buyers and sellers earned *lower* average payoffs in periods in which both subjects opted in than in those in which at least one subject opted out: for observations in which the mechanism was available, average total surplus was 55.3 (35.7 for buyers and 19.6 for sellers), while for dyad-periods in which the mechanism was unavailable, average total surplus was 94.2 (57.4 for buyers and 36.8 for sellers). The average efficiency in periods 11-20 of the control treatment (where the mechanism was never available) was 93.4, which is not significantly different from the average efficiency experienced by dyads who dismiss the mechanism (p-value = 0.48) in a comparison of means. However, it is significantly greater than it is for dyads who keep the mechanism (p-value = 0.03).

Given that both buyers and sellers are worse off with the mechanism, an immediate question arises as to why buyers opt out of the mechanism with greater frequency. One likely answer is that the sellers can always avoid potential states of disagreement by exerting low effort and never challenging the buyer. Thus, a seller can always guarantee a payment at least as high as the SPNE of the game without the mechanism with 100% certainty.

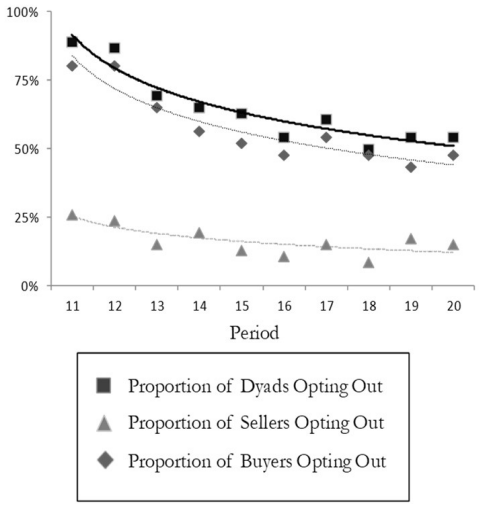
Buyers by contrast must contend with the potential that they will be challenged. Without the mechanism, buyers can guarantee themselves a payoff of 50 by making the lowest possible announcement. With the mechanism, the buyer profit is influenced by (a) the probability that the seller exerts high effort and (b) the probability that the seller will challenge a truthful announcement or a small lie. As both these actions are dependent on the actions of the other player, the mechanism exposes the buyer to uncertainty that he cannot avoid through his choices.[12]

IV. Discussion of SPI-Treatment Results

The data soundly reject SPI Hypotheses 1 and 2. However, the mechanism fails at all behavioral stages in a way that is "internally consistent." If buyers reject counter offers following appropriate challenges of small lies, then sellers have a good reason to shy away from challenging, because it is very costly for them. Yet,

[12]In a previous version of this paper we also reported the results of additional SPI treatments that explored different cost and benefit parameters. In the High-Benefits Treatment we changed the pricing rule such that the buyers had a stronger incentive to tell the truth. In the Low Fine Treatment we reduced the fines but still ensured that all incentive compatibility conditions were met. We hypothesized that a lower fine may reduce the perceived unkindness of a challenge and may thus reduce the buyer's rejection of counteroffers, which may then lead to an increased willingness to challenge among the sellers. Both treatments produced, however, no overall increase in the performance of the SPI mechanism. The results on these mechanism are described in more detail in appendices B2–B4.

(a) Proportion of buyers and sellers opting out of mechanism each period



(b) Buyer and seller outcomes with and without SPI mechanism

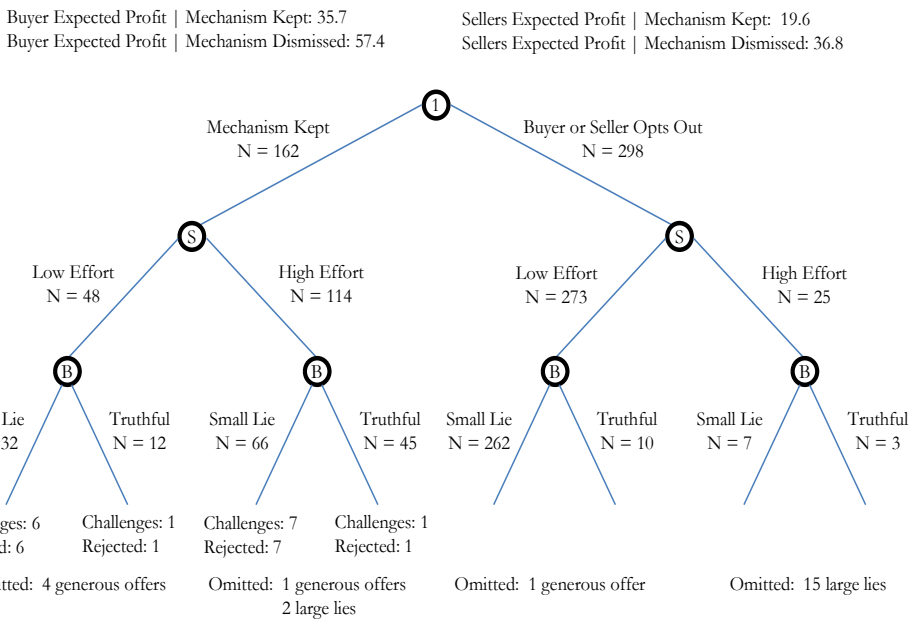


FIGURE 2. BEHAVIOR IN LAST 10 PERIODS (SECOND PHASE) OF SPI TREATMENT

if sellers do not appropriately challenge small lies, then buyers have pecuniary incentives to underreport the value of the good. Indeed, the beliefs data support the above rationale for the failure of the mechanism. Sellers who believe that counter offers following appropriate challenges of small lies will be rejected are significantly less likely to make such challenges. Buyers who believe that they will not be challenged for small lies are considerably more likely to make small lies.

Sellers are also right to believe that buyers will reject counter offers following appropriate challenges of small lies. Although many sellers do not challenge such lies, some do. In these cases, the counter offer is almost always rejected, both parties incur large fines, and no trade takes place. Therefore, the overall pecuniary payoffs generated by the mechanism are negative. On average, parties receive higher pecuniary payoffs trading low quality goods without the mechanism than they receive trading with the mechanism, which explains the observation that the players often do not adopt the mechanism when given the choice.

No matter what their beliefs are, it is payoff maximizing for buyers to accept counter offers in subgames following appropriate challenges of small lies. If buyers acted in their pecuniary interests, they would not reject such counter offers and sellers would not need to fear the high costs of unsuccessful challenges. The mechanism, therefore, would not unravel. Our results indicate that the key to understanding the failure of the mechanism is to understand buyers' willingness to reject counter offers following appropriate challenges of small lies.

A. Do Mistakes Explain the Failure of the SPI Mechanism?

In Appendix B5, we explore whether errors can explain buyer rejection using an Agent Quantal Response Equilibrium (AQRE), which allows subjects to make errors in choosing which pure action to play and that they are more likely to choose pure actions that involve higher expected payoffs. We show there that while the AQRE can match portions of the pattern of play observed, it cannot match the magnitude of rejections. In any QRE model with symmetric noise, a choice that has higher expected utility must be chosen with a higher frequency than one with a lower expected utility. Since accepting an appropriate challenge generates higher returns by construction, the maximum rejection rate that can be predicted is 1/2. Given that 95.5 percent of appropriate challenges were rejected after high effort and a small lie, AQRE on its own has a hard time fully rationalizing the data.[13]

We also conducted a further treatment that introduced an intense training protocol for the purpose of minimizing subjects' mistakes and maximizing their understanding of the logic behind the mechanism. In this **SPI with Intense Training Treatment**, we (i) explicitly explained in the written instructions the

[13]Level-k and other cognitive hierarchy models have a similarly difficult time fitting the extent of rejection by buyers since only type-0 individuals will reject an appropriate challenge.

pecuniary incentives of subjects' counterparties in the trade and (ii) had parties play three unpaid periods and three paid periods against a computerized opponent that was programmed to play the SPNE actions as if they had selfish preferences.

The detailed results of the intense training treatments are described in Appendix B5. Although the intense training protocol caused an improvement in the functioning of the SPI mechanism — sellers choose high effort levels more often and challenged small lies after high effort more frequently — the qualitative results still resemble those previously reported in Section III. In 29 percent of the cases, the buyers underreport the true value of the good. The sellers refrain from challenging small lies in 48 percent of the cases and buyers reject challenges in 58 percent of the cases. Because the mechanism still generates a substantial number of disagreements, the parties are worse off under the mechanism compared to a control treatment without the mechanism. As a consequence, the mechanism was not adopted in the majority of the cases in Phase 2 (i.e., periods $11-20$) of the experiment.

B. The Role of Retaliatory Preferences in the SPI Mechanism

Having ruled out mistakes as the primary explanation for rejections of counteroffers, we now consider whether a preference for retaliation can rationalize the observed behavior. In the SPI mechanism, after the buyer's lie has been challenged, the buyer must immediately pay a fine F_B. The buyer is then presented with two options. He can either buy the good (receiving $v - \hat{p}(\hat{v}) - F_B$) and "reveal" that he has lied, or he can choose not to buy the good (receiving $-F_B$) and "reveal" that he has told the truth. In the former case, the seller receives F_S as a reward and $\hat{p}(\hat{v})$ as compensation for the good. In the latter case, he receives $-F_S$. The private cost to the buyer of choosing the latter is $v - \hat{p}(\hat{v})$, but the cost to the seller is $\hat{p}(\hat{v}) + 2F_S$. If the buyer receives a psychic reward of $\psi_B \lambda_B$ (which we explain below in more detail) for destroying a unit of the seller's payoff as punishment for a perceived unkind act, he will reject the counteroffer if the following condition holds:

$$\psi_B \lambda_B \left[\hat{p}(\hat{v}) + 2F_S\right] \geq v - \hat{p}(\hat{v}).$$

The left-hand side of this inequality measures the buyer's non-pecuniary benefit from rejecting the counter offer and reducing the seller's payoff, while the right-hand side measures the buyer's pecuniary cost of doing so. For small lies, this pecuniary cost can be very small so that only modest preferences for retaliation are necessary to induce the buyer to reject a counter offer after an appropriate challenge.[14]

The non-pecuniary benefit $\psi_B \lambda_B$ in the discussion above was exogenous. However, in Appendix A3, we adapt Dufwenberg and Kirchsteiger's (2004) (hereafter,

[14]For example, a buyer who is challenged after a small lie of 240 must give up only 15 ECU to destroy 745 ECU from the seller. This implies that the buyer must be willing to give up just over \$0.02 to reduce the seller's payoff by \$1.

DK) solution concept, sequential reciprocity equilibrium (hereafter, SRE) to our setting. Following DK, we assume that the buyer and seller have commonly known intensions-based reciprocal preferences. We assume that players care positively about their own pecuniary payoffs and, if they perceive hostility, negatively about the other player's pecuniary payoffs. Player i's actions at each stage are chosen to maximize his pecuniary payoffs, π_i, minus the product of a retaliation factor and player j's pecuniary payoffs: $\pi_i - \psi_i \lambda_i \pi_j$. The retaliation factor $\psi_i \lambda_i$ depends on his retaliatory type ψ_i, which is the strength of his innate preference for negative reciprocity, as well as on how aggrieved he is, λ_i, which captures his perception of the other player's hostility.

We modify the solution concept of DK in two ways. Motivated by the "contracts as reference points" literature, which suggests that individuals form beliefs about their payoffs based on the contract they sign, we use the payoff generated under the efficient outcome as the reference payoff (e.g., in our main experiment, it would be 70 for each player). We believe that this reference point is plausible since both players know what pricing rule the mechanism design is trying to implement and are likely to be aggrieved if they receive a smaller payoff than they would under that pricing rule due to an action of the other party.

By choosing the efficient outcome as the reference point and setting the payoffs of the buyer and seller to be equal on the subgame-perfect-equilibrium path, our experiment leaves little scope for positive reciprocity to influence the outcome of the game. For example, the only direct way for a buyer to be "kind" is to make a generous announcement (i.e., one that is above the true value). Such an action would have no efficiency consequences, as it would only lead to a zero-sum transfer from the buyer to the seller. In our main treatment, such transfers lead to disadvantageous inequity and are never observed.[15] Similarly, sellers also have little scope to be "kind" to the buyer, since a high effort choice is already built into the reference point, and sellers are therefore already "expected" to provide high effort and not challenge a truthful or generous announcement of the buyer. Following the approach of Dufwenberg, Smith, & Van Essen (2011), we therefore restrict our attention to the case where players have only negative reciprocity, and we bound a player's aggrievement level $\lambda_i \in [0, 1]$ at each stage of the game. The upper bound on λ_i normalizes the value of $\psi_i \lambda_i$ so that ψ_i can be interpreted as the amount player i is willing to pay to reduce player j's payoff when he is maximally aggrieved.

Figure 3 characterizes the set of SREs that exist in our main treatment for different retaliatory types of the buyer (ψ_B) and the seller (ψ_S). The figure shows that there are three critical threshold values of the negative reciprocity

[15] As seen in Appendix B2, we do observe some generous offers in the High-Benefits treatment where the buyer receives more of the surplus in equilibrium than the seller. However, in an additional treatment, we find that these generous reports disappear when truthful reports cannot be challenged, suggesting that they are due to a fear of inappropriate challenges rather than altruism or kindness. We also do not observe any evidence of positive reciprocity in our treatments where individuals can opt into the mechanism or in the retaliatory seller mechanism discussed in the next section.

parameters — two for the buyer ($\overline{\psi}_B^{SPI}$ and $\hat{\psi}_B^{SPI}$) and one for the seller ($\overline{\psi}_S^{SPI}$) — that partition the outcome space into three regions that are described in more detail below. The figure is drawn for the specific set of parameters used in the main SPI treatment, but more generally, there always exists Regions I, II and III that are characterized by the three critical threshold values. In particular, for a wide range of parameters, the thresholds satisfy $\overline{\psi}_S^{SPI} > \overline{\psi}_B^{SPI}$ in any γ^{SPI} mechanism that SPE-implements the pricing rule p due to the asymmetric role of fines in the mechanism.

The equilibrium outcomes in the three regions of Figure 3 are characterized as follows:

1) In Region I, truth-telling is not an equilibrium outcome. This region saliently illustrates the asymmetric role of buyer and seller reciprocity because only a small amount of buyer reciprocity ($\psi_B > \overline{\psi}_B^{SPI} = 0.02$) suffices to be in this region unless there is a large amount of seller reciprocity (i.e., $\psi_S > \overline{\psi}_S^{SPI} = 1.27$). In this region, sellers are unwilling to challenge a buyer's lie because they know that the buyer will reject the counteroffer.

2) Truth-telling is the unique equilibrium outcome in Region II, but this requires a large amount of seller reciprocity (i.e., $\psi_S > \overline{\psi}_S^{SPI} = 1.27$) and a limited amount of buyer reciprocity ($\psi_B < \hat{\psi}_B^{SPI} = 0.73$). Intuitively, the asymmetric timing of fines in the SPI mechanism causes the large amount of seller reciprocity necessary to be located in this region; only sellers with a large amount of reciprocity are willing to challenge buyers' lies even when they know buyers will reject the counteroffer. Therefore, in this region the buyers are deterred by seller's reciprocity unless they also have a rather high inclination to reciprocate (i.e., if $\psi_B > \hat{\psi}_B^{SPI} = 0.73$).

3) Finally, truth-telling is part of an equilibrium outcome in some but not all equilibria in Region III, where both players have rather high levels of reciprocity ($\psi_B > \hat{\psi}_B^{SPI} = 0.73$ and $\psi_S > \overline{\psi}_S^{SPI} = 1.27$). In the equilibria involving lies in this region, the seller will challenge the buyer's lie. The reason why lying is nevertheless a part of an SRE outcome is that if buyers also have high retaliatory types, then they may be willing to lie and reject the seller's appropriate challenge because doing so will punish an unkind seller.

Although there are no estimates of the distribution of retaliatory types in our setting, the results from the experimental literature on ultimatum games is consistent with the claim that most sellers will have a $\psi_S < \overline{\psi}_S^{SPI} = 1.27$ and that the majority of buyer–seller dyads are likely to fall into Region I. Translated into an ultimatum-game setting, for example, a responder in the ultimatum game with a retaliatory type of $\overline{\psi}_S^{SPI}$ would reject an offer of 49% of the pie. Such rejections

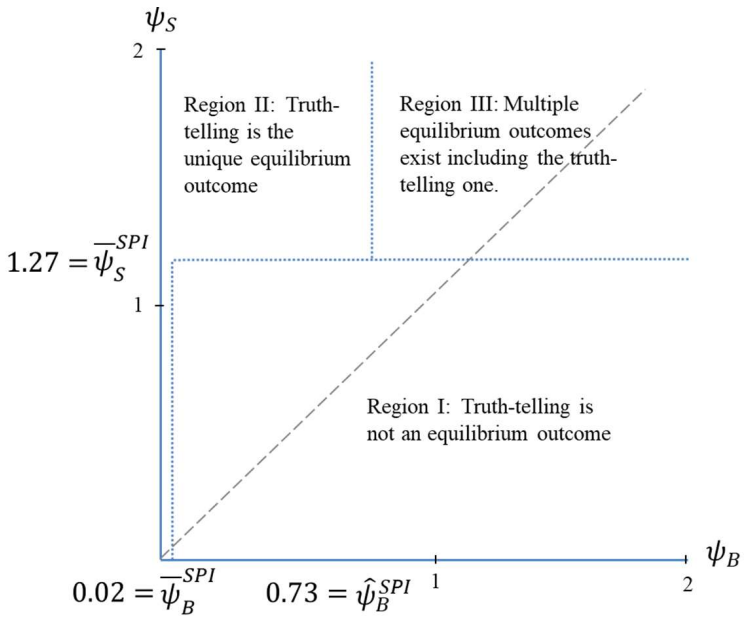


Figure 3. Sequential Reciprocity Equilibrium outcomes in the experimentally implemented SPI mechanism for different buyer (ψ_B) and seller (ψ_S) retaliatory types. The figure is drawn under the assumption that the seller's effort is high, and it considers whether a lie is part of an SRE outcome. Along the forty-five degree line, both parties' retaliatory types are the same.

are extremely rare. This suggests that negative reciprocity can rationalize the main deviations observed in our experiment.[16]

In Appendix A4, we establish a more general result that explores how negative reciprocity impacts SPI mechanisms in general. We consider the set of all mechanisms of the type described in Section I that SPE-implement a non-constant pricing rule p under selfish preferences for a given economic environment. We define a **psychological environment** to be a joint probability distribution over retaliatory types of the buyer and seller, and we assume that players' retaliatory types are independent. The psychological environment is common knowledge, so players agree on the set of feasible retaliatory types as well as on their distribution, and we assume that the realization of a player's retaliatory type is also

[16]Note that if buyers have disadvantageous inequity aversion, they may also reject counter offers that would lead to a large amount of inequity. Thus, in principle, inequity aversion could also explain buyer rejections. However, it does not explain other empirical characteristics of the data. In particular, inequity aversion cannot explain why a fair number of sellers challenge small and moderate size lies, even though they correctly predict that such challenges will be retaliated against. In our experiment, a seller who exerts high effort and ends up with the disagreement payoffs will experience more inequity than if they choose not to challenge a small or moderate lie. Thus, inequity aversion would not lead to challenges by the seller. We have thus concentrated on negative reciprocity, which can rationalize both buyer and seller behavior across all our treatments.

commonly known.

We say that a mechanism γ^{SPI} and pricing rule p are subject to retaliatory implementation failure if it SPE-implements p under selfish preferences, and there exists a psychological environment in which buyer and seller types are drawn from the same distribution and in which, with positive probability, there is no SRE with truth-telling behavior. The following proposition shows SPI mechanisms are subject to retaliatory implementation failure.

Proposition 1. Given an economic environment and a non-constant pricing rule, if γ^{SPI} SPE-implements p, then $\left(\gamma^{SPI}, p\right)$ is subject to retaliatory implementation failure.

The logic behind Proposition 1 is illustrated in Figure 3. As described above, for a wide range of parameters, the thresholds satisfy $\overline{\psi}_B^{SPI} < \overline{\psi}_S^{SPI}$ in any γ^{SPI} mechanism that SPE-implements the pricing rule p due to the asymmetries inherent in the mechanism. In these environments, there exists a point along the diagonal that lies in Region I where truth-telling is not part of any SRE, and therefore $\left(\gamma^{SPI}, p\right)$ is subject to retaliatory implementation failure.

For some economic environments, it may be possible to construct a γ^{SPI} mechanism in which Region I does not occur along the diagonal. When this is the case, it is always possible to construct a psychological environment in which with probability at least 1/4, truth-telling is not an SRE outcome. To do so, consider a distribution in which parties' retaliatory types are drawn independently and are 0 with probability 1/2 and $\psi > \overline{\psi}_B^{SPI}$ with the remaining probability. With probability 1/4, the realization of retaliatory types will be $\psi_B = \psi$ and $\psi_S = 0$, and when this is the case, truth-telling is not a SRE outcome.

In this section, we have assumed that parties' retaliatory types are common knowledge. This assumption allowed us to show that reciprocity, by itself, is sufficient to generate behavior that is consistent with many of our experimental results. In a previous version of our paper, we also considered behavior under the SPI mechanism in a setting in which buyer and seller retaliatory types are drawn from a known distribution but where each player's type is their private information.[17] Incorporating private information in this way allows us to rationalize additional features of our data. In particular, it can help explain why sellers challenge small lies, and counter offers are rejected, even in settings in which parties

[17]The outcomes described in Regions I and II of Figure 3 remain equilibrium outcomes when we relax the assumption that retaliatory types are common knowledge. Doing so requires generalizing the SRE solution concept to allow for private retaliatory types (see Fehr, Powell, & Wilkening (2018) for details). In particular, if retaliatory types are privately known but lie in Region I with probability one, then truth-telling is not an equilibrium outcome. If they lie in the left-most sliver of Region II with probability one (i.e., all buyers have a retaliatory type less than $\overline{\psi}_B^{SPI}$) or they lie in the rest of Region II with probability one (i.e., all sellers have a retaliatory type greater than $\overline{\psi}_S^{SPI}$ and all buyers have a retaliatory type less than $\hat{\psi}_B^{SPI}$), then truth-telling is the unique equilibrium outcome.

typically have moderate retaliatory types.[18] When retaliatory types are private information, a buyer who has a low type and who would accept the counter offer may have an incentive to mimic a high type by lying. Since both low- and high-type buyers lie, the seller may have an incentive to challenge with positive probability. There may therefore exist mixed-strategy equilibria in which (a) buyers regularly tell small lies, (b) sellers occasionally challenge such lies, and (c) buyers frequently retaliate against challenges of small lies. This pattern of play is observed in the main treatment.

V. Towards a Retaliation-Robust Mechanism

One approach to answering the question of whether there is a mechanism that SRE-implements our pricing rule would be to try to develop a truly retaliation-robust class of mechanisms: ones that implement our pricing rule and would do so under any distribution of retaliatory types by eliminating players' desires or abilities to act on their retaliatory preferences. Bierbrauer & Netzer (2016) and Bierbrauer et al. (2017) take this approach in a setting in which players have private information about pecuniary-payoff-relevant states in addition to private information about their retaliatory types. They construct a class of mechanisms under which players cannot unilaterally affect others' pecuniary payoffs, so no player can act on his retaliatory preferences. If such a mechanism implements a social choice function when players do not have preferences for retaliation, then it will do so for any distribution of retaliatory types. Bierbrauer & Netzer (2016) show these mechanisms can partially implement (i.e., do so in some but not necessarily all equilibria) a class of social choice functions that have the "insurance property," that is, they insure the player against others' retaliatory types.

As we discuss in detail in Appendix A7 (Proposition 4), if a social choice function has the insurance property, then any mechanism that implements that social choice function must have two properties. Given any candidate equilibrium of the game induced by that mechanism, it must be the case that (i) a deviation by the buyer cannot impact the payoff of the seller, and (ii) a deviation by the seller cannot impact the payoff of the buyer. Since any action that changes the trade price will impact the payoff of the other party, only constant pricing rules (e.g., a fixed-price contract) satisfy the insurance property in our setting. Such contracts are unable to fully address the hold-up problem in many settings.

However, if constant pricing rules cannot address the hold-up problem, Proposition 1 becomes relevant, i.e., the mechanism is subject to retaliatory implementation failure. This suggests that non-trivial solutions to the hold-up problem will require a priori information on the intensity of negative reciprocity. In other words, it may be possible to mitigate the hold-up problem in many settings only

[18] As illustrated in Figure 3, when retaliatory types are commonly known, such scenarios occur only in Region III, where sellers (buyers) are willing to pay more than \$1.27 (\$0.73) to reduce their counterparty's payoff by \$1. There is substantial empirical evidence that such preferences are rare (Anderson & Putterman 2006; Carpenter 2007; Falk, Fehr & Fischbacher 2005).

if there is a priori information about the intensity of negative reciprocity, and moreover, if it is possible to calibrate a mechanism to this information. Here, we explore one such calibration where we alter our existing mechanism in a way that uses the sellers' retaliatory preferences to our advantage. We propose the following modified mechanism, which we refer to as the retaliatory-seller (RS) mechanism.

Consider the setting described in Section I, and consider the following mechanism:

1) The buyer and seller sign a contract with the arbitrator. The contract specifies (i) an initial price schedule $p(\hat{v}_B)$ at which trade may occur, given an announcement $\hat{v}_B$ the buyer makes in stage 3, (ii) a counter-offer schedule $\hat{p}(\hat{v}_B)$, and a pair of fines F_B and F_S. The initial price schedule corresponds with the pricing rule if $\hat{v}_B = v$.

2) The seller chooses effort e, which determines a distribution over the value of the good $v \in \mathcal{V}$, which is commonly observed by the buyer and seller.

3) The buyer and seller simultaneously announce $\hat{v}_B, \hat{v}_S \in \mathcal{V}$. These announcements are commonly observed by the buyer, the seller, and the arbitrator.

4) If $\hat{v}_B = \hat{v}_S$, then trade occurs at price $p(\hat{v}_B)$, and the game ends. If $\hat{v}_B \neq \hat{v}_S$, then the seller immediately pays a fine F_S and is given the option to challenge the buyer's announcement. If the seller does not challenge, then trade occurs at price $p(\hat{v}_B)$, and the game ends. If the seller challenges, then the buyer pays a fine F_B, and play proceeds.

5) The buyer is given a counter offer $\hat{p}(\hat{v}_B)$. If the buyer accepts the counter offer and buys, he pays $\hat{p}(\hat{v}_B)$ and receives the good, and the seller receives an arbitration reward of F_B by the arbitrator.

6) If the buyer does not buy, the seller gives the good to the arbitrator, and it is destroyed.

A **RS mechanism**, which we will denote by γ^{RS}, is therefore a collection $(\hat{p}(\cdot), F_B, F_S)$ consisting of a counter-offer schedule, a buyer fine, and a seller fine, that is designed to implement pricing rule $p(\cdot)$. The following three conditions are sufficient for the RS mechanism to SPE-implement pricing rule $p(\cdot)$:

(a) **Counter-Offer Condition.** The buyer prefers to accept any counter offer for which he has announced $\hat{v}_B < v$ and reject any counter offer for which he has announced $\hat{v}_B \geq v$.

(b) **Appropriate-Challenge Condition.** If $\hat{v}_B \neq \hat{v}_S$, the seller prefers to challenge announcements $\hat{v}_B < v$ and not challenge announcements $\hat{v}_B \geq v$.

(*c*) **Truth-Telling Condition.** The buyer and seller prefer to announce $\hat{v}_B = \hat{v}_S = v$ rather than to announce any other values.

The first two conditions are similar to the conditions for the SPI mechanism to SPE-implement $p(\cdot)$. As in the SPI mechanism, the counter-offer schedule can be chosen so that the Counter-Offer Condition is satisfied, and the fine F_S can be chosen to satisfy the Appropriate-Challenge Condition. The only condition that differs is the Truth-Telling Condition, which now requires *both* players to announce the true value.

The mechanism is structured so that if Counter-Offer and Appropriate-Challenge Conditions are satisfied, then there is no SPE in which either player announces a value other than v. To see why, note that there is no SPE in which $\hat{v}_B > v$, because then the buyer would prefer to announce $\hat{v}_B = v$, which will not be challenged and would result in a lower price. For a sufficiently high F_B, there is also no SPE in which players do not coordinate their announcements (i.e., $\hat{v}_B \neq \hat{v}_S$) because then the buyer would prefer to deviate by announcing either $\hat{v}_B = \hat{v}_S$, which cannot be challenged, or by announcing $\hat{v}_B = v$, which will not be challenged. And critically, this mechanism does not suffer from the multiple SPE problem: there is no SPE in which players coordinate their announcements on a value other than the true value (i.e., $\hat{v}_B = \hat{v}_S < v$) because then the seller would prefer to announce $\hat{v}_S = v$ and challenge the buyer's announcement.

Having shown that the RS mechanism SPE-implements the pricing rule, we now highlight why it may also SRE-implement that pricing rule. The RS mechanism is similar to the SPI mechanism but restructures the fines so that the seller is fined prior to making his challenge decision. The adjustment of the fine has two effects that are likely to increase challenges. First, being fined is likely to increase the seller's willingness to challenge in cases where the buyer lied and the seller told the truth, since the buyer's action reduces the seller's payoff substantially and will therefore be perceived as unkind. Second, at the time the seller decides to challenge, the seller's fine is sunk in the RS mechanism. In contrast, in the SPI mechanism, whether the seller has to pay a fine depends on the buyer's subsequent action. Therefore the incremental loss associated with challenging and having the counter offer rejected is much lower in the RS mechanism.

In the appendix, we show that reversing the ordering of the fines leads to a larger set of psychological environments for which there exists a truth-telling SRE:

Proposition 2. Given an economic environment and a non-constant pricing rule, if (*i*) γ^{SPI} SPE-implements p, (*ii*) γ^{RS} SPE-implements p, and (*iii*) γ^{SPI} and γ^{RS} use the same counter-offer schedule and fines F_S and F_B, then:

1) There exists a psychological environment in which truth telling is a SRE outcome of the game induced by γ^{RS} but not in the game induced by γ^{SPI}.

2) If truth-telling is a SRE outcome in the game induced by γ^{SPI}, then truth-telling is also a SRE outcome in the game induced by γ^{RS}.

We note that Proposition 2 does not establish a dominance result when it comes to full SRE-implementation (i.e., truth-telling is the outcome for every SRE) because there are psychological environments in which truth-telling is the SRE outcome of every SRE under the SPI mechanism, but there exists a SRE where truth-telling is not the equilibirum outcome under the RS mechanism. This is due to the potential for a buyer and a seller with moderate retaliatory preferences coordinating on a common lie in stage 3 of the RS mechanism. We discuss this issue further in Appendix A6.

A. Testing the Retaliatory Seller Mechanism

Based on the theory discussed above, a RS mechanism can induce truth-telling and high effort for psychological environments where sellers have a moderate level of reciprocity. We test this hypothesis using a "retaliatory seller mechanism" in the **RS Treatment** and the **RS with Intensive-Training Treatment**. In the RS Treatment the standard training protocol was used to make it comparable to our initial SPI Treatment which also used a standard training protocol. In the RS with Intensive-Training Treatment, we used the instensive training protocol where participants play against a computerized opponent prior to Phase I.

To make the treatments as comparable to the original treatments as possible, our *RS* mechanism uses the same price schedule $p(\cdot)$ and counter-offer schedules $\hat{p}(\cdot)$ that we used in the SPI mechanism and was implemented as follows:

1) **Effort Stage**: In the effort stage the seller chooses either high or low effort. Low effort generates a good the buyer values at 120 at a cost of 30. High effort generates a good the buyer values at 260 at a cost of 120.

2) **The Report Stage**: Both parties are informed about the true value of the good. Next, both the buyer and the seller make simultaneous reports about the goods value:

   a) $\hat{v}_S \in \hat{V} = \{100, 120, \ldots, 260, 280, 300\}$

   b) $\hat{v}_B \in \hat{V} = \{100, 120, \ldots, 260, 280, 300\}$

3) **The Verification Stage**: The reports of the buyer and the seller are compared to one another.

   a) If the reports coincide, trade occurs at a price that is based on the agreed upon reports $p(\hat{v}_B) = 70 + 0.75(\hat{v}_B - 100)$.

   b) If the reports do not coincide, the seller is charged a verification fee $F_S = 100$ and enters into the arbitration stage.

4) **The Arbitration Stage**: If the seller enters the arbitration stage, the seller will have the option to <u>continue</u> arbitration or to <u>exit</u> arbitration.

   a) If the seller chooses to continue arbitration, the buyer is charged an arbitration fee of $F_B = 250$ and enters the next stage.

b) If the seller chooses to exit arbitration, the two parties trade at $p(\hat{v}_B) = 70 + 0.75(\hat{v}_B - 100)$.

5) **The Arbitration Response Stage**: If the game enters the arbitration stage, the buyer is given a counter offer that of $\hat{p}(\hat{v}_B) = \hat{v}_B + 5$.

   a) If the buyer accept the counter offer, the seller is given an arbitration reward of F_B and trade occurs at $\hat{p}(\hat{v}_B)$.

   b) Otherwise trade does not occur but the seller still must pay his or her initial production costs.

In comparing the RS Treatment to the SPI Treatment and the RS with Intensive-Training Treatment to the SPI with Intensive-Training Treatment in the first 10 periods where the mechanism was exogenously imposed, we find the following:

RESULT 3: *(a) In phase 1, when the RS mechanism is imposed, the mechanism substantially increases the proportion of sellers who exert high effort and the proportion of truthful reports relative to the SPI mechanism. This relationship holds regardless of the level of training. The RS mechanism with intensive training performs particularly well, with both high effort and truthful reports occurring in roughly 90 percent of cases. (b) In phase 2, when subjects are free to dismiss the mechanism, the RS mechanism also performs significantly better in terms of the share of groups that achieve the efficient outcome and in terms of individual's average earnings under the mechanism. If subjects dismiss the mechanism their average earnings do not differ across treatments.*

Figure 4 compares the proportion of sellers who exert high effort and the proportion of groups where the buyer and the seller were both truthful in the first ten periods (Phase 1) of the four treatments.[19] The 95% confidence interval of each proportion is shown. As can be seen on the left hand side, the seller exerts high effort in 56.5 percent of cases in the SPI Treatment, 69.5 percent of cases in the RS Treatment, 77.5 percent of cases in the SPI with Intensive-Training Treatment, and 91.5 percent of cases in the RS with Intensive-Training Treatment. The difference between the SPI Treatment and the RS Treatment is weakly significant in a simple probit regression where effort choice is regressed on the treatment variable (p-value = 0.06). The difference between the SPI with Intensive-Training Treatment and RS with Intensive-Training Treatment is significant using the same test (p-value = 0.03).

As seen on the right hand side, both the buyer and the seller reported truthfully in 32.4 percent of cases in the SPI Treatment, 52.1 percent of cases in the RS Treatment, 62.8 percent of cases in the SPI with Intensive-Training Treatment, and in 89.1 percent of cases in the RS with Intensive-Training Treatment. Using

[19] In the SPI mechanism, a group is truthful if the buyer announces the true value and the seller does not make an inappropriate challenge. In the RS mechanism, a group is truthful if both the buyer and seller report the true value.

the same probit test described above, the difference between the SPI Treatment and the RS Treatment is significant (p-value < 0.01). Likewise the difference between the SPI with Intensive-Training Treatment and RS with Intensive-Training Treatment is significant (p-value < 0.01).

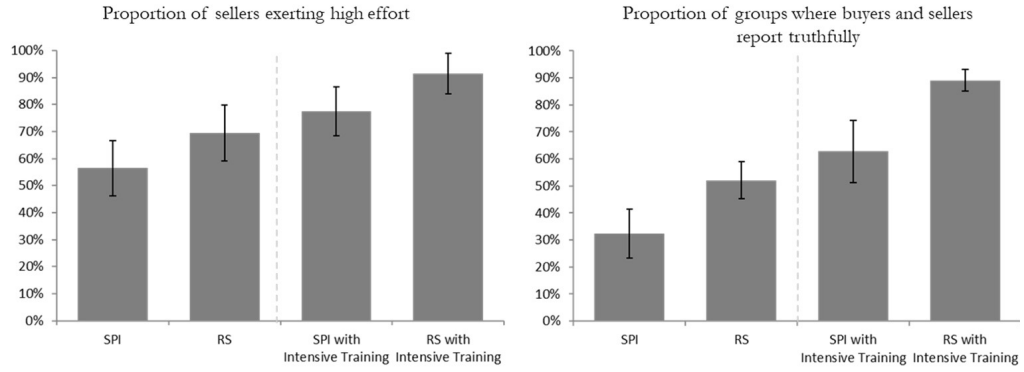


Figure 4. Proportion of sellers exerting high effort and proportion of groups where buyers and sellers report truthfully in periods 1-10. 95 percent confidence intervals shown.

While the RS Treatment has 20 percentage points more truth-telling than the SPI Treatment, it is still lower than one might expect for a mechanism that is predicted to eliminate small lies. In Appendix C2, we graph the distribution of buyer and seller lies separated between cases where the seller exerted high and low effort. As seen there, we find no apparent pattern of small lies and the buyer reports truthfully in 77.9 percent of cases after low effort and in 75.5 percent of cases after high effort. This rate of truth-telling is much higher than those observed by buyers in the SPI Treatment where they told the truth only in 30.5 percent of cases after low effort and 36.9 percent of cases after high effort. However, the seller reports truthfully in only 52.5 percent of cases after low effort and in 75.5 percent of cases after high effort. This rate of truth-telling is much lower than in the SPI Treatment where false challenges by sellers are very rare.

The distribution of reports in the RS Treatment suggests that while the mechanism mitigates the impact of reciprocity on effort provision and small lies, it is more sensitive to mistakes because both the buyer and the seller must make reports. As uncoordinated reports always lead to the seller being fined 100 and also leads to no trade in the majority of cases, the compounded error rate also has a large negative impact on earnings.

In Appendix C2 (Figure C3), we also report the full distribution of reports in the RS with Intensive-Training Treatment. As can be seen there, the additional training eliminates almost all non-truthful reports for buyers and sellers after high effort. In groups where the seller exerts high effort, the buyer reports truthfully

in 93.6 percent of cases and the seller reports truthfully in 98.9 percent of cases. The average earnings in the first 10 periods of the treatment is 109.9. This is significantly higher than the earnings of all other treatments in a pairwise test of means with errors clustered at the buyer level (No mechanism benchmark: p-value $= 0.04$; all other treatments: p-value < 0.01).

Figure 5 reports the proportion of groups that reach the efficient outcome in the first 10 periods (left) and in groups that chose to retain the mechanism in periods 11-20 (right). As can be seen, in the RS with Intensive-Training Treatment, the efficient outcome is achieved in 85 percent of cases in Periods 1-10 and in 91 percent of cases in periods 11-20 where the mechanism was retained. These proportions are significantly greater than in the other treatments using a simple probit regression with a binary variable that is 1 when a group reaches the efficient outcome and 0 otherwise is the dependent variable and this is regressed on the other three treatments (p-value < 0.01 for all treatment-period combinations).

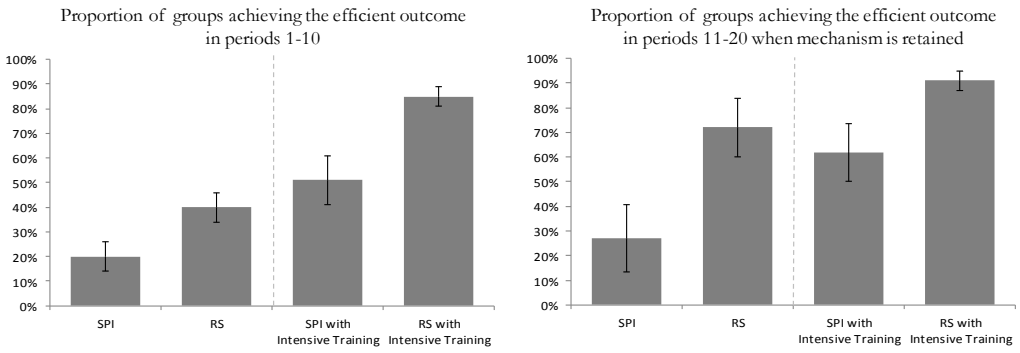


Figure 5. Proportion of groups achieving the efficient outcome. 95 percent confidence intervals shown

Figure 6 reports the average earnings of individual subjects in periods 11-20 for groups that retain the mechanism (left) and for groups that opted out of the mechanism (right). In the RS with Intensive Training treatment, subjects who belonged to a group that retained the mechanism earned 57.7 on average, while subjects who belonged to a group that dismissed the mechanism received 49.0 on average. The difference in average earnings is significant in a simple regression where earnings is regressed on a dummy variable that is one if a group retains the mechanism and zero otherwise (p-value $= .04$). Average earnings in the RS with Intense Training treatment is also significantly higher than average earnings in the SPI with Intense Training for groups that retained the mechanism (p-value $= .02$). Thus, while the RS mechanism does not fully achieve the efficient outcome, it nonetheless improves on efficiency relative to both the SPI mechanism

and the no-mechanism benchmark.

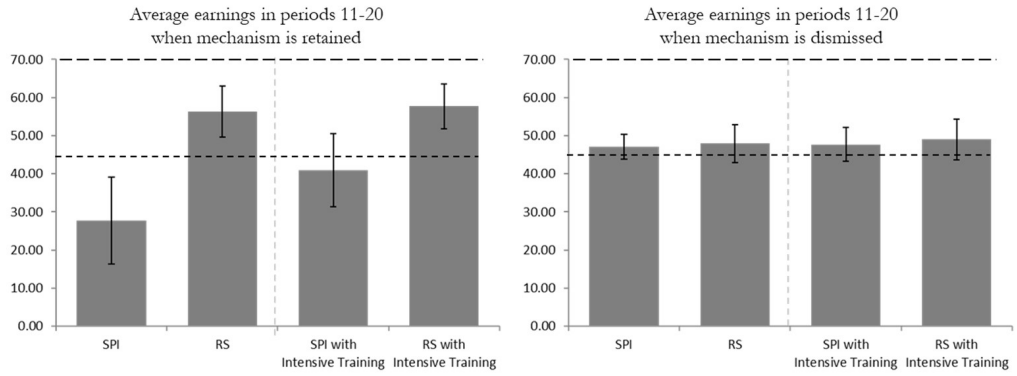


FIGURE 6. AVERAGE EARNINGS OF INDIVIDUAL BUYERS AND SELLERS IN PERIODS 11-20 WHEN THE MECHANISM IS RETAINED AND DISMISSED. THE UPPER DASHED LINE AT 70 SHOWS THE PREDICTED AVERAGE EARNINGS OF A BUYER OR SELLER AT THE EFFICIENT OUTCOME WHERE HIGH EFFORT IS PREDICTED WHILE THE LOWER DASHED LINE AT 45 SHOWS THE AVERAGE EARNINGS OF A BUYER OR SELLER WITHOUT THE MECHANISM WHERE LOW EFFORT IS PREDICTED. 95 PERCENT CONFIDENCE INTERVALS SHOWN.

Looking at the right hand sides of Figure 5 and the left hand side of Figure 6, it is interesting to note that in periods 11-20, the RS Treatment frequently achieves the efficient outcome and has relatively high average earnings.[20] In these groups, truth-telling occurs in 93 percent of cases. This is not significantly different to the truth-telling rate of 95 percent found in the RS with Intensive-Training Treatment suggesting that after some experience, the RS mechanism always performs rather well. In contrast, small lies continue to exist in the SPI with Intensive-Training Treatment and the truth-telling rate is only 74 percent for groups who retain the mechanism in periods 11-20.

Given the high levels of efficiency observed in the RS with Intensive-Training Treatment, one would expect that both parties would be willing to use the mechanism when given the chance to opt-in. However, we find little evidence for this:

RESULT 4: *Despite the very high levels of efficiency observed in the Retaliatory Seller with Intensive-Training Treatment, the proportion of buyers who opt out of the mechanism is still high.*

Figure 7 shows the proportion of sellers (left) and buyers (right) who are willing to opt into the mechanism. As can be seen on the left hand side, sellers opts

[20]Average earnings in the RS treatment is significantly larger in groups where the mechanism is retained relative to groups where it is dismissed using the same specification as above (p-value = .04). Average earnings in the RS treatment is also significantly larger than average earnings in the SPI treatment in groups that retain the mechanism (p-value $< .01$). There is no significant difference in average earnings when groups that retain the mechanism in the RS treatment are compared to groups that retain the mechanism in the RS with Intense Training treatment (p-value = .31).

into the mechanism in 84.3 percent of cases in the RS with Intensive-Training Treatment. This is not significantly different from any of the other treatments. As seen on the right hand side of the figure, buyers opts into the mechanism in 51.8 percent of cases. This opt-in rate is not significantly higher than the opt-in rate observed in the SPI with Intensive-Training Treatment (p-value = 0.29).

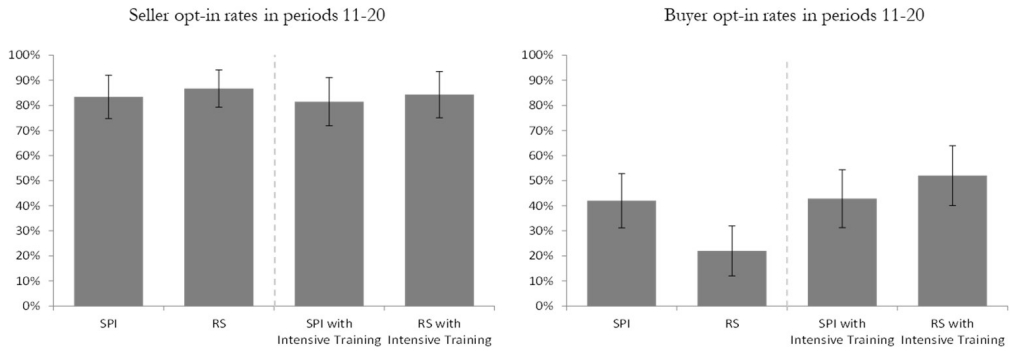


Figure 7. Proportion of sellers and buyers choosing to opt into each mechanism in periods 11-20. 95 percent confidence intervals shown.

The low opt-in rates of buyers appears surprising given that when the arbitrator was retained, sellers chose high effort in 94.8 percent of cases and reports were truthful in 94.5 percent of cases. So perhaps the low acceptance of the mechanism is only a temporary phenomenon. In Appendix C3 we show a figure that illustrates the time path of buyers' and sellers' acceptance of the mechanism in the RS treatment with Intense Training in periods 11-20. The figure shows that there is a slight upwards trend in sellers' acceptance of the mechanism starting from an initial acceptance rate of roughly 82.5 percent in period 11 and ending with an average acceptance rates of 86.0 in the last 5 periods. For buyers we observe a stronger upwards trend from an initial acceptance rate of under 38.5 percent in period 11 and an average acceptance rates of 60.0 in the last 5 periods. This indicates that even during the later periods of the phase buyers have a substantial resistance to the mechanism.

In Appendix C3, we show the time path of buyers' and sellers' earnings in periods 11-20 with and without the mechanism in the RS mechanism with Intense Training. The figure indicates that the sellers were on average better off with the mechanism while the buyers earned roughly the same with and without the mechanism. The reason for this is that (i) a small but non-negligible fraction of trustful sellers provided high effort without the mechanism, and this is often fully exploited by buyers with maximal lies and (ii) in the presence of the mechanism

there was still a small probability of disagreements, which resulted in losses.[21] Thus, by opting out, buyers could eliminate the potential for losses and still had a chance of matching with a trustful seller that could be exploited.

In this environment, where risk-neutral and loss-neutral buyers are basically indifferent between accepting and rejecting the mechanism it takes only a tiny degree of risk or loss aversion to induce buyers to opt out of the mechanism. The potential role of risk/loss aversion is consistent with the fact that individuals who indicated that they are not risk/loss averse in our gambling task — by accepting actuarially fair gambles that involve a 50% chance of a loss — were significantly (p-value $= 0.03$) more likely to participate in the mechanism.

To test whether matching with a trustful seller impacted the buyers likelihood of opting into the mechanism, we calculated the probability of the buyer opting into the mechanism in period t given that the buyer (i) opted into the mechanism in period $t-1$, (ii) opted out and matched with a seller who exerted low effort, and (iii) opted out and matched with a seller who exerted high effort. Buyers who opted into the mechanism in period $t-1$ opted into the mechanism in 90.4 percent of cases while buyers who opted out of the mechanism in period $t-1$ and matched with a seller who put in low effort opted into the mechanism in 18.3 percent of cases. By contrast, buyers who opted out of the mechanism in period $t-1$ and matched with a seller who put in high effort never opted into the mechanism in the next period. The difference in the adoption rate of the mechanism between buyers who match with sellers who put in high effort and buyers who match with sellers who put in low effort is significant in a random effects GLS regression where a buyer's opt-in decision in period t is regressed on a dummy variable that is 1 if the buyer opted out of the mechanism in period $t-1$ and a second dummy variable that is 1 if the buyer opted out of the mechanism in period $t-1$ and the seller nonetheless exerted high effort (p-value < 0.01).

VI. Conclusion

SPI mechanisms have played a key role in the debate over the foundations and the relevance of incomplete-contracting models. If it were indeed possible to make all observable payoff-relevant information verifiable by third parties, the scope for the theory of incomplete contracts would be radically curtailed. In this paper, we examined the performance of SPI mechanisms in the context of a hold-up problem, where they yield complete truth-telling and efficient effort choices if they function as predicted.

In contrast to these predictions, however, we find that under the mechanism, truth-telling occurs in only a minority of the cases. In contrast to the predicted SPE strategies, sellers are often reluctant to challenge the buyers' lies. When they do challenge, the buyers retaliate by rejecting the counter offer. The buyers frequently anticipate the sellers' reluctance to challenge, which makes lying

[21] Sellers exerted high effort in 13 of 28 cases in period 11 when the mechanism was dismissed. In comparison, only 4 of 19 buyers exerted high effort in the first period of the No-Mechanism treatment that we used to benchmark performance in the absence of a mechanism.

worthwhile, and the sellers often anticipate the buyers' retaliatory behavior, which makes refraining from challenging worthwhile. The strong deviations from the predicted SPE are thus not due to failures of backward induction. Instead, they are a rational consequence of buyers' negative reciprocity. Taken together, this pattern of behavior frequently leads to very large monetary losses and, if given the opportunity, the majority of trading pairs opt out of the mechanism.

We show that a slightly modified version of the Sequential Reciprocity Equilibrium (SRE) concept of Dufwenberg & Kirchsteiger (2004) explains the major behavioral patterns. In addition, our theoretical analysis shows that negative reciprocity generally constitutes a fundamental problem for any canonical SPI mechanism because there always exists a distribution of reciprocity preferences such that there is no truth-telling SRE with a positive probability.

A key insight of our theoretical analysis is that a small amount of buyer reciprocity prevents the SPI mechanism from functioning properly, but seller reciprocity could, in principle, restore its truth-telling properties. However, due to the specific timing of the fines in the SPI mechanism, it takes an implausibly large amount of seller reciprocity to achieve this. Based on this insight, we therefore developed an alternative mechanism — the Retaliatory-Seller (RS) mechanism — that reduces the sellers' required reciprocity levels for the existence of truth-telling SRE outcomes.

We also test the new mechanism under our standard training protocol and under an intensive training protocol. Regardless of which protocol we use, the RS mechanism always outperforms the SPI mechanism, and in the RS with Intensive Training Treatment, the new mechanism induces truth-telling by both parties and the efficient outcome in 90 percent of the cases. However, the RS mechanism does not meet the participation constraint of the buyers because they opt into the mechanism only 40-60 percent of the time. This reluctance appears to be due the fact that buyers' expected earnings with the mechanism are not higher than without the mechanism, but in the presence of the mechanism, there was still a small probability of large losses.

We believe that our study provides strong reasons to take reciprocity preferences seriously in mechanism design. Our empirical findings and our theoretical results indicate that reciprocity undermines the functioning of SPI mechanisms. In addition, we have shown that in the hold-up context only fixed price contracts meet the insurance property (i.e, neutralize the impact of reciprocity preferences). Such contracts are however not capable of solving non-trivial hold-up problems. Therefore, mitigating the hold-up problem with the help of mechanisms may in many settings only be possible if a priori information about the intensity of negative reciprocity exists and the mechanisms can be calibrated to this information. We have developed one mechanism that is less vulnerable to reciprocity and show that, with sufficient training opportunities, it performs well in terms of both truth-telling and efficiency. We believe that this shows the high potential value of combining theory and experiments in developing mechanisms that work.

REFERENCES

Acemoglu, Daron, Pol Antras, and Elhanan Helpman. 2007. "Contracts and Technology Adoption." *American Economic Review*, 97(3): 916–943.

Aghion, Philippe, and Patrick Bolton. 1992. "An Incomplete Contracts Approach to Financial Contracting." *The Review of Economic Studies*, 59(3): 473–494.

Aghion, Philippe, Drew Fudenberg, Richard Holden, Takashi Kunimoto, and Olivier Tercieux. 2012. "Subgame-Perfect Implementation Under Value Perturbations." *Quarterly Journal of Economics*, 127(4): 1843–1881.

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." *Journal of the Eurpoean Economic Association*, 16(1): 232–274.

Anderson, Christopher M., and Louis Putterman. 2006. "Do Non-Strategic Sanctions Obey the Law of Demand? The Demand for Punishment in the Voluntary Contribution Mechanism." *Games and Economic Behavior*, 54(1): 1–24.

Andreoni, James, and Hal Varian. 1999. "Pre–Play Contracting in the Prisoners' Dilemma." *Proceedings of the National Acadamy of Science of the United States of America*, 96: 10933–10938.

Antras, Pol. 2003. "Firms, Contracts, and Trade Structure." *Quarterly Journal of Economics*, 118(4): 1375–1418.

Arifovic, Jasmina, and John Ledyard. 2004. "Scaling up Learning Models in Public Good Games." *Journal of Public Economic Theory*, 6(2): 203—-238.

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.

Bartling, Björn, and Nick Netzer. 2016. "An Externality-Robust Auction: Theory and Experimental Evidence." *Games and Economic Behavior*, 97(3): 186–204.

Besley, Timothy, and Maitreesh Ghatak. 2001. "Government Versus Private Ownership of Public Goods." *The Quarterly Journal of Economics*, 116(4): 1343–1372.

Bierbrauer, Felix, and Nick Netzer. 2016. "Mechanism Design and Intentions." *Journal of Economic Theory*, 163(3): 557–603.

Bierbrauer, Felix, Axel Ockenfels, Andreas Pollak, and Désirée Rückert. 2017. "Robust Mechanism Design and Social Preferences." *Journal of Public Economics*, 149(C): 59–80.

Blanco, Mariana, Dirk Engelmann, Alexander K. Koch, and Hans-Theo Normann. 2010. "Belief Elicitation in Experiments: Is There a Hedging Problem?" *Experimental Economics*, 13(4): 412–438.

Blount, Sally. 1995. "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences." *Organizational Behavior and Human Decision Processes*, 63(2): 131–144.

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

Cabrales, Antonio, and Gary Charness. 2010. "Optimal Contracts with Team Production and Hidden Information: An Experiment." *Journal of Economic Behavior & Organization*, 77(2): 163–176.

Cabrales, Antonio, Gary Charness, and Luis C. Corchón. 2003. "An Experiment on Nash Implementation." *Journal of Economic Behavior & Organization*, 51(2): 161–193.

Carpenter, Jeffrey P. 2007. "The Demand for Punishment." *Journal of Economic Behavior and Organization*, 62(4): 522–542.

Chen, Yan, and Charles Plott. 1996. "The Groves–Ledyard Mechanism: An Experimental Study of Institutional Design." *Journal of Public Economics*, 59(3): 335–364.

Chen, Yan, and Fang-Fang Tang. 1998. "Learning and Incentive-Compatible Mechanisms for Public Goods Provision: An Experimental Study." *Journal of Political Economics*, 106(3): 633–662.

Chen, Yi-Chun, Richard Holden, Takashi Kunimoto, Yifei Sun, and Tom Wilkening. 2018. "Getting Dynamic Implementation to Work." accessed on 2018-10-01. Available at: http://tomwilkening.com/.

Cohn, Alain, Ernst Fehr, Benedikt Herrmann, and Frédéric Schneider. 2014. "Social Comparison and Effort Provision: Evidence from a Field Experiment." *Journal of the European Economic Association*, 12(4).

de Clippel, Geoffroy, Kfir Eliaz, and Brian Knight. 2014. "On the Selection of Arbitrators." *American Economic Review*, 104(11): 3434–3458.

Dewatripont, Mathias, and Jean Tirole. 1994. "A Theory of Debt and Equity: Diversity of Securities and Manager-Shareholder Congruence." *The Quarterly Journal of Economics*, 109(4): 1027–54.

Dufwenberg, Martin, Alec Smith, and Matt Van Essen. 2011. "Hold-Up: With a Vengeance." *Economic Inquiry*, 51(1).

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. "A Theory of Sequential Reciprocity." *Games and Economic Behavior*, 47(2): 268–298.

Ederer, Florian, and Ernst Fehr. 2007. "Deception and Incentives: How Dishonesty Undermines Effort Provision." University of Zurich Institute for Empirical Research Working Paper No. 341.

Englmaier, Florian, and Stephen Leider. 2012. "Contractual and Organizational Structure with Reciprocal Agents." *American Economic Journal: Microeconomics*, 4(2): 146–183.

Falk, Armin, and Urs Fischbacher. 2006. "A Theory of Reciprocity." *Games and Economic Behavior*, 54(2): 293–315.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2005. "Driving Forces Behind Informal Sanctions." *Econometrica*, 73(6): 2017–2030.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2008. "Testing Theories of Fairness – Intentions Matter." *Games and Economic Behavior*, 62(1): 287–303.

Falkinger, Josef, Ernst Fehr, Simon Gächter, and Rudolf Winter-Ebrner. 2000. "A Simple Mechanism for the Efficient Provision of Public Goods: Experimental Evidence." *American Economic Review*, 90(1): 247–264.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *The Quarterly Journal of Economics*, 114(3): 817–868.

Fehr, Ernst, and Lorenz Goette. 2007. "Do Workers Work More If Wages Are High? Evidence From a Randomized Field Experiment." *American Economic Review*, 97(1): 298–317.

Fehr, Ernst, Michael Powell, and Tom Wilkening. 2018. "Behavioral Constraints on the Design of Subgame-Perfect Implementation Mechanisms." Available at: http://tomwilkening.com/, accessed on 2018-10-01. Working Paper Version with Perfect Bayesian Retaliation Equilibirum.

Fehr, Ernst, Michael Powell, and Tom Wilkening. 2020. "Data and code for: Behavioral Constraints on the Design of Subgame-Perfect Implementation Mechanisms." https://doi.org/10.3886/E24661V1 *American Economic Review* [publisher], *Inter-university Consortium for Political and Social Research* [distributor].

Fehr, E., S. Gächter, and G. Kirchsteiger. 1997. "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econometrica*, 65(4): 833–860.

Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics*, 10(2): 171–178.

Fischbacher, Urs, Christina M. Fong, and Ernst Fehr. 2000. "Fairness, Errors, and the Power of Competition." *Journal of Economic Behavior & Organization*, 72(1): 527–545.

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

Gromb, Denis. 1993. *Is One Share/One Vote Optimal? Financial Markets Group: LSE Financial Markets Group discussion paper series*, London School of Economics.

Grossman, Sanford J., and Oliver D. Hart. 1986. "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration." *The Journal of Political Economy*, 94(4): 691–719.

Grossman, Sanford J., and Oliver D. Hart. 1988. "One Share-One Vote and the Market for Corporate Control." *Journal of Financial Economics*, 20(1–

2): 175–202.

Güth, Werner, Nadége Marchand, and Jean-Louis Rulliére. 1998. "Équilibration et dépendance du contexte: une évaluation expérimentale du jeu de négociation sous ultimatum." *Revue Économique*, 3(49): 785–94.

Harstad, Ronald M., and Michael Marres. 1982. "Behavioral Explanations of Efficient Public Good Allocations." *Journal of Public Economics*, 19(3): 367–383.

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

Hart, Oliver. 1995. *Firms, Contracts, and Financial Structure.* New York:Oxford University Press, USA.

Hart, Oliver, and John Moore. 1990. "Property Rights and the Nature of the Firm." *Journal of Political Economy*, 98(6): 1119–1158.

Hart, Oliver, and John Moore. 1998. "Default and Renegotiation: A Dynamic Model of Debt." *The Quarterly Journal of Economics*, 113(1): 1–41.

Hart, Oliver, Andrei Shleifer, and Robert W. Vishny. 1997. "The Proper Scope of Government: Theory and an Application to Prisons." *The Quarterly Journal of Economics*, 112(4): 1127–1161.

Healy, Paul J. 2006. "Learning Dynamics for Mechanism Design: An Experimental Comparison of Public Goods Mechanisms." *Journal of Economic Theory*, 129(1): 114–149.

Hoppe, Eva I., and Patrick W. Schmitz. 2011. "Can Contracts Solve the Hold-Up Problem? Experimental Evidence." *Games and Economic Behavior*, 73(1): 186–199.

Katok, Elena, Martin Sefton, and Abdullah Yavaş. 2002. "Implementation by Iterative Dominance and Backward Induction: An Experimental Comparison." *Journal of Economic Theory*, 104(1): 89–103.

Kube, Sebastian, Michel Meréchal, and Clemens Puppe. 2013. "Do Wage Cuts Damage Work Morale? Evidence from a Natural Field Experiment." *Journal of the European Economic Association*, 11(4): 853–870.

Maskin, Eric, and Jean Tirole. 1999. "Unforeseen Contingencies and Incomplete Contracts." *The Review of Economic Studies*, 66(1): 83–114.

Masuda, Takehito, Yoshitaka Okano, and Tatsuyoshi Saijo. 2014. "The Minimum Approval Mechanism Implements the Efficient Public Good Allocation Theoretically and Experimentally." *Games and Economic Behavior*, 83(1): 73–85.

Moore, John, and Raphael Repullo. 1988. "Subgame Perfect Implementation." *Econometrica*, 56(5): 1191–1220.

Netzer, Nick, and André Volk. 2014. "Intentions and Ex-Post Implementation." University of Zurich Mimeo.

Nöldeke, Georg, and Klaus Schmidt. 1995. "Option Contracts and Renegotiation: A Solution to the Hold-Up Problem." *RAND Journal of Economics*, 26(2): 163–179.

Nunn, Nathan. 2007. "Relationship-Specificity, Incomplete Contracts, and the Pattern of Trade." *The Quarterly Journal of Economics*, 122(2): 569–600.

Offerman, Theo. 2002. "Hurting Hurts More Than Helping Helps." *European Economic Review*, 46(8): 1423–1437.

Ponti, Giovanni, Anita Gantner, Dunia López-Pintado, and Robert Mongtgomery. 2003. "Solomon's Dilemma: An Experimental Study on Dynamic Implementation." *Review of Economic Design*, 8(2): 217–239.

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir. 1991. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *American Economic Review*, 81(5): 1068–1095.

Schmidt, Klaus M. 1996*a*. "The Costs and Benefits of Privatization: An Incomplete Contracts Approach." *Journal of Law, Economics, and Organization*, 12(1): 1–24.

Schmidt, Klaus M. 1996*b*. "Incomplete Contracts and Privatization." *European Economic Review*, 40(3–5): 569–579.

Sefton, Martin, and Abdullah Yavaş. 1996. "Abreu-Matsushima Mechanisms: Experimental Evidence." *Games and Economic Behavior*, 16(2): 280–302.

Online Appendix for Behavioral Constraints on the Design of Subgame-Perfect Implementation Mechanisms

Ernst Fehr Michael Powell Tom Wilkening

November 4, 2020

Table of Contents: Appendix

Appendix A: Theory

This appendix has seven sections. Section A1 introduces the key definitions of an economic environment, a pricing rule, and a finite extensive-form mechanism. In Section A2, we then introduce the notion of subgame-perfect equilibrium implementation (SPE-implementation), formally define the set of canonical Moore-Repullo Subgame-Perfect Implementation (SPI) mechanisms, and show that any pricing rule can be SPE-implemented with a SPI mechanism. The proof of this result is constructive and forms the basis for our choice of parameters in our main experiment.

Section A3 introduces the notion of a psychological environment and formally defines our adaptation of Dufwenberg and Kirchsteiger's (2004) sequential reciprocity equilibrium (SRE) concept to our setting. In a SRE, players act at each stage to maximize their own material payoffs minus a scalar times the other player's material payoffs. This scalar is determined by the player's innate retaliatory type as well as on how aggrieved he is at that point. Aggrievement is determined by whether he perceives the other player will act unkindly towards him going forward.

Section A4 applies the notion of a psychological environment and this solution concept to show that for any (non-trivial) pricing rule that can be SPE-implemented with a SPI mechanism, there is a symmetric psychological environment in which there is no SRE in which outcomes always coincide with that pricing rule. This result suggests that implementation mechanisms need to be tailored not only to the economic environment, but also to the underlying psychological environment. We make this argument precise by introducing a notion of SRE implementation.

Section A5 examines the experimental performance of our main mechanism and shows that the key features we see in the data can be understood as outcomes of a SRE. In Section A6, we use this information to construct a new mechanism that addresses what we view as the key weakness of the SPI mechanism: the reluctance of sellers to appropriately challenge false announcements by buyers. We construct a class of mechanisms that we call retaliatory-seller (RS) mechanisms that build off SPI mechanisms but are designed to make sellers aggrieved precisely when they should be challenging the buyer. We show a sense in which the retaliatory-seller mechanism dominates the SPI mechanism and another sense in which it does not. The final section, Section A7, derives implications of Bierbrauer and Netzer's (2016) insurance property for social choice functions in a hold-up setting.

A1. Preliminaries and Definitions

We first introduce several definitions that will be pertinent to our discussion below. An **economic environment** is an array $\mathcal{E} = (\{B, S\}, \mathcal{A}, \mathcal{V}, \pi_B, \pi_S)$ consisting of a set of players $\{B, S\}$, a set of feasible allocations $\mathcal{A}$, where a typical element from $\mathcal{A}$ is a list $a = (q, t_B, t_S)$ consisting of the quantity $q \in \{0, 1\}$ of a good consumed by the buyer, an amount of money $t_B \in \mathbb{R}$ paid by the buyer, and an amount of money $t_S \leq t_B$ received by the seller. The set $\mathcal{V} = \{v_1, \ldots, v_N\} \subset \mathbb{R}$ is a finite set of possible buyer valuations with $v_1 < \cdots < v_N$, and we refer to a typical element $v \in \mathcal{V}$ as a **payoff state**. Players' material payoffs are given by $\pi_B(a) = vq - t_B$ and $\pi_S(a) = t_S$. Finally, we assume that v is common knowledge, and

$v_1 \geq 0$.

A **social choice function** f is a mapping $f : \mathcal{V} \to \{0,1\} \times \mathbb{R} \times \mathbb{R}$ that specifies an allocation for each payoff state. When referring to its constituent parts, we use the notation $f = \left(q^f, t_B^f, t_S^f\right)$. Our analysis will focus on a subset of social choice functions that we call pricing rules. We will refer to a social choice function f as a **pricing rule** if $q^f(v) = 1$ for all $v \in \mathcal{V}$, and $t_B^f(v) = t_S^f(v) \equiv p(v)$ for some nondecreasing, nonnegative function $p(\cdot)$. A pricing rule is summarized completely by p, and we will refer to pricing rule p with the understanding that it corresponds to only a subset of the components of its associated social choice function, since the allocation rule is fixed.

A **finite extensive-form mechanism** (hereafter **mechanism**) is an array $\gamma = (\mathcal{H}, \mathcal{M}_B, \mathcal{M}_S, \mathcal{Z}, g, T)$, which specifies a T-round observable-action extensive-form game with set $\mathcal{H}$ of histories or non-terminal nodes, finite feasible message sets for each player at each non-terminal node, terminal nodes $\mathcal{Z}$, and an outcome function $g : \mathcal{Z} \to \mathcal{A}$ mapping terminal nodes to feasible allocations.

We denote the stage of the mechanism by $t \in \{1, \ldots, T\}$. In stage 1, each player chooses a message m_i^1 from $\mathcal{M}_i^1 = \mathcal{M}_i^1(\emptyset)$. Denote by $\mathcal{M}^1 = \mathcal{M}_B^1 \times \mathcal{M}_S^1$ the set of stage-1 message profiles. In stage t, after observing messages $(m^1, \ldots, m^{t-1})$ chosen in each stage prior to t, each player chooses message $m_i^t \in \mathcal{M}_i^t(m^1, \ldots, m^{t-1})$. Denote $\mathcal{M}^t(m^1, \ldots, m^{t-1}) = \mathcal{M}_B^t(m^1, \ldots, m^{t-1}) \times \mathcal{M}_S^t(m^1, \ldots, m^{t-1})$. A stage-1 history is a vector $h^1 = (v)$, and a stage-t history is a vector $h^t = (v, m^1, \ldots, m^{t-1})$, where $m^1 \in \mathcal{M}^1$, and $m^\tau \in \mathcal{M}^\tau(m^1, \ldots, m^{\tau-1})$. Note that we are assuming that while a history includes the payoff state v, the message set at history h^t cannot differ depending on the realization of v. This is consistent with the assumption that v is nonverifiable. Each terminal node $z = \left(v, m^1, \ldots, m^T\right)$ is associated with a realized message profile $m = \left(m^1, \ldots, m^T\right)$ and, slightly abusing notation, with an outcome $g(m)$ that depends only on the realized message profile.

A2. SPE-Implementable Pricing Rules and SPI Mechanisms

In this section, we will define a class of mechanisms and show that any pricing rule can be implemented with a mechanism from this class. Given a mechanism γ, a strategy profile is a $\sigma = \sigma_B \times \sigma_S$, where σ_i is a mapping from history h^t to a distribution of feasible messages $\mathcal{M}_i^t(h^t)$, where we are slightly abusing notation, since $\mathcal{M}_i^t(h^t)$ depends only on past realized messages and not the payoff state v. Continuation play for player i at history h^t is denoted by $\sigma_i|\, h^t$. The material payoff player i expects to receive, given history h^t, is determined by the distribution over terminal nodes induced by the continuation strategy profile $\sigma|\, h^t$, and we will denote his expected payoff by $\pi_i(\sigma_i|\, h^t, \sigma_j|\, h^t)$.

Let $SPE^\gamma(v)$ be the set of continuation strategy profiles $\sigma^*|\, v$ that form a subgame-perfect equilibrium of the subgame induced by mechanism γ in payoff state v. We will say that a mechanism γ **SPE-implements** pricing rule p if for every $\sigma^*|\, v \in SPE^\gamma(v)$, for any terminal node (v, m^*) reached with positive probability, $f(v) = g(m^*)$. Finally, we will say that pricing rule p is **SPE-implementable** if there exists a mechanism γ that SPE-implements p.

Now consider mechanisms with $T = 3$ that take the following form.

1. The buyer announces $\hat{v} \in \hat{\mathcal{V}}$, where $\mathcal{V} \subset \hat{\mathcal{V}}$ (i.e., $\mathcal{M}_B^1 = \hat{\mathcal{V}}$ and $\mathcal{M}_S^1 = \emptyset$),
2. The seller chooses whether to challenge the announcement ($m_S^2 = C$) or not ($m_S^2 = N$) (i.e., $\mathcal{M}_B^2 = \emptyset$ and $\mathcal{M}_S^2(\hat{v}) = \{C, N\}$). If he does not challenge, the trade occurs at price $p(\hat{v})$, so that $g(m) = (1, p(\hat{v}), p(\hat{v}))$ if $m_2^S = N$.
3. If $m_S^2 = C$, then the buyer pays a fine $F_B \geq 0$ and receives a counter offer: He can choose whether to buy the good at price $\hat{p}(\hat{v})$ ($m_B^3 = Y$) or not ($m_B^3 = N$) (i.e., $\mathcal{M}_B^3(m^1, m^2) = \{Y, N\}$ if $m_S^2 = C$ and $\emptyset$ if $m_S^2 = N$, and $\mathcal{M}_S^3(m^1, m^2) = \emptyset$). If he buys the good, then trade occurs at price $\hat{p}(\hat{v})$, and the seller receives the buyer fine F_B, so that $g(m) = (1, \hat{p}(\hat{v}) + F_B, \hat{p}(\hat{v}) + F_B)$ if $m_B^3 = Y$. If the buyer does not buy the good, then trade does not occur, and the seller also pays a fine F_S, so that $g(m) = (0, F_B, -F_S)$ if $m_B^3 = N$.

We refer to such mechanisms as **canonical Moore-Repullo Subgame-Perfect Implementation (SPI) mechanisms**, and we will denote by Γ^{SPI} the set of such mechanisms. Our first result is that for any pricing rule p, there exists a SPI mechanism $\gamma^{SPI} \in \Gamma^{SPI}$ that SPE-implements p.

Lemma 1 *For any pricing rule p, there is a $\gamma^{SPI} \in \Gamma^{SPI}$ that SPE-implements p.*

Proof of Lemma 1. For this result, it is without loss of generality to set $\hat{\mathcal{V}} = \mathcal{V}$. By construction, the mechanism γ^{SPI} SPE-implements p if and only if, in every subgame-perfect equilibrium, along the equilibrium path, the buyer announces $\hat{v} = v$, and the seller does not challenge. Consider a mechanism γ^{SPI} with the following three properties:

1. $\hat{p}(v_i) \in (v_i, v_{i+1})$ and $\hat{p}(v_N) > v_N$,
2. $\hat{p}(\hat{v}) + F_B - p(\hat{v}) > 0$ for all $\hat{v} \in \mathcal{V}$, and
3. $\hat{p}(v_1) + F_B > p(v_N)$.

We will show that any such mechanism SPE-implements p. In particular, we will show that such a mechanism satisfies the following three conditions, which guarantees that, along the equilibrium path, the buyer announces $\hat{v} = v$, and the seller does not challenge:

1. **Counter-Offer Condition**. The buyer prefers to accept any counter offer for which he has announced $\hat{v} < v$ and to reject any counter offer for which he has announced $\hat{v} \geq v$.
2. **Appropriate-Challenge Condition**. The seller prefers to challenge announcements $\hat{v} < v$ and not challenge announcements $\hat{v} \geq v$.
3. **Truth-Telling Condition**. The buyer prefers to announce $\hat{v} = v$ rather than any $\hat{v} \neq v$.

We refer to a challenge after $\hat{v} < v$ as an **appropriate challenge** and refer to a challenge after $\hat{v} \geq v$ as an **inappropriate challenge**. The counter-offer condition requires that after an appropriate challenge, the counter-offer price is below the value of the good, that is, for each $\hat{v} < v$, $\hat{p}(\hat{v}) < v$. It also requires that after an inappropriate challenge, the counter-offer price is above the value of the good, that is, for each $\hat{v} \geq v$, $\hat{p}(\hat{v}) > v$. These conditions are satisfied, since γ^{SPI} satisfies property (1), so γ^{SPI} satisfies the Counter-Offer Condition.

Next, suppose the seller challenges $\hat{v} \geq v$. Then the buyer will reject the counter offer, and the seller will receive $-F_S$. If the seller does not challenge $\hat{v} \geq v$, then trade will occur at price $p(\hat{v})$, so he prefers not to inappropriately challenge as long as $p(\hat{v}) \geq -F_S$. Similarly, suppose the buyer will accept the counter offer, and the seller will receive $\hat{p}(\hat{v}) + F_B$. If the seller does not challenge $\hat{v} < v$, then trade occurs at price $p(\hat{v})$, so he prefers to appropriately challenge announcement $\hat{v}$ if

$$\hat{p}(\hat{v}) + F_B - p(\hat{v}) > 0$$

for all $\hat{v} < v$, which is satisfied, since γ^{SPI} satisfies property (2). The mechanism γ^{SPI} therefore satisfies the Appropriate-Challenge Condition.

Finally, for the Truth-Telling Condition to be satisfied, the buyer must prefer to announce $\hat{v} = v$ over any other value. If $p(\hat{v})$ is strictly increasing in $\hat{v}$, then overreported values $\hat{v} > v$ will not be challenged but are never optimal for the buyer. If the buyer announces $\hat{v} = v$, he will not be challenged, and he will receive $v - p(v)$. If the buyer announces $\hat{v} < v$, he will be challenged, he will accept the counter offer, and he will receive $v - \hat{p}(\hat{v}) - F_B$. He therefore prefers to announce $\hat{v} = v$ relative to any $\hat{v} < v$ if

$$\hat{p}(\hat{v}) + F_B - p(v) > 0$$

for all $v, \hat{v} \in \mathcal{V}$. Since $\hat{p}(\hat{v})$ and $p(v)$ are increasing in $\hat{v}$ and v, respectively, these inequalities are implied by property (3), so γ^{SPI} satisfies the Truth-Telling Condition. It therefore SPE-implements p. ■

A3. Psychological Environments and Sequential Reciprocity Equilibrium

This section shows how to incorporate retaliatory preferences into the model by augmenting an economic environment with a psychological environment. We will first introduce a couple definitions and then show how to adapt Dufwenberg and Kirchsteiger's (2004) sequential reciprocity equilibrium concept to our setting.

Define a **psychological environment** to be a pair $\mathcal{P} = (\Psi, \mu)$, where $\Psi = \Psi_B \times \Psi_S$ is the set of feasible **retaliatory types**, with typical element (ψ_B, ψ_S), where ψ_B and ψ_S are the buyer's and seller's retaliatory types. The object μ is a joint probability distribution over retaliatory types, and we will assume players' retaliatory types are independent. The psychological environment is common knowledge, as is the realization of players' retaliatory types. An **environment** is a pair $(\mathcal{E}, \mathcal{P})$ consisting of an economic environment and a psychological environment.

Given an environment and a mechanism γ, define histories $h^1 = (v, \psi) \in \mathcal{H}^1$, and $h^t =$

$(v, \psi, m^1, \ldots, m^{t-1}) \in \mathcal{H}^t$, and denote the set of all histories by $\mathcal{H} = \cup_{t=1}^{T} \mathcal{H}^t$ with typical element h. A strategy profile is a $\sigma = \sigma_B \times \sigma_S$, where σ_i is a mapping from h^t to a distribution over player i's feasible messages $\mathcal{M}_i^t(h^t)$ under mechanism γ. Continuation play at h^t is denoted by $\sigma_i|\,\phi_i(h^t)$.

Now that we have defined strategy profiles, we can define players' payoffs. First, to define their expected material payoffs at a specific history $\tilde{h}^t$, suppose player i conjectures player j's strategy to be σ_j^b, where the superscript b denotes player i's first-order belief. At history h^t, his expected material payoffs are therefore $\Pi_i\left(\sigma_i, \sigma_j^b, h^t\right) \equiv \pi_i\left(\sigma_i|\,h^t, \sigma_j^b|\,h^t\right)$.

Player i's expected utility at history h^t is given by the sum of his expected material payoffs and his retaliatory payoffs, which we will now define. Player i's retaliatory payoffs at history h^t have three components: They depend on his retaliatory type ψ_i, his belief about j's expected material payoffs, as well as his aggrievement λ_i. His aggrievement in turn depends on his perception of j's unkindness relative to a reference utility.

To think about j's unkindness, note that j will have some conjecture about what i is going to do going forward. We will say that player j is **acting unkindly** if he knowingly acts in a way that will reduce player i's payoff below a reference payoff. Player i's *perception* of j's unkindness therefore depends on his belief about j's strategy, σ_j^b, as well as his belief about j's belief about his own strategy, which we will denote by σ_i^{bb}, where the superscript bb denotes i's second-order beliefs. Given σ_j^b and σ_i^{bb}, player i's **aggrievement** at history h^t has several components. We will first describe each component, and then we will give the full expression.

First, at history h^t, player i believes player j intends to deliver him an expected payoff of $\pi_i\left(\sigma_i^{bb}|\,h^t, \sigma_j^b|\,h^t\right)$. Next, player i's perception of j's unkindness depends not just on the payoff he perceives j intends to deliver him, but also on what the payoff is relative to a reference payoff. The reference payoff we will use in our adaptation of sequential reciprocity equilibrium will be the payoff player i expects to receive under the pricing rule p in payoff state v: $\pi_i(f(v))$. Our choice of reference point is motivated by the contracts as reference points literature, which suggests that individuals form beliefs about their payoffs that depend on the contract signed. In our setting, players know what pricing rule the mechanism designer is trying to implement, and so we think it is plausible to assume they will be aggrieved if they receive a smaller payoff than they would under that pricing rule.

Finally, we want to normalize player i's aggrievement so that it is between 0 and 1, so that ψ_i can be interpreted as player i's maximum willingness to pay to destroy one unit of player j's material payoff. Given these ingredients, define player i's **aggrievement** at history h^t by

$$\lambda_i\left(\sigma_j^b, \sigma_i^{bb}, h\right) = \min\left\{\frac{\pi_i(f(v)) - \pi_i\left(\sigma_i^{bb}|\,h^t, \sigma_j^b|\,h^t\right)}{\pi_i(f(v)) - \min_{\tilde{\sigma}_j^b|_{h^t}} \pi_i\left(\sigma_i^{bb}|\,h^t, \tilde{\sigma}_j^b|\,h^t\right)}, 0\right\}.$$

It is important to note that, while i's aggrievement depends on his first-order and second-order beliefs, it does not depend directly on his continuation strategy.

Our notion of aggrievement captures the intensity with which player i will act on his retaliatory preferences. We will assume that player i's choices at history h^t are made to maximize his expected material payoff minus the scalar $\psi_i\lambda_i$ times player j's expected ma-

terial payoff, under the assumption that in future rounds, he will continue to play according to the strategy σ_i. That is, he chooses a strategy $\tilde{\sigma}_i|\, h^t$ consisting of a round-t message $\tilde{m}_i^t \in \mathcal{M}_i^t\left(h^t\right)$ followed by $\sigma_i|\, \tilde{h}^{t+1}$, where $\tilde{h}^{t+1} = h^t\tilde{m}^t$ is the concatenation of history h^t with the realization of round-t messages $\tilde{m}^t$, that maximizes

$$\max_{\tilde{\sigma}_i|h^t} U_i\left(\tilde{\sigma}_i, \sigma_j^b, \sigma_i^{bb}, h^t\right) \equiv \max_{\tilde{\sigma}_i|h^t} \Pi_i\left(\tilde{\sigma}_i, \sigma_j^b, h^t\right) - \psi_i\lambda_i\left(\sigma_j^b, \sigma_i^{bb}, h^t\right)\Pi_j\left(\sigma_j^b, \tilde{\sigma}_i, h^t\right).$$

We now define our solution concept.

Definition 1 *A **sequential reciprocity equilibrium** (**SRE**) is a strategy profile σ^* such that for every history $h \in \mathcal{H}$ and player $i \in \{B, S\}$,*

$$\sigma_i^*|h \in \arg\max_{\tilde{\sigma}_i|h} U_i\left(\tilde{\sigma}_i, \sigma_j^*, \sigma_i^*, h\right).$$

Checking whether a strategy profile σ^* is a SRE is conceptually straightforward, albeit tedious. Conceptually, σ^* fully determines the **aggrievement profile** $\lambda^*(\cdot)$, which determines each player's aggrievement at each history h^t. At each history, each player i acts as a "short-run player i" who chooses message $\tilde{m}_i^t$ to maximize his utility, which is given by $\Pi_i - \psi_i, \lambda_i^*\Pi_j$, given that his future self will play according to σ_i^* and given that the other player plays according to σ_j^*. The strategy profile is part of a SRE if at each history, each m_i^{*t} in the support of $\sigma_i^*\left(h^t\right)$ maximizes this utility.

A4. Retaliatory Implementation Failure and SRE Implementation

This section defines the notion of implementation failure under SRE and shows that the SPI mechanisms defined in Appendix A2 are prone to implementation failure. It also defines the notions of ful and partial implementation under the SRE equilibrium concept.

Given an economic environment $\mathcal{E}$ and a **non-constant pricing rule** p such that $p(\cdot)$ is not constant on $\mathcal{V}$, suppose a mechanism γ SPE-implements p. Say that a psychological environment $\mathcal{P}$ is a **symmetric psychological environment** if buyer and seller retaliatory types are identically distributed under $\mathcal{P}$. We will say that the pair (γ, p) is subject to **retaliatory implementation failure** if there exists a symmetric psychological environment $\mathcal{P}$ in which in every SRE, with positive probability, the buyer announces some $\hat{v} \neq v$ for some v. The following proposition shows SPI mechanisms are subject to retaliatory implementation failure.

Proposition 1 *Given an economic environment and a non-constant pricing rule, if γ^{SPI} SPE-implements p, then $\left(\gamma^{SPI}, p\right)$ is subject to retaliatory implementation failure.*

Proof of Proposition 1. Since p is a non-constant pricing rule, there exists $v, v' \in \mathcal{V}$ such that $p(v) < p(v')$. Since γ^{SPI} SPE-implements p, it must have the property that in any SPE, if the payoff state is v', a buyer announcement of $\hat{v} = v < v'$ is challenged by the seller

with positive probability, or else the buyer would prefer to announce $\hat{v}$, and γ^{SPI} would not implement p.

Given v, v', define the following two cut-off values:

$$\begin{aligned}\overline{\psi}_B^{SPI}(v,v') &= \frac{v-\hat{p}(v')}{\hat{p}(v')+F_B+F_S}\\ \overline{\psi}_S^{SPI}(v,v') &= \frac{p(v')+F_S}{v-p(v')+F_B}.\end{aligned}$$

The first object is a critical value of buyer retaliatory preferences above which the buyer will retaliate against a challenge of announcement $\hat{v}=v'$ in payoff state v in every SRE. Note that the numerator is the change in his monetary payoff from rejecting a counter offer, and the denominator is the resulting change in the seller's monetary payoffs if the buyer rejects a counter offer, and at the history at which the buyer has been challenged, $\lambda_B^*=1$ in every SRE.

The second object is a critical value of seller retaliatory preferences above which the seller will challenge an announcement of $\hat{v}=v'$ in payoff state v in every SRE, even if she knows the buyer will retaliate against her challenge with probability one. Note that the numerator is the change in the seller's monetary payoffs from challenging an announcement of v' given that the buyer will retaliate. The denominator is the resulting change in the buyer's monetary payoffs if the seller challenges announcement v' in payoff state v.

Suppose $\mathcal{P}$ is such that, with positive probability, there is a realization (ψ_B,ψ_S) of retaliatory types that satisfies $\psi_B > \overline{\psi}_B^{SPI}(v,v')$ and $\psi_S < \overline{\psi}_S^{SPI}(v,v')$. Given this realization of retaliatory types, in payoff state v, the buyer will retaliate against a challenge of $\hat{v}=v'$, the seller will not challenge such an announcement, and so the buyer will announce $\hat{v}=v'\neq v$ in state v.

For any γ^{SPI}, such a $\mathcal{P}$ exists. To see why, there are two relevant cases. First, suppose $\overline{\psi}_B^{SPI}(v,v') < \overline{\psi}_S^{SPI}(v,v')$ for some v,v'. Then let $\mathcal{P}$ be such that $\psi_B=\psi_S=\psi$ with probability one, where $\overline{\psi}_B^{SPI}(v,v') < \psi < \overline{\psi}_S^{SPI}(v,v')$. Second, suppose that for all v,v', $\overline{\psi}_B^{SPI}(v,v') > \overline{\psi}_S^{SPI}(v,v')$. Fix v,v', and let $\mathcal{P}$ be such that $\psi_B,\psi_S\in\{0,\psi\}$, with $\Pr[\psi_B=\psi]=\Pr[\psi_S=\psi]=1/2$, where $\psi > \overline{\psi}_B^{SPI}(v,v')$. Then with probability $1/4$, $\psi_B=\psi$ and $\psi_S=0$, and so we have $\psi_B > \overline{\psi}_B^{SPI}(v,v')$ and $\psi_S < \overline{\psi}_S^{SPI}(v,v')$, in which case the buyer will announce $\hat{v}=v'$ in payoff state v. ■

The proof of Proposition 1 shows that for any $\gamma^{SPI}\in\Gamma^{SPI}$ that SPE-implements a non-constant pricing rule p, if the buyer's retaliatory type is sufficiently high, and the seller's retaliatory type is sufficiently low, then there exists a payoff state in which the buyer lies, and the seller never challenges that lie. This implies that there exists a symmetric psychological environment in which for some realizations of retaliatory types, and in some payoff states, the buyer does not announce the truth in any SRE.

Proposition 1 is a somewhat negative result for SPI mechanisms, but it naturally raises the question of whether there are other mechanisms that implement a given pricing rule when players have retaliatory preferences. To make this question precise, we will define

what it means for a mechanism to implement a pricing rule when players have retaliatory preferences.

Given an environment $(\mathcal{E}, \mathcal{P})$ and a mechanism γ, let SRE^{γ} be the set of SRE strategy profiles σ^* under mechanism γ, and let $SRE^{\gamma}(v, \psi)$ be the set of associated continuation strategies $\sigma^*|(v, \psi) \equiv (\sigma_B^*|(v, \psi), \sigma_S^*|(v, \psi))$ given payoff state v and retaliatory types ψ. We will say that a mechanism γ **SRE-implements** a pricing rule p if, *for every* $\sigma^*|(v, \psi) \in SRE^{\gamma}(v, \psi)$, for any terminal node (v, ψ, m^*) reached with positive probability under $\sigma^*|(v, \psi)$, $f(v) = g(m^*)$. Additionally, we will say that a mechanism γ **SRE-partially implements** a pricing rule p if *there exists* a $\sigma^*|(v, \psi) \in SRE^{\gamma}(v, \psi)$ in which, for any terminal node (v, ψ, m^*) reached with positive probability under $\sigma^*|(v, \psi)$, $f(v) = g(m^*)$. Finally, we will say that a pricing rule p is **SRE-implementable** if there exists a mechanism γ that SRE-implements p, and we will say that p is **SRE-partially implementable** if there exists a mechanism γ that SRE-partially implements p. These definitions imply that in a psychological environment with $\Psi = \{(0, 0)\}$, a pricing rule p is SRE-implementable if and only if p is SPE-implementable.

Given pricing rule p, the fact that γ^{SPI} mechanisms are subject to retaliatory implementation failure does not imply that for a given psychological environment, p is not SRE-implementable. Rather, it suggests that mechanisms that implement p in one environment need not implement p in another psychological environment, holding fixed the economic environment. And as a practical matter, it suggests that mechanisms should be tailored to the psychological environment if there is to be any hope of implementing a particular pricing rule. We take this lesson, coupled with the results from our main experiment, as the motivation for our re-design in Section A6.

We conclude this section with a brief comment on SRE-implementation. First, it is an open and important question whether there are any classes of mechanisms Γ for which (a) any pricing rule p can be SPE-implemented with a mechanism $\gamma \in \Gamma$, and (b) for any mechanism γ that SPE-implements p, γ SRE-implements p in every psychological environment $\mathcal{P}$. In other words, are there any truly retaliation-robust SPE-implementation mechanisms? The analysis of Bierbrauer and Netzer (2016) suggests an affirmative answer to a narrower version of this question. In particular, it suggests that there is a class of pricing rules p for which one can construct mechanisms that SRE-partially implement them and in which players have no ability to act on their retaliatory preferences. We show, however, in Section A7 that the conditions on p required for such a result preclude pricing rules that motivate important kinds of bilateral relationship-specific investments that more general pricing rules can motivate.

A5. Diagnosing the Failure of the SPI Mechanism

Our experimental and survey results suggest several important features of subject behavior under the mechanism: (1) buyers retaliate against appropriate challenges with very high probability, (2) sellers do not always challenge small lies, and (3) buyers regularly tell small lies. In this section, we will show that these features are consistent with SRE. We discuss at the end of this section how incorporating private information about retaliatory types into our analysis can help explain additional findings, but we refer the interested reader to Fehr, Powell, and Wilkening (2018) for the details.

We will consider the environment from our main experiment and describe the outcomes that are consisent with SRE when the value of the good is $v = 260$. Recall that the initial-price schedule as a function of the buyer's announcement is $p(\hat{v}) = 70 + 0.75(\hat{v} - 100)$, the counter-offer schedule is $\hat{p}(\hat{v}) = \hat{v} + 5$, the fines are set at $F_B = F_S = 250$, and the set of possible announcements is $\{100, \ldots, 300\}$. Define the following three cutoffs:

$$\begin{aligned}
\overline{\psi}_B^{SPI} &= \frac{260 - \hat{p}(240)}{\hat{p}(240) + F_B + F_S} = \frac{260 - 245}{245 + 250 + 250} = \frac{3}{149} \\
\hat{\psi}_B^{SPI} &= \frac{260 - p(260) + F_B}{p(260) + F_S} = \frac{260 - 190 + 250}{190 + 250} = \frac{8}{11} \\
\overline{\psi}_S^{SPI} &= \frac{p(240) + F_S}{260 - p(240) + F_B} = \frac{175 + 250}{260 + 175 + 250} = \frac{85}{67}.
\end{aligned}$$

The following lemma characterizes the set of SRE outcomes when $v = 260$ as a function of the realization of retaliatory types and forms the basis for Figure 3 in the main text.

Lemma 2 *Suppose $v = 260$. Then the following are true:*

(i.) If $\psi_B < \overline{\psi}_B^{SPI}$ or if $\psi_B < \hat{\psi}_B^{SPI}$ and $\psi_S > \overline{\psi}_S^{SPI}$, then $\hat{v} = 260$ in every SRE;

(ii.) If $\psi_B > \overline{\psi}_B^{SPI}$ and $\psi_S < \overline{\psi}_S^{SPI}$, then there is no SRE in which $\hat{v} = 260$;

(iii.) If $\psi_B > \hat{\psi}_B^{SPI}$ and $\psi_S > \overline{\psi}_S^{SPI}$, then there are multiple SRE outcomes, including one in which $\hat{v} = 260$.

Proof of Lemma 2. Define the following three functions for $v' < v$:

$$\overline{\psi}_B(v, v') = \frac{v - \hat{p}(v')}{\hat{p}(v') + F_B + F_S}; \quad \hat{\psi}_B(v) = \frac{v - p(v) + F_B}{p(v) + F_S}; \quad \overline{\psi}_S(v, v') = \frac{p(v') + F_S}{v - p(v') + F_B}.$$

Note that these values satisfy $\overline{\psi}_B(260, 240) = \overline{\psi}_B^{SPI}$, $\hat{\psi}_B^{SPI}(260) = \hat{\psi}_B^{SPI}$, and $\overline{\psi}_S(260, 240) = \overline{\psi}_S^{SPI}$. We first establish several useful preliminary results. First, for any $v' < v$, in any SRE, the buyer will retaliate against an appropriate challenge if $\psi_B > \overline{\psi}_B(v, v')$. To see why, note that following an appropriate challenge, $\lambda_B^* = 1$, and he receives a payoff of $-F_B - \psi_B(-F_S)$ if he rejects the counter offer and $v - \hat{p}(v') - F_B - \psi_B(\hat{p}(v') + F_B)$ if he accepts the counter offer. The cutoff $\overline{\psi}_B(v, v')$ is the value at which these two payoffs are equal.

Second, suppose $\psi_B > \overline{\psi}_B(v, v')$ so that in every SRE, the buyer will retaliate against an appropriate challenge of $v' < v$. Then the seller will challenge nevertheless if $\psi_S \geq \overline{\psi}_S(v, v')$. As the buyer will retaliate against an appropriate challenge, at the history at which the seller decides whether to challenge an announcement of v', we have that $\lambda_S^* = 1$. He therefore receives a payoff of $-F_S - \psi_S(-F_B)$ if he challenges and $p(v') - \psi_S(v - p(v'))$ if he does not. The cutoff $\overline{\psi}_S(v, v')$ is the value at which these two payoffs are equal.

The first part of part $(i.)$ of the lemma is straightforward. If $\psi_B < \overline{\psi}_B(v, v')$, then the buyer will not retaliate against an appropriate challenge of v', so the seller will prefer to challenge him. He will therefore not announce v'. Moreover, $\overline{\psi}_B(v, v')$ is decreasing in $v' < v$, so if $\psi_B < \overline{\psi}_B^{SPI}$, then the buyer will not lie in any SRE. '

Next, suppose $\psi_B > \overline{\psi}_B^{SPI}$ and $\psi_S > \overline{\psi}_S^{SPI}$. Then if there is an SRE in which the buyer announces v with probability $1 - b^*$ for some $b^* > 0$, then there is an SRE in which the buyer announces v with probability $1 - b^*$ and $\hat{v} = 240$ with probability b^*. Moreover, there exists an SRE in which $b^* > 0$ only if $\psi_B > \hat{\psi}_B^{SPI}$. To see why, consider an SRE in which the buyer lies with probability b^*. Following a lie, he will be challenged, and he will reject the counter offer, receiving a monetary payoff of $-F_B$. Following a truthful announcement, he will not be challenged, and he will receive a monetary payoff of $v - p(v)$. Neither of these payoffs depend on the particular lie the buyer tells, so it is without loss of generality to focus on SREs in which the buyer announces $\hat{v} = 240$ with probability b^*. In such an SRE, the buyer's aggrievement at the announcement stage will be

$$\lambda_B^* = \frac{v - p(v) - (1 - b^*)(v - p(v)) - b^*(-F_B)}{v - p(v) - (-F_B)} = b^*.$$

In such an SRE, if the buyer tells the truth, he receives a payoff of $v - p(v) - \psi_B b^* p(v)$. If he lies, he receives a payoff of $-F_B - \psi_B b^* (-F_S)$. For $\psi_B < \hat{\psi}_B^{SPI}$, the buyer always strictly prefers to tell the truth, so it must be the case that $b^* = 0$. For $\psi_B > \hat{\psi}_B^{SPI}$, the buyer is indifferent between announcing v and $\hat{v} = 240$ if $b^* = \hat{\psi}_B^{SPI}/\psi_B$. This result implies that if $\overline{\psi}_B^{SPI} < \psi_B < \hat{\psi}_B^{SPI}$, and $\psi_S > \overline{\psi}_S^{SPI}$, then there is no SRE in which the buyer lies with positive probability, establishing the second part of part $(i.)$ of the lemma.

For part $(ii.)$ of the lemma, note that if $\hat{v} = v'$ is a profitable deviation from a truth-telling SRE for some $v' < v$, then so is $\hat{v} = 240$. To see why, consider a truth-telling SRE. At the initial node, the buyer's aggrievement is $\lambda_B^* = 0$, so he will be willing to deviate and lie only if doing so increases his material payoffs, given the continuation strategies specified by the SRE. He will therefore only be willing to lie if he will not be challenged. Since $\overline{\psi}_S(v, v')$ is increasing in v' and $\overline{\psi}_B(v, v')$ is decreasing in v', if he will not be challenged following an announcement of v', he will not be challenged following an announcement of 240. Therefore, if $\hat{v} = v'$ is a profitable deviation from a truth-telling SRE, so is $\hat{v} = 240$, so it is necessary to check whether $\hat{v} = 240$ is a profitable deviation for the buyer. Indeed, in the region described in part $(ii.)$ of the lemma, with $\psi_S < \overline{\psi}_S^{SPI}$ and $\psi_B > \overline{\psi}_B^{SPI}$, $\hat{v} = 240$ is a profitable deviation, so truth-telling cannot be part of an SRE.

For part $(iii.)$ of the lemma, our argument for part $(i.)$ of the lemma established that in this region, there is an SRE in which the buyer lies with strictly positive probability. It remains to argue that truth-telling is also an SRE outcome. Consider a truth-telling SRE. At the initial node, the buyer's aggrievement is $\lambda_B^* = 0$, so he will be willing to deviate and lie only if doing so increases his material payoffs. But since $\psi_B > \overline{\psi}_B^{SPI}$ and $\psi_S > \overline{\psi}_S^{SPI}$, for any lie, the seller will challenge, and the buyer will retaliate, so the buyer's material payoff must be lower following a lie. There is therefore no profitable deviation, and truth-telling is an SRE outcome.■

Lemma 2 characterizes the equilibrium outcomes in the different regions of Figure 3 in the main text. It shows that when the seller's retaliatory type is less than one, SREs involve truth-telling by the buyer only if $\psi_B < 3/149 \approx 0.02$. In other words, if the buyer is willing to sacrifice more than two cents in order to reduce the seller's material payoffs by one dollar, then there is no SRE in which the buyer tells the truth when $v = 260$. The

lemma also shows that when this is the case, in any SRE in which the buyer retaliates against challenges following $\hat{v} = 240$, the seller will never challenge such an announcement. In such psychological environments, therefore, SREs can rationalize lying and retaliation by the buyer as well as reluctance to challenge by the seller.

Lemma 2 also shows that when $\psi_B > 8/11 \approx 0.73$ and $\psi_S > 85/67 \approx 1.27$, there are multiple outcomes consistent with SRE behavior. This result echoes the result of Rabin (1993) that when material payoffs are small relative to psychological payoffs, equilibrium outcomes roughly coincide with the set of strategy profiles that deliver both parties very low payoffs or very high payoffs. For these outcomes to arise in equilibrium, the seller has to be willing to sacrifice at least \$1.27 in material payoffs to reduce the buyer's material payoffs by one dollar, which in the experimental literature documenting retaliatory behavior is a preference that is rarely observed.

This lemma also shows that when retaliatory types are common knowledge, it is challenging to explain why the seller would be willing to challenge a small lie by the buyer, an outcome we see in our main treatment roughly 20 percent of the time the buyer makes a small lie. In the appendix of Fehr, Powell, and Wilkening (2018), we show in this setting that if parties have private information about their retaliatory types, there are natural equilibria that involve small lies, occasional challenges, and frequent retaliation on the equilibrium path. This result holds even when parties tend to have moderate retaliatory types.

A6. The Retaliatory-Seller Mechanism

As we argued in the previous section, many features of the experimental results from our main treatments are consistent with SRE outcomes in a psychological environment in which players have retaliatory preferences. As a constructive matter, we are interested in whether in such a psychological environment, there exists a mechanism γ that both SPE-implements the pricing rule from our experiment and SRE-implements.

One of the key weaknesses of the SPI mechanism in our setting is that sellers are reluctant to challenge small lies. Our goal is to address this weakness by constructing a mechanism under which, if sellers have similar retaliatory types as buyers, we can use their retaliatory preferences to improve their propensity to challenge small lies. The idea of our construction is to make a small change to our baseline mechanism that makes sellers aggrieved exactly when they *should* be challenging the buyer. To do so, we will add a simultaneous announcement by the seller to the announcement stage, and we will charge the seller a fine if his announcement differs from the buyer's.

To be specific, consider mechanisms with $T = 3$ that take the following form.

1. The buyer and seller simultaneously announce $\hat{v}_B, \hat{v}_S \in \mathcal{V}$ (i.e., $\mathcal{M}_B^1 = \mathcal{V}$ and $\mathcal{M}_S^1 = \mathcal{V}$). If the announcements agree, then trade occurs at price $p(\hat{v}_B)$, so that $g(m) = (1, p(\hat{v}_B), p(\hat{v}_B))$ if $\hat{v}_B = \hat{v}_S$.
2. If the announcements disagree, the seller must pay a fine F_S, and he chooses whether to challenge the buyer's announcement ($m_S^2 = C$) or not ($m_S^2 = N$) (i.e., $\mathcal{M}_B^2 = \emptyset$ and $\mathcal{M}_S^2(\hat{v}) = \{C, N\}$). If the seller does not challenge, then trade occurs at price $p(\hat{v}_B)$, so that $g(m) = (1, p(\hat{v}_B), p(\hat{v}_B) - F_S)$ if $m_2^S = N$.

3. If $m_S^2 = C$, then the buyer pays a fine F_B and receives a counter offer: He can choose whether to buy the good at price $\hat{p}(\hat{v})$ ($m_B^3 = Y$) or not ($m_B^3 = N$) (i.e., $\mathcal{M}_B^3(m^1, m^2) = \{Y, N\}$ if $m_S^2 = C$ and $\emptyset$ if $m_S^2 = N$, and $\mathcal{M}_S^3(m^1, m^2) = \emptyset$). If the buyer buys the good, then trade occurs at price $\hat{p}(\hat{v}_B)$, and the seller receives the fine F_B, so that $g(m) = (1, \hat{p}(\hat{v}_B) + F_B, \hat{p}(\hat{v}_B) + F_B - F_S)$ if $m_B^3 = Y$. If the buyer does not buy the good, then trade does not occur, so that $g(m) = (0, F_B, -F_S)$ if $m_B^3 = N$.

We refer to such mechanisms as **retaliatory-seller mechanisms**, and we will denote by Γ^{RS} the set of such mechanisms. It is straightforward to show that for any pricing rule p, there exists a retaliatory-seller mechanism $\gamma^{RS} \in \Gamma^{RS}$ that SPE-implements p, and the specific mechanism we describe in Section 6.2 SPE-implements the specific pricing rule used in our experiment.

We will now establish a partial dominance result, showing a sense in which the retaliatory-seller mechanism induces truth-telling in a broader range of psychological environments than does the SPI mechanism. To do so, we will compare two mechanisms, one SPI mechanism and one retaliatory-seller mechanism, that have the same buyer and seller fines and arbitration schedules. To this end, denote a $\gamma^{SPI} \in \Gamma^{SPI}$ mechanism with buyer fine F_B, seller fine F_S, and arbitration schedule $\hat{p}(\cdot)$ by $\gamma^{SPI}(F_B, F_S, \hat{p})$. Similarly, denote a $\gamma^{RS} \in \Gamma^{RS}$ mechanism with buyer fine F_B, seller fine F_S, and arbitration schedule $\hat{p}(\cdot)$ by $\gamma^{RS}(F_B, F_S, \hat{p})$. Take a pricing rule p, and suppose $\gamma^{SPI}(F_B, F_S, \hat{p})$ SPE-implements p, and so does $\gamma^{RS}(F_B, F_S, \hat{p})$. We will say that $\gamma^{RS}(F_B, F_S, \hat{p})$ **SRE-partially dominates** $\gamma^{SPI}(F_B, F_S, \hat{p})$ if the following two conditions are satisfied:

1. If $\gamma^{SPI}(F_B, F_S, \hat{p})$ SRE-partially implements p in psychological environment $\mathcal{P}$, then so does $\gamma^{RS}(F_B, F_S, \hat{p})$.
2. There exists a psychological environment $\mathcal{P}$ in which $\gamma^{RS}(F_B, F_S, \hat{p})$ SRE-partially implements p, but $\gamma^{SPI}(F_B, F_S, \hat{p})$ does not.

For the purposes of establishing the partial dominance result, it will be useful to define the following cutoffs, given a pair of values $v, v' \in \mathcal{V}$:

$$\begin{aligned}
\overline{\psi}_B^{SPI}(v, v') &= \frac{v - \hat{p}(v')}{\hat{p}(v') + F_B + F_S};\ \overline{\psi}_S^{SPI}(v, v') = \frac{p(v') + F_S}{v - p(v') + F_B} \\
\overline{\psi}_B^{RS}(v, v') &= \frac{v - \hat{p}(v')}{\hat{p}(v') + F_B};\ \overline{\psi}_S^{RS}(v, v') = \frac{p(v')}{v - p(v') + F_B}.
\end{aligned}$$

The next proposition shows that $\gamma^{RS}(F_B, F_S, \hat{p})$ SRE-partially dominates $\gamma^{SPI}(F_B, F_S, \hat{p})$.

Proposition 2 *Fix the buyer fine, F_B, the seller fine, F_S, and the arbitration schedule $\hat{p}$. Consider a pricing rule p for which both $\gamma^{SPI}(F_B, F_S, \hat{p})$ and $\gamma^{RS}(F_B, F_S, \hat{p})$ SPE-implement p. Then $\gamma^{RS}(F_B, F_S, \hat{p})$ SRE-partially dominates $\gamma^{SPI}(F_B, F_S, \hat{p})$.*

Proof of Proposition 2. We want to show the conditions under which in payoff state v, there is an SRE of $\gamma^{SPI}(F_B, F_S, \hat{p})$ in which the buyer announces $\hat{v} = v$ and the conditions under which there is an SRE of $\gamma^{RS}(F_B, F_S, \hat{p})$ in which both parties announce $\hat{v}_B = \hat{v}_S$. We

will first describe the set of conditions under which truth-telling is an SRE outcome in the SPI mechanism and the RS mechanism. Then we will compare these two sets of conditions. To this end, take an arbitrary v, and consider a $v' < v$ to be a candidate deviation at the announcement stage.

Truth-telling in the SPI mechanism. The proof of Lemma 2 can be extended to show that there is an SRE in which the buyer announces $\hat{v} = v$ as long as for all $v' < v$, either $\psi_B \leq \overline{\psi}_B^{SPI}(v, v')$ or $\psi_S \geq \overline{\psi}_S^{SPI}(v, v')$. When $\psi_B \leq \overline{\psi}_B^{SPI}(v, v')$, the buyer will accept the counter offer if challenged, so the seller will challenge if the buyer announces $v' < v$, and so the buyer will announce v. When $\psi_S \geq \overline{\psi}_S^{SPI}(v, v')$, then even if the buyer will reject the counter offer if challenged, the seller will challenge an announcement $v' < v$, and so again, the buyer will announce v.

Truth-telling in the RS mechanism. As with the SPI mechanism, the buyer's retaliation behavior as a function of his retaliatory type is characterized by a cutoff. If the buyer has announced $\hat{v}_B = v' < v$ with $\hat{v}_B \neq \hat{v}_S$ and been challenged, he will accept the counter offer if $\psi_B \leq \overline{\psi}_B^{RS}(v, v')$ and reject it if $\psi_B \geq \overline{\psi}_B^{RS}(v, v')$. To see why, note that if he is challenged, $\lambda_B^* = 1$. If he accepts the counter offer, he receives utility $v - \hat{p}(v') - F_B - \psi_B(\hat{p}(v') - F_S + F_B)$. If he rejects the counter offer, he receives utility $-F_B - \psi_B(-F_S)$. The value $\overline{\psi}_B^{RS}(v, v')$ equates these two expressions.

Similarly, the seller's challenging behavior as a function of his retaliatory type is characterized by a cutoff. If the buyer is sure to retaliate, then when deciding whether to challenge, the seller's aggrievement is $\lambda_S^* = 1$. He will challenge if $\psi_S \geq \overline{\psi}_S^{RS}(v, v')$. To see why, note that if he challenges and the buyer retaliates, then he receives utility $-F_S - \psi_S(-F_B)$. If he does not challenge, then he receives utility $p(v') - F_S - \psi_S(v - p(v'))$. The value $\overline{\psi}_S^{RS}(v, v')$ equates these two expressions.

Putting these two results together, there is an SRE in which $\hat{v}_B = \hat{v}_S = v$ as long as for all $v' < v$, either $\psi_B \leq \overline{\psi}_B^{RS}(v, v')$ or $\psi_S \geq \overline{\psi}_S^{RS}(v, v')$. Suppose $\hat{v}_S = v$. We will ask whether the buyer wants to deviate and announce $\hat{v}_B = v'$. Paralleling the argument in the SPI mechanism, when $\psi_B \leq \overline{\psi}_B^{RS}(v, v')$, the buyer will accept the counter offer if challenged, so the seller will challenge if the buyer announces $v' < v$, and so the buyer will announce $\hat{v}_B = v$. When $\psi_S \geq \overline{\psi}_S^{RS}(v, v')$, then even if the buyer will reject the counter offer if challenged, the seller will challenge an announcement $v' < v$, and so again, the buyer will announce $\hat{v}_B = v$. If $\hat{v}_B = v$, then the seller's best response is to announce $\hat{v}_S = v$.

Comparison between the SPI mechanism and the RS mechanism. Let $\hat{\Psi}^{SPI}$ be the set of (ψ_B, ψ_S) such that for all v and $v' < v$, $\psi_B \leq \overline{\psi}_B^{SPI}(v, v')$ or $\psi_S \geq \overline{\psi}_S^{SPI}(v, v')$. Then truth-telling is part of an SRE under $\gamma^{SPI}(F_B, F_S, \hat{p})$ if and only if $(\psi_B, \psi_S) \in \hat{\Psi}^{SPI}$. Similarly, let $\hat{\Psi}^{RS}$ be the set of (ψ_B, ψ_S) such that for all v and $v' < v$, $\psi_B \leq \overline{\psi}_B^{RS}(v, v')$ or $\psi_S \geq \overline{\psi}_S^{RS}(v, v')$. Then truth-telling is part of an SRE under $\gamma^{RS}(F_B, F_S, \hat{p})$ if and only if $(\psi_B, \psi_S) \in \hat{\Psi}^{RS}$. Finally, note that $\overline{\psi}_B^{SPI}(v, v') < \overline{\psi}_B^{RS}(v, v')$ and $\overline{\psi}_S^{SPI}(v, v') > \overline{\psi}_S^{RS}(v, v')$ for all v and all $v' < v$. This implies that $\hat{\Psi}^{SPI} \subsetneq \hat{\Psi}^{RS}$, so $\gamma^{RS}(F_B, F_S, \hat{p})$ SRE-partially dominates $\gamma^{SPI}(F_B, F_S, \hat{p})$. ■

Proposition 2 shows that if we fix $(F_B, F_S, \hat{p})$ and p, the associated the retaliatory-seller mechanism SRE-partially implements p in a larger class of psychological environments than does the SPI mechanism. If, however, we consider full implementation rather than partial implementation, such a result does not hold. Specifically, we will say that $\gamma^{RS}(F_B, F_S, \hat{p})$ SRE dominates $\gamma^{SPI}(F_B, F_S, \hat{p})$ if the following two conditions are satisfied:

1. If $\gamma^{SPI}(F_B, F_S, \hat{p})$ SRE-implements p in psychological environment $\mathcal{P}$, then so does $\gamma^{RS}(F_B, F_S, \hat{p})$.
2. There exists a psychological environment $\mathcal{P}$ in which $\gamma^{RS}(F_B, F_S, \hat{p})$ SRE-implements p, but $\gamma^{SPI}(F_B, F_S, \hat{p})$ does not.

The next proposition shows that $\gamma^{RS}(F_B, F_S, \hat{p})$ does not SRE dominate $\gamma^{SPI}(F_B, F_S, \hat{p})$ by constructing a counter example.

Proposition 3 *There exists a $(F_B, F_S, \hat{p})$ and a pricing rule p for which $\gamma^{SPI}(F_B, F_S, \hat{p})$ and $\gamma^{RS}(F_B, F_S, \hat{p})$ SPE-implement p, and $\gamma^{RS}(F_B, F_S, \hat{p})$ does not SRE dominate $\gamma^{SPI}(F_B, F_S, \hat{p})$.*

Proof of Proposition 3. Suppose $\mathcal{V} = \{240, 260\}$, and take $F_B = 250$, $F_S = 100$, $\hat{p}(\hat{v}) = \hat{v} + 5$, and $p(240) = 175$ and $p(260) = 190$. Take (ψ_B, ψ_S) such that $\overline{\psi}_B^{RS}(260, 240) < \psi_B < \hat{\psi}_B^{SPI}(260, 240)$ and $\overline{\psi}_S^{SPI}(260, 240) < \psi_S < \frac{F_S}{p(260)-p(240)}\overline{\psi}_S^{SPI}(260, 240)$. Then the following are true:

1. Truthtelling is part of every SRE under $\gamma^{SPI}(F_B, F_S, \hat{p})$ and
2. There is an SRE under $\gamma^{RS}(F_B, F_S, \hat{p})$ in which $\hat{v}_B = \hat{v}_S = 240$ in payoff state $v = 260$.

The first claim follows directly from Lemma 2. For the second claim, let us consider the mechanism $\gamma^{RS}(F_B, F_S, \hat{p})$, and suppose the payoff state is $v = 260$. Suppose $\hat{v}_B = 240$ and $\hat{v}_S = 260$. Then, since $\psi_B > \overline{\psi}_B^{RS}(260, 240)$, the buyer will reject the counter offer, and since $\psi_S > \overline{\psi}_S^{SPI}(260, 240) > \overline{\psi}_S^{RS}(260, 240)$, the seller will challenge nevertheless. At the announcement stage, the buyer's aggrievement under this candidate equilibrium is $\lambda_B^* = 0$, and the seller's aggrievement is

$$\lambda_S^* = \frac{p(260) - p(240)}{p(260) - [p(260) - F_S]} = \frac{p(260) - p(240)}{F_S},$$

where this expression holds because the worst payoff that the buyer can deliver to the seller when he announces $\hat{v}_S = 240$ is to announce $\hat{v}_B = 260$, in which case the seller will not challenge, so the seller will receive $p(260) - F_S$.

Given the seller's aggrievement level at the announcement stage, he will therefore be willing to announce $\hat{v}_S = 240$ when $\hat{v}_B = 240$ as long as

$$p(240) - \psi_S \lambda_S^* (260 - p(240)) > -F_S - \psi_B \lambda_S^* (-F_B)$$

or

$$\psi_S < \frac{1}{\lambda_S^*} \frac{p(240) + F_S}{260 - p(240) + F_B} = \frac{F_S}{p(260) - p(240)} \overline{\psi}_S^{SPI}(260, 240).$$

We therefore have that in payoff state $v = 260$, $\hat{v}_B = \hat{v}_S = 240$ is part of an SRE under $\gamma^{RS}(F_B, F_S, \hat{p})$, but $\hat{v}_B = 240$ is not part of an SRE under $\gamma^{SPI}(F_B, F_S, \hat{p})$.■

Proposition 3 shows that for a given psychological environment, there may exist non-truthtelling SREs in which both parties coordinate on making an untruthful announcement under the retaliatory-seller mechanism, while only truth-telling is an SRE outcome under the SPI mechanism.

A7. The Insurance Property and Fixed-Price Contracts

This section establishes the implications of Bierbrauer and Netzer's (2016) insurance property for social choice functions in a hold-up setting with commonly known payoff states. We first describe a more general economic environment in which the seller's costs as well as the buyer's value can take on multiple values. An economic environment with different costs is an array $\mathcal{E} = (\{B, S\}, \mathcal{A}, \mathcal{C}, \mathcal{V}, \pi_B, \pi_S)$ defined as in Appendix A1, except that it also includes a set of possible seller costs $\mathcal{C} = \{c_1, \ldots, c_M\}$ with $c_1 > \cdots > c_M$, and players' material payoffs are given by $\pi_B(a) = vq - t_B$ and $\pi_S(a) = t_S - cq$. A payoff state is a pair $\theta \equiv (c, v)$, where $\theta \in \Theta \equiv \mathcal{C} \times \mathcal{V}$. To introduce the appropriate notation, assume each player privately observes a signal $\theta_i \in \Theta$. For our purposes, we will assume that both players observe the payoff state without noise: $\theta_B = \theta_S = (c, v)$.

In this setting, a social choice function f is a mapping $f : \Theta^2 \to \{0, 1\} \times \mathbb{R} \times \mathbb{R}$ that specifies an allocation for each pair (θ_B, θ_S), where $\theta_B = (c_B, v_B)$ and $\theta_S = (c_S, v_S)$. When referring to its constituent parts, we use the notation $f = \left(q^f, t_B^f, t_S^f\right)$. We say that a social choice function f has **no marginal externalities on the buyer** if in payoff state θ, the associated direct mechanism has the property that

$$q^f\left(\theta, \hat{\theta}_S\right) v + t_B^f\left(\theta, \hat{\theta}_S\right)$$

is independent of $\hat{\theta}_S \in \Theta$, and it has **no marginal externalities on the seller** if the associated direct mechanism has the property that

$$t_S^f\left(\hat{\theta}_B, \theta\right) - q^f\left(\hat{\theta}_B, \theta\right) c$$

is independent of $\hat{\theta}_B \in \Theta$. A social choice function that has no marginal externalities on either the seller or the buyer satisfies what Bierbrauer and Netzer (2016) refers to as the **insurance property**. The insurance property therefore implies that whether the buyer buys the good and at what price is independent of the seller's private information, and it also implies that whether the seller sells and at what price is independent of the buyer's private information.

We will say that f is a **fixed-price contract** if it is budget balanced (i.e., $t_B^f\left(\hat{\theta}_B, \hat{\theta}_S\right) = t_S^f\left(\hat{\theta}_B, \hat{\theta}_S\right)$ for all $\hat{\theta}_B, \hat{\theta}_S \in \Theta$), and the price the buyer pays depends on the payoff state only inasmuch as the payoff state affects the quantity traded: $t_B^f\left(\hat{\theta}_B, \hat{\theta}_S\right) = \tilde{t}_B^f\left(q^f\left(\hat{\theta}_B, \hat{\theta}_S\right)\right)$. We will say that such a social choice function is an **option-to-buy contract** if it is a fixed-price

contract, and $q^f\left(\hat{\theta}_B, \hat{\theta}_S\right)$ is independent of $\hat{\theta}_S$. We will say that a social choice function is an **option-to-sell contract** if it is a fixed-price contract, and $q^f\left(\hat{\theta}_B, \hat{\theta}_S\right)$ is independent of $\hat{\theta}_B$. We will say that a social choice function is **constant** if it is a fixed-price contract, and $q^f\left(\hat{\theta}_B, \hat{\theta}_S\right)$ is independent of both $\hat{\theta}_B$ and $\hat{\theta}_S$.

Proposition 4 *Suppose f satisfies the insurance property, truth-telling, and budget balance. Then f is a fixed-price contract. If $|\mathcal{C}| = 1$, then f is an option-to-buy contract. If $|\mathcal{V}| = 1$, then f is an option-to-sell contract. If $|\mathcal{C}|, |\mathcal{V}| > 1$, then f is constant.*

Proof of Proposition 4. Since f satisfies the insurance property, it has no marginal externalities on the buyer. We can therefore write $q^f\left(\theta, \hat{\theta}_S\right) = q_B(\theta)$ and $t_B^f\left(\theta, \hat{\theta}_S\right) = t_B(\theta)$ for all θ. Buyer truth-telling then requires that for all $\theta = (c, v)$, $\theta' = (c', v')$,

$$q_B(\theta)\, v - t_B(\theta) \geq q_B(\theta')\, v - t_B(\theta'),$$

which implies the monotonicity condition

$$(q_B(\theta) - q_B(\theta'))(v - v') \geq 0.$$

Next, to show that the price the buyer pays depends on the payoff state only inasmuch as it affects quantity, suppose there are two payoff states θ, θ' for which $q_B(\theta) = q_B(\theta')$. Then

$$\begin{aligned} q_B(\theta)\, v - t_B(\theta) &\geq q_B(\theta')\, v - t_B(\theta') \\ q_B(\theta')\, v' - t_B(\theta') &\geq q_B(\theta)\, v' - t_B(\theta) \end{aligned}$$

implies that $t_B(\theta) = t_B(\theta')$. Thus, f is a fixed-price contract, which establishes the first part of the proposition.

Since f has no marginal externalities on the buyer, buyer truth-telling requires that

$$q_B(c, v)\, v + \tilde{t}_B(q_B(c, v)) = q_B(c', v)\, v + \tilde{t}_B(q_B(c', v))$$

for all $v \in \mathcal{V}$ and $c, c' \in \mathcal{C}$. For q_B to depend nontrivially on c, it must therefore be the case that $|\mathcal{V}| = 1$.

If we go through the same exercise but instead use the fact that f has no marginal externalities on the seller, then we have the monotonicity condition

$$(q_S(\theta) - q_S(\theta'))(c - c') \leq 0,$$

and $q_S(\theta) = q_S(\theta')$ implies $t_S(\theta) = t_S(\theta')$, so again, f must be a fixed-price contract. And again, q_S can depend nontrivially on v only if seller costs take on a single value, that is $|\mathcal{C}| = 1$.

These results imply that if $|\mathcal{C}| = 1$, then q can depend nontrivially on v and therefore is an option-to-buy contract. If $|\mathcal{V}| = 1$, then q can depend nontrivially on c and is therefore an option-to-sell contract. If $|\mathcal{C}|$, $|\mathcal{V}| > 1$, then q cannot depend nontrivially on either c or v and is therefore constant.■

Proposition 4 illustrates how the insurance property limits the set of social choice functions to fixed-price contracts for which at most one party gets to choose whether or not to trade. The insurance property therefore constrains the types of incentives that can be provided to the parties to make relationship-specific investments. In particular, in a two-sided hold-up problem with $|\mathcal{C}|, |\mathcal{V}| > 1$, no social choice function that satisfies the insurance property can provide incentives for either party to make relationship-specific cross investments. Note that the insurance property does not, however, imply that parties cannot be provided with incentives to make relationship-specific self investments.

Appendix B: Additional Analyses and Treatments

B1. The Role of Beliefs

In this appendix, we explore the role of a subject's beliefs in shaping his or her decisions under the mechanism. If sellers believe that counter offers following an appropriate challenge of a small lie will be rejected, they will be reluctant to challenge such announcements. Likewise, if buyers believe that small lies will not be challenged, they ought to be willing to underreport the value of the good. We find evidence that sellers and buyers have these beliefs, and that sellers and buyers who have these beliefs act accordingly.

Result B.1 (*a*) *Most buyers believe that being challenged for a small lie is unlikely or will never occur. Buyers who have these beliefs are more likely to lie than those who believe that sellers will challenge them.* (*b*) *Most sellers believe that a challenge of a small lie is likely to be rejected or will always be rejected. Sellers who believe that their challenges will be rejected are significantly less likely to challenge a small lie.*

Recall that in each period, we elicited the buyers' beliefs about the likelihood of being challenged for each potential announcement using a 4-point Likert scale (Never/Unlikely/ Likely/Always). Figure B1 shows the proportion of buyers who indicated "Never" or "Unlikely" for each announcement after the seller exerts high effort. 82 percent of buyers believe that announcements of 240 are never challenged or are unlikely to be challenged, and 66 percent believe that an announcement of 220 is never challenged or is unlikely to be challenged. Similar results hold following low effort choices where 53 percent of buyers believe that the seller is "Unlikely" to challenge or will "Never" challenge an announcement of 100. These results suggest that buyers correctly forecast that many sellers are reluctant to challenge a small lie.

To better understand the role that beliefs have in buyers announcements we look at the decision of the buyer to make a small lie based on his belief about being challenged after such lies. Table B1 reports the results of a probit regression where the dependent variable is 1 if a buyer makes a small lie and 0 if the buyer makes a truthful announcement. We report regressions for choices after high effort in regressions (1) and (2), choices after low effort in regressions (3) and (4), and choices after both high and low effort in regressions (5) and (6).

In regressions (1), the small lie indicator is regressed on the belief that an announcement

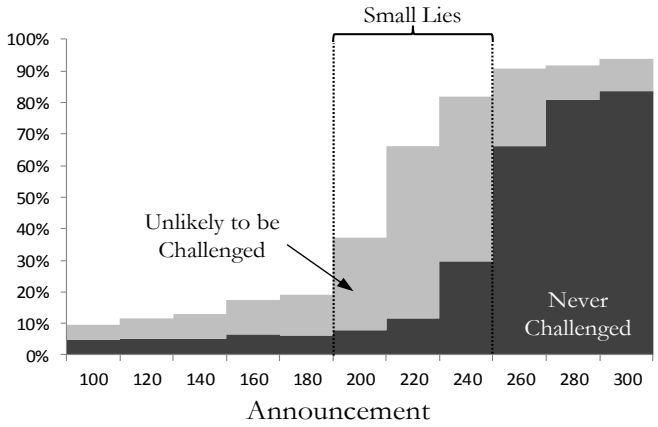


Figure B1: Proportion of buyers who believe that a given announcement will "Never" be challenged or is "Unlikely" to be challenged after observing high effort.

of 240 — the smallest possible lie — will be challenged in cases where high effort occurs.[1] Likewise, in regression (3), the indicator for small lies is regressed on the belief that an announcement of 100 will be challenged in the case of low effort. We combine these beliefs in regression (5). To allow for potential non-linearities in the belief data we treat buyers' beliefs as categorical data and split the 4-point Likert scale into a series of dummy variables. We use the category "Never" as the omitted dummy category.

Beliefs about the likelihood of being challenged are a good predictor of the buyers likelihood of making a small lie. Based on the marginal effects of a probit regression, buyers are 36.6 percentage points less likely to lie after high effort if they believe that being challenged is "Likely" relative to individuals who believe that this will "Never" occur. Likewise, they are 56.2 percentage points less likely to make a small lie after low effort if they believe that being challenged is "Likely." The probability of making a small lie is decreasing as an individual's belief moves to more pessimistic categories suggesting a monotonic relationship between beliefs and announcements.

As can be seen by referring back to Figure B1, while most buyers believe that truthful announcements will "Never" be challenged, a small subset of buyers have more pessimistic beliefs. As the decision to make a small lie is based on the expected value of lying relative to the expected value of telling the truth, such pessimistic beliefs should increase the likelihood of buyers to make a small lie. To test for this relationship, we extend the probit regression in equations (2), (4), and (6) to also include beliefs about being challenged after a truthful announcement. As expected, individuals are more likely to lie as they become more pessimistic about being challenged after a truthful announcement. Thus optimistic beliefs about being challenged after a lie and pessimistic beliefs about being challenged after a truthful announcement appear to influence the buyers announcement decision.

Turning to the beliefs of sellers, 71.6 percent (62.3 percent) of sellers who are confronted with a small lie after high (low) effort believe that an appropriate challenge will "Never" be

[1]We used the belief on 240 to keep the high and low effort regressions the same. Alternative specifications that use combined measures from announcements of 200, 220, and 240 have similar coefficients and predictive power.

Table B1: Probit Regression of Small Lies by Buyers

	High Effort		Low Effort		Combined	
	(1)	(2)	(3)	(4)	(5)	(6)
Buyer's Belief that Seller Will Challenge Smallest Lie:						
"Unlikely"	-0.242 **	-0.320 ***	-0.297 *	-0.404 ***	-0.245 **	-0.336 ***
	(0.116)	(0.116)	(0.174)	(0.187)	(0.098)	(0.102)
"Likely"	-0.366 **	-0.549 ***	-0.562 ***	-0.685 ***	-0.404 ***	-0.600 ***
	(0.154)	(0.147)	(0.159)	(0.123)	(0.109)	(0.094)
"Always"	-0.487 ***	-0.614 ***	-0.639 ***	-0.934 ***	-0.491 ***	-0.704 ***
	(0.160)	(0.104)	(0.151)	(0.029)	(0.118)	(0.063)
Buyer's Belief that Seller will Challenge a Truthful Announcement:						
"Unlikely"	-	0.232 **	-	0.180	-	0.228 ***
		(0.109)		(0.126)		(0.083)
"Likely"	-	0.359 ***	-	0.193 *	-	0.271 ***
		(0.099)		(0.114)		(0.073)
"Always"	-	0.225	-	0.633 ***	-	0.358 ***
		(0.226)		(0.061)		(0.068)
Psuedo R^2	0.061	0.100	0.148	0.220	0.076	0.116
Observations	237	237	183	183	420	420

Marginal effects from a probit regression are reported in the table where the dependent variable is 1 if the buyer makes a small lie and 0 if the buyer makes a truthful announcement. Standard errors in parentheses, clustered by individual. The omitted category is Seller "Never" Challenges. Regressions (1) and (2) restrict the sample to periods where High effort is chosen. Regressions (3) and (4) restrict the sample to periods where Low effort is chosen. *,**,*** denote siginificance at the 10%, 5% and 1%-level, respectively.

accepted or is "Unlikely" to be accepted. Thus, sellers also correctly forecast that buyers are likely to reject appropriate challenges.

As with buyers, sellers are not only correctly forecasting that appropriate challenges will be rejected, they appear to use these beliefs to guide their decisions. Table B2 reports the marginal effects of a probit regression where we regress an indicator for the seller's challenge decision on his beliefs. Data in these regressions are restricted to cases where the buyer makes a small lie and are divided into the low-effort case, the high-effort case, and the combined case. As can be seen in column (1), sellers who exert high effort and believe that it is "Likely" that their challenge will be accepted are 81.7 percentage points more likely to challenge than sellers who believe that their challenge will "Never" be accepted. Similarly, sellers who exert low effort and believe that their challenge is "Likely" to be accepted are 39.1 percentage points more likely to challenge than sellers who believe that their challenge will "Never" be accepted.

Taken together, our belief data suggests that individuals are correctly predicting deviations from the SPI predictions in later stages of the game and are responding to these beliefs in a consistent manner.

Table B2: Probit Regression of Challenges by Sellers After A Small Lie

	High Effort (1)	Low Effort (2)	Combined (3)
Sellers Belief: Acceptance of Appropriate Challenge "Unlikely"	0.083	0.165	0.108
	(0.131)	(0.131)	(0.088)
Sellers Belief: Acceptance of Appropriate Challenge "Likely"	0.817 ***	0.391 ***	0.604 ***
	(0.089)	(0.121)	(0.089)
Sellers Belief: Appropriate Challenge "Always" Accepted	0.678 ***	0.504 ***	0.586 ***
	(0.155)	(0.187)	(0.111)
Psuedo R^2	0.110	0.471	0.252
Observations	122	141	263

Marginal effects from a probit regression are reported in the table where the dependent variable is 1 if the seller challenges a small lie and 0 if the seller doesn't challenge. Standard errors in parentheses, clustered by individual. The omitted category is Buyer "Never" Accepts. Regression (1) restricts the sample to periods with High Effort and a Small Lie. Regression (2) restricts the sampe to periods with Low Effort and a Small Lie. *,**,*** denote significance at the 10%, 5%, 1%-level, respecively.

B2. High-Benefits Treatment

Under the SPI hypothesis, the appropriate-challenge condition predicts that sellers always challenge a lie and never challenge a truthful or generous offer. As was seen in panel (b) of Figure **??**, the sellers do not behave in accordance with this condition, because small lies are not challenged frequently.

While the appropriate-challenge condition is violated, the likelihood that the seller will challenge is decreasing in the size of the buyer's announcements. Thus, the empirical distribution of sellers' challenges continues to satisfy at least one central property of the original appropriate-challenge condition: small lies are more likely to be challenged than truthful announcements. We take advantage of this property in the following High-Benefits treatment.

The decision for a buyer to make a truthful announcement or a small lie is based on the buyer's expected utility for telling the truth relative to the expected utility of lying. This implies that any change in the SPI mechanism that increases the utility of truth-telling relative to small lies has the potential of inducing the buyer to make a truthful announcement.

A buyer is less likely to be challenged after a truthful announcement than a small lie. This implies that if the value that a buyer receives when he is *not* challenged increases by a constant across all potential announcements, the expected value of announcing a truthful announcement will increase by more than the expected value of announcing a small lie. For example, if a buyer believes that a small lie will be challenged 50 percent of the time and a truthful announcement will never be challenged, then an increase in the value of not being challenged of 10 will increase the expected value of the small lie by 5 $(10 * .5)$ and increase the value of truth telling by 10.

In the High-Benefits treatment we make precisely this type of shift in the value of not

Table B3: Correspondence Between Announcement, Prices, and Outcomes in High-Benefits Treatment

			Low Effort (True Value = 120, Cost of Effort = 30)			High Effort (True Value = 260, Cost of Effort = 120)		
Value Announced $\hat{v}$	Price Offered to Seller $p(\hat{v})$	Counter-Offer Price $\hat{p}(\hat{v})$	Buyer's Surplus if No Challenge Occurs	Seller's Surplus if No Challenge Occurs	Buyer's Net Profit of Accepting Counter Offer	Buyer's Surplus if No Challenge Occurs	Seller's Surplus if No Challenge Occurs	Buyer's Net Profit of Accepting Counter Offer
100	50	105	70	20	15	210	-70	155
120	65	125	55	35	-5	195	-55	135
140	80	145	40	50	-25	180	-40	115
160	95	165	25	65	-45	165	-25	95
180	110	185	10	80	-65	150	-10	75
200	125	205	-5	95	-85	135	5	55
220	140	225	-20	110	-105	120	20	35
240	155	245	-35	125	-125	105	35	15
260	170	265	-50	140	-145	90	50	-5
280	185	285	-65	155	-165	75	65	-25
300	200	305	-80	170	-185	60	80	-45

Grey boxes in the "Buyer's Net Profit if No Challenge Occurs" columns show announcements for which a selfish buyer would accept the counter offer if challenged. A selfish buyer will make the lowest possible announcement that is not challenged. This will be an announcement of 260 after high effort and 120 after low effort. Thus the SPNE with selfish players in this treatment is the same as the Main treatment.

being challenged by decreasing the initial-price schedule $p(\hat{v})$ uniformly across all announcements. The structure of this treatment is just as in the SPI Treatment except that we decrease the price $p(\hat{v})$ by 20:

$$p(\hat{v}) = 50 + .75\,(\hat{v} - 100)\,.$$

As the change involves a constant shift in the initial-price schedule, it does not affect the predictions from the SPI hypothesis. This can be seen in Table B3, which summarizes the payoffs for each potential choice within the treatment. However, holding the challenge probabilities of the seller fixed, the treatment is predicted to increase the value of announcements where the buyer believes there is a low probability of being challenged relative to announcements where the buyer believes there is a high probability of being challenged. We thus expect more truthful announcements, fewer small lies, and (by backward induction) a higher proportion of sellers exerting high effort.

The High-Benefits treatment consisted of two sessions with 26 subjects in each session, and we find the following:

Result B.2 *The High-Benefits Treatment has a larger proportion of sellers who exert high effort than the SPI Treatment. It also has fewer small lies and sellers are more likely to challenge these lies. However, buyers still retaliate against most challenges, leading to inefficiency. Thus, although the High-Benefits Treatment improves the efficiency of the mechanism relative to the SPI Treatment, the mechanism's efficiency still remains very low.*

Figure B2 displays the results for the High-Benefits Treatment with data aggregated across all 10 periods: The left-hand side of the figure follows the pattern of play after the seller selects low effort ($N = 66$) while the right-hand side of the figure follows the pattern of play following high effort ($N = 194$). Directly comparable to Figure 1, panel (a) shows

the distribution of announcements, panel (b) shows the likelihood of a challenge after each announcement, and panel (c) shows the frequency that a challenge is accepted or rejected.

Comparing the proportion of sellers who exert high effort in the SPI and High-Benefits Treatments, the High-Benefits Treatment has a larger proportion of sellers who choose high effort. In the SPI Treatment, sellers select high effort in only 260 out of 460 observations (57 percent), while sellers in the High-Benefits treatment choose high effort in 194 out of 260 observations (75 percent). This difference is significant in a simple probit regression where effort choice is regressed on the treatment variable (p-value $= 0.02$).

Controlling for the difference in effort levels, the High-Benefits Treatment also has significantly fewer lies than in the SPI Treatment. Panel (a) shows that small lies occur in only 11 out of 66 cases after low effort (17 percent) and 30 out of 194 cases after high effort (16 percent). These small lie rates are very low relative to the SPI Treatment where lies occurred 61 percent of the time after low effort and 54 percent of the time after high effort. The difference in the propensity to make small lies between the two treatments is statistically significantly different in two separate probit regressions — one for low effort and one for high effort — where a binary variable that is 1 for a small lie and 0 for a truthful announcement is regressed on the treatment variable (p-value < 0.01 for both regressions).

Interestingly, unlike the SPI Treatment, buyers in the High-Benefits Treatment frequently make generous announcements, $\hat{v} > v(e)$. For example, after high effort, buyers make generous announcements in 38 percent of the cases. The large proportion of these generous offers suggests a new deviation from the SPNE hypothesis that did not occur in the SPI Treatment. We return to this issue when we discuss the beliefs data below.

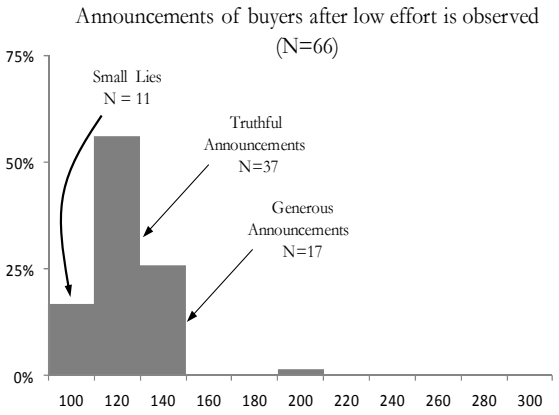
Looking at Panel (b) and comparing it to the SPI Treatment, sellers are much more likely to challenge small lies in the High-Benefits Treatment: following high effort, announcements of 240 are challenged 58 percent of the time as compared to 18 percent of the time in the SPI Treatment. These differences are statistically significant, based on a probit regression of an indicator that is 1 if the seller challenges and 0 otherwise on the treatment variable (p-value < 0.01).

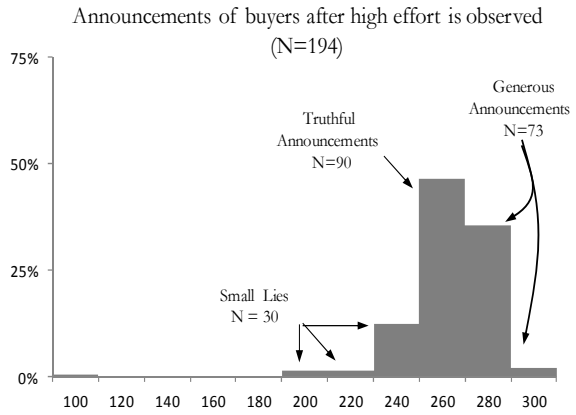
Despite the apparent increase in effort and decrease in small lies, retaliation is still frequent in our data. Panel (c) shows that buyers reject the vast majority of legitimate challenges after both high and low effort (80 percent after high; 75 percent after low), just as in the SPI Treatment. Thus, while the High-Benefits treatment increases truth-telling and the proportion of appropriate challenges, it does not reduce retaliation.

Taken together, the High-Benefits treatment has a larger proportion of truthful announcements and higher effort than the SPI Treatment. However, the losses that occur due to disagreement in early periods of the experiment are larger than the gains that occur from improvements in effort and therefore the mechanism continues to reduce overall pecuniary payoffs. Looking at the first five periods of the experiment, for example, the average total surplus of a dyad pair is -7.9. Relative to the guaranteed gains of 90 for a pair without the mechanism and the potential surplus of 140 with the mechanism, the realized gains from the mechanism of $\frac{-7.9-90}{140-90} = -196\%$ is strongly negative. The mechanism performs better in periods 6–10 where the average total surplus of a dyad pair is 97.9 (a realized gain of 16 percent).

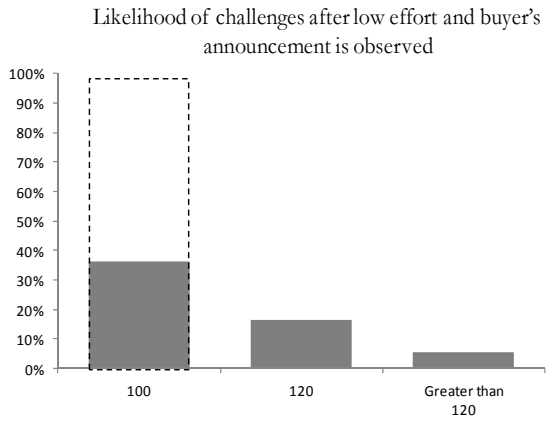
Given that players realize positive gains toward the end of the first phase of the experiment, we might expect that buyers and sellers are more likely to opt into the mechanism in

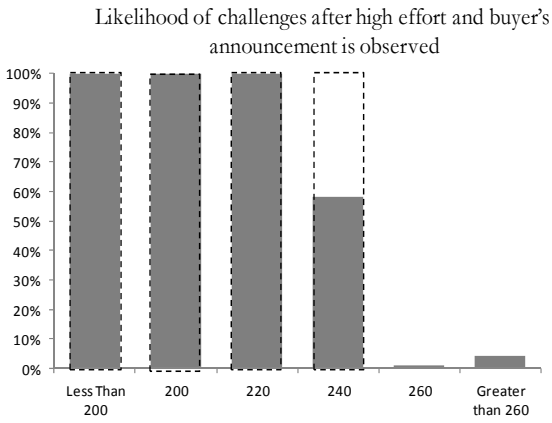
(a) Distribution of announcements after low and high effort





(b) Likelihood of a challenge after each announcement





(c) Number of challenges accepted and rejected

Number of challenges accepted and rejected
after low effort, given announcement and a seller challenge

Announcement	Challenge Accepted	Challenge Rejected
100	1	3
120	0	6
140	0	1

Grey boxes are predicted action by SPI Hypothesis

Number of challenges accepted and rejected
after high effort, given announcement and a seller challenge

Announcement	Challenge Accepted	Challenge Rejected
Less than 200	1	0
200	1	2
220	0	3
240	3	11
260	0	1
Greater than 260	0	3

Grey boxes are predicted action by SPI Hypothesis

Figure B2: Pattern of Play in High-Benefits Treatment

this treatment. However, we find no significant difference in the overall opt-out rates in the second phase of the experiment.

Result B.3 *In a majority of cases, the parties do not adopt the mechanism. This is largely due to the buyers' dismissals of the mechanism which stems from the buyers' high propensity to render the mechanism unprofitable by making generous announcements. Generous announcements are more likely to be made by buyers who believe that truthful announcements may be challenged.*

Panel (a) of Figure B3 shows the opt-out behavior of buyers and sellers over the ten periods of the treatment. As can be seen, the buyer's opt-out rate is 81 percent in period 11 and converges to 62 percent by period 20. The buyer's average opt-out rate of 65 percent is higher but not significantly different from the buyer's average opt-out rate of 58 percent in the SPI Treatment (p-value = 0.46). The seller's opt-out rates in the High-Benefits Treatment is low at 3.4 percent, suggesting that the high opt-out rate is primarily due to the dismissal of the mechanism by buyers.

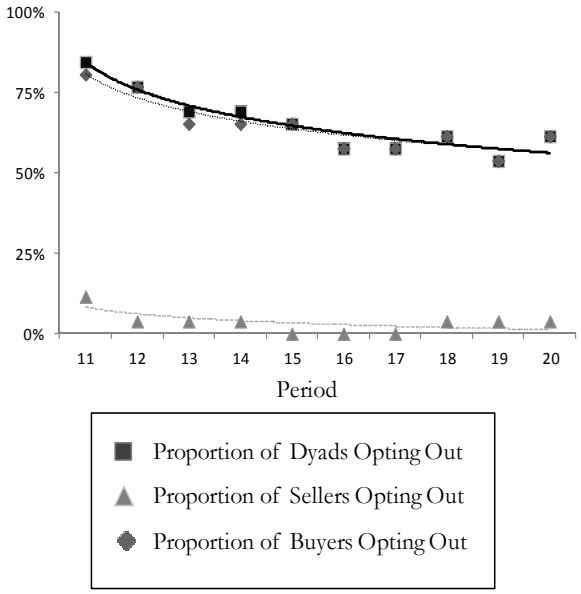
As with the SPI Treatment, in periods without the mechanism, the hold-up problem is unresolved. As seen in panel (b), when either party opts out of the mechanism, 154 out of 171 sellers exert low effort. In 134 of these cases the buyer announces $\hat{v} = 100$. Of the 17 observations where high effort is observed, the buyer announces a $\hat{v} \leq 180$ in 9 of them.

For those periods in which both subjects opted in, high effort is observed in 79 out of 89 cases. Buyers who keep the mechanism make truthful announcements in 46 cases, generous offers in 25 cases, and small lies in only 8 cases. The increase in truthful announcements and generous offers results in only 2 challenges and raises the overall average surplus of a buyer and seller pair to 108.8 relative to 95.0 when the arbitrator is dismissed. However, the increase in average efficiency is enjoyed primarily by the sellers; looking at buyers' profits in isolation, buyers' expected profits actually decrease from 76.9 when the mechanism is dismissed to 71.1 when the mechanism is kept. Thus the decrease in the seller's opt-out rate and the lack of change in the buyer's opt-out rate can be explained in part by an asymmetric return on the mechanisms adoption.

The asymmetric return to the adoption of the mechanism is due primarily to the buyers' generous announcements. Relative to the SPNE without the mechanism where low effort is exerted and the buyer announces a value of 100, the SPNE with the mechanism available leads to an increase in the buyer's payoff of 20 and an increase in the seller's payoff of 30. When a buyer makes a generous offer, however, he effectively transfers a large portion of the potential gains from the mechanism back to the seller. These transfers make the mechanism unattractive to buyers from an expected value standpoint.

Why do the buyers behave in a manner that renders the mechanism unprofitable for them? One likely reason for buyers' generous offers is that they have pessimistic beliefs about the likelihood of challenges by the seller after a truthful announcement. While sellers challenge truthful announcements very rarely (1 out of 90 cases after high effort; 6 out of 37 cases after low effort), a buyer who believes that truthful announcements may be challenged may choose to make a generous offer as a way of reducing the probability of a challenge. Our belief data support the hypothesis that buyers have such pessimistic beliefs. In comparison to the distribution of beliefs in the SPI Treatment where 66 percent of buyers believed that

(a) Proportion of buyers and sellers opting out of mechanism each period



(b) Buyer and seller outcomes with and without SPI mechanism

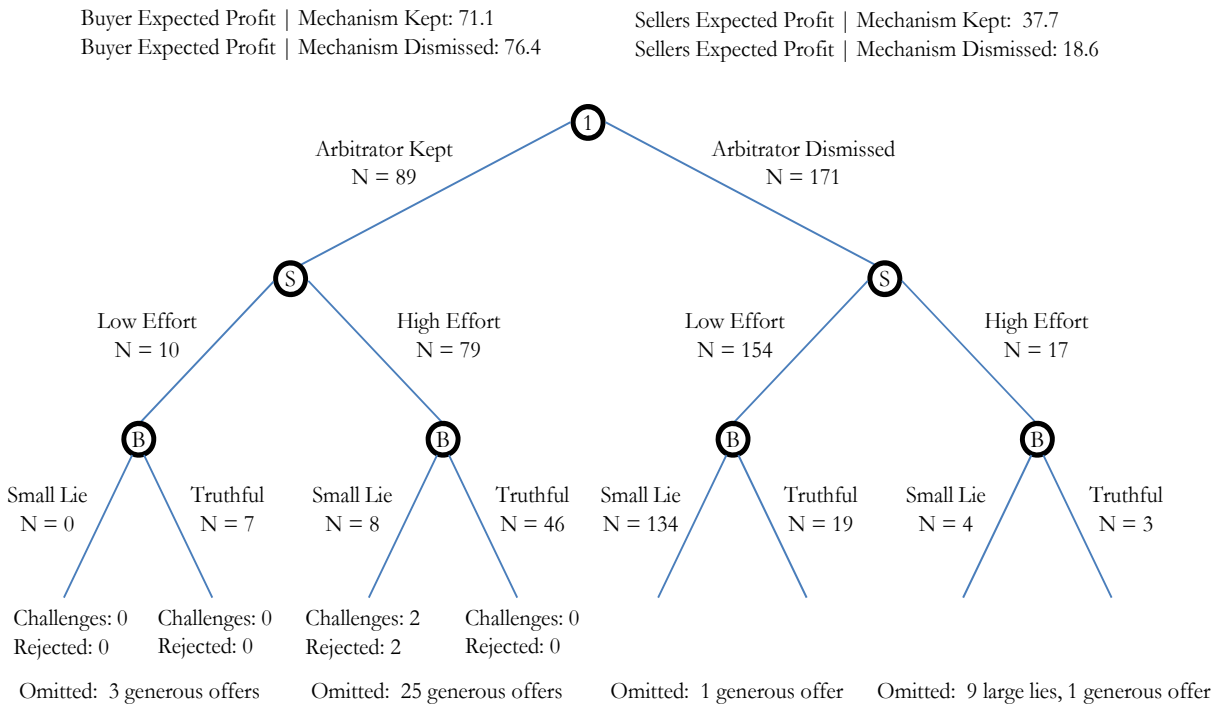


Figure B3: Behavior in Last 10 Periods of High-Benefits Treatment

a truthful announcement would never be challenged after high effort, only 39 percent of buyers in the High Benefits Treatment believe that truthful announcements would never be challenged.

The shift in pessimism and the fear of inappropriate challenges in the High-Benefits treatment was not expected when we designed the treatment but it is consistent with buyers believing that at least some sellers dislike unequal allocations of surplus. Unlike the SPI Treatment where buyers and sellers enjoyed an equal split of surplus along the equilibrium path, the High-Benefits treatment reduces the price that occurs without a challenge and gives the buyer a payoff of 90 while the seller receives 50. If buyers believe that sellers have a distaste for such unequal allocations, they may make generous offers which lead to more equitable surplus splits. Thus buyers' beliefs about the distribution of other-regarding preferences in the population could explain the fear of inappropriate challenges.[2]

To better understand the role that beliefs have in making generous announcements we look at how decisions of buyers to make generous announcements depend on his belief about being challenged after truthful announcements. Table B4 reports the results of a probit regression where the dependent variable is 1 if a buyer makes a generous offer and 0 if the buyer makes a truthful announcement. We regress this generous offer variable on the buyer's belief about being challenged after a truthful announcement. Column (1) restricts the sample to high effort, column (2) restricts the sample to low effort, and column (3) uses the combined sample.

Beliefs about the likelihood of being challenged are a good predictor of the buyer's likelihood of making a generous announcement. Based on the marginal effects of a probit regression, buyers are 68.6 percentage points more likely to make a generous offer after high effort if they believe that being challenged is "Likely" relative to individuals who believe that challenges of truthful announcements will "Never" occur. Likewise, they are 99.5 percentage points more likely to make a generous offer after low effort if they believe that truthful announcements are "Likely."

In aggregate, the High-Benefits treatment does indeed increase the probability of truthful announcements and decrease the probability of small lies. However, the buyers' pessimistic beliefs regarding the potential of being challenged leads them to make generous offers which shift surplus away from the buyer and toward the seller. This shift in surplus eliminates the buyers' incentives to use the mechanism and ultimately leads buyers to dismiss the mechanism when the mechanism is voluntary.

B3. Low-Fine Treatment

While the High-Benefits treatment improved truth-telling and increased the challenging of small lies, it did not directly attempt to deal with violations in the counter-offer condition. In this section we look at how reductions in the fine F might reduce the buyers desire to reciprocate and potentially improve the performance of the mechanism.

The large fine in the SPI Treatment was chosen as we were interested in testing the general application of the SPI mechanism to a broad set of social choice functions. As many

[2]Note that buyers themselves do not appear to care about equity. When the mechanism does not exist generous offers are detected in only 2 of 171 cases.

Table B4: Probit Regression of Generous Announcements by Buyer

	High Effort (1)	Low Effort (2)	Combined (3)
Buyers Belief that Seller Will Challenge Truthful Announcement:			
"Unlikely"	0.483 ***	0.929 ***	0.417 ***
	(0.136)	(0.049)	(0.135)
"Likely"	0.686 ***	0.995 ***	0.678 ***
	(0.081)	(0.002)	(0.080)
"Always"	0.503 ***	0.965 ***	0.498 ***
	(0.190)	(0.016)	(0.166)
Psuedo R^2	0.253	0.249	0.195
Observations	164	55	219

Marginal effects from a probit regression are reported in the table where the dependent variable is 1 if buyer makes a generous announcement and 0 if buyer makes a truthful announcement. Standard errors in parentheses, clustered by individual. The ommited category is Seller "Never" Challenges. Regression (1) restricts the sample to observations with High Effort. Regressions (2) restricts the sample to observations with Low Effort. *,**,*** denote significance at the 10%, 5%, 1%-level, respecively.

applications hinge on the assumption that fines can be made arbitrarily large, we selected a fine that was large as we expected this to increase the incentives of buyers to be truthful. However, for the particular hold-up problem explored in the experiment, a smaller fine could also implement the first best in theory. If the mechanism functions better with a smaller fine, then our results would suggest that subgame-perfect implementation may work for problems where the fines can be kept low but may be unsuitable for cases where they are required to be very high.

There are a number of reasons to suspect that the buyer's retaliation factor may be increasing in F. First, as F goes up, the buyer's losses due to a challenge increase. If the buyer's return for retaliation scales with the amount he is harmed by a challenge, reducing F should reduce his incentive to retaliate. Second, as F goes up, the amount that the buyer can hurt the seller by retaliating also increases. Thus, when the fine is lower, the amount of the sellers profit that can be destroyed by retaliation is declining. Taken together, this may well imply that a lower fine is associated with lower psychological returns to retaliation.

To explore whether a reduced fine reduces retaliation and improves the sellers' incentives to challenge small lies, we ran an additional **Low-Fine Treatment** in which we used the same initial-price and counter-offer schedules as the High-Benefits treatment, but with the fine set at 80 rather than 250. Payoffs for this treatment are the same as in Table B3. The resulting mechanism still satisfies the Counter-Offer, Appropriate-Challenge, and Truth-Telling conditions. Our Low-Fine treatment consists of two sessions, each with 20 subjects. We find the following.

Result B.4 *In the Low-Fine treatment, sellers' effort choices and the buyer's likelihood*

of making a small lie or a truthful announcement are similar to the High-Benefits Treatment. However, following high effort, a large proportion of buyers make the lowest possible announcement, $\hat{v} = 100$. These "maximal lies" are more frequent among buyers who are averse to gambles and who fear inappropriate challenges. Sellers always challenge maximal lies and buyers who are challenged after a maximal lie almost always accept the counter offer. Sellers almost always challenge small lies and buyers still retaliate against the majority of these challenges.

Figure B4 displays the results for the Low-Fine treatment with data aggregated across all 10 periods. The figure shows that sellers exert high effort in 158 out of 200 cases (79 percent), a rate that is similar to the effort rates found in the High-Benefits treatment (75 percent). The small difference in these effort rates is not significantly different in a regression of effort choice on the treatment dummy (p-value = 0.55).

Panel (a) shows that buyers make a small lie in only 16 out of 158 cases after high effort and 11 out of 42 times after low effort. The aggregate small lie rate of 14 percent is similar to that found in the High-Benefits treatment (16 percent) and not significantly different in a probit regression where a dummy, which is 1 when an individual makes a small lie and 0 when he makes any other announcement, is regressed on the treatment dummy (p-value = 0.64). Buyers make truthful announcements in 23 of 42 cases after low effort and 46 of 158 cases after high effort. This aggregate truth-telling rate of 35 percent is lower than the 49 percent found in the high benefits treatment, but not significantly different using the same specification as above (p-value = 0.14).

There are, however, striking differences in the announcement distribution between the Low-Fine Treatment and the High-Benefits treatment. After high effort, buyers in the Low-Fine treatment make maximal lies in 65 out of 158 cases (41 percent) and make generous offers in only 25 out of 158 cases (16 percent). This contrasts strongly with the maximal lie rate of 1 percent and generous offer rate of 38 percent in the High Benefits Treatment. We discuss these maximal lies in detail after describing actions in the other stages of the game.

Seller's challenge rates in the Low-Fine treatment are very high, with all small lies and all maximal lies challenged after high effort and 82 percent of small lies challenged after low effort. The challenge rates of lies is significantly higher than the High-Benefits treatment in a probit regression where sellers' challenges are regressed on the treatment effect and the sample is restricted to lies or small lies (all lies: p-value < 0.01; small lies: p-value $0 < .01$). The challenge rate of truthful announcements is higher in the Low-Fine treatment, but not significantly different using the same probit specification with the sample restricted to truthful announcements (p-value = 0.11).

Looking at the acceptance rate of counter offers shown in panel (c), in 65 of the 68 case where the buyer made large lies and were challenged, the buyer accepted the counter offer. Looking at the beliefs of the subset of 65 buyers who made maximal lies, 69 percent believed they would "Always" be challenged and the remaining 31 percent believed they were "Likely" to be challenged. Thus, it appears that individuals who made these maximal lies expected to be challenged and expected to receive the payoff of 75 from this action.

Challenges of small lies are rejected in 9 out of 16 cases after high effort and in 7 out of 9 cases after high effort. While the aggregate rejection rate of challenges after small lies of 64 percent is 15.2 percentage points lower than the High-Benefits treatment, the difference

Table B5: The Relationship Between Maximal Lies and Aversion to Gambles.

	Averse to Fair Gambles	*Accept Fair Gambles*
Truthful Announcement	34	12
Maximal Lies	64	1

is not significant in a probit regression that regresses the acceptance rate of small lies on the treatment (p-value $= 0.29$). This suggests that retaliation has not been fully resolved in this treatment.

Why do the buyers in the Low-Fine treatment lie so often maximally? As with the generous offers in the High-Benefits Treatment, a likely reason for maximal lies is a fear that a truthful announcement would be challenged. An individual who makes a truthful announcement and will reject an inappropriate challenge will receive 90 if he is not challenged and -80 if he is challenged. By contrast, even if a maximal lie is always challenged, an individual making a maximal lie is guaranteed a profit of at least 75. As this is equal to the value an individual will get for making a generous offer of 280 after high effort and not being challenged, an individual who fears that a truthful announcement will be challenged has strong incentives to make a maximal lie.

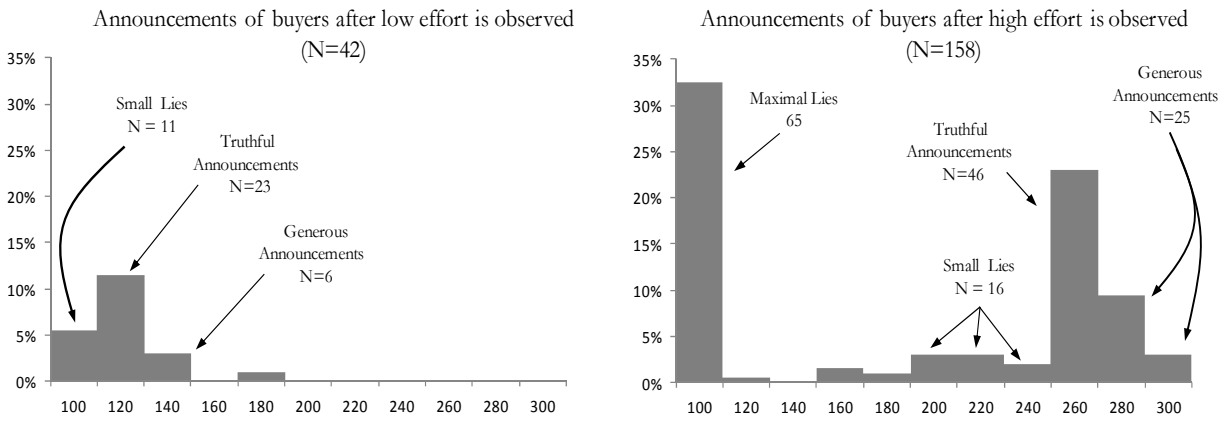
The hypothesis that fear of inappropriate challenges leads to maximal lies is supported by two pieces of evidence. First, for buyers who believed that they would never be inappropriately challenged, maximal lies occur in 28 percent of the observations. For buyers who believe that inappropriate challenges were "Unlikely," "Likely," or would "Always" occur, maximal lies occurred in 48 percent of the observations. Thus, those with higher beliefs of being inappropriately challenged were substantially more likely to make maximal lies.

The hypothesis is further corroborated by relating the likelihood of a subject to make a maximal lie to our secondary measure of aversion to gambles. Using data from our follow-up lottery treatment, we divided subjects into two categories: those who accepted the gamble the 50-50 gamble of winning \$12 or losing \$10 and those who rejected it. Table B5 shows the number of observations in which sellers exerted high effort and buyers announced either a maximal lie or the truth. buyers who do not exhibit an aversion to fair gambles are more likely to announce truthfully than to make a maximal lie, while those who are averse to fair gambles are more likely to make a maximal lie. These differences are significant in a probit regression where we regress a binary variable that is 1 if the buyer makes a maximal lie after high effort and 0 if the buyer makes a truthful announcement after high effort on a binary variable of risk preferences that is 1 if the buyer accepts the gamble and 0 if he rejects it (p-value < 0.01).

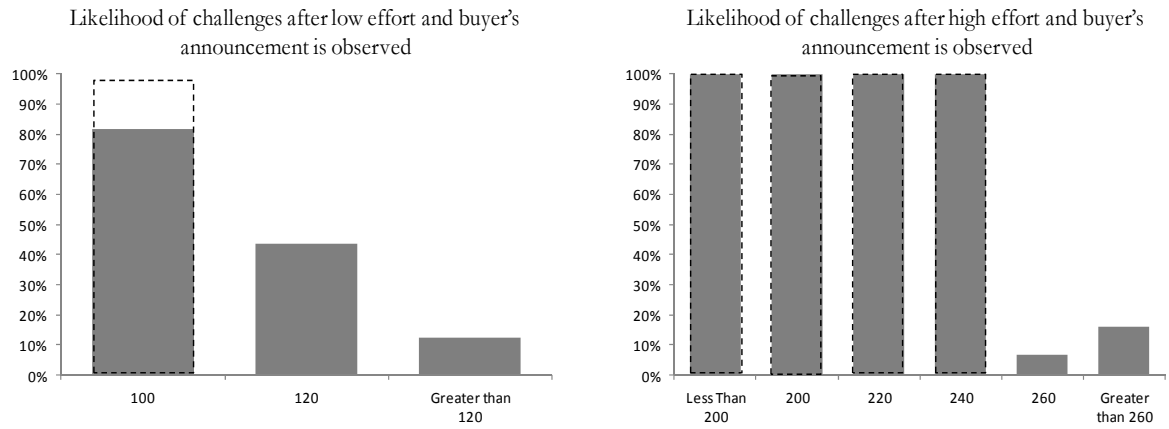
As with the High-Benefits treatment, buyers in this treatment take strategic actions that shift surplus from buyers to sellers in the mechanism due to the fear of inappropriate challenges. We would thus expect similar opt-in and opt-out behavior in the second part of the experiment.

Result B.5 *Buyers in the Low-Fine Treatment opt out of the mechanism in the majority of cases and in similar proportions as seen in the High-Benefits treatment. This aversion to the mechanism appears to be due to a fear that sellers will challenge truthful announcements.*

(a) Distribution of announcements after low and high effort



(b) Likelihood of a challenge after each announcement



(c) Number of challenges accepted and rejected

Number of challenges accepted and rejected
after low effort, given announcement and a seller challenge

Announcement	Arbitration Accepted	Arbitration Rejected
100	2	7
120	0	10
140	1	0

Grey boxes are predicted action when buyers do not retaliate

Number of challenges accepted and rejected
after high effort, given announcement and a seller challenge

Announcement	Arbitration Accepted	Arbitration Rejected
Less than 200	68	3
200	4	2
220	1	5
240	2	2
260	0	3
Greater than 260	0	4

Grey boxes are predicted action when buyers do not retaliate

Figure B4: Pattern of Play in Low-Fine Treatment

Buyer opt-out behavior is almost identical to that in the High-Benefits treatment with opt-out rates converging to 60 percent from above with an initial opt-out rate of 85 percent. The average opt-out rate of 64 percent in the Low-Fine treatment is not significantly different to the average opt-out rate of 65 percent in the High-Benefits treatment (p-value = 0.96) Sellers's opt-out rate of 4 percent is also not significantly different to the opt-out rate in the High-Benefits treatment (p-value = 0.97). Buyers who retain the mechanism have an average return of 59.1 while buyers who opt out of the mechanism have an average return of 74.1. This loss of profit from buyers who retain the mechanism is due primarily to maximal lies and generous offers which transfer surplus to seller.

Taken together, the Low-Fine treatment shares strong similarities to the High-Benefits treatment. Many buyers who fear that truthful announcements will be challenged make maximal lies which guarantee a payoff of 75 rather than making truthful announcements. This deviation transfers profit from the buyer to the seller thereby eliminating their monetary incentive to enter into the mechanism.

B4. The No-False-Challenge Treatment

In the High-Benefits treatment, we found that a fear of inappropriate challenges was a potential driver for the buyers' generous announcements. Here we report on an additional control treatment that eliminates the ability of sellers to challenge buyers when he has made a truthful announcement. Such a mechanism would not be feasible in practice, because it requires that the sellers action space following an announcement depends on whether the announcement was truthful. However, here it helps to understand the extent to which deviations from truth-telling are due to a fear of inappropriate challenges.

In the follow-up **No-False-Challenge Treatment**, we use an identical parametrization to the High-Benefits Treatment but augment the mechanism with the following rules: if after observing low effort the buyer announces the true value of 120, he cannot be challenged, and the game ends. Likewise, after observing high effort, if the buyer announces the true value of 260, he cannot be challenged, and the game ends. We conducted 3 sessions of the No-False-Challenge Treatment with 22, 24, and 26 subjects respectively in these sessions.

Figure B5 shows the proportion of generous and truthful announcements in the High-Benefits treatment and the No-False-Challenge treatments for both low and high effort along with 95 percent confidence intervals clustered by individual. As can be seen, after both high and low effort, there is a dramatic decrease in generous offers and a significant increase in truthful announcements in the No-False-Challenge treatment. The treatment effects is also significant in a probit regression that regresses a binary variable that is 1 if an individual makes a generous announcement and 0 if an individual makes a truthful offer on the treatment (p-value < 0.01, errors clustered by individual).

Sellers' challenge behavior is similar in the two treatments with 59 percent of small lies being challenged in the High-Benefits treatment and 59 percent of small lies being challenged in the No-False-Challenge treatment. The buyers' willingness to reject the challenges of small lies are also similar with 79 percent of challenges being rejected in the High-Benefits treatment and 87 percent of challenges being rejected in the No-False-Challenge treatment. Neither difference is significant (Sellers Challenge Behavior: p-value = 0.97; Buyers Rejection Behavior: p-value = 0.55).

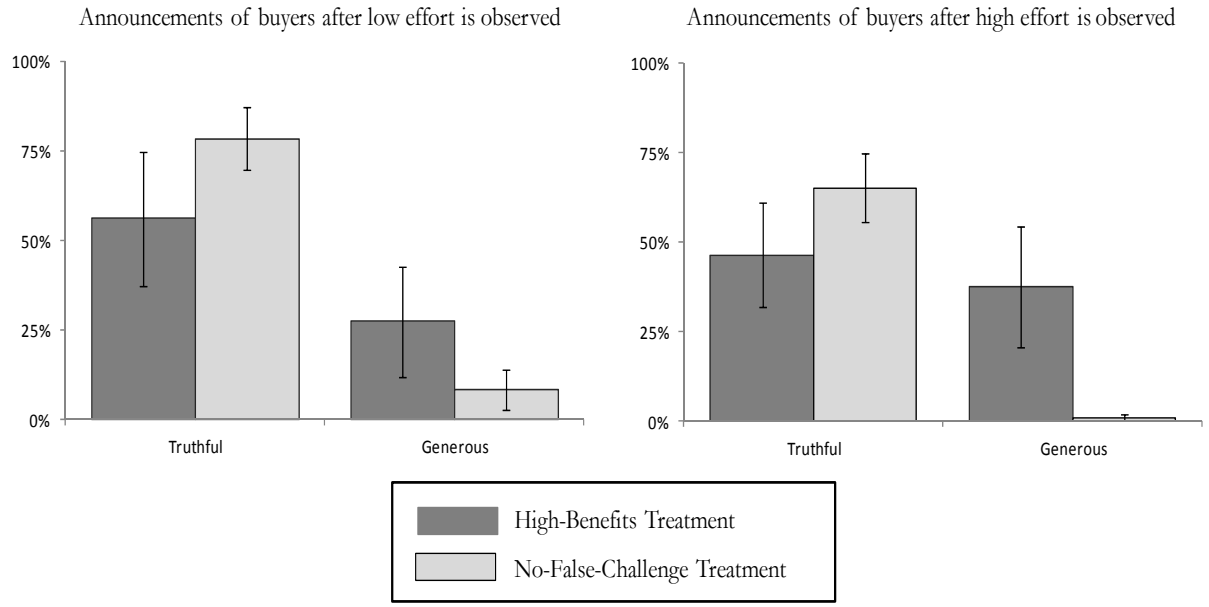


Figure B5: Comparison of Truthful Announcements and Generous Announcements in the High-Benefits and the No-False-Challenge Treatments

Given that there are less generous offers in the No-False-Challenge treatment, one might conjecture that individuals would be less likely to opt out of the mechanism. This turns out not to be the case: While buyer's opt-out rate declines from 65 percent in the High-Benefits treatment to 52 percent in the No-False-Challenge treatment, seller's opt-out rate increases from 4 percent to 10 percent. Thus, on net, the overall increase in retention rates is small (66 percent vs 58 percent) and not significant (p-value $= 0.46$).

Overall, the No-False-Challenge treatment supports the conjecture that a fear of being challenged after an appropriate challenge is a major cause of generous announcements in the High Benefits treatment. We find that the No-False-Challenge Treatment eliminates generous offers in periods where high effort occurs and significantly increases truthful announcements by the buyers. However, the proportion of buyers opting into the mechanism improves only slightly and the proportion of sellers opting into the mechanism decreases. This suggests that it is hard to satisfy both parties concerns about the mechanism simultaneously. We leave further study of mechanisms such as this one (what John Moore (1992) calls "simple sequential mechanisms") to future research.

Appendix B5. The SPI with Intense Training Treatment

A natural hypothesis for the observed pattern of play in the SPI treatment is that subjects make errors in choosing which pure action to play and that they are more likely to choose pure actions that involve higher expected payoffs. In extensive-form games, a useful way to model such errors is with an Agent Quantal Response Equilibrium (AQRE). AQRE is similar to a standard quantal response model with the additional assumption that at a given

decision node, the player determines the expected payoff of each action by treating their future self as an independent player with a known probability distribution over actions.

In an AQRE, the rejection of counter offers after small lies can be partially explained by noting that the expected utility of accepting and rejecting a challenge are similar. Relative to larger lies (where the difference between accepting and rejecting a challenge is large), AQRE predicts that buyers are more likely to reject challenges after a small lie. Forecasting the errors of buyers, sellers may be less likely to challenge small lies. Likewise, buyers who correctly forecast sellers reluctance to challenge may be more likely to make small lies. Thus, the introduction of errors can generate deviations that are directionally consistent with a major feature of the data.

While the structure of AQRE can match portions of the pattern of play, it cannot match the magnitude of rejections. In any QRE model with symmetric noise, a choice that has higher expected utility must be chosen with higher frequency than one with a lower expected utility. Since accepting an appropriate challenge generates higher returns by construction, the maximum rejection rate that can be predicted is 1/2. Given that 94.4 percent of appropriate challenges were rejected after high effort and a small lie, AQRE on its own has a hard time fully rationalizing the data. Level-k and other cognitive hierarchy models have a similarly difficult time fitting the extent of rejection by buyers since only type-0 individuals will reject an appropriate challenge.

Although AQRE itself cannot explain the large number of rejections, mistakes and reciprocity could potentially interact in subtle ways. For example, noisy behavior increases the likelihood that buyers experiment with non-truthful announcements. If these buyers find that small lies are not challenged, they are likely to continue to make them and their behavior will look similar to the reciprocal types. Alternatively, an individual who enters into arbitration due to a mistake may be more upset by a challenge than an individual who lies due to strategic considerations. This implies that the observed willingness to retaliate may depend on the propensity of buyers and sellers to make mistakes.

To help separate noise from reciprocity, we ran an additional **SPI with Intense-Training Treatment** consisting of 4 sessions and 80 subjects. This treatment used the same mechanism and parametrization as the SPI Treatment, but extended the instructions phase of the experiment for the purpose of minimizing subjects' mistakes and maximizing their understanding of the logic behind the mechanism. The intense training protocol went beyond the typical way of making subjects familiar with the payoff structure of a game. In our original instructions for the SPI mechanism (i.e., the standard training protocol) we thoroughly explained the mechanics of the mechanism and the payoff consequences of different sequences of actions. However, the mechanisms have some complexity such that mistakes may still occur — in particular, mistakes in understanding the counterparties' pecuniary incentives. The intense training protocol was therefore designed to minimize subjects' mistakes and maximize the understanding of both their own pecuniary incentives and *the pecuniary incentives of their counterparty at each stage of the mechanism.*

We achieved this with two additional features. *First*, we explicitly explained in the written instructions the pecuniary incentives of subjects' counterparties in the trade. For example, the buyers were explicitly informed that if they announce the true value of the good and are willing to reject counteroffers if the seller nevertheless challenged their truthful report, it is in the seller's pecuniary interests to refrain from challenging them. Likewise, the

sellers were explicitly informed that if they challenge a buyer's lie, then it is in the buyer's pecuniary interest to accept the counteroffer.

Second, before subjects played against a human partner, they played for six periods against a computerized opponent that was programmed to play the SPNE actions as if they had selfish preferences. By playing an opponent who maximizes the pecuniary return, subjects learned to understand the pecuniary incentives of their opponents in a practical way. The first three of these periods were unpaid while periods 4, 5, and 6 were paid. Note that by playing both unpaid and paid periods against the computer, we first gave subjects the opportunity to experiment with potential strategies against an opponent that always punished lies and false challenges and avoided cases where a player was mistakenly rewarded for deviating from the SPNE. Further, it allowed players to experiment without affecting the beliefs of human partners.

Following the computer rounds, subjects were reminded that from now on (i.e., in Phase 1), they were no longer playing against a computer and that they would be matched with a different person in the room for each of the next 10 periods. All other parts of the instructions were the same as the SPI Treatment.

The intense training protocol produced the following results.

Result B.6 *The SPI with Intense-Training Treatment has a larger proportion of sellers who exert high effort than the SPI Treatment. It also has fewer small lies and sellers are more likely to challenge these lies. However, small lies remain common and buyers still retaliate against most challenges, leading to inefficiency. Thus, although the SPI with Intense-Training Treatment improves the efficiency of the mechanism relative to the SPI Treatment, the mechanism's efficiency still remains low.*

Figure B6 displays the results of the SPI with Intense-Training Treatment with data aggregated across the 10 periods of Phase 1. The left hand side of the figure follows the pattern of play after sellers selects low effort ($N = 90$) while the right hand side of the figure follows the pattern of play following high effort ($N = 310$). Directly comparable to Figure 1, panel (a) shows the distribution of announcements, panel (b) shows the likelihood of a challenge after each announcement, and panel (c) shows the frequency that a challenge is accepted or rejected.

Under the intense training protocol a larger proportion of sellers chooses high effort compared to the standard training protocol. In the SPI treatment with standard training, sellers select high effort in only 260 out of 460 observations (57 percent), while sellers in the SPI treatment with intense training choose high effort in 310 out of 400 observations (78 percent). This difference is significant in a simple probit regression where effort choice is regressed on the treatment variable (p-value = 0.01).

Controlling for the difference in effort levels, the SPI with Intense-Training Treatment also has significantly fewer small lies than the SPI Treatment. Panel (a) shows that small lies occur in 28 out of 90 cases after low effort (31 percent) and in 58 out of 310 cases after high effort (19 percent). These small lie rates are low relative to the SPI Treatment where lies occurred 61 percent of the time after low effort and 54 percent of the time after high effort.[3] However, the lie rate in the SPI with Intense-Training Treatment is still high relative to the predictions of no lies made in SPI Hypothesis 1.

[3]The difference in the propensity to make small lies between the two treatments is statistically significantly

Looking at the right side of panel (b), sellers who exert high effort in the SPI with Intense-Training Treatment challenge small lies 72 percent of the time. This is significantly higher than the challenge rate of 26 percent observed in the SPI Treatment with standard training based on a simple probit regression where a binary variable that is 1 for a challenge and zero for a no challenge, is regressed on the treatment variable (p-value = 0.01). As seen on the left side of panel (b), sellers who exert low effort in the SPI with Intense-Training Treatment challenge small lies only 11 percent of the time. This is not significantly lower than the challenge rate of 22 percent observed in the SPI Treatment (p-value = 0.10).

Despite the apparent increase in effort and decrease in small lies, retaliation is still frequent in our data. Panel (c) shows that buyers reject a large proportion of legitimate challenges after high and low effort (49 percent after high; 100 percent after low), just as in the SPI Treatment with standard training. Thus, while the SPI with Intensive-Training Treatment increases truth-telling and the proportion of appropriate challenges, it does not reduce retaliation. In addition, small lies are still relatively common and the high challenge rate leads to a large number of disagreements that continue to reduce overall pecuniary payoffs. The average payoff of a buyer-seller pair was only 54.5, well below the guaranteed gains of 90 for a pair without the mechanism and the potential surplus of 140 that could be achieved with an efficient mechanism. Normalizing the actual gain generated by the mechanism by the predicted gain of the mechanism, the realized gain from the mechanism is $(54.5 - 90)/(140 - 90) = -71\%$. There is also no improvement in efficiency over time. The average payoff for a group in periods 1–5 was 62.0 while the average payoff for groups in periods 6–10 was 47.0. The average payoff for a group in periods 1–5 was 62.0 while the average payoff for groups in periods 6–10 was 47.0.

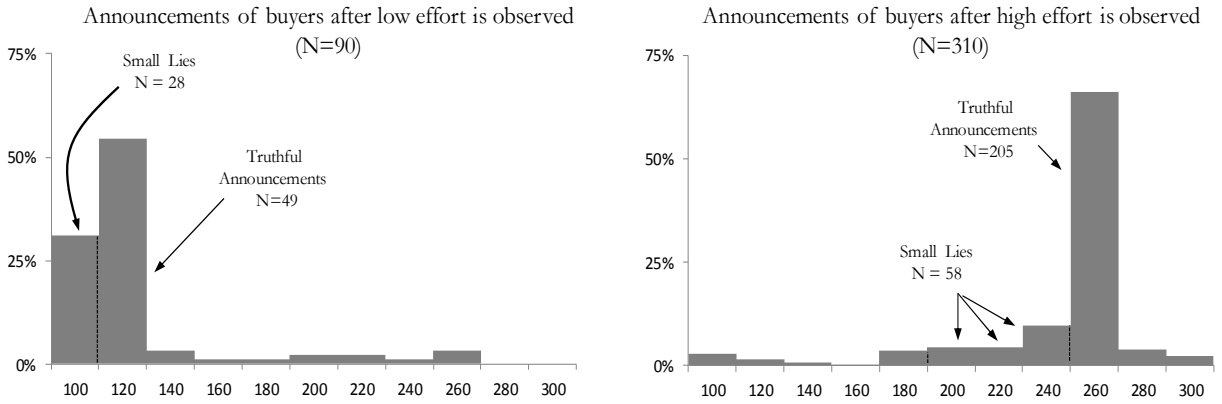
As with the SPI Treatment, buyers and sellers in the SPI with Intense-Training Treatment earn less with the mechanism than is guaranteed without the mechanism. We would thus expect similar opt-in and opt-out behavior between the two treatments.

Result B.7 *In the majority of cases, the parties do not adopt the mechanism in the SPI with Intense-Training Treatment. This is largely due to buyers opting out of the mechanism. There is no significant difference in opt-out rates between the SPI Treatment and the SPI with Intesive-Training Treatment.*

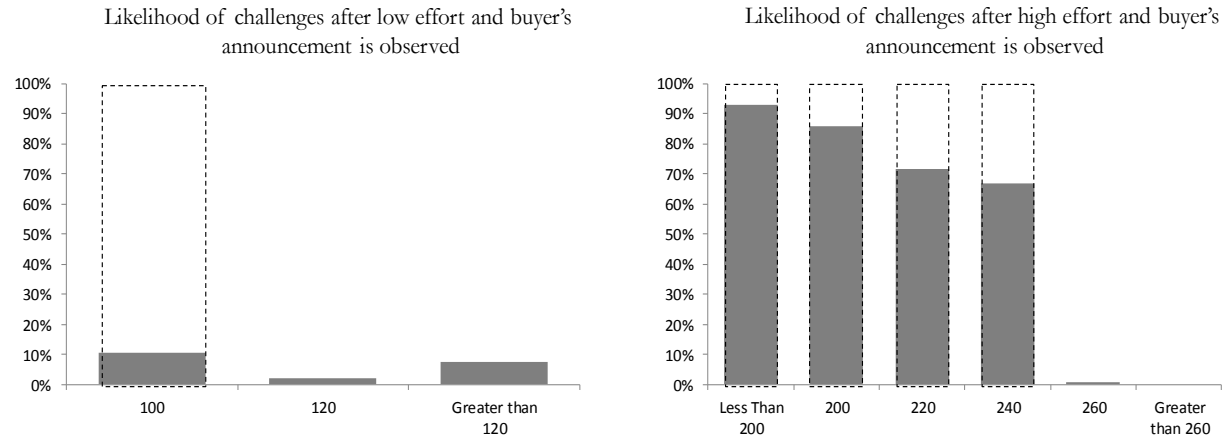
Buyers opt-out of the mechanism 57 percent of the time while sellers opt-out 19 percent of the time. These opt-out rates are not significantly different to the buyers' (58 percent) and sellers' (16 percent) opt-out rates in the SPI treatment with standard training (based on a simple probit regression that regresses the opt-in rate on the treatment (p-value = 0.96 for the buyer; p-value = 0.76 for the seller). Buyers who retain the mechanism have an average return of 38.7 while buyers who opt out of the mechanism have an average return of 56.1. In groups where the mechanism is retained, small lies are still reasonably common and occur in 14 out of 36 cases after low effort (39 percent) and in 13 out of 105 cases after high effort (12 percent). Disagreements that occurred after these small lies were the main reason for the reduced profits for the buyers.

different in two separate probit regressions — one for low effort and one for high effort — where a binary variable that is 1 for a small lie and 0 for a truthful announcement is regressed on the treatment variable (p-value < 0.01 for low; p-value < 0.01 for high).

(a) Distribution of announcements after low and high effort



(b) Likelihood of a challenge after each announcement



(c) Number of challenges accepted and rejected

Number of challenges accepted and rejected after low effort, given announcement and a seller challenge

Announcement	Challenge Accepted	Challenge Rejected
100	0	3
120	0	1
Greater Than 120	0	1

Grey boxes are predicted action by SPI hypothesis

Number of challenges accepted and rejected after high effort, given announcement and a seller challenge

Announcement	Challenge Accepted	Challenge Rejected
Less than 200	16	10
200	9	3
220	5	5
240	5	15
260	1	2

Grey boxes are predicted action by SPI hypothesis

Figure B6: Pattern of Play in First 10 Periods of SPI with Intense-Training Treatment

Overall, the extended training appears to reduce the propensity of sellers to lie and increases the probability that small lies will be challenged. However, small lies are still frequent enough that the average return of using the mechanism is negative. When given the opportunity, a large proportion of buyers and a small proportion of sellers continue to opt-out of the mechanism.

B6. Personality Measures of Reciprocity

In a previous version of the paper, we explored the implications of private information regarding the reciprocity types of the buyers and sellers. As noted in the main text, when types are private information, low reciprocity-type buyers who would accept the counter offer may try to mimic a high reciprocity type by lying. Since both low- and high-type buyers lie, the sellers may have an incentive to challenge with positive probability. This leads to equilibria in which (a) buyers regularly tell small lies, (b) sellers occasionally challenge such lies, and (c) buyers frequently retaliate against challenges of small lies. This pattern of play was observed in the main treatment.

This section offers further evidence that private information regarding the reciprocity types of buyers and sellers is generating the pattern of play observed in the SPI treatment. We test for a between-subject correlation between a measure of preferences for negative reciprocity and the propensity to make a small lie using data from the Personal Norms of Reciprocity (PNR) survey we conducted two weeks prior to the SPI treatment.[4]

Based on the predictions of the Perfect Bayesian Retaliation Equilibrium we developed in the previous draft, the relationship between negative reciprocity and small lies is expected to be weakly monotonic but potentially non-linear. This is due to two forces that exist in heterogenous models but not in models with a single type. First, in the absence of strategic incentives to mimic other types, the decision to lie is based on a set of threshold conditions where individuals with similar levels of reciprocity will pool on the same announcements. This will lead to discrete jumps in announcements over the type distribution. Second, in any equilibrium where sellers are reluctant to challenge, less reciprocal buyers will want to pretend to be more reciprocal. This mimicry will lead to mixing which implies even non-reciprocal types will lie with positive probability.

Given this potential non-linear relationship, we construct a binary measure of negative reciprocity that is less sensitive to non-linearities in the relationship between negative reciprocity and small lies. The measure is constructed as follows: we first generate a negative reciprocity score constructed by applying principal-component analysis to the PNR survey using the procedures outlined in Perugini et al. (2003). Individuals who are more negatively reciprocal score higher on this measure. We then divide these scores at the median to construct a binary variable that is 0 for less reciprocal individuals and 1 for more reciprocal

[4]We concentrate on the decision to make a small lie rather than the decision to accept or reject counter offers, because the likelihood of being challenged is conditional on the announcement and, as shown below, the announcement is influenced by reciprocity. Thus, the buyers being challenged are a non-random sample. Further, as was seen in panel (c) of Figure **??**, buyers reject the counter offer in 56 of 64 cases after a small lie. We thus have very little variation in acceptance and rejection behavior that could be used to differentiate between types.

individuals.[5]

Table B6 shows the marginal effects of the negative reciprocity measures in an extension of the probit regressions performed in Table B1. As in the earlier regression, the independent variable is a binary variable that is 1 if an individual makes a small lie in the period and 0 if the individual makes a truthful announcement. The regression includes controls for beliefs about (i) the likelihood of being challenged after a truthful announcement and (ii) the likelihood of being challenged after a small lie. These beliefs are coded as categorical data in the same way as in Appendix B2.

Column (1) reports the marginal impact of negative reciprocity on the likelihood of making a small lie in periods where high effort occurs. As can be seen in column (1) individuals who are above the median of the negative reciprocity score are 28.5 percentage points more likely to make a small lie relative to those below the median, a difference that is significant (p-value < 0.01). Column (2) reports the marginal impact of negative reciprocity on the likelihood of making a small lie in periods when Low effort occurs. As in the High effort case, the impact of reciprocity on the propensity to lie is positive. However, it is not significant.

Pooling the data after high and low effort, column (3) shows that negative reciprocity has a significant impact on the likelihood of a small lie in the full sample. Across both high and low effort, individuals who are above the median of the negative reciprocity score are 21.9 percentage points more likely to make a small lie relative to those below the median, a difference that is significant (p-value $= 0.02$).

Table B6: Probit Regression of Small Lies by Buyers

	High Effort (1)	**Low Effort (2)**	**Combined (3)**
Negative Reciprocity Above Median	0.285 ***	0.125	0.219 **
	(0.107)	(0.121)	(0.090)
Controls			
Buyer's Beliefs: Challenges of Smallest Lie	Yes	Yes	Yes
Buyer's Beliefs: Challenges of Truthful Announcements	Yes	Yes	Yes
Psuedo R^2	0.162	0.237	0.152
Observations	230	180	410

Marginal effects from a probit regression are reported in the table. Standard errors in parentheses, clustered by individual. The omitted category is Seller "Never" Challenges. Regression (1) restricts the sample to periods where High effort is chosen. Regression (2) restricts the sample to periods where Low effort is chosen. *,**,*** denote significance at the 10%, 5%, 1%-level, respecively.

We might also expect a strong relationship between the seller's willingness to challenge and his level of negative reciprocity. However, as discussed in the main text, sellers preferences for reciprocity must be very strong in order to be willing to challenge a buyer. Thus, we would predict that the relationship between reciprocity and challenges is likely to be

[5]The results of this section are robust to alternative linear specifications of the negative reciprocity score as well as specifications that use the disaggregated negative reciprocity questions from the survey.

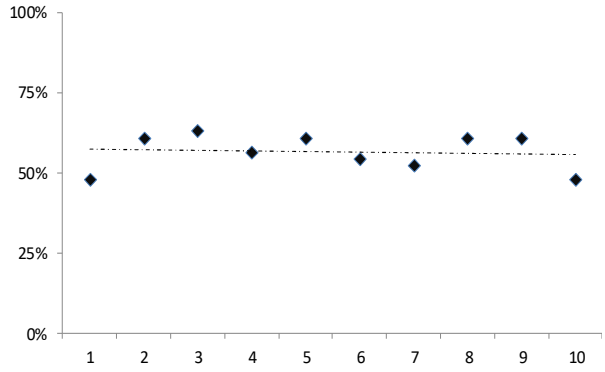
weak. This is indeed the case: extending the probit regression in Table B2 to include negative reciprocity shows that sellers with negative reciprocity scores above the median are not significantly more likely to challenge after high effort (p-value = 0.77), low effort (p-value = 0.83), or in the combined sample (p-value = 0.64).

Appendix C: Additional Figures

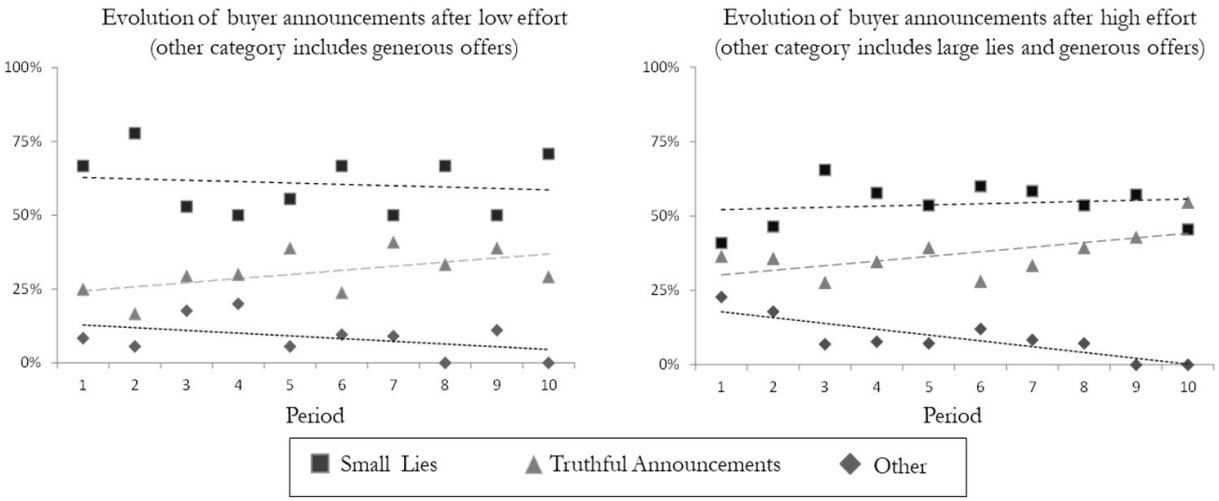
C1. Additional Figures from SPI Treatment

C2. Additional Figures from RS Treatment (Phase 1)

(a) Proportion of sellers exerting high effort in each period



(b) Likelihood of a small lie, truthful announcement, and other announcement in each period



(c) Proportion of small lies challenged each period

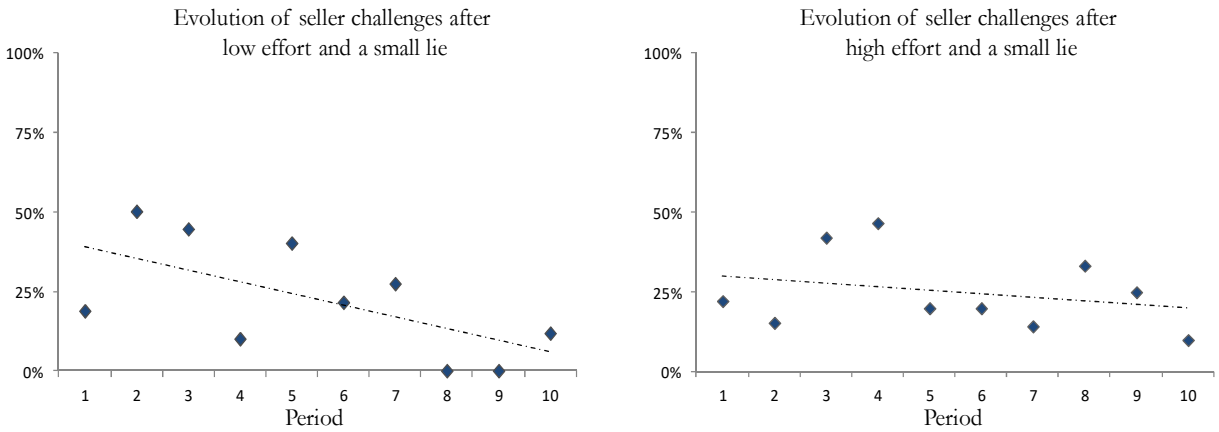
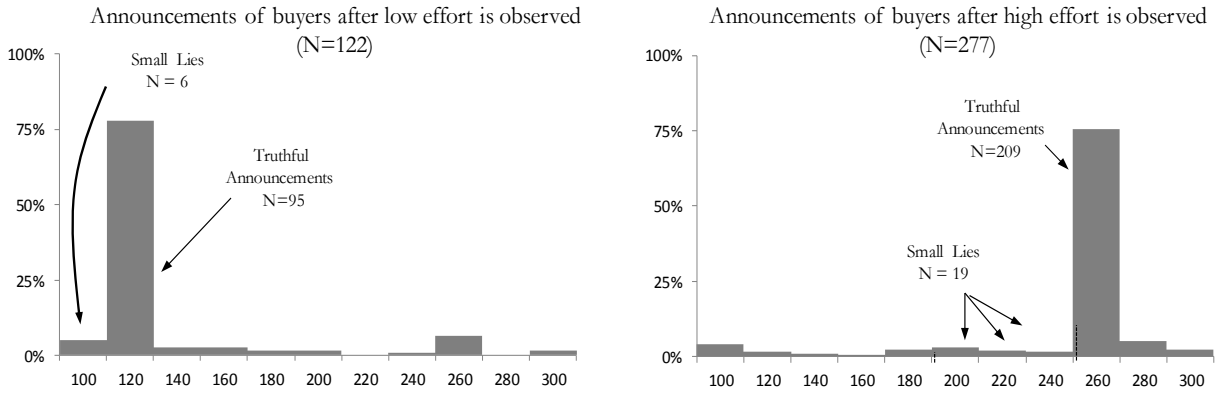
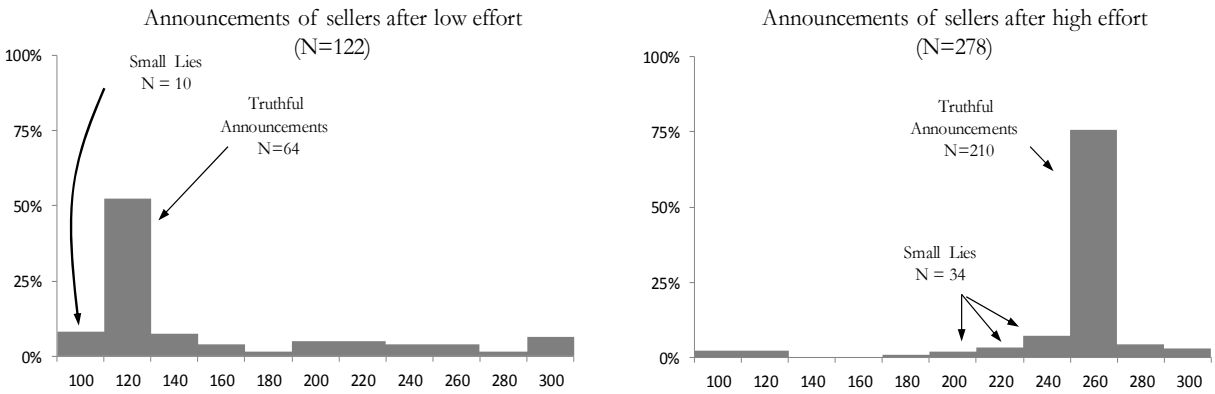


Figure C1: Evolution of Play in First 10 Periods of SPI Treatment

(a) Distribution of buyers' announcements after low and high effort



(b) Distribution of sellers' announcements after low and high effort



(c) Outcomes of groups where buyer and seller reports do not coincide

Outcomes of groups where seller effort is low and buyer and seller reports do not coincide

Cause of Arbitration	Arbitration Stopped by Seller	Arbitration Continued and Buyer Accepts	Arbitration Continued and Buyer Rejects
Buyer Underreport	0	4	2
Other	30	5	32

Grey boxes are predicted outcomes of SPNE with selfish types

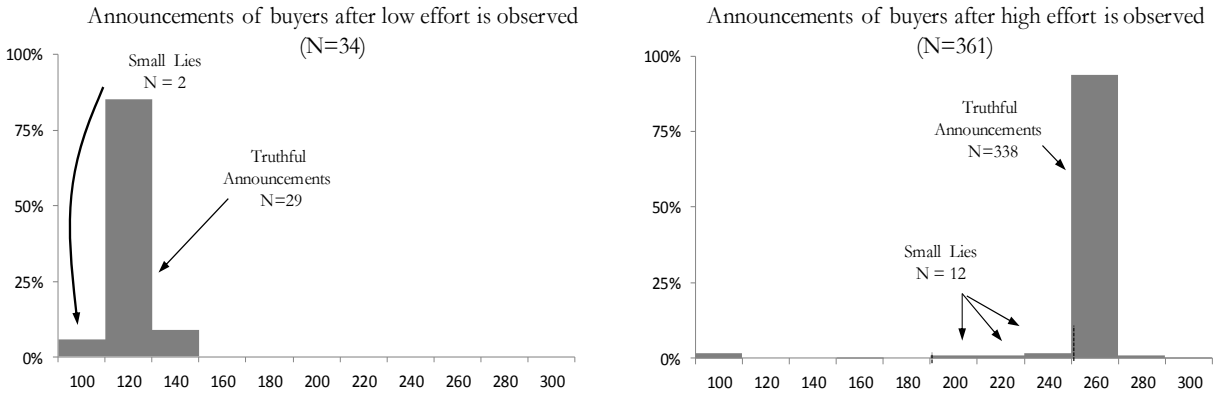
Outcomes of groups where seller effort is high and buyer and seller reports do not coincide

Cause of Arbitration	Arbitration Stopped by Seller	Arbitration Continued and Buyer Accepts	Arbitration Continued and Buyer Rejects
Buyer Underreport	3	33	11
Other	32	1	31

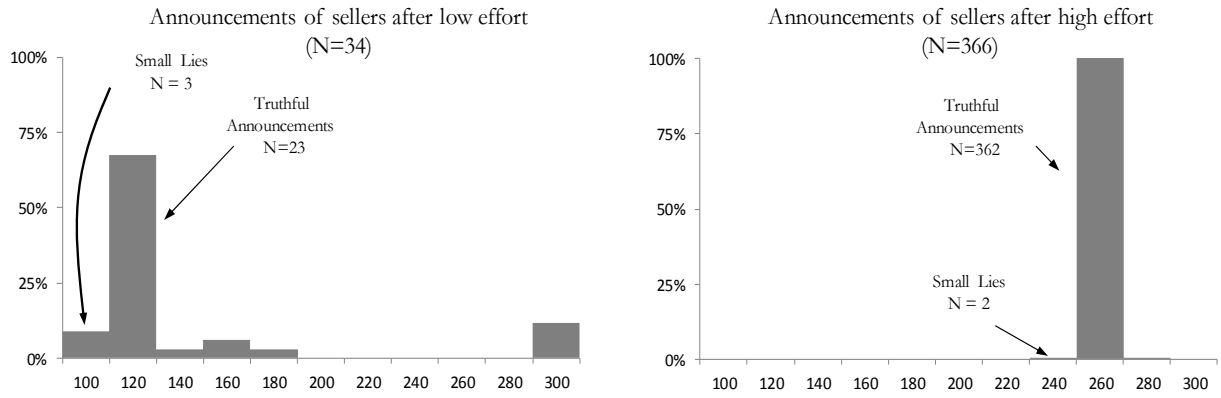
Grey boxes are predicted outcomes of SPNE with selfish types

Figure C2: Pattern of Play in First 10 Periods of RS Treatment

(a) Distribution of buyers' announcements after low and high effort



(b) Distribution of sellers' announcements after low and high effort



(c) Outcomes of groups where buyer and seller reports do not coincide

Outcomes of groups where seller effort is low and buyer and seller reports do not coincide

Cause of Arbitration	Arbitration Stopped by Seller	Arbitration Continued and Buyer Accepts	Arbitration Continued and Buyer Rejects
Buyer Underreport	0	0	2
Other	5	2	7

Grey boxes are predicted outcomes of SPNE with selfish types

Outcomes of groups where seller effort is high and buyer and seller reports do not coincide

Cause of Arbitration	Arbitration Stopped by Seller	Arbitration Continued and Buyer Accepts	Arbitration Continued and Buyer Rejects
Buyer Underreport	1	15	3
Other	4	2	2

Grey boxes are predicted outcomes of SPNE with selfish types

Figure C3: Pattern of Play in First 10 Periods of RS with Intensive-Training Treatment

C3. Additional Figures from RS Treatment (Phase 2)

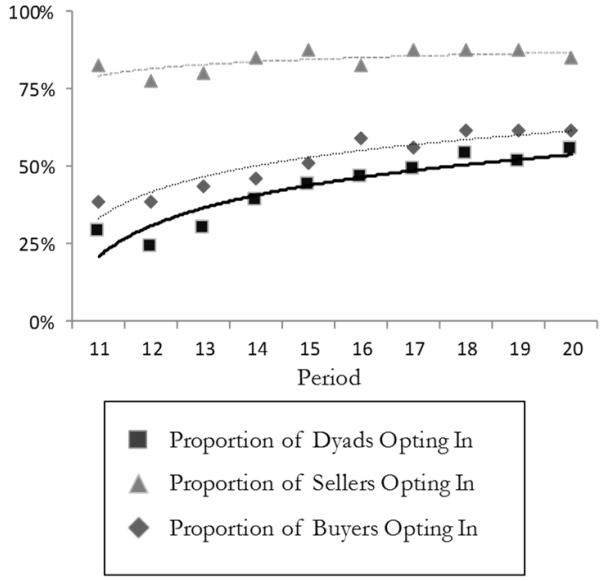


Figure C4: Proportion of Buyers and Sellers Opting Into the Mechanism in Periods 11–20 of RS with Intense-Training Treatment

References

Fehr, Ernst, Michael Powell, and Tom Wilkening. 2018. "Behavioral Constraints on the Design of Subgame-Perfect Implementation Mechanisms." Available at: http://tomwilkening.com/, accessed on 2018-10-01. Working Paper Version with Perfect Bayesian Retaliation Equilibirum.

Moore, John. 1992. "Implementation, Contracts, and Renegotiation in Environments with Complete Information." In *Advances in Economic Theory: Sixth World Congress Volume I.* , ed. John-Jacques Laffont, 182–282. Cambridge University Press.

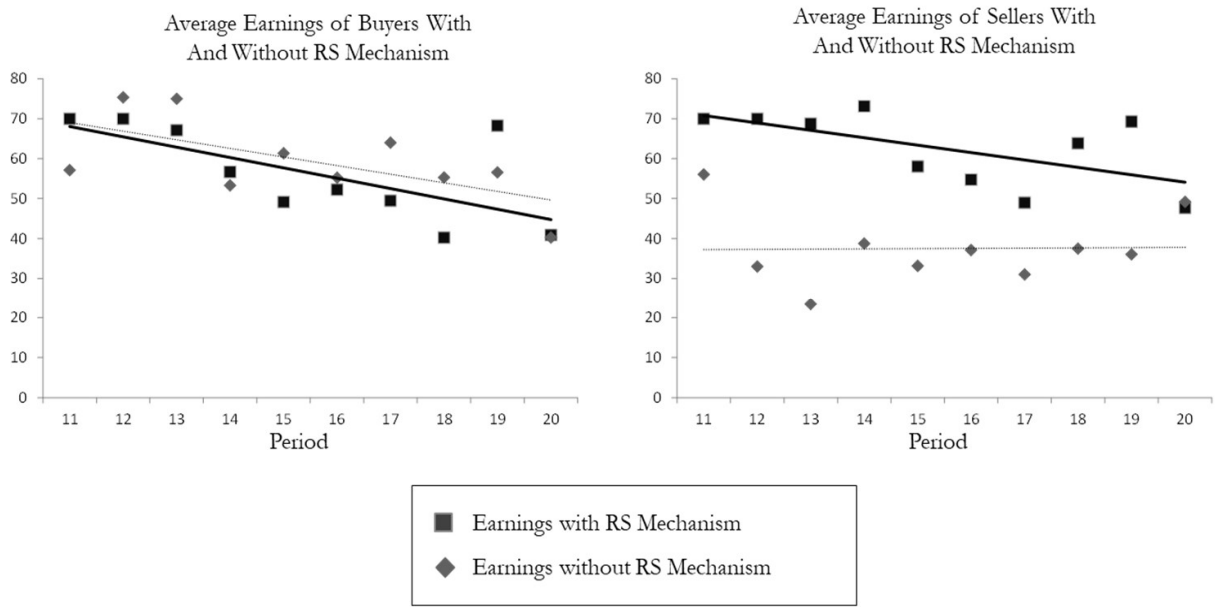


Figure C5: Average Earnings of Buyers and Sellers in Periods 11–20 of RS with Intense-Training Treatment With and Without the Mechanism