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

Ernst Fehr Michael Powell Tom Wilkening

February 17, 2021

Table of Contents: Appendix

Appendix A: Theory

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

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

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

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

Throughout Appendix A, we focus on the part of the game tree in which the seller's effort choice has already been made. This focus allows us to speak directly to the implementation problem and simplifies the analysis. It is also without loss of generality under the sequential reciprocity equilibrium solution concept given our specification of parties' references payoffs, which as we will show in Section A3 depends directly on the value of the good to the buyer. For alternative specifications of parties' reference payoffs, parties' retaliatory motives may depend on the seller's effort choice.

A1. Preliminaries and Definitions

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

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

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

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

A2. SPE-Implementable Pricing Rules and SPI Mechanisms

In this section, we will define a class of mechanisms and show that any pricing rule can be implemented with a mechanism from this class. Given a mechanism γ, a strategy profile is a $\sigma = \sigma_B \times \sigma_S$, where σ_i is a mapping from history h^t to a distribution of feasible messages $\mathcal{M}_i^t(h^t)$, where we are slightly abusing notation, since $\mathcal{M}_i^t(h^t)$ depends only on past realized

messages and not the payoff state v. Continuation play for player i at history h^t is denoted by $\sigma_i|\, h^t$. The material payoff player i expects to receive, given history h^t, is determined by the distribution over terminal nodes induced by the continuation strategy profile $\sigma|\, h^t$, and we will denote his expected payoff by $\pi_i\left(\sigma_i|\, h^t, \sigma_j|\, h^t\right)$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A3. Psychological Environments and Sequential Reciprocity Equilibrium

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

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

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

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

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

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

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

Finally, we want to normalize player i's aggrievement so that it is between 0 and 1, so that ψ_i can be interpreted as player i's maximum willingness to pay to destroy one unit of player j's material payoff. Given these ingredients, define player i's **aggrievement** at

history h^t by

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

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

Our notion of aggrievement captures the intensity with which player i will act on his retaliatory preferences. We will assume that player i's choices at history h^t are made to maximize his expected material payoff minus the scalar $\psi_i\lambda_i$ times player j's expected material payoff, under the assumption that in future rounds, he will continue to play according to the strategy σ_i. That is, he chooses a strategy $\tilde{\sigma}_i|\, h^t$ consisting of a round-t message $\tilde{m}_i^t \in \mathcal{M}_i^t\left(h^t\right)$ followed by $\sigma_i|\, \tilde{h}^{t+1}$, where $\tilde{h}^{t+1} = h^t\tilde{m}^t$ is the concatenation of history h^t with the realization of round-t messages $\tilde{m}^t$, that maximizes

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

We now define our solution concept.

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

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

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

A4. Retaliatory Implementation Failure and SRE Implementation

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

Given an economic environment $\mathcal{E}$ and a **non-constant pricing rule** p such that $p\left(\cdot\right)$ is not constant on $\mathcal{V}$, suppose a mechanism γ SPE-implements p. Say that a psychological environment $\mathcal{P}$ is a **symmetric psychological environment** if buyer and seller retaliatory types are identically distributed under $\mathcal{P}$. We will say that the pair (γ, p) is subject to **retaliatory implementation failure** if there exists a symmetric psychological environment $\mathcal{P}$ in

which in every SRE, with positive probability, the buyer announces some $\hat{v} \neq v$ for some v. The following proposition shows SPI mechanisms are subject to retaliatory implementation failure.

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

Proof of Proposition 1. Since p is a non-constant pricing rule, there exists $v, v' \in \mathcal{V}$ such that $p(v) < p(v')$. Since γ^{SPI} SPE-implements p, it must have the property that in any SPE, if the payoff state is v', a buyer announcement of $\hat{v} = v < v'$ is challenged by the seller with positive probability, or else the buyer would prefer to announce $\hat{v}$, and γ^{SPI} would not implement p.

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

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

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

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

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

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

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

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

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

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

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

A5. Diagnosing the Failure of the SPI Mechanism

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

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

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

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

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

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

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

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

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

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

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

Second, suppose $\psi_B > \overline{\psi}_B(v, v')$ so that in every SRE, the buyer will retaliate against an appropriate challenge of $v' < v$. Then the seller will challenge nevertheless if $\psi_S \geq \overline{\psi}_S(v, v')$.

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

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

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

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

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

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

For part $(iii.)$ of the lemma, our argument for part $(i.)$ of the lemma established that in this region, there is an SRE in which the buyer lies with strictly positive probability. It remains to argue that truth-telling is also an SRE outcome. Consider a truth-telling SRE. At the initial node, the buyer's aggrievement is $\lambda_B^* = 0$, so he will be willing to deviate and lie only if doing so increases his material payoffs. But since $\psi_B > \overline{\psi}_B^{SPI}$ and $\psi_S > \overline{\psi}_S^{SPI}$, for

any lie, the seller will challenge, and the buyer will retaliate, so the buyer's material payoff must be lower following a lie. There is therefore no profitable deviation, and truth-telling is an SRE outcome.■

Lemma 2 characterizes the equilibrium outcomes in the different regions of Figure 3 in the main text. It shows that when the seller's retaliatory type is less than one, SREs involve truth-telling by the buyer only if $\psi_B < 3/149 \approx 0.02$. In other words, if the buyer is willing to sacrifice more than two cents in order to reduce the seller's material payoffs by one dollar, then there is no SRE in which the buyer tells the truth when $v = 260$. The lemma also shows that when this is the case, in any SRE in which the buyer retaliates against challenges following $\hat{v} = 240$, the seller will never challenge such an announcement. In such psychological environments, therefore, SREs can rationalize lying and retaliation by the buyer as well as reluctance to challenge by the seller.

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

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

A6. The Retaliatory-Seller Mechanism

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

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

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

1. The buyer and seller simultaneously announce $\hat{v}_B, \hat{v}_S \in \mathcal{V}$ (i.e., $\mathcal{M}_B^1 = \mathcal{V}$ and $\mathcal{M}_S^1 = \mathcal{V}$). If the announcements agree, then trade occurs at price $p(\hat{v}_B)$, so that $g(m) = (1, p(\hat{v}_B), p(\hat{v}_B))$ if $\hat{v}_B = \hat{v}_S$.

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

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

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

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

1. If $\gamma^{SPI}(F_B, F_S, \hat{p})$ SRE-partially implements p in psychological environment $\mathcal{P}$, then so does $\gamma^{RS}(F_B, F_S, \hat{p})$.

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

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

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

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

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

Proof of Proposition 2. We want to show the conditions under which in payoff state v, there is an SRE of $\gamma^{SPI}(F_B, F_S, \hat{p})$ in which the buyer announces $\hat{v} = v$ and the conditions under which there is an SRE of $\gamma^{RS}(F_B, F_S, \hat{p})$ in which both parties announce $\hat{v}_B = \hat{v}_S$. We will first describe the set of conditions under which truth-telling is an SRE outcome in the SPI mechanism and the RS mechanism. Then we will compare these two sets of conditions. To this end, take an arbitrary v, and consider a $v' < v$ to be a candidate deviation at the announcement stage.

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

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

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

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

Comparison between the SPI mechanism and the RS mechanism. Let $\hat{\Psi}^{SPI}$ be the set of (ψ_B, ψ_S) such that for all v and $v' < v$, $\psi_B \leq \overline{\psi}_B^{SPI}(v, v')$ or $\psi_S \geq \overline{\psi}_S^{SPI}(v, v')$. Then

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

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

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

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

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

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

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

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

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

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

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

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

or

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

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

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

A7. The Insurance Property and Fixed-Price Contracts

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

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

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

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

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

is independent of $\hat{\theta}_B \in \Theta$. A social choice function that has no marginal externalities on either the seller or the buyer satisfies what Bierbrauer and Netzer (2016) refers to as the **insurance property**. The insurance property therefore implies that whether the buyer

buys the good and at what price is independent of the seller's private information, and it also implies that whether the seller sells and at what price is independent of the buyer's private information.

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

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

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

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

which implies the monotonicity condition

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

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

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

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

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

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

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

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

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

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

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

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

Appendix B: Additional Analyses and Treatments

B1. The Role of Beliefs

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

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

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

To better understand the role that beliefs have in buyers announcements we look at the decision of the buyer to make a small lie based on his belief about being challenged after

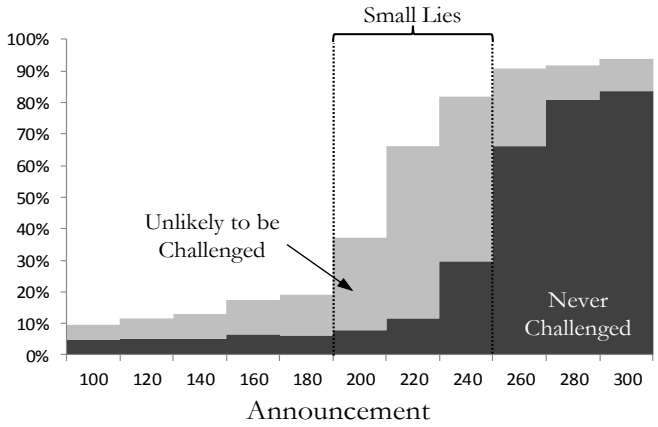


Figure B1: Proportion of buyers who believe that a given announcement will "Never" be challenged or is "Unlikely" to be challenged after observing high effort.

such lies. Table B1 reports the results of a probit regression where the dependent variable is 1 if a buyer makes a small lie and 0 if the buyer makes a truthful announcement. We report regressions for choices after high effort in regressions (1) and (2), choices after low effort in regressions (3) and (4), and choices after both high and low effort in regressions (5) and (6).

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

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

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

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

more pessimistic about being challenged after a truthful announcement. Thus optimistic beliefs about being challenged after a lie and pessimistic beliefs about being challenged after a truthful announcement appear to influence the buyers announcement decision.

Table B1: Probit Regression of Small Lies by Buyers

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

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

Turning to the beliefs of sellers, 71.6 percent (62.3 percent) of sellers who are confronted with a small lie after high (low) effort believe that an appropriate challenge will "Never" be accepted or is "Unlikely" to be accepted. Thus, sellers also correctly forecast that buyers are likely to reject appropriate challenges.

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

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

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

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

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

B2. High-Benefits Treatment

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

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

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

A buyer is less likely to be challenged after a truthful announcement than a small lie. This implies that if the value that a buyer receives when he is *not* challenged increases by a constant across all potential announcements, the expected value of announcing a truthful announcement will increase by more than the expected value of announcing a small lie. For example, if a buyer believes that a small lie will be challenged 50 percent of the time and a

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

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

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

truthful announcement will never be challenged, then an increase in the value of not being challenged of 10 will increase the expected value of the small lie by 5 $(10 * .5)$ and increase the value of truth telling by 10.

In the High-Benefits treatment we make precisely this type of shift in the value of not being challenged by decreasing the initial-price schedule $p(\hat{v})$ uniformly across all announcements. The structure of this treatment is just as in the SPI Treatment except that we decrease the price $p(\hat{v})$ by 20:

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

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

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

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

Figure B2 displays the results for the High-Benefits Treatment with data aggregated across all 10 periods: The left-hand side of the figure follows the pattern of play after the

seller selects low effort ($N = 66$) while the right-hand side of the figure follows the pattern of play following high effort ($N = 194$). Directly comparable to Figure 1, panel (a) shows the distribution of announcements, panel (b) shows the likelihood of a challenge after each announcement, and panel (c) shows the frequency that a challenge is accepted or rejected.

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

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

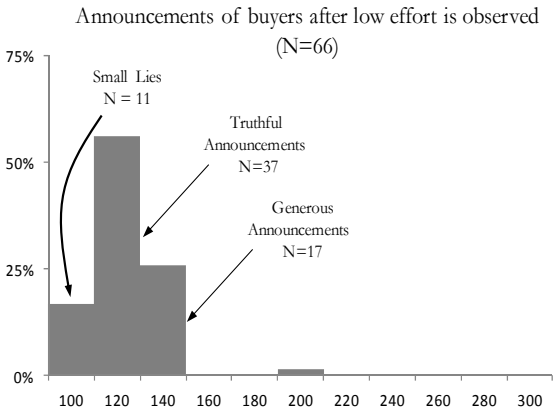
Interestingly, unlike the SPI Treatment, buyers in the High-Benefits Treatment frequently make generous announcements, $\hat{v} > v(e)$. For example, after high effort, buyers make generous announcements in 38 percent of the cases. The large proportion of these generous offers suggests a new deviation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

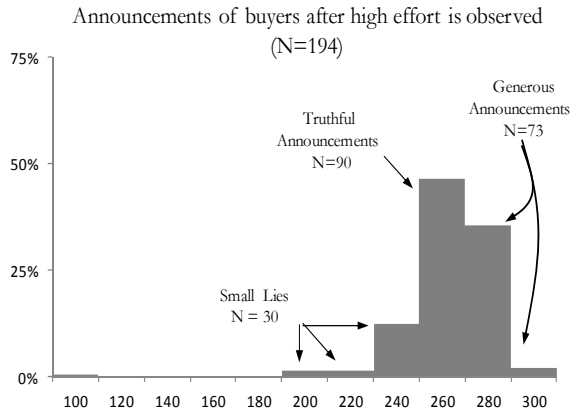
Looking at Panel (b) and comparing it to the SPI Treatment, sellers are much more likely to challenge small lies in the High-Benefits Treatment: following high effort, announcements of 240 are challenged 58 percent of the time as compared to 18 percent of the time in the SPI Treatment. These differences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

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

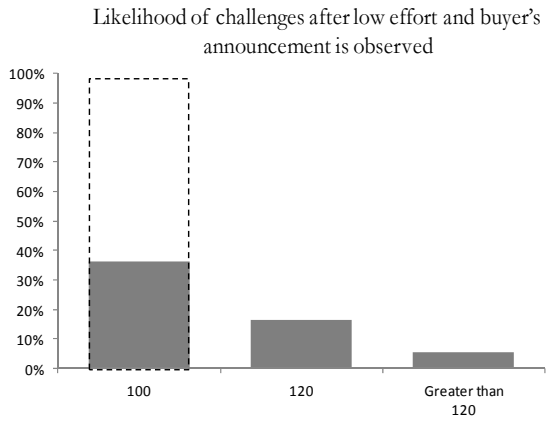
Taken together, the High-Benefits treatment has a larger proportion of truthful announcements and higher effort than the SPI Treatment. However, the losses that occur due to disagreement in early periods of the experiment are larger than the gains that occur from improvements in effort and therefore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

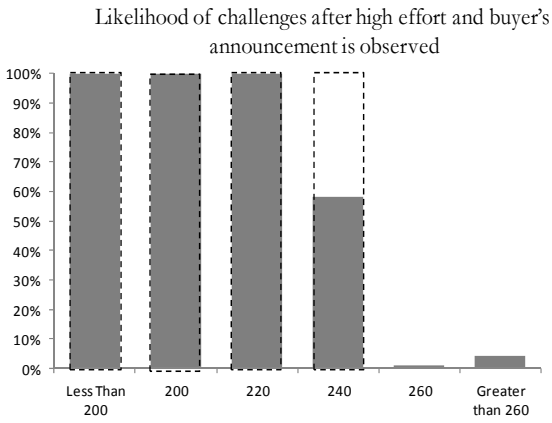
(a) Distribution of announcements after low and high effort





(b) Likelihood of a challenge after each announcement





(c) Number of challenges accepted and rejected

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

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

Grey boxes are predicted action by SPI Hypothesis

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

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

Grey boxes are predicted action by SPI Hypothesis

Figure B2: Pattern of Play in High-Benefits Treatment

Given that players realize positive gains toward the end of the first phase of the experiment, we might expect that buyers and sellers are more likely to opt into the mechanism in this treatment. However, we find no significant difference in the overall opt-out rates in the second phase of the experiment.

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

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

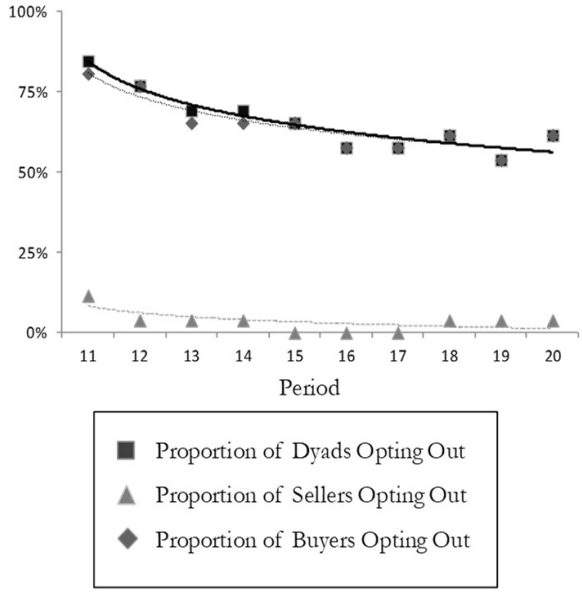
As with the SPI Treatment, in periods without the mechanism, the hold-up problem is unresolved. As seen in panel (b), when either party opts out of the mechanism, 154 out of 171 sellers exert low effort. In 134 of these cases the buyer announces $\hat{v} = 100$. Of the 17 observations where high effort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

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

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

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

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



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

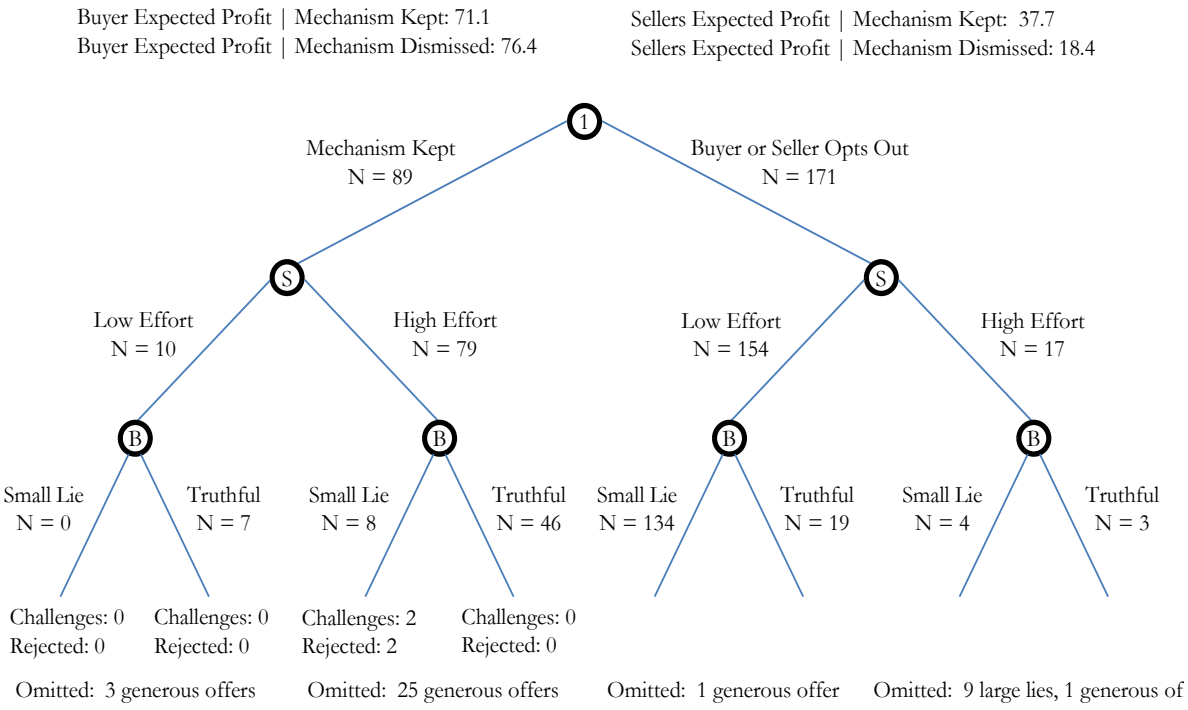


Figure B3: Behavior in Last 10 Periods of High-Benefits Treatment

belief data support the hypothesis that buyers have such pessimistic beliefs. In comparison to the distribution of beliefs in the SPI Treatment where 66 percent of buyers believed that a truthful announcement would never be challenged after high effort, only 39 percent of buyers in the High Benefits Treatment believe that truthful announcements would never be challenged.

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

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

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

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

B3. Low-Fine Treatment

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

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

Table B4: Probit Regression of Generous Announcements by Buyer

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

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

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

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

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

Result B.4 *In the Low-Fine treatment, sellers' effort choices and the buyer's likelihood of making a small lie or a truthful announcement are similar to the High-Benefits Treatment. However, following high effort, a large proportion of buyers make the lowest possible announcement, $\hat{v} = 100$. These "maximal lies" are more frequent among buyers who are averse to gambles and who fear inappropriate challenges. Sellers always challenge maximal lies and buyers who are challenged after a maximal lie almost always accept the counter offer. Sellers almost always challenge small lies and buyers still retaliate against the majority of these challenges.*

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

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

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

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

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

Challenges of small lies are rejected in 9 out of 16 cases after high effort and in 7 out of 9 cases after high effort. While the aggregate rejection rate of challenges after small lies of

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

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

64 percent is 15.2 percentage points lower than the High-Benefits treatment, the difference is not significant in a probit regression that regresses the acceptance rate of small lies on the treatment (p-value $= 0.29$). This suggests that retaliation has not been fully resolved in this treatment.

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

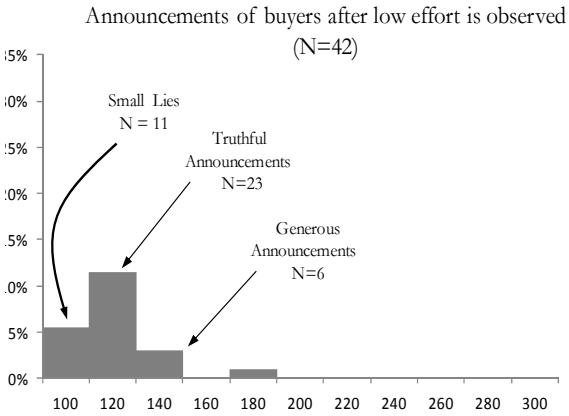
The hypothesis that fear of inappropriate challenges leads to maximal lies is supported by two pieces of evidence. First, for buyers who believed that they would never be inappropriately challenged, maximal lies occur in 28 percent of the observations. For buyers who believe that inappropriate challenges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

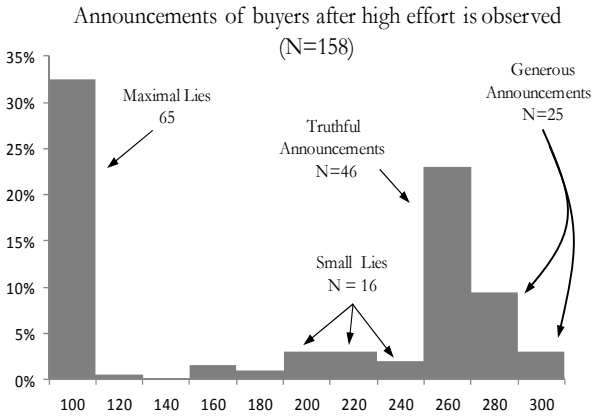
The hypothesis is further corroborated by relating the likelihood of a subject to make a maximal lie to our secondary measure of aversion to gambles. Using data from our follow-up lottery treatment, we divided subjects into two categories: those who accepted the gamble the 50-50 gamble of winning \$12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

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

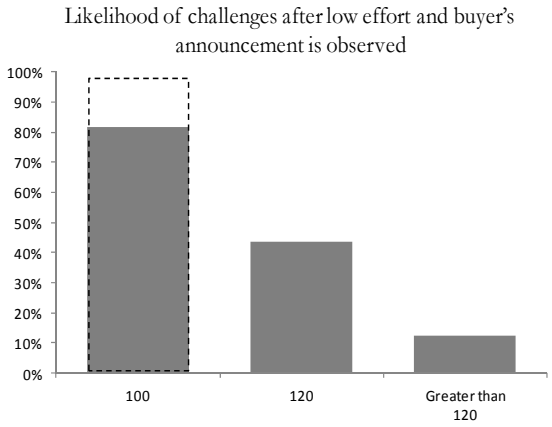
Result B.5 *Buyers in the Low-Fine Treatment opt out of the mechanism in the majority of cases and in similar proportions as seen in the High-Benefits treatment. This aversion to the*

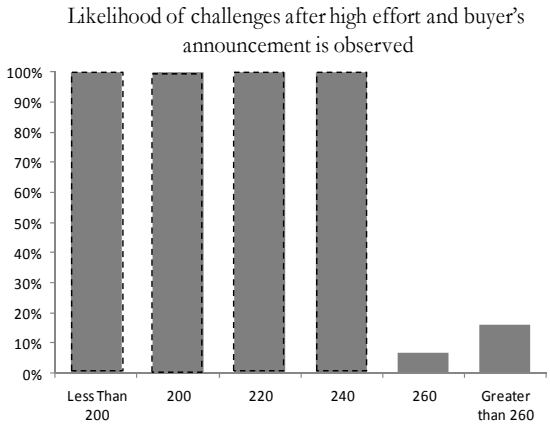
(a) Distribution of announcements after low and high effort





(b) Likelihood of a challenge after each announcement





(c) Number of challenges accepted and rejected

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

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

Grey boxes are predicted action when buyers do not retaliate

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

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

Grey boxes are predicted action when buyers do not retaliate

Figure B4: Pattern of Play in Low-Fine Treatment

mechanism appears to be due to a fear that sellers will challenge truthful announcements.

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

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

B4. The No-False-Challenge Treatment

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

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

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

Sellers' challenge behavior is similar in the two treatments with 59 percent of small lies being challenged in the High-Benefits treatment and 59 percent of small lies being challenged in the No-False-Challenge treatment. The buyers' willingness to reject the challenges of small lies are also similar with 79 percent of challenges being rejected in the High-Benefits treatment and 87 percent of challenges being rejected in the No-False-Challenge treatment.

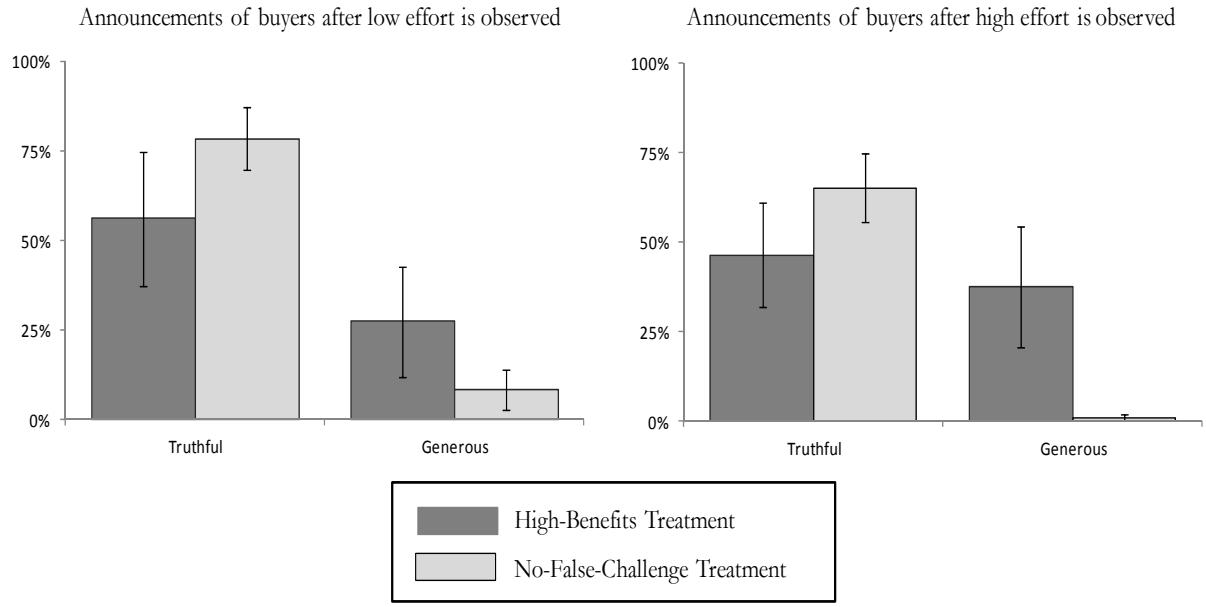


Figure B5: Comparison of Truthful Announcements and Generous Announcements in the High-Benefits and the No-False-Challenge Treatments

Neither difference is significant (Sellers Challenge Behavior: p-value $= 0.97$; Buyers Rejection Behavior: p-value $= 0.55$).

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

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

Appendix B5. The SPI with Intense Training Treatment

A natural hypothesis for the observed pattern of play in the SPI treatment is that subjects make errors in choosing which pure action to play and that they are more likely to choose pure actions that involve higher expected payoffs. In extensive-form games, a useful way

to model such errors is with an Agent Quantal Response Equilibrium (AQRE). AQRE is similar to a standard quantal response model with the additional assumption that at a given decision node, the player determines the expected payoff of each action by treating their future self as an independent player with a known probability distribution over actions.

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

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

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

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

We achieved this with two additional features. *First*, we explicitly explained in the written instructions the pecuniary incentives of subjects' counterparties in the trade. For example, the buyers were explicitly informed that if they announce the true value of the

good and are willing to reject counteroffers if the seller nevertheless challenged their truthful report, it is in the seller's pecuniary interests to refrain from challenging them. Likewise, the sellers were explicitly informed that if they challenge a buyer's lie, then it is in the buyer's pecuniary interest to accept the counteroffer.

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

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

The intense training protocol produced the following results.

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

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

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

Controlling for the difference in effort levels, the SPI with Intense-Training Treatment also has significantly fewer small lies than the SPI Treatment. Panel (a) shows that small lies occur in 28 out of 90 cases after low effort (31 percent) and in 58 out of 310 cases after high effort (19 percent). These small lie rates are low relative to the SPI Treatment where lies occurred 61 percent of the time after low effort and 54 percent of the time after high

effort.[3] However, the lie rate in the SPI with Intense-Training Treatment is still high relative to the predictions of no lies made in SPI Hypothesis 1.

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

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

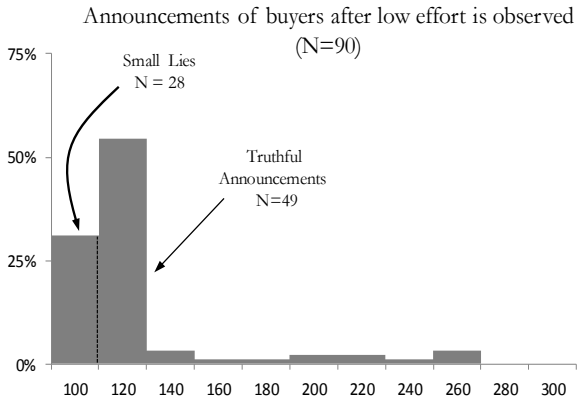
As with the SPI Treatment, buyers and sellers in the SPI with Intense-Training Treatment earn less with the mechanism than is guaranteed without the mechanism. We would thus expect similar opt-in and opt-out behavior between the two treatments.

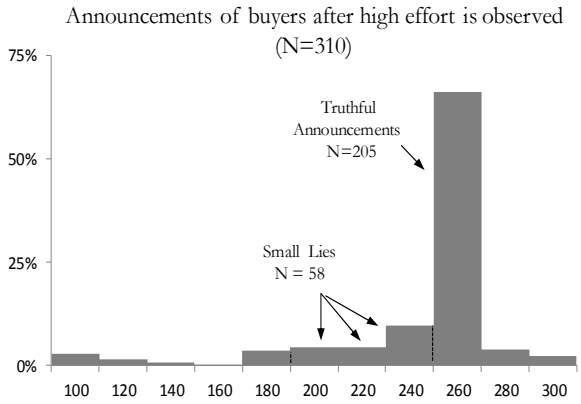
Result B.7 *In the majority of cases, the parties do not adopt the mechanism in the SPI with Intense-Training Treatment. This is largely due to buyers opting out of the mechanism. There is no significant difference in opt-out rates between the SPI Treatment and the SPI with Intesive-Training Treatment.*

Buyers opt-out of the mechanism 57 percent of the time while sellers opt-out 19 percent of the time. These opt-out rates are not significantly different to the buyers' (58 percent) and sellers' (16 percent) opt-out rates in the SPI treatment with standard training (based on a simple probit regression that regresses the opt-in rate on the treatment (p-value = 0.96 for the buyer; p-value = 0.76 for the seller). Buyers who retain the mechanism have an average return of 38.7 while buyers who opt out of the mechanism have an average return of 56.1. In groups where the mechanism is retained, small lies are still reasonably common and occur

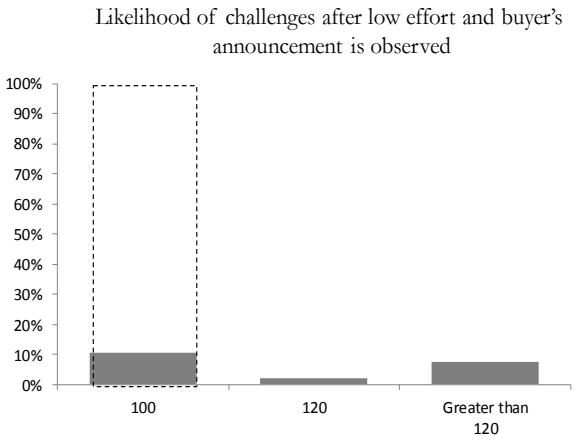
[3]The difference in the propensity to make small lies between the two treatments is statistically significantly different in two separate probit regressions — one for low effort and one for high effort — where a binary variable that is 1 for a small lie and 0 for a truthful announcement is regressed on the treatment variable (p-value < 0.01 for low; p-value < 0.01 for high).

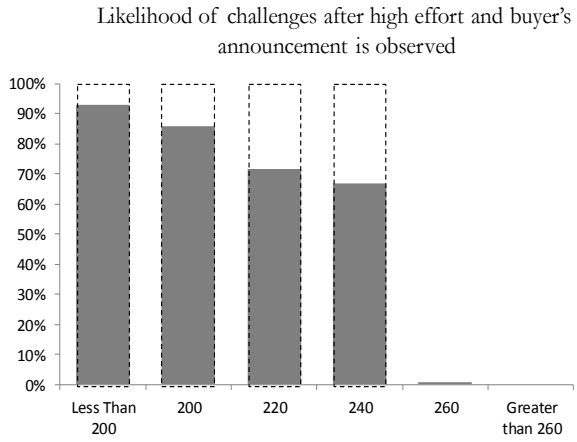
(a) Distribution of announcements after low and high effort





(b) Likelihood of a challenge after each announcement





(c) Number of challenges accepted and rejected

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

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

Grey boxes are predicted action by SPI hypothesis

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

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

Grey boxes are predicted action by SPI hypothesis

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

in 14 out of 36 cases after low effort (39 percent) and in 13 out of 105 cases after high effort (12 percent). Disagreements that occurred after these small lies were the main reason for the reduced profits for the buyers.

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

B6. Personality Measures of Reciprocity

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

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

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

Given this potential non-linear relationship, we construct a binary measure of negative reciprocity that is less sensitive to non-linearities in the relationship between negative reciprocity and small lies. The measure is constructed as follows: we first generate a negative reciprocity score constructed by applying principal-component analysis to the PNR survey using the procedures outlined in Perugini et al. (2003). Individuals who are more negatively

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

reciprocal score higher on this measure. We then divide these scores at the median to construct a binary variable that is 0 for less reciprocal individuals and 1 for more reciprocal individuals.[5]

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

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

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

Table B6: Probit Regression of Small Lies by Buyers

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

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

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

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

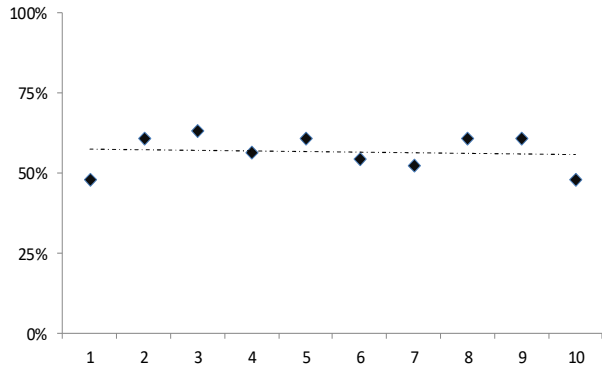
we would predict that the relationship between reciprocity and challenges is likely to be weak. This is indeed the case: extending the probit regression in Table B2 to include negative reciprocity shows that sellers with negative reciprocity scores above the median are not significantly more likely to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

Appendix C: Additional Figures

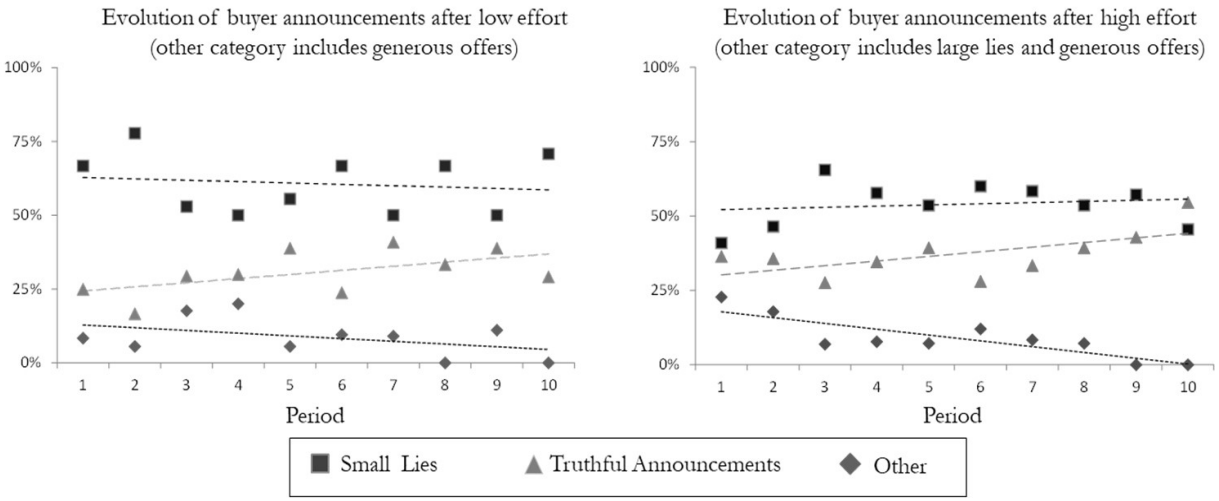
C1. Additional Figures from SPI Treatment

C2. Additional Figures from RS Treatment (Phase 1)

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



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



(c) Proportion of small lies challenged each period

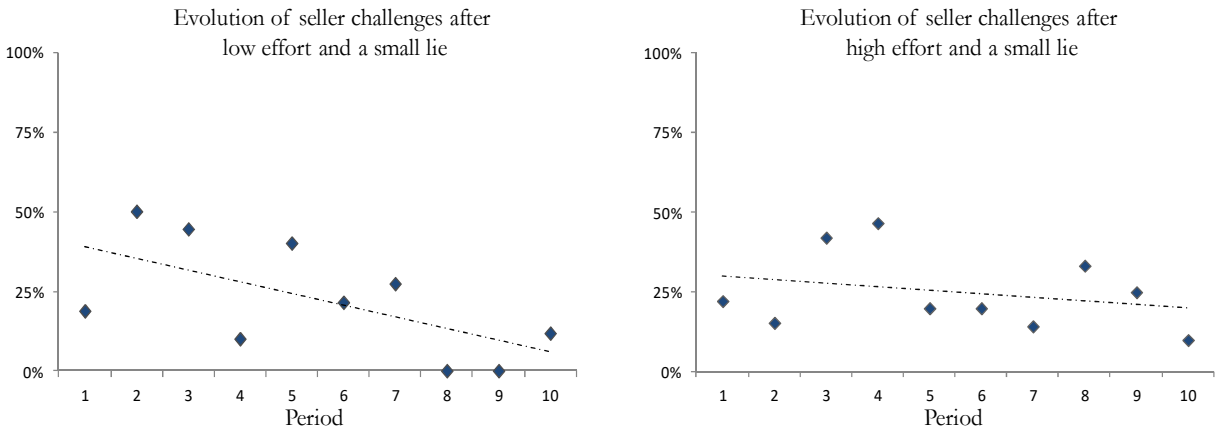
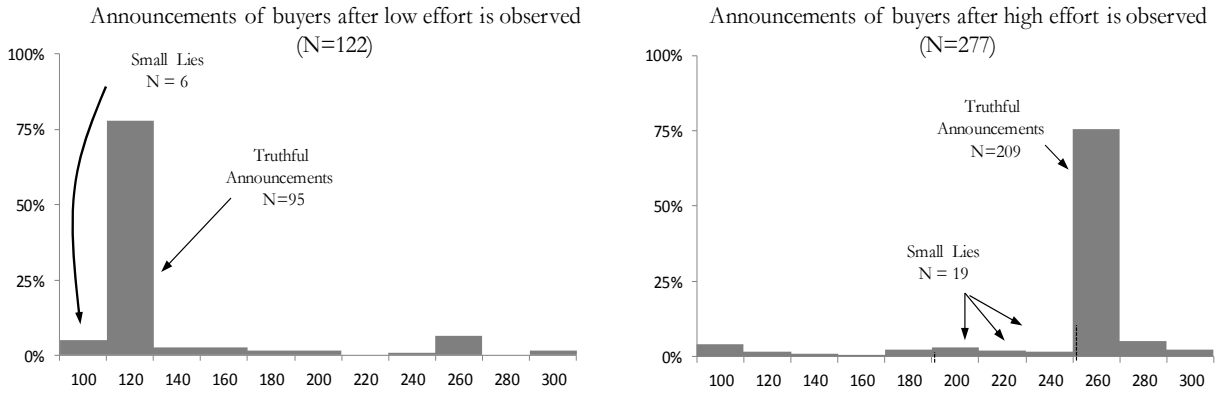
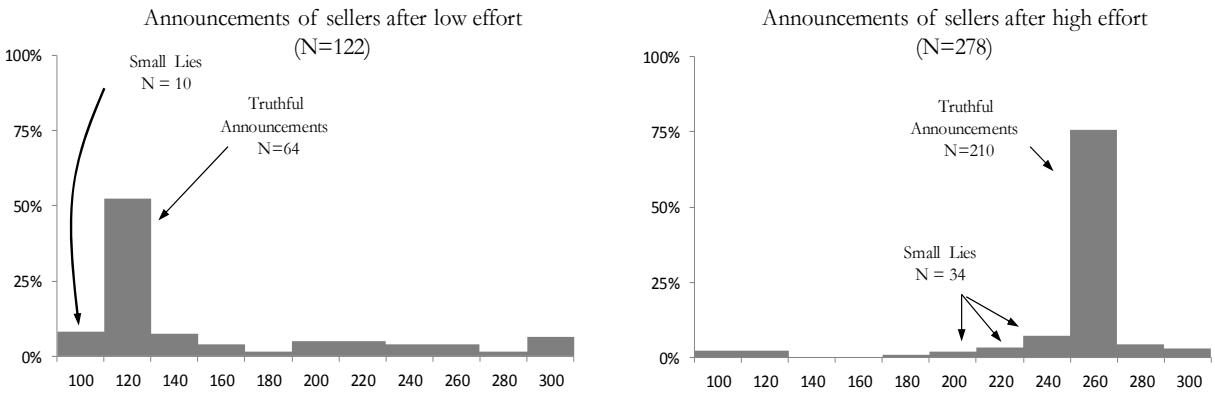


Figure C1: Evolution of Play in First 10 Periods of SPI Treatment

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



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



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

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

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

Grey boxes are predicted outcomes of SPNE with selfish types

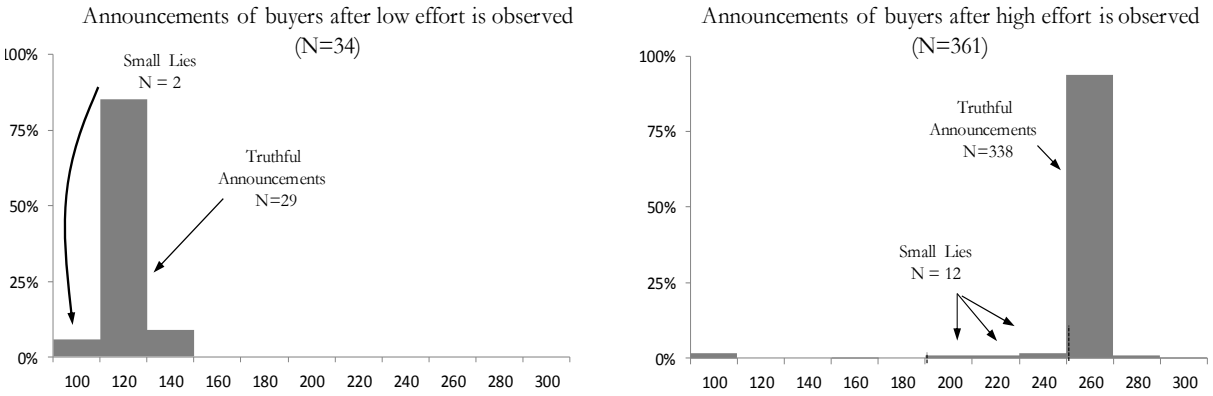
Outcomes of groups where seller effort is high and buyer and seller reports do not coincide

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

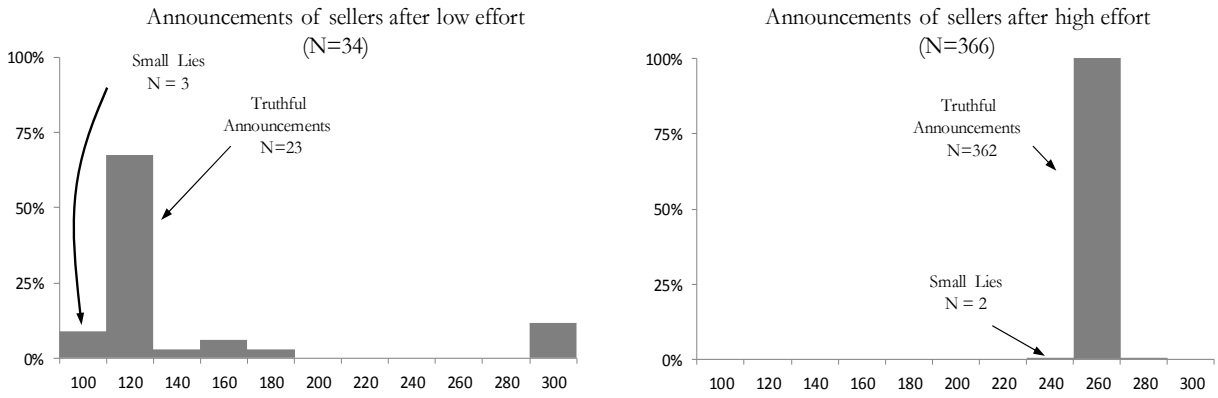
Grey boxes are predicted outcomes of SPNE with selfish types

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

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



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



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

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

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

Grey boxes are predicted outcomes of SPNE with selfish types

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

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

Grey boxes are predicted outcomes of SPNE with selfish types

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

C3. Additional Figures from RS Treatment (Phase 2)

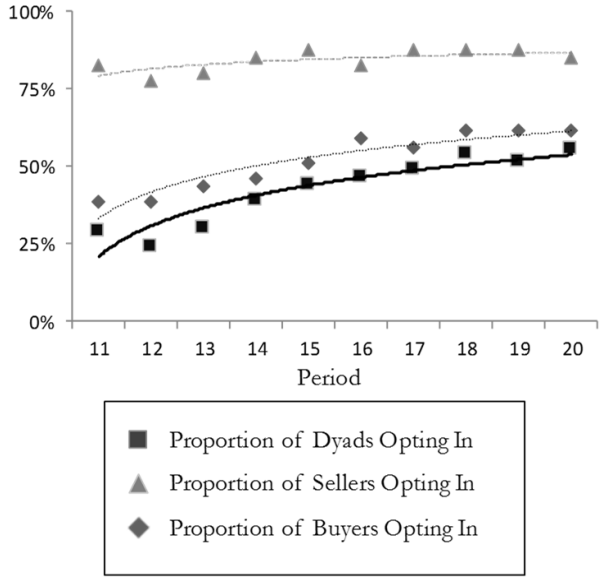


Figure C4: Proportion of Buyers and Sellers Opting Into the Mechanism in Periods 11–20 of RS with Intense-Training Treatment

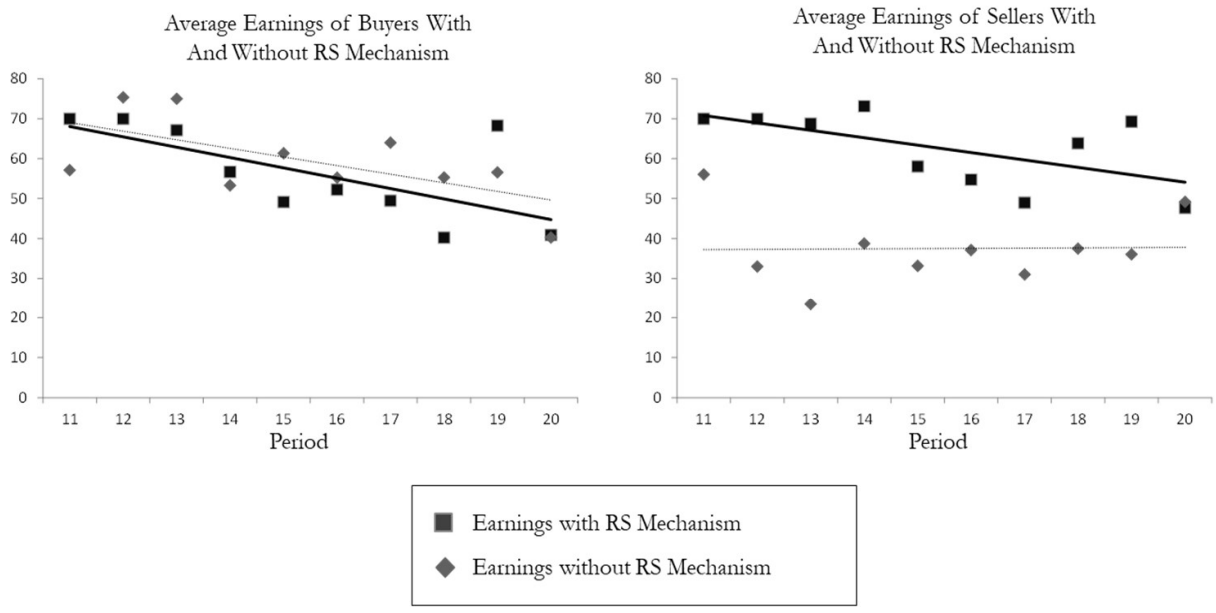


Figure C5: Average Earnings of Buyers and Sellers in Periods 11–20 of RS with Intense-Training Treatment With and Without the Mechanism