Behavioral constraints on the design of subgame-perfect implementation mechanisms

Fehr, Ernst ; Powell, Michael ; Wilkening, Tom

DOI: https://doi.org/10.1257/aer.20170297

American Economic Review 2021, 111(4): 1055–1091
https://doi.org/10.1257/aer.20170297

Behavioral Constraints on the Design of Subgame-Perfect Implementation Mechanisms[†]

By Ernst Fehr, 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* C92, D44, D82, D86, D91)

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 nonverifiable information implies that third-party enforcement of state-contingent contracts is infeasible and that formal contracting is ineffective.

*Fehr: Department of Economics, Zurich University (email: ernst.fehr@econ.uzh.ch); Powell: Strategy Department, Kellogg School of Management, Northwestern University (email: mike-powell@kellogg.northwestern.edu): Wilkening: Department of Economics, The University of Melbourne (email: Tom.Wilkening@unimelb.edu.au). Roland Bénabou was the coeditor for this article. 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 100018_140734\1) and the European Research Council, Adavanced Grant on "Foundations of Economic Preferences."

[†]Go to https://doi.org/10.1257/aer.20170297 to visit the article page for additional materials and author disclosure statements.

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

The tractable nature of models using the assumption of observable but nonverifiable 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 nonverifiable stands on controversial theoretical foundations. Building on work by Moore and Repullo (1988), Maskin and 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-nonverifiable 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 and Tirole (1999) that is designed to resolve the hold-up problem in bilateral exchange with observable but nonverifiable 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 ensuring buyer truth-telling 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 counter-offer that the buyer will accept only if he lied. If the buyer accepts the counter-offer 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 counter-offer (i.e., fail 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 sizable 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 suggest 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, and Kirchsteiger 1997; Offerman 2002; Falk, Fehr, and Fischbacher 2008) and field evidence also points toward the importance of this motive (e.g., Kube, Maréchal, and Puppe 2013; Cohn et al. 2014). However, theories of social preferences and reciprocity (e.g., Fehr and Schmidt 1999; Falk and Fischbacher 2006; Dufwenberg, Smith, and Van Essen 2013) as well as experimental evidence (e.g., Roth et al. 1991; Fischbacher, Fong, and Fehr 2000; Güth, Marchand, and 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 toward the importance of negative reciprocity for SPI mechanisms, we apply (a slightly adapted version of) the Sequential Reciprocity Equilibrium (SRE) concept of Dufwenberg and 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 counter-offers 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 counter-offers by challenging buyers' lies. In other words, reciprocal sellers could, in principle, discourage buyers from lying by threatening to challenge lies, even if they know that buyers will reject the subsequent counter-offer. 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 counter-offers. 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 counter-offer, 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 counter-offer, implying that much stronger reciprocity motives are required to challenge a lie compared to rejecting a counter-offer.

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 one-quarter 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

[2] We also considered the approach pursued by Bierbrauer and Netzer (2016) and Bierbrauer et al. (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.

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 counter-offers) 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.

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 microfoundation 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 and Charness 2010, Englmaier and Leider 2012, Netzer and Volk 2014), mechanism design (Bierbrauer and Netzer 2016, Bartling and Netzer 2016, Bierbrauer et al. 2017), and implementation (de Clippel, Eliaz, and 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, and Van Essen 2013). The interesting study by de Clippel, Eliaz, and 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

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

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 and Yavaş (1996) study extensive-form Abreu-Matsushima mechanisms that vary in the number of stages and find that incentive-compatible mechanisms with 8 and 12 stages perform worse than a mechanism with 4 stages that is not incentive compatible. Katok, Sefton, and Yava¸s (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 et al. (2018), which tests the theoretical predictions of Aghion et al. (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 noncontractible. 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 $v - p$.

[4] An extensive experimental literature also exists looking at efficiency of implementation mechanisms in the public goods provision problem. Chen and Plott (1996), Chen and Tang (1998), and Healy (2006) study learning dynamics in public good provision mechanisms. Andreoni and Varian (1999) and Falkinger et al. (2000) study two-stage compensation mechanisms that build on work from Moore-Repullo (1988), while Harstad and Marrese (1981, 1982); Attiyeh, Franciosi, and Isaac (2000); Arifovic and Ledyard (2004); and Bracht, Figuières, and Ratto (2008) study the voluntary contribution game, Groves–Ledyard, and Falkinger mechanisms, respectively. Masuda, Okano, and 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 papers mentioned above give subjects the opportunity to voluntarily select into the mechanism.

[5] Chen et al. (2018) 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 good's value to the buyer is *observable* to both parties but *nonverifiable* 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 nonverifiable. However, Maskin and 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 and Repullo (1988). In particular, consider the following class of subgame-perfect-implementation (SPI) mechanisms that is designed to implement a non-decreasing pricing rule $p(v)$:

(i) The buyer and seller sign a contract with a third party, whom we will call the arbitrator. The contract specifies (a) an **initial-price schedule** $p(\hat{v})$ at which trade may occur, given an announcement $\hat{v}$ that the buyer makes in stage (iii), and (b) 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.

(ii) 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.

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

(iv) 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.

(v) 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.

(vi) If the buyer rejects the counter-offer, 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.

An **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 $p(\cdot)$. 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 online 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 nonnegative 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 nonnegative 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:

(i) **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.

(ii) **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 (a) the true value for each potential effort choice, (b) small lies below each true value, and (c) generous offers above each true value. We discuss this choice of announcement space in Section IIA.

(iii) **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.

(iv) **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 seller's 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* $\hat{v} < v$.

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-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	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

Notes: Bolded values 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.

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 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 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 and 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 11 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 and 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

[6] The experimental literature on implementation (e.g., Cabrales, Charness, and Corchón 2003; Aghion et al. 2018) and contract theory (e.g., Sanchez-Pages and Vorsatz 2007, Ederer and 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.

[7] Hoppe and 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.

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.

B. *Experimental Protocol*

The experiments were run in the Experimental Economics Laboratory at the University of Melbourne between May and September 2009 and between November 2017 and February 2018. Experiments were conducted using z-Tree (Fischbacher 2007). All 520 subjects participating in the SPI Treatment and follow-up treatments (described in Sections IV and V) were undergraduate students at the university and were randomly invited from a pool of more than 5,000 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 and 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

[8] All data and code for this paper are available in Fehr, Powell, and Wilkening (2021).

[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.

which the arbitrator is exogenously available) as well as the rules 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 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. There were no bankruptcies in the SPI treatment and a total of five bankruptcies in all other

[10] See Blanco et al. (2010) for a discussion of hedging.

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.

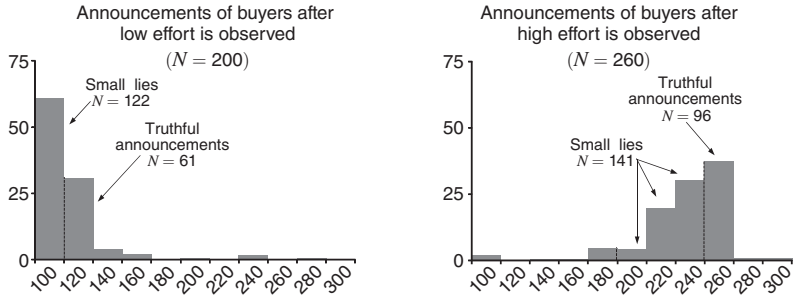
Figure 1 displays the patterns of play we observed in the first ten periods of the experiment. The left column examines play following low effort ($N = 200$), and the right column examines play following high effort ($N = 260$). Panel A summarizes the buyers' announcement decisions, panel B summarizes 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 buyers lied in the majority of observations: 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.

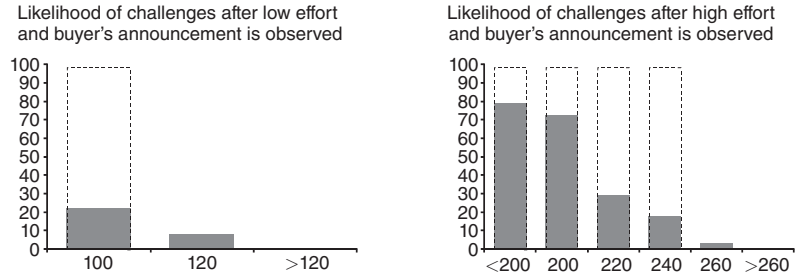
Panel B shows the proportion of sellers who challenge each announcement $\hat{v}$. SPI Hypothesis 1 predicts that sellers challenge 100 percent of the time after a lie and never challenge after a truthful or generous announcement. In the data, the challenge probability for small lies is less than 30 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.

Panel A. Distribution of announcements after low and high effort (percent)



Panel B. Likelihood of a challenge after each announcement (percent)



Panel 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

Gray 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
<200	7	8
200	2	6
220	0	15
240	0	14
260	0	3

Gray boxes are predicted action by SPI hypothesis

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

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 $(7.2 - 90)/(140 - 90) = -166$ percent.

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 online Appendix Section C1, we show how effort, announcements, and

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

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 online Appendix Section 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 suggest 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 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.

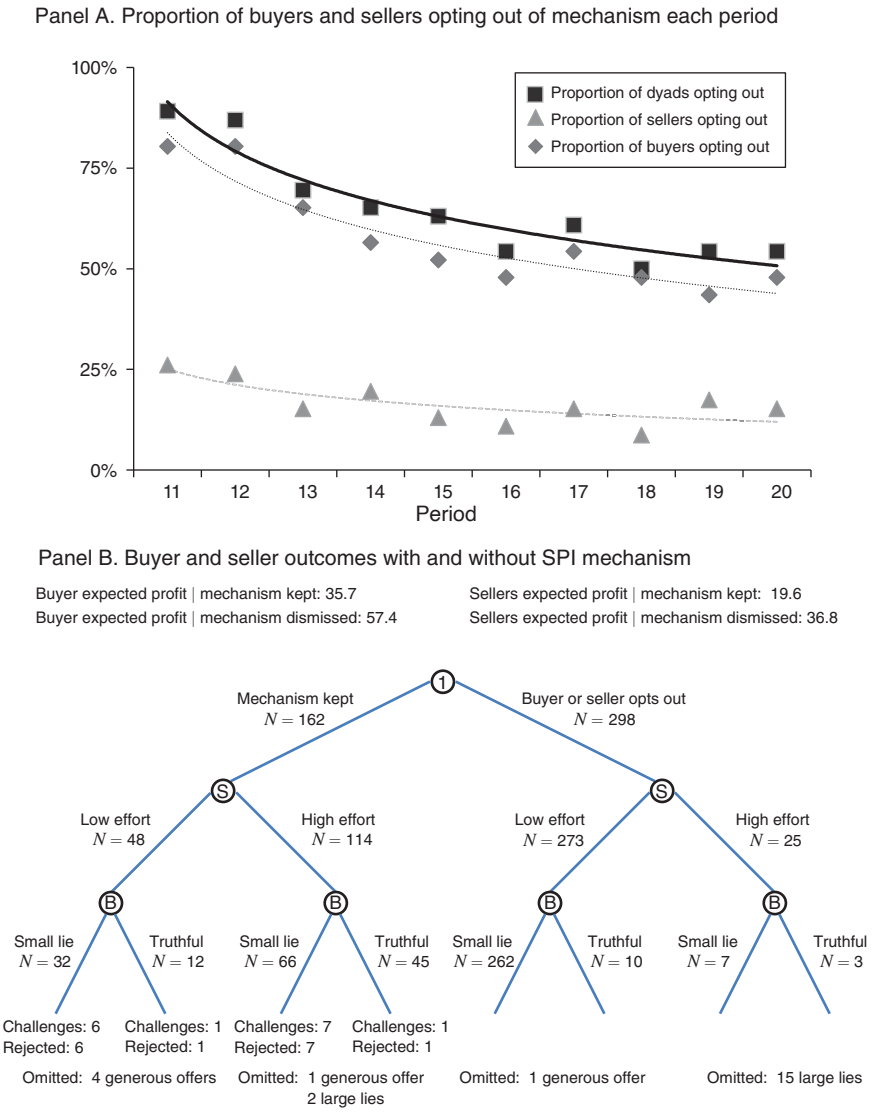


FIGURE 2. BEHAVIOR IN LAST 10 PERIODS (SECOND PHASE) OF SPI TREATMENT

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 percent 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 (i) the probability that the seller exerts high effort and (ii) 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]

[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 counter-offers, 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 mechanisms are described in more detail in online Appendix Sections B2–B4.

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, if sellers do not appropriately challenge small lies, then buyers have pecuniary incentives to under-report the value of the good. Indeed, the beliefs data support the rationale above 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 online Appendix Section 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**

[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.

Treatment, we (i) explicitly explained in the written instructions the 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 online Appendix Section 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 under-report 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 counter-offers, 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 counter-offer 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 nonpecuniary 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 nonpecuniary benefit $\psi_B \lambda_B$ in the discussion above was exogenous. However, in online Appendix Section A3, we adapt Dufwenberg and Kirchsteiger's (2004)—henceforth, DK—solution concept, sequential reciprocity equilibrium (SRE) to our setting. Following DK, we assume that the buyer and seller have commonly known intentions-based reciprocal preferences. We assume that players care positively

[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.

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, and Van Essen (2013), 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 parameters, two for the buyer ($\bar{\psi}_B^{SPI}$ and $\hat{\psi}_B^{SPI}$) and one for the seller ($\bar{\psi}_S^{SPI}$), 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 $\bar{\psi}_S^{SPI} > \bar{\psi}_B^{SPI}$ in any γ^{SPI} mechanism that SPE-implements the pricing rule p due to the asymmetric role of fines in the mechanism.

[15] As seen in Appendix Section 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.

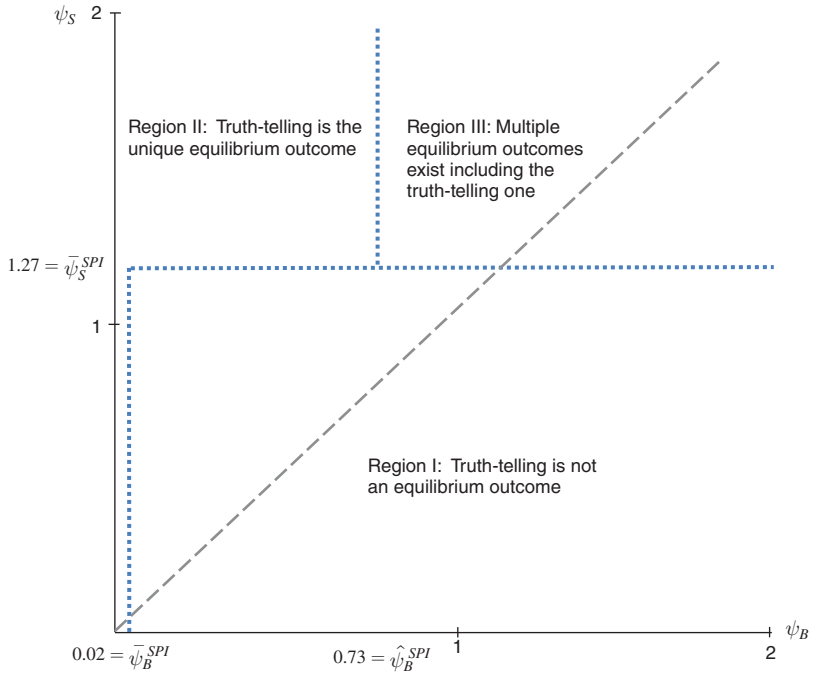


FIGURE 3. SEQUENTIAL RECIPROCITY EQUILIBRIUM OUTCOMES IN THE EXPERIMENTALLY IMPLEMENTED SPI MECHANISM FOR DIFFERENT BUYER (ψ_B) AND SELLER (ψ_S) RETALIATORY TYPES

Notes: 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 45-degree line, both parties' retaliatory types are the same.

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

(i) 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 ($\psi_B > \bar{\psi}_B^{SPI} = 0.02$) suffices to be in this region unless there is a large amount of seller reciprocity (i.e., $\psi_S > \bar{\psi}_S^{SPI} = 1.27$). In this region, sellers are unwilling to challenge a buyer's lie because they know that the buyer will reject the counter-offer.

(ii) Truth-telling is the unique equilibrium outcome in Region II, but this requires a large amount of seller reciprocity (i.e., $\psi_S > \bar{\psi}_S^{SPI} = 1.27$) 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 counter-offer. Therefore, in this region the buyers are deterred by sellers' reciprocity unless they also have a rather high inclination to reciprocate (i.e., if $\psi_B > \hat{\psi}_B^{SPI} = 0.73$).

(iii) 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 $\psi_S > \bar{\psi}_S^{SPI} = 1.27$). 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 $\psi_S < \bar{\psi}_S^{SPI} = 1.27$ 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 $\bar{\psi}_S^{SPI}$ would reject an offer of 49 percent of the pie. Such rejections are extremely rare. This suggests that negative reciprocity can rationalize the main deviations observed in our experiment.[16]

In online Appendix Section 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 nonconstant 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 commonly known.

We say that a mechanism γ^{SPI} and pricing rule p are subject to retaliatory implementation failure if γ^{SPI} 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 nonconstant pricing rule, if γ^{SPI} SPE-implements p, then (γ^{SPI}, p) 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 $\bar{\psi}_B^{SPI} < \bar{\psi}_S^{SPI}$ 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

[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.

Region I where truth-telling is not part of any SRE, and therefore (γ^{SPI}, p) 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 $\psi > \bar{\psi}_B^{SPI}$ 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 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 (i) buyers regularly tell small lies, (ii) sellers occasionally challenge such lies, and (iii) buyers frequently retaliate against challenges of small lies. This pattern of play is observed in the main treatment.

V. Toward 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 and 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

[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, and Wilkening 2018 for details). In particular, if retaliatory types are privately known but lie in Region I with probability 1, then truth-telling is not an equilibrium outcome. If they lie in the left-most sliver of Region II with probability 1 (i.e., all buyers have a retaliatory type less than $\bar{\psi}_B^{SPI}$) or they lie in the rest of Region II with probability 1 (i.e., all sellers have a retaliatory type greater than $\bar{\psi}_S^{SPI}$ and all buyers have a retaliatory type less than $\hat{\psi}_B^{SPI}$), then truth-telling is the unique equilibrium outcome.

[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 and Putterman 2006; Carpenter 2007; Falk, Fehr, and Fischbacher 2005).

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 and 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 online Appendix Section 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 nontrivial 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 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:

(i) The buyer and seller sign a contract with the arbitrator. The contract specifies (a) an initial price schedule $p(\hat{v}_B)$ at which trade may occur, given an announcement $\hat{v}_B$ the buyer makes in stage (iii), (b) 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 $\hat{v}_B = v$.

(ii) 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.

(iii) 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.

(iv) 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.

(v) 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.

(vi) 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 online 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 nonconstant 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:*

(a) *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}.

(b) *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 equilibrium 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 (iii) of the RS mechanism. We discuss this issue further in online Appendix Section 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 intensive 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:

(i) **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.

(ii) **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) $\hat{v}_B \in \hat{V} = \{100, 120, \ldots, 260, 280, 300\}$.

(iii) **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.

(iv) **The Arbitration Stage:** If the seller enters the arbitration stage, the seller will have the option to *continue* arbitration or to *exit* 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 $p(\hat{v}_B) = 70 + 0.75(\hat{v}_B - 100)$.

(v) **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.

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

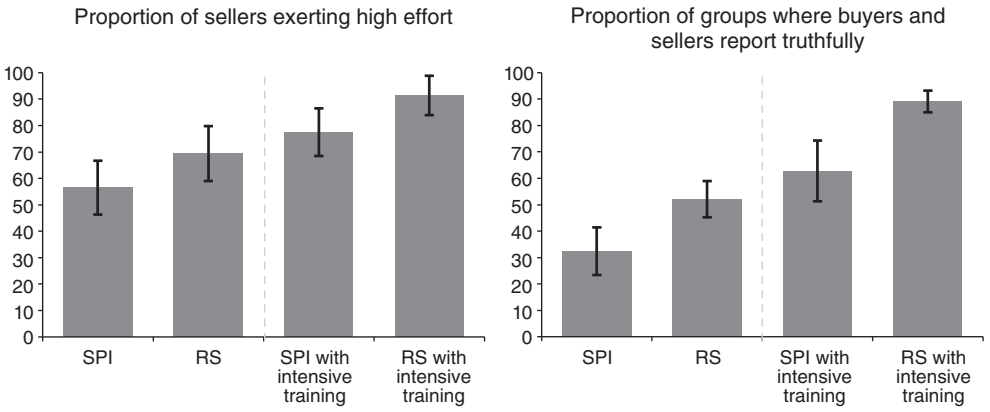


FIGURE 4. PROPORTION OF SELLERS EXERTING HIGH EFFORT AND PROPORTION OF GROUPS WHERE BUYERS AND SELLERS REPORT TRUTHFULLY IN PERIODS 1–10 (PERCENT)

Note: 95 percent confidence intervals shown.

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 percent 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 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).

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

[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.

to eliminate small lies. In online Appendix Section 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 online Appendix 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 nontruthful 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 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 1 if a group retains the mechanism and 0 otherwise (p-value $=$ 0.04). 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 $=$ 0.02). 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.

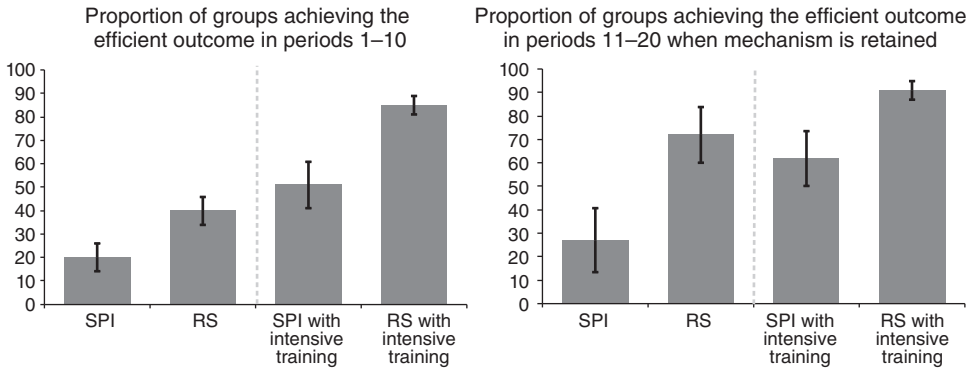


FIGURE 5. PROPORTION OF GROUPS ACHIEVING THE EFFICIENT OUTCOME (PERCENT)

Note: 95 percent confidence intervals shown.

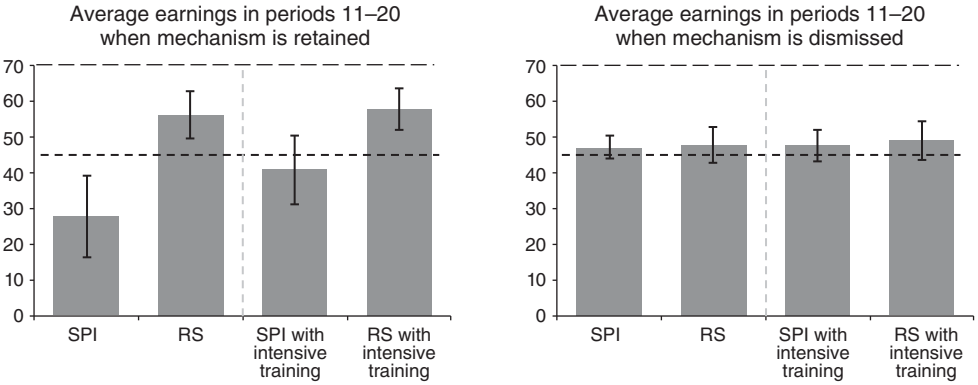


FIGURE 6. AVERAGE EARNINGS OF INDIVIDUAL BUYERS AND SELLERS IN PERIODS 11–20 WHEN THE MECHANISM IS RETAINED AND DISMISSED (PERCENT)

Notes: 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

[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 $= 0.04$). 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 < 0.01). 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 $= 0.31$).

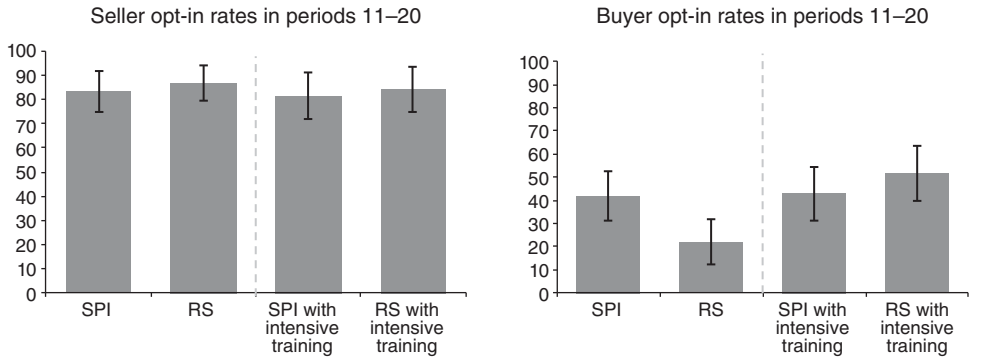


FIGURE 7. PROPORTION OF SELLERS AND BUYERS CHOOSING TO OPT INTO EACH MECHANISM IN PERIODS 11–20 (PERCENT)

Note: 95 percent confidence intervals shown.

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 opt 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 opt 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 $= 0.29$).

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 online Appendix Section 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 online Appendix Section 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 nonnegligible 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 percent 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 buyer's 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.

[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.

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 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 and 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 nontrivial 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 Antràs, and Elhanan Helpman. 2007. "Contracts and Technology Adoption." *American Economic Review* 97 (3): 916–43.

Aghion, Philippe, and Patrick Bolton. 1992. "An Incomplete Contracts Approach to Financial Contracting." *Review of Economic Studies* 59 (3): 473–94.

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 European Economic Association* 16 (1): 232–74.

Aghion, Philippe, Drew Fudenberg, Richard Holden, Takashi Kunimoto, and Olivier Tercieux. 2012. "Subgame-Perfect Implementation under Value Perturbations." *Quarterly Journal of Economics* 127 (4): 1843–81.

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 Academy of Science of the United States of America* 96: 10933–38.

Antràs, Pol. 2003. "Firms, Contracts, and Trade Structure." *Quarterly Journal of Economics* 118 (4): 1375–418.

Arifovic, Jasmina, and John Ledyard. 2004. "Scaling up Learning Models in Public Good Games." *Journal of Public Economic Theory* 6 (2): 203–38.

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

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." *Quarterly Journal of Economics* 116 (4): 1343–72.

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): 412–38.

Blount, Sally. 1995. "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences." *Organizational Behavior and Human Decision Processes* 63 (2): 131–44.

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): 163–76.

Cabrales, Antonio, Gary Charness, and Luis C. Corchón. 2003. "An Experiment on Nash Implementation." *Journal of Economic Behavior & Organization* 51 (2): 161–93.

Carpenter, Jeffrey P. 2007. "The Demand for Punishment." *Journal of Economic Behavior and Organization* 62 (4): 522–42.

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

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): 633–62.

Chen, Yi-Chun, Richard Holden, Takashi Kunimoto, Yifei Sun, and Tom Wilkening. 2018. "Getting Dynamic Implementation to Work." http://tomwilkening.com/ (accessed October 1, 2018).

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): 877–98.

de Clippel, Geoffroy, Kfir Eliaz, and Brian Knight. 2014. "On the Selection of Arbitrators." *American Economic Review* 104 (11): 3434–58.

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

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

Dufwenberg, Martin, Alec Smith, and Matt Van Essen. 2013. "Hold-Up: With a Vengeance." *Economic Inquiry* 51 (1): 896–908.
Ederer, Florian, and Ernst Fehr. 2007. "Deception and Incentives: How Dishonesty Undermines Effort Provision." University of Zurich Institute for Empirical Research Working Paper 341.
Englmaier, Florian, and Stephen Leider. 2012. "Contractual and Organizational Structure with Reciprocal Agents." *American Economic Journal: Microeconomics* 4 (2): 146–83.
Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2005. "Driving Forces behind Informal Sanctions." *Econometrica* 73 (6): 2017–30.
Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2008. "Testing Theories of Fairness: Intentions Matter." *Games and Economic Behavior* 62 (1): 287–303.
Falk, Armin, and Urs Fischbacher. 2006. "A Theory of Reciprocity." *Games and Economic Behavior* 54 (2): 293–315.
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): 247–64.
Fehr, E., S. Gächter, and G. Kirchsteiger. 1997. "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econometrica* 65 (4): 833–60.
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." http://tomwilkening.com/ (accessed October 1, 2018).
Fehr, Ernst, Michael Powell, and Tom Wilkening. 2021. "Replication Data for: Behavioral Constraints on the Design of Subgame-Perfect Implementation Mechanisms." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E24661V1.
Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114 (3): 817–68.
Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10 (2): 171–78.
Fischbacher, Urs, Christina M. Fong, and Ernst Fehr. 2000. "Fairness, Errors, and the Power of Competition." *Journal of Economic Behavior & Organization* 72 (1): 527–45.
Greiner, Ben. 2015. "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE." *Journal of the Economic Science Association* 1 (1): 114–25.
Gromb, Denis. 1993. "Is One Share/One Vote Optimal?" LSE Financial Markets Group Discussion Paper.
Grossman, Sanford J., and Oliver D. Hart. 1986. "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration." *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 Marrese. 1981. "Implementation of Mechanism by Processes: Public Good Allocation Experiments." *Journal of Economic Behavior & Organization* 2 (2): 129–51.
Harstad, Ronald M., and Michael Marrese. 1982. "Behavioral Explanations of Efficient Public Good Allocations." *Journal of Public Economics* 19 (3): 367–83.
Hart, Oliver. 1995. *Firms, Contracts, and Financial Structure.* New York: Oxford University Press.
Hart, Oliver, and John Moore. 1990. "Property Rights and the Nature of the Firm." *Journal of Political Economy* 98 (6): 1119–58.
Hart, Oliver, and John Moore. 1998. "Default and Renegotiation: A Dynamic Model of Debt." *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." *Quarterly Journal of Economics* 112 (4): 1127–61.
Healy, Paul J. 2006. "Learning Dynamics for Mechanism Design: An Experimental Comparison of Public Goods Mechanisms." *Journal of Economic Theory* 129 (1): 114–49.
Hoppe, Eva I., and Patrick W. Schmitz. 2011. "Can Contracts Solve the Hold-Up Problem? Experimental Evidence." *Games and Economic Behavior* 73 (1): 186–99.
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): 853–70.

Maskin, Eric, and Jean Tirole. 1999. "Unforeseen Contingencies and Incomplete Contracts." *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): 1191–20.

Netzer, Nick, and André Volk. 2014. "Intentions and Ex-Post Implementation." Unpublished.

Nöldeke, Georg, and Klaus Schmidt. 1995. "Option Contracts and Renegotiation: A Solution to the Hold-Up Problem." *RAND Journal of Economics* 26 (2): 163–79.

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

Offerman, Theo. 2002. "Hurting Hurts More than Helping Helps." *European Economic Review* 46 (8): 1423–37.

Perugini, Marco, Marcello Gallucci, Fabio Presaghi, and Anna Paola Ercolani. 2003. "The Personal Norm of Reciprocity." *European Journal of Personality* 17 (4): 251-83.

Ponti, Giovanni, Anita Gantner, Dunia López-Pintado, and Robert Montgomery. 2003. "Solomon's Dilemma: An Experimental Study on Dynamic Implementation." *Review of Economic Design* 8 (2): 217–39.

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): 1068–95.

Sánchez-Pagés, Santiago, and Marc Vorsatz. 2007. "An Experimental Study of Truth-Telling in a Sender-Receiver Game." *Games and Economic Behavior* 61 (1): 86–112.

Schmidt, Klaus M. 1996a. "Incomplete Contracts and Privatization." *European Economic Review* 40 (3–5): 569–79.

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

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