ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Fehr, Ernst; Powell, Michael; Wilkening, Tom

Working Paper

Handing out guns at a knife fight: Behavioral limitations of subgame-perfect implementation

Working Paper, No. 171

Provided in Cooperation with:
Department of Economics, University of Zurich

Suggested Citation: Fehr, Ernst; Powell, Michael; Wilkening, Tom (2014) : Handing out guns at a knife fight: Behavioral limitations of subgame-perfect implementation, Working Paper, No. 171, University of Zurich, Department of Economics, Zurich, https://doi.org/10.5167/uzh-98147

This Version is available at:
https://hdl.handle.net/10419/111229

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 171

Handing Out Guns at a Knife Fight: Behavioral Limitations of Subgame-Perfect Implementation

Ernst Fehr, Michael Powell and Tom Wilkening

August 2014

Handing Out Guns at a Knife Fight: Behavioral Limitations of Subgame-Perfect Implementation

Ernst Fehr Michael Powell Tom Wilkening*

August 12, 2014

Abstract

The assumption that payoff-relevant information is observable but not verifiable is important for many core results in contract, organizational and institutional economics. However, subgame-perfect implementation (SPI) mechanisms – which are based on off-equilibrium arbitration clauses that impose fines for lying and the inappropriate use of arbitration – can be used to render payoff-relevant observable information verifiable. Thus, if SPI mechanisms work as predicted they undermine the foundations of important economic results based on the observable but non-verifiable assumption. Empirical evidence on the effectiveness of SPI mechanisms is, however, scarce. In this paper we show experimentally that SPI mechanisms have severe behavioral limitations. They induce retaliation against legitimate uses of arbitration and thus make the parties reluctant to trigger arbitration. The inconsistent use of arbitration eliminates the incentives to take first-best actions and leads to costly disagreements such that individuals – if given the choice – opt out of the mechanism in the majority of the cases. Incentive compatible redesigns of the mechanism solve some of these problems but generate new ones such that the overall performance of the redesigned mechanisms remains low. Our results indicate that there is little hope for SPI mechanisms to solve verifiability problems unless they are made retaliation-proof and, more generally, robust to other-regarding preferences.

Keywords: Implementation Theory, Incomplete Contracts, Experiments
JEL Classification Codes: D23, D71, D86, C92

---

*Ernst Fehr: Department of Economics, Zurich University, Bluemlisalpstrasse 10, CH-8006 Zurich. E-mail: ernst.fehr@econ.uzh.ch. Michael Powell: Department of Management and Strategy, Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208-2001. E-mail: mike-powell@kellogg.northwestern.edu. Tom Wilkening: Department of Economics, The University of Melbourne, Economics and Commerce Building, Victoria, 3010, Australia. E-mail: Tom.Wilkening@unimelb.edu.au.
We thank James Bland and Hans Zhu for excellent research assistance. We also thank Mathias Dewatripont, Martin Dufwenberg, Greg Fischer, Robert Gibbons, Lorenz Goette, Jean Tirole, Christian Zehnder, and seminar participants at the National University of Singapore, the University of Technology, Sydney, and the 2013 Asian Pacific ESA for helpful comments. Funding for these experiments was supplied by the Faculty of Business and Economics at the University of Melbourne and is gratefully acknowledged.

> "To the extent that [existing institutions] do not replicate the performance of [subgame-perfect implementation], one must ask why the market for institutions has not stepped into the breach, an important unresolved question."
>
> -Eric Maskin (2002)

1 Introduction

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

The tractable nature of models using the assumption of observable but non-verifiable information has made them an essential tool for evaluating trade-offs in institutional design. The assumption has been used to understand property rights and firm boundaries (Grossman & Hart 1986; Hart & Moore 1990; Hart 1995), the optimal scope of governments (Hart, Shleifer, & Vishny 1997; Besley & Ghatak 2001), problems of privatization (Schmidt 1996a, 1996b), the control of insiders by outsiders through voting rights (Grossman & Hart 1988; Gromb 1993) and financial contracts (Aghion & Bolton 1992; Dewatripont & Tirole 1994; Hart & Moore 1998), and patterns of international trade and technology adoption (Antras 2003; Nunn 2007; Acemoglu, Antras, & Helpman 2007).

In addition, the observable but non-verifiable assumption often underpins second-best approaches to relational-contracting models in organizational economics (MacLeod & Malcomson 1989; Baker, Gibbons, & Murphy 1994, 2002; Levin 2003) and to efficiency-wage models of motivation (MacLeod & Malcomson 1998; Board & Meyer-ter-Vehn 2014) where signals of past behavior are mutually observable (so that players can coordinate on punishment and rewards) but not verifiable by third parties.

However, despite its widespread influence, the assumption that payoff-relevant information is observable but non-verifiable stands on controversial theoretical foundations. Building on work by Moore & Repullo (1988), Maskin & Tirole (1999) show that if parties commonly observe payoff-relevant information, there often exists an auxiliary extensive form mechanism that can credibly make this information verifiable under subgame perfection. In particular, they point out that there exists a subgame-perfect implementation (SPI) mechanism that

is capable of ensuring truthful revelation of mutually known, payoff-relevant information as part of the unique subgame perfect equilibrium. Therefore, even though payoff-relevant information may not be directly observable by third parties, truthful revelation of information via the mechanism allows for indirect verification, which should make first-best outcomes attainable.

Maskin and Tirole's critique of the microfoundation of incomplete contracting models that use the observable but non-verifiable information assumption is troubling in its implications. Comparing the effectiveness of second-best institutional arrangements under incomplete contracts is moot when a mechanism exists that is capable of achieving first-best outcomes. On the other hand, the limited use of implementation mechanisms in environments with observable but non-verifiable information leads one to question the assumptions of Maskin and Tirole's critique itself. Or to put it like Maskin (2002): why does the market not step in and implement institutions that replicate the performance of efficient SPI mechanisms? If that were the case, first best outcomes in many collective action, agency and institutional design problems could be achieved.

In this paper, we explore Maskin's question by studying the performance and adoption of an SPI mechanism described in Maskin & Tirole (1999) that is designed to resolve the hold-up problem in bilateral exchange with observable but non-verifiable ex-ante effort. We chose this environment as it is the original environment for which observable but non-verifiable information has been assumed, and because it is one of the canonical models in the theory of the firm.

We find that, although the mechanism is predicted to induce perfect truth-telling and high effort levels, it often fails to do either. In addition, the mechanism frequently has disastrous welfare consequences for the involved parties because their behavior under the mechanism deviates strongly from the predicted subgame-perfect equilibrium and these deviations lead to the imposition of sizeable fines on both parties. Due to the frequently negative welfare consequences of the mechanism, the trading parties opt out of the mechanism in the majority of the cases when given the chance to do so. We believe these findings provide a plausible explanation for the absence of the mechanism in real-world environments.

Why though, does the mechanism perform so badly relative to the theoretical predictions? Answering this question is important because it may help theorists to design mechanisms that are immune to the behavioral forces that render Maskin and Tirole's SPI mechanism so inefficient. To answer this question, we briefly need to describe the essential components of the SPI mechanism.

In our experiment, a seller is selling a good to a buyer and may provide costly ex-ante effort to increase the value of the good. Effort and the value of the good are commonly

known to the trading parties, but they are not verifiable by a third-party court. This implies that the two parties cannot write a contract that conditions payments on effort or the value of the good and hence, any effort made by the seller is prone to hold up.

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

The crux of the implementation problem, then, is to give the seller the power to challenge announcements, but to prevent "he said, she said" scenarios wherein the seller challenges the buyer's announcement when in fact he has told the truth. A key property of the SPI mechanism is to set up an institution that combines the seller's challenge with a fine for the buyer as well as a test: if the seller challenges the buyer's announcement, the buyer is immediately fined and given a counter offer that he will only accept if he in fact originally told a lie (for example, if the true value of the good is 200 and he announced that he valued the good at 100, offer to sell it to him for 105). If the buyer accepts and thus reveals that he was lying, the mechanism rewards the seller for appropriately challenging the buyer. If, however, the buyer rejects the challenge, the mechanism also fines the seller and no trade occurs.

Since the value of the good is common knowledge between the buyer and the seller, the seller will only challenge if he knows the buyer will fail the test, because otherwise the seller will be fined. Because the buyer understands that the seller has the incentive to only challenge incorrect announcements, he will make a truthful announcement. Truth telling is therefore part of the unique subgame-perfect Nash equilibrium of the announcement game and can be used to promote first-best effort levels.

We liken the SPI mechanism to handing out guns at a knife fight: the capability of escalating potential losses from conflict can resolve disagreement without a shot ever being fired. However, "handing out guns at a knife fight" will only lead to peaceful encounters if important conditions are met. For example, a key requirement for the incentive to challenge is that the sellers believe that buyers who lied will indeed accept the counter offer; if buyers reject counter offers, the seller faces large costs from being fined. Likewise, a key requirement for the buyer to have the incentive to tell the truth is that he expects the seller to challenge him if he lies. Contrary to the prediction that the mechanism induces perfect truth-telling

and high effort levels, the mechanism frequently fails to achieve either. In addition, sellers often do not challenge buyers' lies and buyers reject counter offers in the large majority (80%) of the cases, leading to disastrous welfare consequences. Handing out guns at a knife fight often escalates conflict rather than bringing about efficient agreements and peace.

While the mechanism fails at all behavioral stages, buyers' and sellers' action profiles and beliefs suggest that deviations from the equilibrium predictions are systematic and internally consistent. Buyers typically underreport the good's value by a small amount. They correctly believe that these "small lies" are unlikely to be challenged. Sellers correctly believe that the buyers are likely to reject the counter offers associated with the challenging of small lies and, therefore, the sellers challenge such lies infrequently. Finally, buyers who have been challenged after a small lie indeed reject the counter offer in the vast majority of cases, which leads to no trade and to both parties having to pay large fines.

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

There are two classes of explanations for why buyers reject these counter offers. The first is that these buyers are simply making random mistakes in choosing between two actions when their private cost of mistakes is not very high. However, random errors cannot explain why the buyers' rejection rate is so high because the random mistake hypothesis also implies that buyers accept counter offers when theory predicts they should reject them. The second explanation is that buyers' choices reflect more than just their pecuniary payoffs: buyers may view challenges of small lies as an unkind act and retaliate against these acts by rejecting the counter offer and inflicting costs on the sellers. In other words, the rejection of counter offers is driven by negative reciprocity motives, which have consistently been shown to be important in laboratory games (e.g., Blount 1995; Offerman 2002; Falk, Fehr & Fischbacher 2008) and field settings (e.g., Dohmen, Falk, Huffman, & Sunde 2008; Kube, Meréchal, & Puppe 2013; Cohn, Fehr, Hermann, & Schneider 2011). It seems quite likely that the buyers view the challenge of a false announcement, in particular if it involved only a small amount of underreporting, as a hostile act, because it triggers the imposition of a large fine on the buyer. However, large fines are a crucial part of SPI mechanisms, because they ensure in theory (i.e., in the absence of non-selfish motives like negative reciprocity) that all incentive compatibility conditions are met.

In the appendix, we introduce negative reciprocity motives into the model and show

that there exist sequential-reciprocity equilibria with behavioral patterns similar to those observed in our experiment. In particular, if subjects are motivated by reciprocity the buyers are willing to reject counter offers after small lies even if they have weak preferences for reciprocity, while the sellers shy away from challenging such announcements even if they have strong preferences for reciprocity.[1] In addition we provide more direct empirical evidence that negative reciprocity plays a role. Two weeks prior to the experiment, we measured subjects' propensity for negative reciprocity with the "reciprocity questionnaire" developed by Perugini, Gallucci, Presaghi, & Ercolani (2002). Buyers who have a strong propensity for negative reciprocity are significantly more willing to make small lies. This makes perfect sense, because these subjects knew that they were willing to reject counter offers after small lies and were likely to be confident that sellers would not challenge such lies.

The failure of the mechanism and the associated behavioral patterns prompted us to run additional treatments where we tried to improve the performance of the mechanism. In response to the high proportion of small lies, we implemented a treatment where the buyers have a stronger incentive to tell the truth.[2] A by-product of this change is that the buyers now earn a larger share of the surplus in equilibrium. These changes in incentives lead to a decrease in small lies and an increase in challenges of such lies. However, we still observed a large number of rejections of counter offers after challenges with the associated bad consequences that both parties pay fines and no trade takes place. Interestingly, this treatment also generates over-reports from some buyers who fear that truthful announcements will be challenged. Overall, this improved mechanism leads to small welfare gains after an initial learning period (during which large welfare losses occurred), but if the trading parties are given the chance to voluntarily opt out of the mechanism, only 35% of the buyer-seller pairs keep the mechanism. Thus, taken together the improved mechanism also does not perform in a promising way.

While the SPI literature is based on the assumption that fines can be unlimited and that performance is likely to improve with fine size, large fines may lead to large losses by the

[1] In equilibrium, reciprocal sellers anticipate this rejection of counter offers, and thus view small lies as unkind. However, the impact of their negative reciprocity on the action profile is muted due to an important asymmetry in the mechanism. When the seller decides whether to retaliate against the buyers' anticipated rejection of a counter offer he can still avoid paying the fine by refraining from the challenge. This implies that negatively reciprocating an unkind action by challenging is very costly for him. In contrast, when the buyer decides whether to reject a counter offer, the fine has already been imposed on him and thus does not count as a cost of rejection. For buyers, retaliation is much cheaper, implying that relatively weak reciprocity motives may already trigger the buyers' rejection of counter offers.

[2] This has been achieved by lowering the initial-price offer that is associated with the buyers' announcement while keeping the counter offer in case of a challenge the same as before. In this treatment, the buyers have a stronger incentive to tell the truth because if they are not challenged they have to pay a lower initial price.

buyer after a challenge and may be a driver of reciprocal behavior. Thus, in principle, a reduction of the fine could reduce buyer retaliation and promote efficiency. Therefore, we implemented a further treatment in which we reduced the fine substantially relative to the main treatment, but we still ensured that all incentive compatibility conditions were met. We hypothesized that a lower fine may reduce the perceived unkindness of a challenge and may thus reduce the buyers' rejection of counter offers, which may then lead to an increased willingness to challenge among the sellers. As a consequence, both parties may display a higher willingness to adopt the mechanism in the presence of smaller, yet still incentive compatible, fines. However, behavior in the low-fine treatment does not lead to a substantial improvement of the mechanism. The reduction of the fine leads to a substantial share of large underreports, followed by an acceptance of the counter offer. This strategy makes sense for a buyer who believes that truth-telling after high effort is also challenged; it is then better to report the lowest possible value of the good – which ensures a low counter-offer price – and accept the fine for being challenged. However, it leads to an aggregate truth-telling rate of only 34.5%. In addition, the buyers' rejection of counter offers after small underreports is still very high (65%). Thus, the mechanism with smaller, yet incentive compatible, fines is also not performing well and, as a consequence, the parties keep the mechanism only in 36% of the cases if they have the chance to opt out.

Taken together, our results suggest that the design of real-world mechanisms must account for the non-pecuniary characteristics of subjects' preferences. Subgame-perfect implementation mechanisms designed under the assumption that participants are self-interested perform poorly and are abandoned by participants. Our findings suggest that reciprocity and other-regarding preferences may cripple proposed mechanisms in many settings. Viable real world mechanisms need to be robust to the retaliatory inclinations of the people and to beliefs about other players' envy and retaliatory propensities. Our findings thus provide exactly the kind of evidence that was asked for in the review paper in Moore (1992) who stated: "Implementation theory should, I believe, be largely driven by applications; and in principle each application should bring with it some assumption about how the agents in that specific situation will plausibly behave." Our results provide the type of information that is needed to assess whether mechanisms are viable and induce truth-telling, and if they fail to do so, why they fail. In addition, they have potentially important implications for Maskin and Tirole's critique of the micro-foundation of incomplete contracts that is based on the observable but not verifiable assumption. The severe behavioral limitations of subgame perfect implementation mechanisms documented in this paper may be one of the reasons why the market for institutions has not stepped in and why incomplete contracts are so frequent.

Apart from speaking to the debate on the micro-foundation of incomplete contracts and the justifiability of the observable but not verifiable assumption, our paper is also related to the theoretical literature on the role of reciprocity in contract (Englmaier & Leider 2012; Netzer & Volk 2014) and mechanism design (Bierbrauer & Netzer 2012[3]), as well as to the experimental literature that examines how negative reciprocity affects behavior in settings with a hold up problem (e.g., Dufwenberg, Smith, & Essen 2011). However, none of these papers empirically studies and identifies the behavioral limitations of subgame perfect implementation mechanisms.

Finally, our paper also contributes to the experimental literature on implementation. Sefton & Yavas (1996) study extensive-form Abreu-Matsushima mechanisms that vary in the number of stages and find that incentive-compatible mechanisms with 8 and 12 stages perform worse than a mechanism with four stages that is not incentive compatible. Katok, Sefton, & Yavas (2002) study both simultaneous and sequential versions of the Abreu-Matsushima mechanism and conclude that individuals use only a limited number of iterations of dominance and steps of backward induction. Based on these papers, we limited our attention to mechanisms that required only two levels of backward induction.[4] Our paper is also related to the recent experimental work of Aghion, Fehr, Holden, & Wilkening (2014), which tests the theoretical predictions of Aghion, Fudenberg, Holden, Kunimoto, & Tercieux (2012). The theory paper shows that the absence of common knowledge about the state of nature limits the performance of SPI mechanisms, and the experimental paper confirms this prediction. This paper is thus complementary to the current paper which demonstrates that subjects' propensity to retaliate, and subjects' beliefs about others' social preferences and retaliatory tendencies thoroughly undermines SPI mechanisms. In addition, the current paper directly studies how the SPI mechanism contributes to the solution of a hold-up problem, whether subjects would be better off with the mechanism, and whether they voluntarily accept the mechanism.

[3]Bierbrauer & Netzer (2012) use the Bayes-Nash fairness equilibrium as solution concept and provide conditions under which a mechanism can be made robust to intention-based reciprocity. They show that two-party problems such as the one studied here cannot be made fully robust to reciprocity since the insurance property cannot be satisfied. The insurance property means that players cannot affect the other players' payoffs by unilaterally deviating from truth-telling.

[4]An extensive experimental literature also exists looking at efficiency of implementation mechanisms in the public goods provision problem. Chen & Plott (1996), Chen & Tang (1998), and Healy (2006) highlight the importance of supermodularity in Groves–Ledyard mechanisms in improving public goods provision. Andreoni & Varian (1999) and Falkinger, Fehr, Gachter, & Winter-Ebmer (2000) study two-stage compensation mechanisms that build on work from Moore-Repullo (1988), while Harstad & Marese (1981, 1982), Attiyeh, Franciosi, & Isaac (2000), Arifovic & Ledyard (2004), and Bracht, Figuieres, & Ratto (2008) study the voluntary contribution game, Groves–Ledyard, and Falkinger mechanisms respectively. Masuda, Okano & Saijo (2014) study approval mechanisms and emphasize the need for implementation mechanisms to be robust to multiple reasoning processes and behavioral assumptions.

2 Subgame-Perfect Implementation

We begin with a description of a simplified version of the Maskin and Tirole mechanism and highlight how it can solve the classic hold-up problem when effort is non-contractible. A seller (S) and buyer (B) bargain over the production and exchange of a good. S can choose an effort level $e \in \{e_L, e_H\}$, with $e_H > e_L$, to produce a good for B that has outside value for the seller of $\underline{v}$. Effort is privately costly to S but increases the value of the good to B. Let v_L and v_H denote B's valuation when S chooses e_L and e_H. Assume throughout that $v_H - e_H > v_L - e_L$ so that high effort is socially efficient.

The good's value to B is *observable* to both parties but *non-verifiable* by a court. To highlight the hold-up problem, assume that after S's effort choice has been sunk, B makes a take-it-or-leave-it offer to S, resulting in a trade price of $p = \underline{v}$. Since the trade price does not depend on S's effort choice, S has no incentive to choose high effort even though doing so would be socially efficient. Consequently, both parties would prefer a trade price that is more sensitive to the actual value of the good, as such a price schedule would provide incentives for S to choose high effort. Formal contracts written directly on this value cannot be used. However, Maskin and Tirole suggest an indirect contract that can achieve the first-best outcome if the parties use the following mechanism, which is based on Moore & Repullo (1988):

1. B and S sign a contract with a third party, whom we will call the arbitrator. The contract specifies *(i)* an **initial-price schedule** $p(\hat{v})$ at which trade may occur, given an announcement $\hat{v}$ that B makes in stage 3 and *(ii)* a **counter-offer schedule** $\hat{p}(\hat{v})$ and fine F, which may jointly be used to mediate disagreement and will be discussed below. Note that both p and $\hat{p}$ are based only on B's announcement, which can be made publicly observable (and therefore verifiable).

2. S chooses effort e to produce a good of value $v(e) \in V$, which is commonly observed by B and S.

3. B announces $\hat{v} \in \hat{V} = \{v_0, v_1, \ldots, v_N\}$, where $v_0 \leq v_L < v_H \leq v_N$ and $V \subset \hat{V}$. $\hat{v}$ is observable to S and the arbitrator.

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

5. B is given a counter offer $\hat{p}(\hat{v})$. If B accepts the counter offer and buys, he pays $\hat{p}(\hat{v})$ and receives the good, and S is paid F by the arbitrator.

6. If B does not buy, S gives the good to the arbitrator, and it is destroyed. Additionally, S must also pay a fine F to the arbitrator.

A **SPI mechanism**, which we will denote by μ, is therefore a collection $\left(p(\cdot), \hat{p}(\cdot), \hat{V}, F\right)$ consisting of an initial-price schedule, a counter-offer schedule, an announcement space, and a fine level. The logic of this mechanism is that the initial-price and counter-offer schedules are constructed so that if B and S are commonly known to be sequentially rational, B never has an incentive to announce a $\hat{v} \neq v(e)$. For this to be true, the SPI mechanism μ must satisfy three conditions.

(a) **Counter-Offer Condition.** B must prefer to accept any counter offer for which he has announced $\hat{v} < v(e)$ and reject any counter offer for which he has announced $\hat{v} \geq v(e)$.

(b) **Appropriate-Challenge Condition.** S must prefer to challenge announcements $\hat{v} < v(e)$ and not challenge announcements $\hat{v} \geq v(e)$.

(c) **Truth-Telling Condition.** B must prefer to announce $\hat{v} = v(e)$ rather than $\hat{v} \neq v(e)$.

We refer to a challenge after $\hat{v} < v(e)$ as an **appropriate challenge** and refer to a challenge after $\hat{v} \geq v(e)$ 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, i.e., for each $\hat{v} < v(e)$, $\hat{p}(\hat{v}) < v(e)$. It also requires that after an inappropriate challenge the counter offer price is above the value of the good, i.e., for each $\hat{v} \geq v(e)$, $\hat{p}(\hat{v}) \geq v(e)$. The counter-offer condition implies that

$$\hat{p}(v_i) \in [v_i, v_{i+1}) \tag{1}$$

for all $i < N$. Note that the difference between $\hat{p}(v_i)$ and v_i cannot be larger than the difference between v_{i+1} and v_i. Thus the maximum return to accepting a counter offer is constrained by the coarseness of the announcement space.

Under a counter-offer schedule that satisfies the Counter-Offer Condition, B will reject counter offers following inappropriate challenges and will accept counter offers following appropriate challenges, so the Appropriate-Challenge Condition is satisfied if S has an incentive to challenge only in cases where B will accept such a challenge (i.e., when B lied). The return to S of appropriately challenging an announcement $\hat{v} < v(e)$ relative to not challenging is

$$\begin{aligned} \Delta_S(\hat{v}, F) &= \hat{p}(\hat{v}) + F - e - (p(\hat{v}) - e) \\ &= \hat{p}(\hat{v}) + F - p(\hat{v}). \end{aligned} \tag{2}$$

Thus, the Appropriate-Challenge Condition is satisfied if $\Delta_S(\hat{v}, F) > 0$ for all $\hat{v} \in \hat{V}$. This will be the case as long as the counter-offer price exceeds the initial price for each announcement (i.e., $\hat{p}(\hat{v}) > p(\hat{v})$ for all $\hat{v}$) and F is positive.

Finally, for the Truth-Telling Condition to be satisfied, B must prefer to report the actual value over any other value. If $p(\hat{v})$ is strictly increasing in $\hat{v}$, then overreported values $\hat{v} > v(e)$ (which by (b) will not be challenged by S) are never optimal for B. B's pecuniary returns to announcing the actual value $v(e)$ relative to an underreported value $\hat{v} < v(e)$ are

$$\begin{aligned} \Delta_B(\hat{v}, v, F) &= v - p(v) - (v - \hat{p}(\hat{v}) - F) \\ &= \hat{p}(\hat{v}) + F - p(v). \end{aligned} \tag{3}$$

If $\Delta_B(\hat{v}, v, F) > 0$ is satisfied for all $v \in V, \hat{v} \in \hat{V}$, then the Truth-Telling Condition is satisfied. Since $\hat{p}(\hat{v})$ and $p(v)$ are strictly increasing in $\hat{v}$ and v, respectively, $\Delta_B(\hat{v}, v, F) > 0$ is guaranteed for all $\hat{v}$ and v if

$$\hat{p}(v_0) + F \geq p(v_H). \tag{4}$$

The fine F is thus set to ensure that making the lowest announcement possible is not more profitable than telling the truth.

Maskin and Tirole show that if F is allowed to be arbitrarily large, initial-price and counter-offer schedules can be constructed such that a unique SPNE exists in which the Counter-Offer, Appropriate-Challenge, and Truth-Telling Conditions hold, and B always announces v truthfully in this SPNE. Because announcements are verifiable, the price schedule $p(\hat{v})$ can then act as a complete contract, restoring S's incentives to choose high effort. Since total surplus is increased when S chooses high effort, prices can be set so that both players prefer to subject themselves to the mechanism rather than to go without it.

3 Experimental Design

In section 3.1, we describe the SPI stage game we implement experimentally and highlight the patterns of play that SPI theory predicts. Section 3.2 discusses several of our experimental design features, and section 3.3 discusses our experimental protocol as well as secondary tasks that we used to measure heterogeneity in preferences for negative reciprocity and aversion to gambles.

Table 1: Correspondence Between Announcement, Prices, and Outcomes in Main 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	70	105	50	40	15	190	-50	155
120	85	125	35	55	-5	175	-35	135
140	100	145	20	70	-25	160	-20	115
160	115	165	5	85	-45	145	-5	95
180	130	185	-10	100	-65	130	10	75
200	145	205	-25	115	-85	115	25	55
220	160	225	-40	130	-105	100	40	35
240	175	245	-55	145	-125	85	55	15
260	190	265	-70	160	-145	70	70	-5
280	205	285	-85	175	-165	55	85	-25
300	220	305	-100	190	-185	40	100	-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. As these are the true values, this mechanism induces truth telling.

3.1 Main Treatment

At the center of our experimental design are two phases of the SPI mechanism described in section 2. Both phases are computerized and vary only in the rules governing the mechanism's adoption.

Phase 1: In the initial 10 periods of each session, a seller S is perfect-stranger matched with a buyer B and chooses an effort level for the production of a tradeable good. S has two possible effort choices. Low effort generates a good B values at 120 at a cost of 30 to S. High effort generates a good B values at 260 at a cost of 120 to S. B observes the actual value of the good.

B then announces $\hat{v} \in \hat{V} = \{100, 120, \ldots, 260, 280, 300\}$. Note that $\hat{V}$ includes (i) the true value for each potential effort choice, (ii) small lies below each true value, and (iii) generous offers above each true value. We discuss this choice of announcement space in Section 3.3.

The announcement $\hat{v}$ affects both the price offered to B, $p(\hat{v})$, and the counter offer made by the arbitrator, $\hat{p}(\hat{v})$, should S challenge $\hat{v}$. The price function is given by

$$p(\hat{v}) = 70 + 0.75(\hat{v} - 100),$$

which is shown in column 2 of Table 1 for all possible announcements.

S may then accept the price $p(\hat{v})$ or challenge $\hat{v}$ by calling the arbitrator, who makes a counter offer to B. If S challenges $\hat{v}$, B immediately pays a fine of $F = 250$ and the

counter-offer price is

$$\hat{p}(\hat{v}) = \hat{v} + 5.$$

If B accepts the counter offer, trade occurs at $\hat{p}(\hat{v})$, and S receives $F = 250$ from the arbitrator. If B rejects the counter offer, no trade occurs, and S is also fined $F = 250$. Note that the cost of production has been sunk at the production phase. Under this parametrization, S has the incentive to challenge $\hat{v} < v(e)$ if he expects B to accept the counter offer if and only if he has lied, since

$$\Delta_S(\hat{v}, F) = \hat{p}(\hat{v}) + F - p(\hat{v}) = 0.25\hat{v} + 260 > 0.$$

B always has the incentives to tell the truth if he anticipates a challenge otherwise, since

$$\Delta_B(100, 260, F) = 165 > 0.$$

Phase 2: In periods $11 - 20$, B and S are given the choice to opt in or opt out of the mechanism prior to S's effort choice. We framed opting out of the mechanism as "dismissing the arbitrator," so that opting in is the status quo. If B and S opt in, they are informed that the arbitrator is available, and play continues as in the first ten periods. If either party opts out, the game is identical to the game in the first phase, except that S may not challenge $\hat{v}$, and trade must occur at price $p(\hat{v})$. Both parties are informed about whether the arbitrator is available but are not informed about the dismissal decision of the other party. This implies that if a subject opts out, he cannot determine whether his counterparty opted in or out.

The mechanism μ, summarized in Table 1, satisfies the Counter-Offer, Appropriate-Challenge, and Truth-Telling Conditions. As a result, the unique subgame-perfect equilibrium involves the following pattern of play.

SPI Hypothesis 1. The path of play under the SPI mechanism involves high effort, a truthful announcement of 260, and no challenges. If S challenges an announcement of $\hat{v}$, B accepts the counter offer if and only if $\hat{v} < v(e)$.

We refer to the behavior described in SPI Hypothesis 1 as **efficient truth-telling behavior** and the resulting outcome as the **efficient outcome**. Note that in this equilibrium B earns 70 and S earns 70. If either party opts out of the mechanism in the second phase, the arbitrator is not available, and B will make the lowest possible announcement, $\hat{v} = 100$, regardless of the true value. S has no incentive to choose high effort in this case and will therefore choose low effort. Consequently, the SPNE payoffs if either party opts out are 50 for the buyer and 40 for the seller, so both parties are better off with the mechanism than

without it. The SPNE thus predicts:

SPI Hypothesis 2. B and S opt into the mechanism in periods $11 - 20$.

3.2 Discussion of Design Features

As the goal of our experiment is to assess the plausibility of using SPI mechanisms in real-world contracting environments, we make a number of design choices that can be divided into roughly two categories: features that make the mechanism easier to implement experimentally and features that broaden the applicability of the mechanism to richer settings.

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

Our choice of initial-price and counter-offer schedules is intended to favor the SPNE, in which B's and S's receive an equal payoff of 70. Our expectation is that preferences for equity, for which there is substantial evidence in laboratory experiments, makes the SPNE more salient. Additionally, in designing the counter-offer schedule, we chose a parametrization that gives most of the possible surplus from the counter offer to B. This increases B's value of accepting a challenge, again biasing play in favor of the SPI hypotheses. We also transferred the entire fine F to the seller in the case of a successful challenge to maximize S's expected value to challenging.

Finally, to ensure that B has strict incentives to adopt the mechanism in the second phase, we give B some of the surplus generated from efficient effort. Absent the mechanism, the unique SPNE involves S choosing $e = 30$ and B announcing $\hat{v} = 100$, yielding payoffs of 50 to B and 40 to S. If the mechanism induces efficient behavior, B's gains from adopting it are 20, and S's gains are 30.

Moore and Repullo show that in a broad class of environments that they refer to as "economic environments," any social choice function can be implemented using a three-stage mechanism. In simpler environments, some social choice functions can be implemented using two-stage mechanisms. For example, in our environment, the efficient outcome can be implemented using a two-stage "option contract" (see, for example, Nöldeke & Schmidt

(1995)).[5] We deliberately explore the performance of a three-stage mechanism in our simple environment with one-sided hold-up and no uncertainty, because if such mechanisms fail to work well in a simple environment, they are even more likely to fail in the more complex environments that necessitate their use.

The specific three-stage mechanism we chose is richer than is necessary to generate efficient truth-telling behavior as the unique SPNE in our simple setting. For example, a mechanism μ' with the more coarse announcement space $\hat{V}' = \{120, 260\}$ but otherwise identical initial-price and counter-offer schedules and fees to μ would still generate efficient truth-telling behavior in our environment. However, in a richer environment with more possible effort choices for the seller (and therefore a richer space of valuations), the mechanism μ' would be unable to generate truth-telling behavior, and if there are different states of the world in which different effort choices are jointly desirable, such a mechanism would not yield efficiency. Enriching μ' to include more potential announcements therefore expands the applicability of the mechanism but also allows B to make small lies about the value of the good. Of course, in the unique SPNE of the mechanism, these additional possible announcements are irrelevant.

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

3.3 Experimental Protocol

The experiments were run in the Experimental Economics Laboratory at the University of Melbourne in May and September of 2009 and were conducted using z-Tree (Fischbacher (2007)). All 280 subjects used in the Main Treatment and follow-up treatments (described in Section 5) were undergraduate students at the University and were randomly invited from a pool of more than 3000 volunteers using ORSEE (Greiner (2004)). Session sizes varied from 20 to 26.

Two weeks prior to the experiment, subjects participated in a Personal Norms of Reciprocity (PNR) survey developed by Perugini et. al. (2003). This survey consisted of 27 questions related to a subject's inclination to punish hostile or reward kind acts. Using principal-components analysis, these questions were combined into orthogonal measures of

[5] Hoppe & Schmitz (2011) experimentally study simple single-price option contracts in a one-sided hold-up environment and find promising efficiency improvements even when renegotiation is allowed.

positive and negative reciprocity for each subject. Subjects earned \$10 for the survey and a \$10 show-up fee, which were used to insulate individuals from bankruptcy. The reciprocity measures were done at the sign-up point rather than during the lab session in order to mitigate demand effects that might occur from running the main treatment and survey sequentially.

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

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

The six lotteries varied in the amount X that could be lost, where $X \in \{4, 6, 8, 10, 12, 14\}$. One of the six gambles was randomly selected at the end of the experiment and paid.[6] These lotteries enable us to construct a measure of heterogeneity in the willingness to accept actuarially fair gambles. Discussion of the lottery task can be found in Fehr & Goette (2007) and Fehr, Herz & Wilkening (2013).

Following the lottery task, subjects were assigned the role of B or S, which was fixed for the duration of the experiment. Subjects were then asked to read the instructions and answer a series of practice questions that were checked by the experimenter. These instructions explained the first phase of the experiment (in which the arbitrator is exogenously available) as well as the rules regarding random matching and payment. The instructions were accompanied by a detailed payment chart showing the price and counter offer for each announcement as well as the payment to B and S for each potential outcome of the game. The instructions explicitly explained how to read this chart, and subjects were required to work through examples of play with announcements of 180 and 260 to ensure that everyone understood the incentives of B and S after a truthful announcement and a lie.

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

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

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

The choice of unpaid beliefs for our main experiment were based on three considerations. First, we wanted to have a full set of belief information including beliefs about counterfactual actions. In order to elicit these beliefs in an incentive compatible way, we would have had to use the strategy method for eliciting S's challenges and B's acceptance or rejection decision. Given that the solution concept of subgame perfection is such an important part of the implementation mechanism, we were averse to using the strategy method at interior nodes. Second, we felt explaining an additional belief elicitation mechanism would take attention away from the main experiment. Third, in games where both beliefs and action are compensated, risk averse individuals may find it optimal to hedge risk by stating beliefs which differ from their true estimates.[7]

The large fine size in the main treatment opened up the possibility that subjects could go bankrupt. As such, the protocol for bankruptcy was made explicit to all subjects. Subjects began the experiment with a \$10 show-up fee and the \$10 from the online survey. If a subject accumulated \$10 in losses, their money from the online survey payment was liquidated, and they received a warning. If they lost all \$20 of their initial endowment, they were removed from the experiment. Over the main treatment and all follow-up sessions, two subjects were removed from the experiment. In these cases, the lab manager took over the terminal and played the SPNE equilibrium path actions as the B and exerted high effort and never challenged any announcement as a S.[8]

4 Experimental Results of the Main Treatment

We describe the results of the Main Treatment in this section. Section 4.1 explores SPI Hypothesis 1, examining play in the first ten periods of the experiment. Section 4.2 uses data on beliefs to interpret some of the results in Section 4.1. Section 4.3 explores SPI Hypothesis

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

[8] All observations with the lab manager are excluded from the analysis. Bankruptcies occurred in one session of the Main treatment and one session of the No-False-Challenge treatment. Removing these sessions does not significantly affect the results.

2. For purposes of categorizing data, we define to $\hat{v} < v(e)$ as a **lie**, $v(e) - 60 \leq \hat{v} < v(e)$ as a **small lie**, $\hat{v} = v(e)$ as a **truthful announcement**, and $\hat{v} > v(e)$ as a **generous announcement**. We continue to define an **appropriate challenge** as a challenge of a lie and an **inappropriate challenge** as a challenge of a truthful announcement or a generous announcement.

4.1 Behavior Under the Mechanism

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

Result 1 (a) *In a majority of cases B's make small lies,* (b) *the large majority of these lies are not challenged by the S's,* (c) *the B's reject counter offers in most cases, and* (d)*, the mechanism does not induce high effort in many cases. On average, (e) the parties would be better off without the mechanism.*

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

Panel (a) shows that in the majority of observations, B's lied: following high (low) effort, only 38% (30%) of B's announce the actual value of the good, while 54% (61%) make small lies. Downward lies are increasingly less frequent the larger they are.

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

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

Finally, average surplus in periods $1-10$ of the experiment for a buyer and seller pair was only 7. To put this into perspective, average total surplus under the efficient outcome

is 140. Average total surplus under the unique SPNE when the mechanism is unavailable is 90. The realized gains from this mechanism relative to the maximum potential gains are then $\frac{7-90}{140-90} = -166\%$.

While the results in Figure 1 are presented as the aggregate of all 10 periods, there is very little change in the pattern of play when looked at on a period by period basis. Figure 2 shows how effort, announcements and challenges of small lies evolve over the first ten periods. As can be seen in panel (a), the proportion of sellers exerting high effort is relatively stable over time with roughly 55% of the sellers exerting high effort each period.

Panel (b) shows the proportion of small lies, truthful announcements, and large lies or generous offers over time. The proportion of small lies is stable and constitute roughly 55 percent of observations. The proportion of truthful announcements is increasing while the proportion of individuals making other announcements (i.e., large lies or generous offers) decreases rapidly. Only small lies and truthful announcements are observed by period 10.

While small lies are stable, the likelihood of a seller challenging such an announcement is actually decreasing over time. As can be seen in panel (c), which shows the proportion of small lies being challenged each period, sellers are very unlikely to challenge a small lie in later periods. This reluctance makes sense given that the likelihood a buyer accepts a counter offer is very low. This implies that the mechanism is actually moving away from the truth-telling equilibrium since sellers are becoming more reluctant to challenge over time.

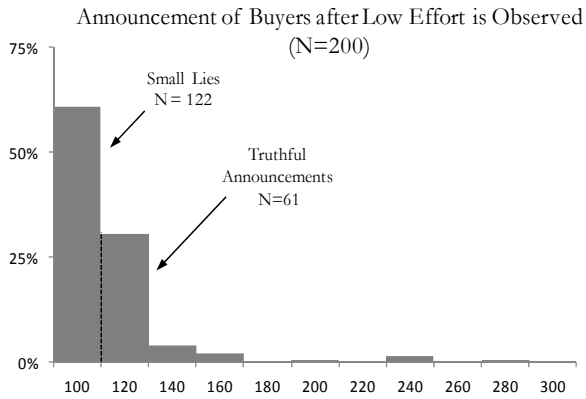
4.2 The Role of Beliefs

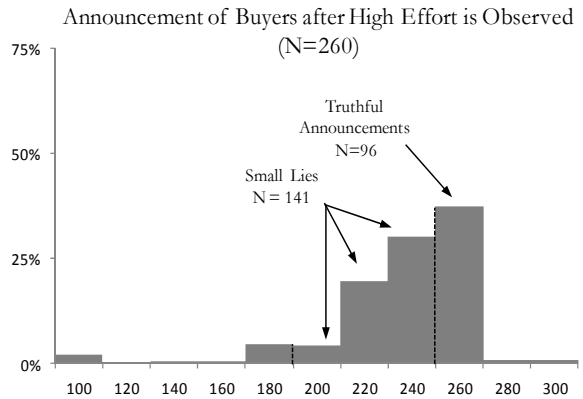
We next explore the role of a subject's beliefs in shaping his or her decisions under the mechanism. If S's believe that counter offers following an appropriate challenge of a small lie will be rejected, they will be reluctant to challenge such announcements. Likewise, if B's believe that small lies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

Result 2 (*a*) *Most B's believe that being challenged for a small lie is unlikely or will never occur. B's who have these beliefs are more likely to lie than those who believe that S's will challenge them.* (*b*) *Most S's believe that a challenge of a small lie is likely to be rejected or will always be rejected. S's who believe that their challenges will be rejected are significantly less likely to challenge a small lie.*

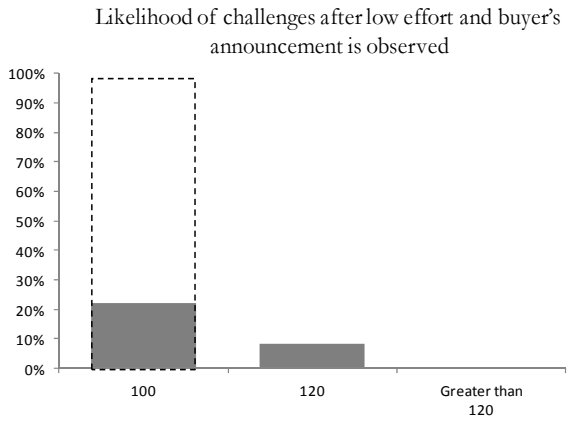
Recall that in each period, we elicited B's beliefs about the likelihood of being challenged for each potential announcement using a 4-point likert scale (Never/Unlikely/Likely/Always).

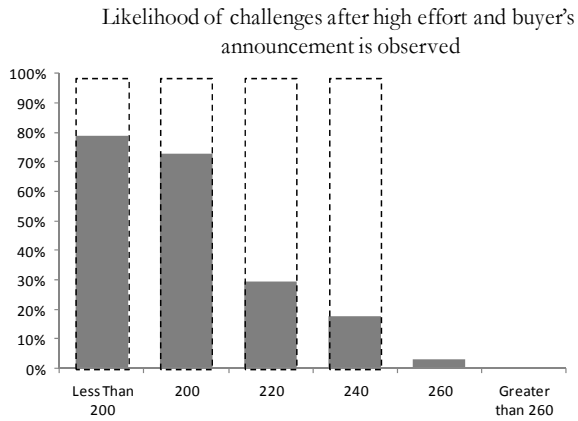
(a) Distribution of Announcements after Low and High Effort





(b) Likelihood of a challenge after each announcement





(c) Number of challenges accepted and rejected

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

Announcement	Challenge Accepted	Challenge Rejected
100	6	21
120	0	5

Grey boxes are predicted action by SPI hypothesis

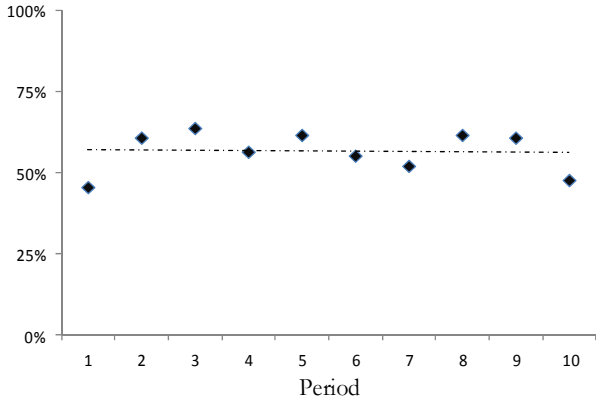
Number of challenges accepted and rejected
after high effort, given announcement and a seller challenge

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

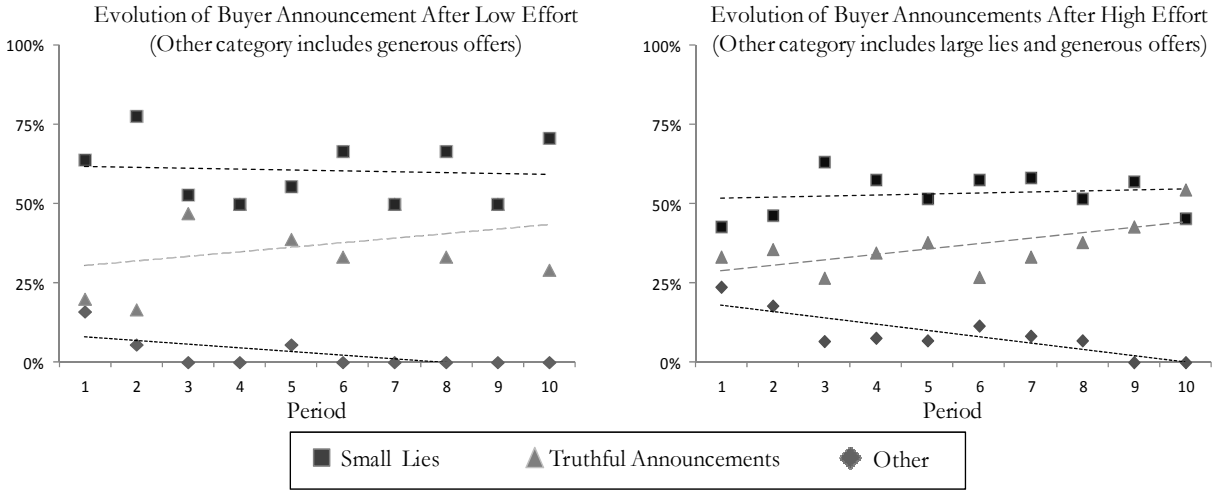
Grey boxes are predicted action by SPI hypothesis

Figure 1: Pattern of Play in First 10 Periods of Main Treatment

(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

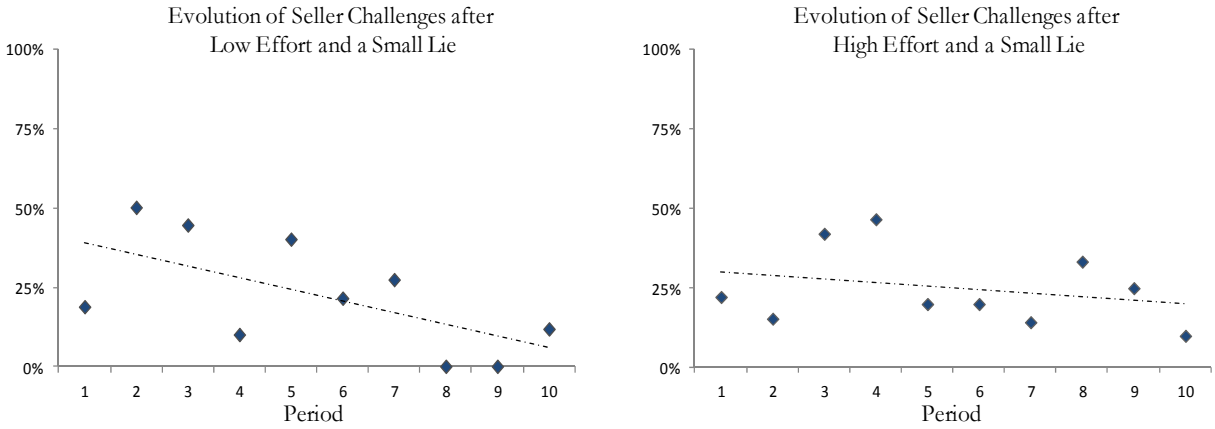


Figure 2: Evolution of Play in First 10 Periods of Main Treatment

Figure 3 shows the proportion of B's who indicated "Never" or "Unlikely" for each announcement after the seller exerts high effort. 82% of B's believe that announcements of 240 are never challenged or are unlikely to be challenged, and 66% believe that an announcement of 220 is never challenged or is unlikely to be challenged. Similar results hold following low effort choices where 52% of B's believe that S is "Unlikely" to challenge or will "Never" challenge an announcement of 100. These results suggest that B's correctly forecast that many S's are reluctant to challenge a small lie.

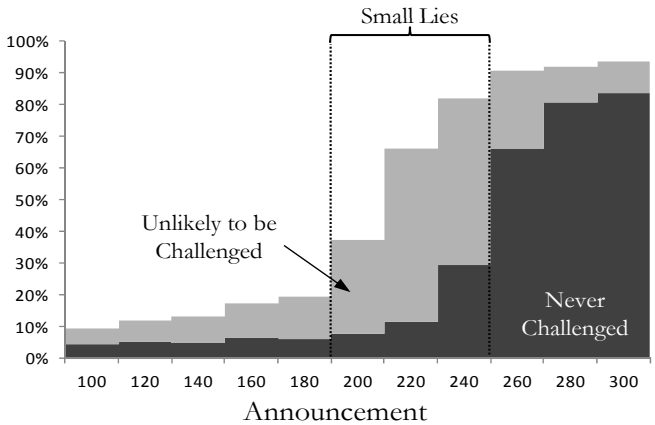


Figure 3: Proportion of buyers who believe that a given announcement will "Never" be challenged or is "Unlikely" to be challenged after observing high effort.

To better understand the role that beliefs have in B's announcements we look at the decision of B's to make small lies based on his belief about being challenged after such lies. Table 2 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.[9] 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 B's 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. Errors are clustered at the individual level for all regressions.

[9]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.

Beliefs about the likelihood of being challenged are a good predictor of B's likelihood of making a small lie. Based on the marginal effects of a probit regression, B's 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 3, while most B's believe that truthful announcements will "Never" be challenged, a small subset of B's 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 B's to make a small lie. To test for this relationship, we extend the probit regression in equations (2), (4), and (6) to also include beliefs about being challenged after a truthful announcement. As expected, individuals are more likely to lie as they become more pessimistic about being challenged after a truthful announcement. Thus optimistic beliefs about being challenged after a lie and pessimistic beliefs about being challenged after a truthful announcement appear to influence B's announcement decision.

Turning to the beliefs of S's, 72% (62%) of S's 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, S's also correctly forecast that B's are likely to reject appropriate challenges.

As with B's, S's are not only correctly forecasting that appropriate challenges will be rejected, they appear to use these beliefs to guide their decisions. Table 3 reports the marginal effects of a probit regression where we regress an indicator for S'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), S's who exert high effort and believe that it is "Likely" that their challenge will be accepted are 39.1 percentage points more likely to challenge than S's who believe that their challenge will "Never" be accepted. Similarly, S's who exert low effort and believe that their challenge is "Likely" to be accepted are 81.7 percentage points more likely to challenge than S's who believe that their challenge will "Never" be accepted.

Taken together, our belief data suggests that individuals are correctly predicting deviations from the SPI predictions in later stages of the game and are responding to these beliefs in a consistent manner. The consistency in the data suggests that the model may be missing an important force which exerts a systematic influence on beliefs and behavior. We return

Table 2: 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 ommited category is Seller "Never" Challenges. Regression (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.

Table 3: Probit Regression of Challenges of Sellers After A Small Lie

	High Effort (1)	Low Effort (2)	Combined (3)
Sellers Belief: Acceptance of Appropriate Challenge "Unlikely"	0.165	0.083	0.108
	(0.131)	(0.131)	(0.088)
Sellers Belief: Acceptance of Appropriate Challenge "Likely"	0.391 ***	0.817 ***	0.604 ***
	(0.121)	(0.089)	(0.089)
Sellers Belief: Appropriate Challenge "Always" Accepted	0.504 ***	0.678 ***	0.586 ***
	(0.187)	(0.155)	(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. Regressions (2) restricts the sampe to periods with Low Effort and a Small Lie. *,**,*** denote significance at the 10%, 5%, 1%-level, respecively.

to this issue after reporting the results from the second phase of the experiment.

4.3 Selection of the Mechanism

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

Result 3 *A majority of dyads opt out of the mechanism. Although the proportion of S's who choose high effort is greater when the mechanism exists, both B's and S's are better off when the mechanism is unavailable than when it is available.*

Panel (a) of Figure 4 shows the opt-out behavior for buyers and sellers over the last 10 periods of the experiment. On average, 65% of dyads have at least one subject choosing to opt out of the mechanism. While this opt-out rate is decreasing over periods 11-15, the opt-out rate continues to be high, with at least 50% of groups opting out of the mechanism in every period. B's are much more likely to opt out of the mechanism (as they did in 58% of the cases) than S's are. The latter opt out of the mechanism in only 16% of the cases.

In the unique SPNE of the game without the mechanism available, the hold-up problem is predicted to be unresolved: S's are predicted to choose low effort and B's are predicted

to make the smallest possible announcement. As can be seen on the right hand side of panel (b), these predictions hold true. When either party opts out of the mechanism, 273 out of 298 sellers exert low effort. In 262 of these cases, B announces $\hat{v} = 100$. Of the 25 observations where the seller put in high effort, B was truthful in only 3 cases, made a small lie in 7 cases, and made the maximal lie of $\hat{v} = 100$ in 15.

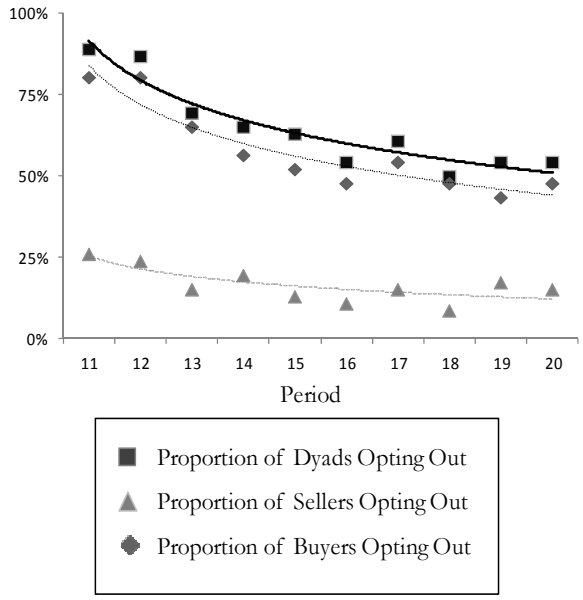
For those periods in which both subjects opted in, we conjectured that the mechanism would perform better than it did in the first phase of the experiment, since opting into the mechanism ought to serve as a positive signal to the other subject in the dyad. From the perspective of effort, this conjecture appears to hold; 114 out of 162 sellers (70%) who had access to the mechanism exerted high effort in periods 11-20 whereas high effort was observed in only 260 out of 460 cases (56.5%) in the first 10 periods. However, when the mechanism is kept, buyers still make small lies in 32 out of 48 cases (66%) after low effort and in 66 out of 114 cases (57%) after high effort. These lying rates are similar to the first 10 periods where the rate of small lies was 61% after low effort and 54% after high effort. Across both effort levels, small lies were challenged in only 13 out of 98 cases (13%), a rate that is similar to periods 8-10.

Empirically, both B's and S's earned *lower* average payoffs in periods in which both subjects opted in than in those in which at least one subject opted out: for observations in which the mechanism was available, average total surplus was 56.4 (36.8 for B's and 19.6 for S's), while for dyad-periods in which the mechanism was unavailable, average total surplus was 94.1 (57.4 for B's and 36.7 for S's).

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

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

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



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

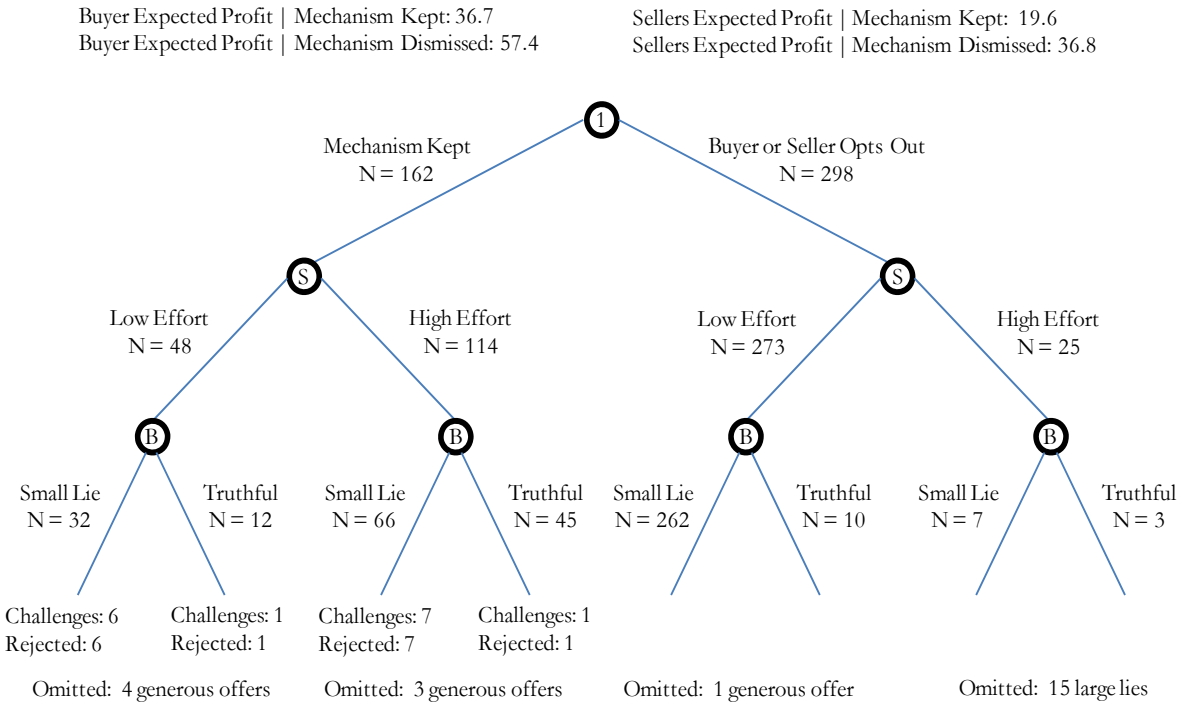


Figure 4: Behavior in Last 10 Periods (Second Phase) of Main Treatment

5 Discussion of Main-Treatment Results

The data soundly reject SPI Hypotheses 1 and 2. However, the mechanism fails at all behavioral stages in a way that is "internally consistent." If B's reject counter offers following appropriate challenges of small lies, then S's have a good reason to shy away from challenging, because it is very costly for them. Yet, if S's do not appropriately challenge small lies, then B's have pecuniary incentives to underreport the value of the good. Indeed, the beliefs data support the above rationale for the failure of the mechanism. S's who believe that counter offers following appropriate challenges of small lies will be rejected are significantly less likely to make such challenges. B's who believe that they will not be challenged for small lies are considerably more likely to make small lies.

Further, S's are right to believe that B's will reject counter offers following appropriate challenges of small lies. Although many S's do not challenge such lies, some do. In these cases, the counter offer is almost always rejected, both parties incur large fines, and no trade takes place. Therefore, the aggregate welfare gains generated by the mechanism are negative. On average, it would be better for the parties to trade the good at low quality without the mechanism than to adopt the mechanism, which explains the observation that the players often do not adopt the mechanism when given the choice.

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

5.1 Exploring potential reasons for rejections of counter offers following appropriate challenges.

Mistakes— A natural initial hypothesis for the observed pattern of play 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 B's, S's may be less likely to challenge small lies. Likewise, B's who correctly forecast S's 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% of appropriate challenges were rejected after high effort and a small lie, AQRE on its own has a hard time fully rationalizing the data.[10]

Retaliation— A second plausible hypothesis is that B's have a preference for retaliation. In our mechanism, after B's announcement has been challenged, B must immediately pay a fine F. B is then presented with two options. He can either buy the good (receiving $v - \hat{p}(\hat{v}) - F$) and "reveal" that he has told a lie, or he can choose not to buy the good (receiving $-F$) and "reveal" that he has told the truth. In the former case, S receives F as a reward and $\hat{p}(\hat{v})$ as compensation for the good. In the latter case, he receives $-F$. The private cost to B of choosing the latter is $v - \hat{p}(\hat{v})$, but the cost to S is $\hat{p}(\hat{v}) + 2F$. If B receives a psychic reward of $r(\hat{v}, v, F)$ for punishing a perceived unkind act, he will do so whenever

$$r(\hat{v}, v, F) \geq v - \hat{p}(\hat{v}). \tag{6}$$

The right-hand side of this inequality may be very small when $\hat{v}$ is close to v. Thus for small lies, B may retaliate.

Further, given that only modest preferences for retaliation are necessary to induce B not to buy the good following an appropriate challenge, S may prefer to challenge B's announcement only in the most egregious of circumstances: small lies may go unchallenged due to a fear of retaliation. Finally, B's may anticipate this fear. When deciding whether or not to announce $\hat{v} = v(e)$, if B is forward-looking and recognizes that S may not challenge small lies out of fear of retaliation, then he has an incentive to tell a lie.

The $r(\hat{v}, v, F)$ representation of preferences for retaliation is reduced form. However, using a slight modification of Dufwenberg and Kirchsteiger (2004), we show in the appendix that a modest preference for negative reciprocity can support a sequential reciprocity equi-

[10] Level-k and other cognitive hierarchy models have a similar difficult time fitting the extent of rejection by B's since only type-0 individuals will reject an appropriate challenge.

librium in which B makes small lies, and S does not challenge these lies. We further demonstrate that when B's preferences for negative reciprocity are drawn from a distribution, and the realization is B's private information, even a B who is fully self-interested may anticipate S's fear of challenging and also lie. This anticipation of fear may ultimately lead to a partial-pooling equilibrium in which S challenges with small probability, and disagreements are common. The losses generated in equilibrium may be enough for parties to opt out of the mechanism or for S to always choose low effort (and thus minimize his own private costs) under the mechanism.

5.2 External Measures of Negative Reciprocity

Two weeks prior to the experiment, we collected data on individual preferences for negative reciprocity using the Personal Norms of Reciprocity (PNR) survey. If negative reciprocity is the key force responsible for the prevalence of lies, we should see a between-subject correlation between a measure of preferences for negative reciprocity and the propensity to make a small lie.

Based on a simple model with reciprocity and no noise discussed in the appendix, the relationship between negative reciprocity and small lies is expected to be weakly monotonic but potentially non-linear.[11] Given this potential non-linear relationship, we construct a binary measure of negative reciprocity that is less sensitive to non-linearities in the relationship between negative reciprocity and small lies. The measure is constructed as follows: we first generate a negative reciprocity score constructed by applying principal-component analysis to the PNR survey using the procedures outlined in Perugini et. al. (2003). Individuals who are more negatively reciprocal score higher on this measure. We then divide these scores at the median to construct a binary variable that is 0 for less reciprocal individuals and 1 for more reciprocal individuals.

Table 4 shows the marginal effects of the negative reciprocity measures in an extension of the probit regressions performed in Table 2. 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

[11] 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 S's are reluctant to challenge, less reciprocal B's 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.

data in the same way as they are done in the previous regressions.

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 $< .01$). 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 $= .015$).

Table 4: 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 ommited 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 S's willingness to challenge and his level of negative reciprocity. However, as described in the appendix, S preferences for reciprocity must be very strong in order to be willing to challenge a B that will surely retaliate. Thus, we would instead predict that negative reciprocity is a weak force in the decisions of the sellers. This is indeed the case: extending the probit regression in Table 3 to include negative reciprocity shows that S's with negative reciprocity scores above the median are not significantly more likely to challenge after high effort (p-value $= .770$), low effort (p-value $= .832$), or in the combined sample (p-value $= .640$).

Taken together, preferences for retaliation appear to be important for rationalizing the observed patterns of play. In the next section, we will take this idea seriously and explore a number of additional treatments that seek to improve the mechanism's performance in the presence of subjects with preferences for retaliation.

6 Toward a Retaliation-Robust Mechanism

The Maskin and Tirole mechanism is based on three conditions that must be satisfied in order for the mechanism to function: the counter-offer condition, the appropriate-challenge condition and the truth-telling condition. Sections 4 and 5 provide evidence that the counter-offer condition fails due to reciprocity leading to further violations in the appropriate challenge condition and the truth-telling condition.

In this section we run two additional treatments aimed at improving the mechanism. The first treatment, the **High-Benefits Treatment**, makes changes to the initial-price schedule that increases the expected value of truth-telling for B's relative to small lies. By making truthful reports more attractive it is predicted that the truth-telling condition will be satisfied for a larger subset of B's even if the other two conditions do not fully subscribe to the SPI hypothesis.

In the second treatment, the **Low-Fine Treatment**, our goal is to reduce the degree to which B's want to retaliate by reducing the fine F. If B's propensity to retaliate is related to the magnitude of his losses due to a challenge, or the extent to which rejecting the challenge reduces S's profit, reducing the fine should reduce the willingness of some B's to reject appropriate challenges. Our second treatment thus targets the counter-offer condition and aims to increase truth-telling indirectly.

6.1 High-Benefits Treatment

Under the SPI hypothesis, the appropriate-challenge condition predicts that S's always challenge a lie and never challenge a truthful or generous offer. As was seen in panel (b) of Figure 1, the seller's do not behave in accordance with this condition, because small lies are challenged infrequently.

While the appropriate-challenge condition is violated, the likelihood that S will challenge is decreasing in the size of B's announcements. Thus, the empirical distribution of S's 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 B'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 B to make a truthful report.

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 B believes that a small lie will be challenged 50% of the time and a truthful announcement will never be challenged, then an increase in the value of not being challenged of 10 will increase the expected value of the small lie by 5 $(10 * .5)$ and increase the value of truth telling by 10.

In the High Benefits treatment we make precisely this type of shift in the value of not being challenged by decreasing the initial-price schedule $p(v)$ uniformly across all announcement. The structure of this treatment is just as in the main 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 5, 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 S's exerting high effort.

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

Result 4 *The High-Benefits Treatment has a larger proportion of S's who exert high effort than the Main Treatment. It also has fewer small lies and S's are more likely to challenge these lies. However, B's still retaliate against most challenges, leading to inefficiency. Thus, although the High-Benefits Treatment improves the efficiency of the mechanism relative to the Main Treatment, the mechanism's efficiency still remains very low.*

Figure 5 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 S selects low

Table 5: 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.

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 S's who exert high effort in the Main and High-Benefits treatments, the High-Benefits treatment has a larger proportion of S's who choose high effort. In the Main Treatment, S's select high effort in only 260 out of 460 observations (56.5%), while S's in the High-Benefits treatment choose high effort in 194 out of 260 observations (74.6%). This difference is significant in a simple regression where effort choice is regressed on the treatment variable with errors clustered at the individual level (p-value = 0.018). Similar results hold for a rank-sum test where an observation is the proportion of time that a subject chooses high effort (p-value = 0.037).

Controlling for the difference in effort levels, the High-Benefits Treatment also has significantly fewer lies than in the Main Treatment. Panel (a) shows that small lies occur in only 11 out of 66 cases after low effort (16.7%) and 30 out of 194 cases after high effort (15.5%). These small lie rates are very low relative to the Main Treatment where lies occurred 61.0% of the time after low effort and 54.2% 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 $< .01$ for both regressions with errors clustered by individual). Alternatively, clustered versions of the Mann-Whitney rank sum test developed by

Datta and Satten (2005) yield similar results (p-value < 0.01 after both high and low effort). Interestingly, unlike the Main Treatment, B's in the High-Benefits Treatment now make generous announcements, $\hat{v} > v(e)$. For example, after high effort buyers make generous announcements in 37.6% of the cases. The large proportion of these generous offers suggests a new deviation from the SPNE model that did not occur in the Main Treatment. We return to this issue when we discuss the beliefs data below.

Looking at Panel (b) and comparing it to the main treatment, S's are much more likely to challenge small lies in the High-Benefits Treatment: following high effort, announcements of 240 are challenged 58.3% of the time as compared to 17.7% of the time in the Main Treatment. These differences are statistically significant, based on a probit regression of an indicator that is 1 if S challenges and 0 otherwise on the treatment variable, with errors clustered at the individual level (p-value < 0.01).

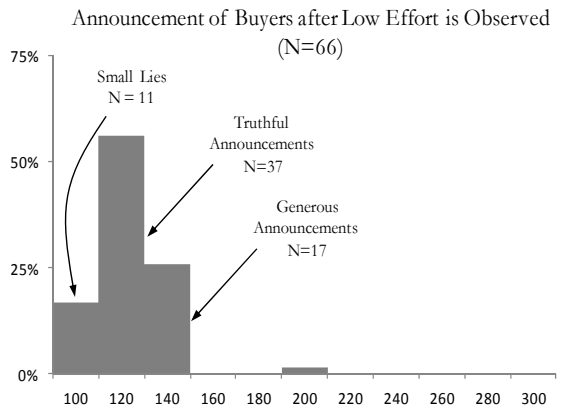
Despite the apparent increase in effort and decrease in small lies, retaliation is still frequent in our data. Panel (c) shows that B's reject the vast majority of legitimate challenges after both high and low effort (80% after high; 75% after low), just as in the Main Treatment. Thus, while the High-Benefits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

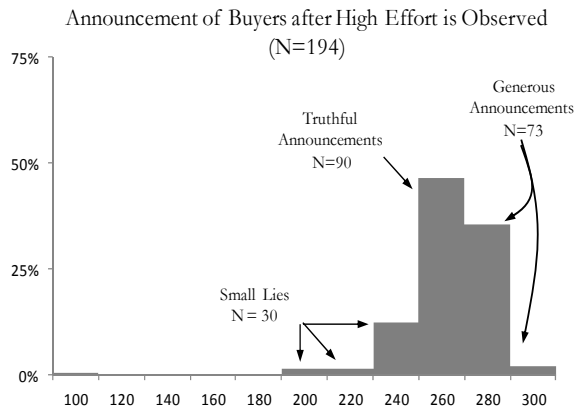
Taken together, the High-Benefits treatment has a larger proportion of truthful announcements and higher effort than the Main Treatment. However, the losses that occur due to disagreement in early periods of the experiment are larger than the gains that occur from improvements in effort and therefore the mechanism continues to reduce overall welfare. 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 $\frac{-7.9-90}{140-90} = -195.8\%$ 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 15.8%) but these gains cannot offset the early losses that arise from disagreement.

Given the improvement in efficiency, the greater level of truth telling, and the reduction in disagreements over time, we might expect that B's and S's 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.

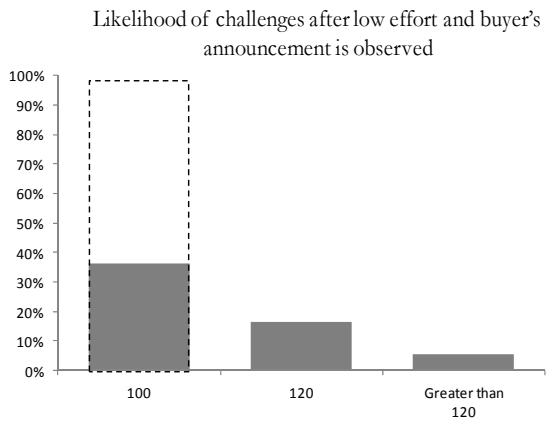
Result 5 *In a majority of cases, the parties do not adopt the mechanism. This is largely due to B's dismissal of the mechanism which stems from Bs' high propensity to render the mechanism unprofitable by making generous announcements. Generous announcements are caused by Bs' beliefs that truthful announcements are challenged.*

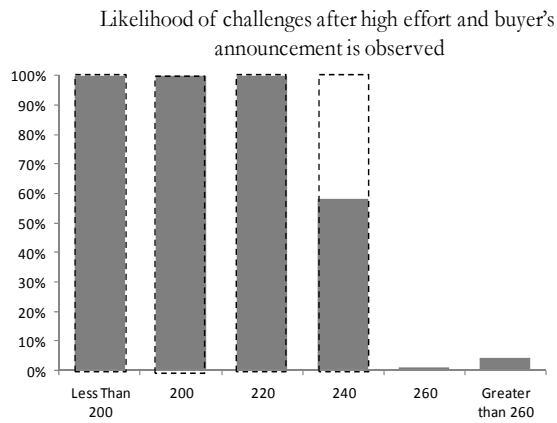
(a) Distribution of Announcements after Low and High Effort





(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	0	1
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 5: Pattern of Play in High-Benefits Treatment

Panel (a) of Figure 6 shows the opt-out behavior of B's and S's over the ten periods of the treatment. As can be seen, B's opt out rate is 80.8% in period 11 and converges to 57.7% by period 20. The average opt-out rate of 64.6% is higher but not significantly different from the average opt-out rate of 57.6% in the Main treatment (p-value = 0.181). S's opt-out rates in the High-Benefits Treatment is low at 3.4%, suggesting that the high opt-out rate is primarily due to the dismissal of the mechanism by B's.

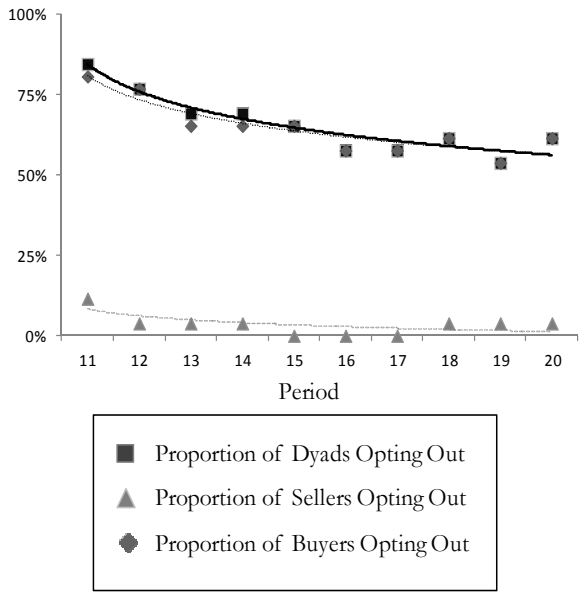
As with the Main Treatment, in periods without the mechanism, the hold-up problem is unresolved. As seen in panel (b), when either party opts out of the mechanism, 154 out of 171 S's exert low effort. In 134 of these cases B announces $\hat{v} = 100$. Of the 17 observations where high effort is observed, B makes a maximal lie of $\hat{v} = 100$ 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 108.4 relative to 95.0 when the arbitrator is dismissed. However, the increase in average efficiency is enjoyed primarily by the sellers; looking at B's profits in isolation, B's expected profits actually decrease from 76.4 when the mechanism is dismissed to 71.1 when the mechanism is kept. Thus the decrease in S's opt-out rate and the lack of change in B'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 B's generous announcements. Relative to the SPNE without the mechanism where low effort is exerted and B announces a value of 100, the SPNE with the mechanism available leads to an increase in B's payoffs of 20 and an increase in S payoffs of 30. When B makes a generous offer, however, he effectively transfers a large portion of the potential gains from the mechanism back to S. These transfers make the mechanism unattractive to B's from an expected value standpoint.

Why do the B's behave in a manner that renders the mechanism unprofitable for them? One likely reason for B's generous offers is that they have pessimistic beliefs about the likelihood of challenges by the seller after a truthful announcement. While S challenge truthful announcements very rarely (1 out of 90 cases after high effort; 6 out of 37 cases after low effort), a B who believes that truthful announcements may be challenged may choose to make a generous offer as a way of reducing the probability of a challenge. Our belief data support the hypothesis that B's have such pessimistic beliefs. In comparison to the distribution of beliefs in the Main treatment where 66.1% of B's believed that a truthful announcement would never be challenged, only 41.7% of B's in the High Benefits Treatment

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



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

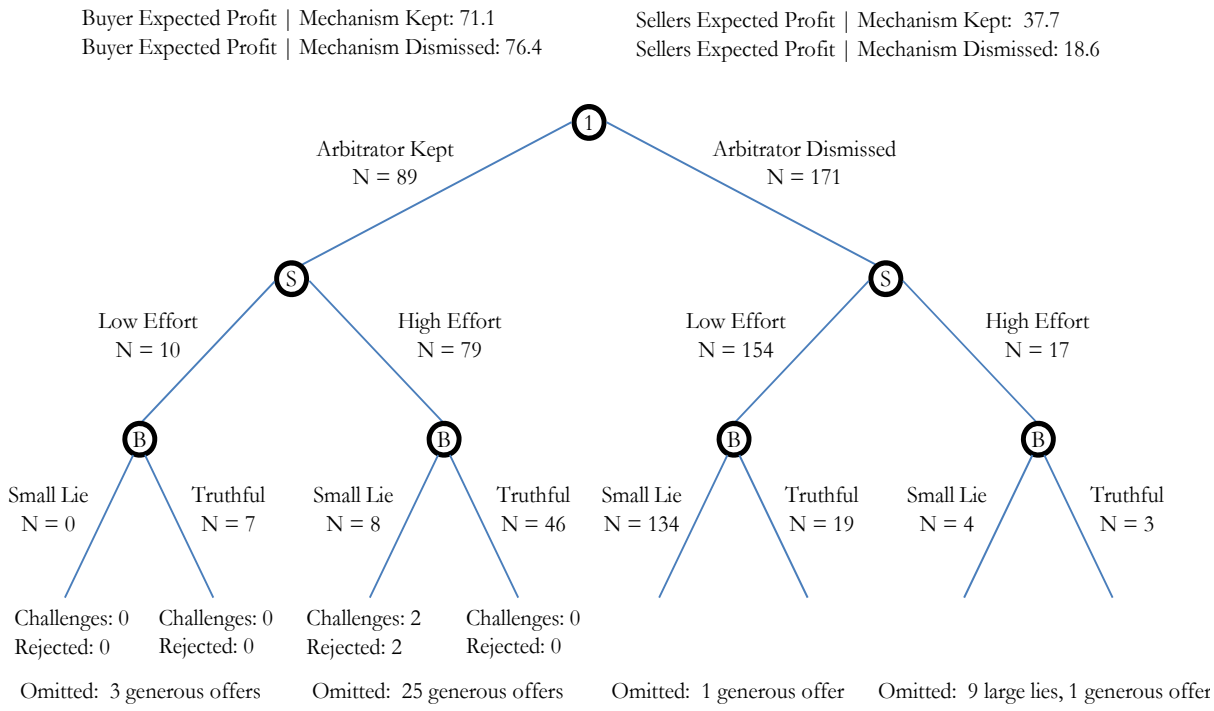


Figure 6: Behavior in Last 10 Periods of High-Benefits Treatment

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 B's believing that at least some S's dislike unequal allocations of surplus. Unlike the Main treatment where B's and S's enjoyed an equal split of surplus along the equilibrium path, the High-Benefits treatment reduces the price that occurs without a challenge and gives B a payoff of 90 while S receives 50. If B's feel that S's have a distaste for such unequal allocations, they may make generous offers which lead to more equitable surplus splits. Thus B's beliefs about the distribution of other-regarding preferences in the population could explain the fear of inappropriate challenges.[12]

To better understand the role that beliefs have in making generous announcements we look at how decisions of B's to make generous announcements depends on his belief about being challenged after truthful announcements. Table 6 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 B'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 B's likelihood of making a generous announcement. Based on the marginal effects of a probit regression, B's 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."

If the fear of inappropriate challenges is a main driver of generous announcements and the dismissal of the mechanism, then eliminating the ability of S to challenge truthful announcements should eliminate generous offers and improve uptake of the mechanism. In the appendix, we report on an additional **No-False-Challenge Treatment** where we use an identical parametrization to the High-Benefits Treatment but augment the mechanism to not allow truthful announcements to be challenged. We find that this treatment eliminates generous offers in periods where high effort occurs and significantly increases truthful announcements by B's. However, the proportion of B's opting into the mechanism improves only slightly and the proportion of S's opting into the mechanism decreases. This suggests

[12]Note that B's 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 6: 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.

that it is hard to satisfy both parties concerns about the mechanism simultaneously.

In aggregate, the High-Benefits treatment does indeed increase the probability of truthful announcements and decrease the probability of small lies. However, B's 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 B's incentives to use the mechanism and ultimately leads B's to dismiss the mechanism when the mechanism is voluntary.

6.2 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 B's desire to reciprocate and potentially improve the performance of the mechanism.

The large fine in the Main treatment was chosen as we were interested in testing the general application of the Maskin and Tirole 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 B's 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 B's return to retaliation, $r(\hat{v}, v, F)$, may be increasing in F. First, as F goes up, B's losses due to a challenge increase. If B'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 B can hurt S for retaliating also increases. Thus, when the fine is lower, the amount of S's 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 S's 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 5. 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 6 *In the Low-Fine treatment, S's effort choices and B'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 B's make the lowest possible announcement,* $\hat{v} = 100$. *These "maximal lies" are more frequent among B's who are averse to gambles and who fear inappropriate challenges. S's almost always challenge small lies and B's still retaliate against the majority of these challenges.*

Figure 7 displays the results for the Low-Fine treatment with data aggregated across all 10 periods. The figure shows that S's exert high effort in 158 out of 200 cases (79.0%), a rate that is similar to the effort rates found in the High-Benefits treatment (74.6%). The small difference in these effort rates is not significantly different in a regression of effort choice on the treatment dummy with errors clustered at the individual level (p-value = 0.549).

Panel (a) shows that B's 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 13.5% is similar to that found in the High-Benefits treatment (15.7%) 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 = .646, clustered by individual). B's 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 34.5% is lower than the 48.8% found in the high benefits treatment, but not significantly different using the same specification as above ($p = .147$, clustered by individual).

There are, however, striking differences in the announcement distribution between the Low-Fine Treatment and the High-Benefits treatment. After high effort, B's in the Low-Fine treatment make maximal lies in 65 out of 158 cases (41.1%) and make generous offers in only 25 out of 158 cases (15.8%). This contrasts strongly with the maximal lie rate of 0.5% and generous offer rate of 37.6% in the High Benefits Treatment. We discuss these maximal lies in detail after describing actions in the other stages of the game.

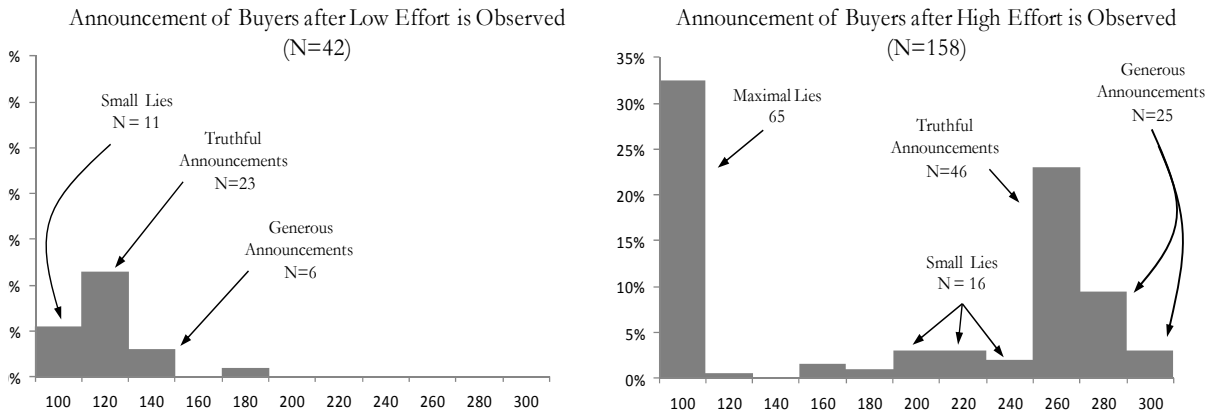
Seller's challenge rates in the Low-Fine treatment are very high, with all small lies and all maximal lies challenged after high effort and 81.8% of small lies challenged after low effort. The challenge rates of lies is significantly higher than the High-Benefits treatment in a probit regression where S's challenges are regressed on the treatment effect and the sample is restricted to lies or small lies with errors clustered at the individual level (all lies: p-value $< .01$; small lies: p-value $< .01$). 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 $= .110$) or both truthful and generous announcements (p-value $= .074$).

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

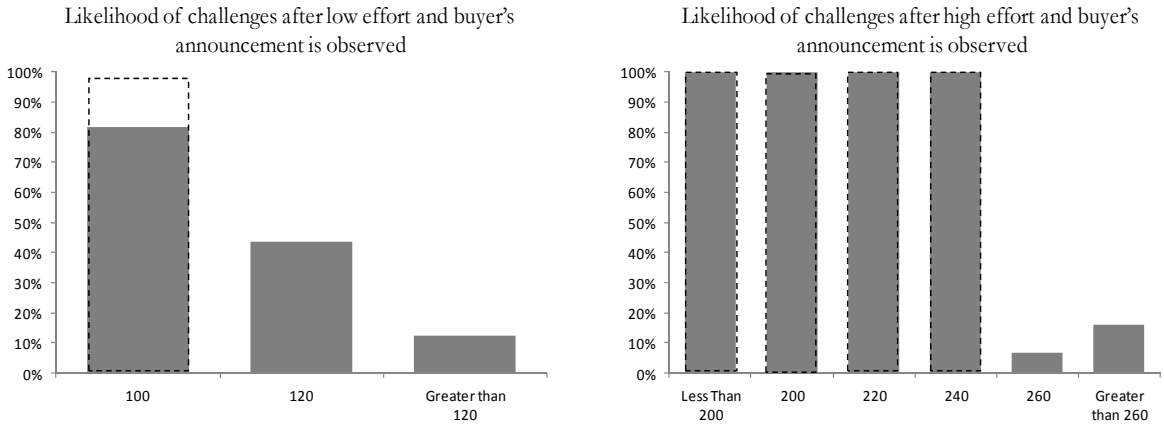
Challenges of small lies are rejected in 9 out of 16 cases after high effort and in 7 out of 9 cases after high effort. While the aggregate rejection rate of challenges after small lies of 64% is 15.2 percentage points lower than the High-Benefits treatment, the difference is not significant in a probit regression that regresses the acceptance rate of small lies on the treatment (p-value $= 0.232$, errors clustered by individual). This suggests that retaliation has not been fully resolved in this treatment.

Why do the B's 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

(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 7: Pattern of Play in Low-Fine Treatment

Table 7: The Relationship Between Maximal Lies and Aversion to Gambles.

	Averse to Fair Gambles	*Accept Fair Gambles*
Truthful Announcement	39	12
Maximal Lies	62	1

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 B's who believed that they would never be inappropriately challenged, maximal lies occur in 28% of the observations. For B's who believe that inappropriate challenges were "Unlikely," "Likely," or would "Always" occur, maximal lies occurred in 46% 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 accept all gambles for which the expected value of the gamble was above the certain payout of zero and those who rejected at least one of these gambles. Table 7 shows the number of observations in which S's exerted high effort and B's announced either a maximal lie or the truth. B's 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 all positive expected value gamble in the lottery game and 0 if he rejects any of them (p-value $< .01$, errors clustered by individual).

As with the High-Benefits treatment, B's in this treatment take strategic actions that shift surplus from B to S 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 7 *B's in the Low-Fine Treatment opt out of the mechanism in the majority of cases and in similar proportions as seen in the High-Benefits treatment. This aversion to the mechanism appears to be due to a fear that S's will challenge truthful announcements.*

B's opt-out behavior is almost identical to that in the High-Benefits treatment with opt-

out rates converging to 60% from above with an initial opt-out rate of 85%. The average opt-out rate of 64.0% in the Low-Fine treatment is not significantly different to the average opt-out rate of 64.6% in the High-Benefits treatment. S's opt-out rate of 3.5% is also not significantly different to the opt-out rate in the High-Benefits treatment. B's who retain the mechanism have an average return of 59.1 while B's who opt out of the mechanism have an average return of 74.1. This loss of profit from B's who retain the mechanism is due primarily to maximal lies and generous offers which transfer surplus to S.

Taken together, the Low-Fine treatment shares strong similarities to the High-Benefits treatment. Many B's 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 B to S thereby eliminating their monetary incentive to enter into the mechanism.

7 Conclusion

SPI mechanisms have played a key role in the debate about the foundations and the relevance of incomplete contracts. If it were indeed possible to make all observable payoff-relevant information verifiable by third parties, the scope for incomplete contract theory would be radically curtailed. In this paper, we examined the performance of SPI mechanisms in the context of a hold-up problem where they yield complete truth-telling and efficient effort choices if they function as predicted. Because of their efficiency-enhancing properties both parties are predicted to be better off in the presence of the mechanism and thus be willing to adopt it voluntarily.

However, in contrast to these predictions we find that under the mechanism truth-telling typically occurs only in the minority of the cases. We document systematic deviations of the parties' actual behavior from the predicted strategies: sellers are often reluctant to challenge the buyers' lies. When they do challenge, the buyers retaliate to this by rejecting the counter-offer. The buyers frequently anticipate the sellers' reluctance to challenge, which makes lying a worthwhile strategy, and the sellers often anticipate the buyers' retaliatory behavior, which makes refraining from challenging rational. Taken together, this frequently leads to very bad welfare consequences and – if given the chance – the majority of the trading pairs opt out of the mechanism.

In response to these deviations we attempted to improve the mechanism by increasing the buyers' incentives for truth-telling (in the High Benefits treatment) or by decreasing their propensity for rejecting counter-offers (in the Low Fine treatment). However, these changes did not lead to major performance improvements. The mechanism still did not induce truth-

telling in the majority of the cases and the welfare achieved under the mechanism remained rather low. While the incentive compatible redesigns of the mechanism alleviated some of the problems – for example by decreasing under-reporting in the High Benefits treatment – they also generated new ones. In a substantial number of cases the buyers feared that even truthful announcements would be challenged. This fear led buyers to over-report the value of the good and/or to make maximal lies, which transferred most of the additional surplus generated from inducing high effort to the seller. The shift of surplus from buyer to seller eliminated the buyers' incentives to use the mechanism and ultimately led the buyers to dismiss the mechanism in periods where it's usage was voluntary.

Our paper explicitly documents several vulnerabilities of subgame-perfect implementation. First, the parties' willingness to retaliate to payoff-decreasing behaviors by the other party points to the importance of social preferences and negative reciprocity for the functioning of SPI mechanisms. Recall that in our experiments the buyers could impose a large loss on the sellers by rejecting the counteroffer at a low cost to themselves. The buyers made frequent use of this opportunity which is one of the key reasons for the unraveling of the SPI mechanism. In this context, it is important to point out that the low-cost of generating large losses for other parties is not an arbitrary design choice but instead an essential characteristic of subgame-perfect implementation in most environments – a point emphasized by Baliga & Sjöström (2008).[13] Thus, social preferences are likely to severely limit the efficacy of the Maskin-Tirole mechanism and all other mechanisms with the feature that decisions of one side of the market are used to punish or reward actions from the other side of the market.

While the first point stresses the vulnerability of SPI mechanisms to the existence of social preferences, the second point stresses their vulnerability to irrational beliefs. Recall that many buyers in the High Benefits and the Low Fine treatment feared that truthful announcements will be challenged although such challenges almost never occurred. These fears manifested themselves in different ways, depending on the size of the fine. In the Low Fine treatment where the fine was small and the incentive compatibility constraint was almost binding, the buyers frequently made maximal lies as a way of limiting the potential damage that irrational partners might cause. In the High Benefits treatment where fines were large, the buyers were willing to give up significant rents in an attempt to placate their partner. Both approaches eliminated the surplus buyers gained by implementing the first best outcome and led buyers to abandon the mechanism.

Our research opens two general avenues for future study. First, in relation to implemen-

[13] Baliga and Sjöström write: "By changing his strategy, an agent can have a large impact on another agent's payoff without materially changing his own. Such mechanisms may have little hope of practical success if agents are inclined to manipulate each others' payoffs due to feelings of spite or kindness."

tation, it is an interesting theoretical question to derive optimal implementation mechanisms in the presence of social preferences and negative reciprocity. Our work in this paper already suggests that the distribution of payments and the size of fines are important components in adoption, but further work is needed to more broadly understand implementation with intention-based and other types of social preferences. Recent theoretical work by Bierbrauer & Netzer (2014) shows that the design of optimal mechanisms can be heavily influenced by intentions-based preferences.

Second, irrational beliefs about other parties' also appears to be a major force in the functioning of the mechanism considered here. This strategic risk has been given limited attention in the theoretical literature. We conjecture that one reason simple contracts may be the norm for exchange is that they limit the exposure of individuals to strategic risk.

References

Acemoglu, D., P. Antras, and E. Helpman (2007): "Contracts and Technology Adoption," *The American Economic Review*, 97(3), 916–943.

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

Aghion, P., E. Fehr, R. Holden, and T. Wilkening (2014): "Subgame Perfect Implementation under Approximate Common Knowledge: Evidence from a Laboratory Experiment," *Working Paper*.

Aghion, P., D. Fudenberg, R. Holden, T. Kunimoto, and O. Tercieux (2012): "Subgame-Perfect Implementation Under Value Perturbations," *Quarterly Journal of Economics*, 127(4), 1843–1881.

Andreoni, J., and H. Varian (1999): "Pre–play contracting in the Prisoners' Dilemma," *Proceedings of the National Acadamy of Science of the United States of America*, 96, 10933–10938.

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

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

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

Baker, G., R. Gibbons, and K. J. Murphy (1994): "Subjective Performance Measures in Optimal Incentive Contracts," *The Quarterly Journal of Economics*, 109(4), 1125–1156.

——— (2002): "Relational Contracts and the Theory of the Firm," *The Quarterly Journal of Economics*, 117(1), 39–84.

Baliga, S., and T. Sjöström (2008): "Mechanism Design (New Developments)," in *The New Palgrave Dictionary of Economics. Second Edition.*, ed. by S. N. Durlauf, and L. E. Blume. Palgrave Macmillan.

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

Bierbrauer, F., and N. Netzer (2014): "Mechanism Design and Intentions," *University of Zurich Department of Economics Working Paper, No. 66*.

Blanco, M., D. Engelmann, A. K. Koch, and H. T. Normann (2010): "Belief Elicitation in Experiments: Is There a Hedging Problem?," *Experimental Economics*, 13(4), 412–438.

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

Board, S., and M. Meyer-ter Vehn (2014): "Relational Contracts in Competitive Labor Markets," *Working Paper*, pp. 1–55.

Bolton, G. E., and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90(1), 166–193.

Bracht, J., C. Figuires, and M. Ratto (2008): "Relative Performance of Two Simple Incentive Mechanisms in a Public Goods Experiment," *Journal of Public Economics*, 92(12), 54 – 90.

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

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

Cohn, A., E. Fehr, B. herrmann, and F. Schneider (2014): "Social Comparison and Effort Provision: Evidence from a Field Experiment," *Hournal of the European Economic Association*, 12(4).

Cox, J. C., D. Friedman, and V. Sadiraj (2008): "Revealed Altruism," *Econometrica*, 76(1), 31–69.

Datta, S., and G. A. Satten (2005): "Rank-Sum Tests for Clustered Data," *Journal of the American Statistical Association*, 100(471), 908–915.

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

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2008): "Representative Trust and Reciprocity: Prevelence and Determinants," *Economic Inquiry*, 46(1), 84–90.

Donner, A., and N. Klar (2000): *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold, London.

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

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

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

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

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

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

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

Fehr, E., H. Herz, and T. Wilkening (2013): "The Lure of Authority: Motivation and Incentive Effects of Power," *The American Economic Review*, 103(4), 1325–59.

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

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

Greiner, B. (2004): "The Online Recruitment System ORSEE 2.0 - A Guide for the Organization of Experiments in Economics," Working Paper Series in Economics 10, University of Cologne, Department of Economics.

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

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

——— (1988): "One share-one vote and the market for corporate control," *Journal of Financial Economics*, 20(1–2), 175–202.

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

——— (1982): "Behavioral explanations of efficient public good allocations," *Journal of Public Economics*, 19(3), 367–383.

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

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

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

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

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

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

Katok, E., M. Sefton, and A. Yavas (2002): "Implementation by Iterative Dominance and Backward Induction: An Experimental Comparison," *Journal of Economic Theory*, 104(1), 89–103.

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

Levin, J. (2003): "Relational Incentive Contracts," *The American Economic Review*, 93(3), 835–857.

MacLeod, B., and J. Malcomson (1989): "Implicit Contracts, Incentive Compatibility, and Involuntary Unemployment," *Econometrica*, 57(2), 447–480.

——— (1998): "Motivation and Markets," *The American Economic Review*, 88(3), 388–411.

Maskin, E. (2002): "On Indescribable Contingencies and Incomplete Contracts," *European Economic Review*, 46(4), 725–733.

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

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

Moore, J. (1992): *Advances in Economic Theory: Sixth World Congress Volume I*chap. Implementation, contracts, and renegotiation in environments with complete information, pp. 182–282. Cambridge University Press.

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

Netzer, N., and A. Volk (2014): "Intentions and ex-post implementation," Mimeo.

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

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

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

Perugini, M., M. Gallucci, F. Presaghi, and A. P. Ercolani (2002): "The Personal Norm of Reciprocity," *European Journal of Personality*, 17(4), 251–283.

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

——— (1996b): "Incomplete contracts and privatization," *European Economic Review*, 40(3–5), 569–579.

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

8 Appendix A: Retaliation Equilibrium

This section uses an extensive-form psychological games framework to study the expected effects of retaliation on equilibrium play. As in Dufwenberg et. al. (2011), we use a simplified version of Dufwenberg and Kirchsteiger (2004) which explicitly excludes kindness and concentrates solely on the effects of negative reciprocity. We start by studying a version of the model where B's and S's are homogeneous in their level of negative reciprocity. We then study the properties of a mixed strategy sequential-reciprocity equilibrium that exists when there is heterogeneity in buyers' preferences for negative reciprocity. The purpose of the appendix is to provide a potential other-regarding preference framework that is consistent with the empirical observation that retaliation occurs. Alternative reciprocity models in the spirit of Falk and Fischbacher (2006) or Cox et. al. (2008) can generate similar equilibrium predictions. Likewise, outcome-based models in which players care about the distribution of payments, such as Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) can generate predictions consistent with some of our findings.

8.1 The Baseline Model

We initially consider the exogenously imposed Maskin and Tirole mechanism when B's and S's preferences for negative reciprocity are common knowledge. Let H be the set of histories that lead to a subsequent subgame and let A_i be the set of behavioral strategies of $i \in \{B, S\}$ where each strategy assigns a probability distribution over the set of possible choices of i for each history $h \in H$. Define $A = \prod_i A_i$ as the set of potential strategies and let $\pi_i : A \to \mathbb{R}$ denote the assignment of endnodes to monetary payoffs.

Similar to Dufwenberg and Kirchsteiger (2004), we assume that individuals care not only about their monetary payoff but also care about retaliating against perceived unkind acts. To this end, we augment an individual's utility function with negative reciprocity payoffs. The reciprocity payoff depends on i's beliefs about other players' strategies and beliefs. Let $B_{ij} = A_j$ be the set of possible beliefs of player i about the strategy of player j. Let $C_{iji} = B_{ji} = A_i$ be the set of possible beliefs of player i about the beliefs of player j's beliefs about player i's own strategy.

As B's and S's beliefs about their counterparty's strategy may change as play proceeds, it is important to track how each player's behavior, beliefs, unkindness, and perception of others' unkindness differ across histories. Let $a_i(h)$ be the (updated) strategy that prescribes the same choices as a_i except for the choices that define history h which are made with probability one. Further define $b_{ij}(h)$ as the updated beliefs about the strategies of the other party and $c_{iji}(h)$ as the updated beliefs about the beliefs of the other party regarding player i's strategy.

For a given history h, an individual i's expected utility is given by:

$$U_i(a_i(h), b_{ij}, c_{iji}) = \pi_i(a_i(h), b_{ij}(h)) + \theta_i \cdot \kappa_{ij}(a_i(h), b_{ij}(h)) \cdot \bar{\lambda}_{iji}(b_{ij}(h), c_{iji}(h))$$

where $\pi_i(a_i(h), b_{ij}(h))$ is i's monetary payoff and $\theta_i \kappa_{ij} \bar{\lambda}_{iji}$ reflects his psychological utility from retaliation. The retaliation term is made up of three components: $\theta_i \geq 0$ is a sensitivity parameter, which reflects the relative size of monetary utility and psychologi-

cal utility. $\kappa_{ij}(a_i(h), b_{ij}(h))$ is i's unkindness for taking action profile $a_i(h)$ and is based on the expected monetary payoff of player j in the current subgame relative to other subgames that might have occurred if i had taken different actions in previous nodes. Finally, $\bar{\lambda}_{iji}(b_{ij}(h), c_{iji}(h))$ is player i's perception about the unkindness of player j given his current and future actions.

In order to assess the unkindness of his potential strategies at a given history h, an individual assesses the expected return that the other party receives from a given strategy profile, $\pi_j(b_{ij}(h))$ relative to a reference payoff $\pi_j^{ref}(b_{ij}(h))$ in a linear fashion:

$$\kappa_{ij}(a_i(h), b_{ij}(h)) = \pi_j(a_i(h), b_{ij}(h)) - \pi_j^{ref}(b_{ij}(h)).$$

The reference payoff, $\pi_j^{ref}(b_{ij}(h))$, may be exogenous or based on the expected returns of different potential actions in future subgames. For example, the reference point may be the payments along the truth-telling path, the "equitable" splitting of the surplus 50/50 between B and S (as in Falk and Fischbacher (2006)), or (as in Dufwenberg and Kirchsteiger (2004)), a function of the material payoffs that player i believes j would receive if i had chosen a different action profile. In the calibration exercises, we assume that the reference point is the payments on the truth-telling path. However, our only restriction is that $\pi_j^{ref}(b_{ij}(h)) > \min\{\pi_j(a_i(h), b_{ij}(h)) | a_i(h) \in A_i\}$, so that there are at least some states for which an action is perceived as unkind.

Having defined unkindness, we now define retaliation.[14] An individual forms beliefs about the perceived unkindness of the other party based on their actions and his beliefs about the actions the other party would have taken over all nodes. As before, define $\pi_j^{ref}(c_{iji}(h))$ as an exogenous or endogenous reference point about equitable payoffs and define:

$$\lambda_{iji}(b_{ij}(h), c_{iji}(h)) = \pi_i(b_{ij}(h), c_{iji}(h)) - \pi_j^{ref}(c_{iji}(h)).$$

We restrict attention to the case where an individual forms negative beliefs about the other party and assume that utility is formed over $\bar{\lambda}_{iji} = \min\{0, \lambda_{iji}(b_{ij}(h), c_{iji}(h))\}$, since unkind acts are more salient in our experiment, and the experimental evidence for preferences for reciprocity is not as strong as for negative reciprocity.

A retaliation equilibrium is defined as follows:

Definition 1 *The profile $a^* = (a_i^*)_{i\in\{B,S\}}$ is a* ***retaliation equilibrium*** *if for all $i \in \{B, S\}$ and for each history h it holds that*

1. $a_i^*(h) \in \arg\max_{a_i \in A_i(h,a^*)} U_i(a_i(h), b_{ij}(h), c_{iji}(h))$
2. $b_{ij} = a_j^*$
3. $c_{iji} = a_i^*$

[14] Dufwenberg et. al. (2011) call this vengeance. As we have used retaliation in the paper and differ from their model in the choice of reference points, we have stuck with our own terminology.

8.2 Retaliation Equilibrium with a Single Buyer Type

Having defined our model of retaliation, we now study the retaliation equilibria of the Maskin and Tirole game when B's and S's preferences for negative reciprocity are known. Let θ_B and θ_S be, respectively, B's and S's sensitivity parameters. Further assume that there is common knowledge about the utility function of B and S and thus that θ_B and θ_S are commonly known.

Recall that the timing on Maskin and Tirole's SPI mechanism is as follows:

1. S chooses $e \in \{e_L, e_H\}$, generating a value $v(e)$
2. B observes v and announces $\hat{v}$
3. S can challenge (or not) $\hat{v}$
4. B can accept (or reject) the counter offer $\hat{p}(\hat{v})$

In specifying an equilibrium, we begin with B's decision to accept/reject a counter offer for all histories in which this subgame is reached.

Lemma 1 *Consider the history $h = \{e, \hat{v}, challenge\}$ with $\hat{v} < v$. Then, for each announcement $\hat{v}$, there exists a θ_B^* such that for all $\theta_B \geq \theta_B^*$, B will reject the counter offer.*

Proof. Consider S's decision at history $h = \{e, \hat{v}\}$. If S does not challenge, B's payoff is $v - p(\hat{v})$. If S does challenge, B's payoff is either $-F$ in the case of rejection or $v - \hat{p}(\hat{v}) - F$ in the case of acceptance. By design, when $\hat{v} < v$, B's payoff is lower after a challenge than (i) the truth-telling path and (ii) the payment he would receive in the case that no challenge occurred. Thus, for any possible belief S has about B's next action, challenging is interpreted as an unkind act by S and $\lambda_{BSB} < 0$. It follows that B will increase his utility by decreasing the returns of S.

At the last node, rejecting the challenge is a less kind act than accepting the challenge. Thus, at history $\hat{h} = \{e, \hat{v}, challenge\}$, B will reject if:

$$\theta_B \lambda_{BSB} \left[\kappa_{BS}\left(reject|\hat{h}\right) - \kappa_{BS}\left(accept|\hat{h}\right)\right] \geq \pi_B\left(accept|\hat{h}\right) - \pi_B\left(reject|\hat{h}\right).$$

For all lies, the right-hand side of this inequality is positive and $\kappa_{BS}\left(reject|\hat{h}\right) - \kappa_{BS}\left(accept|\hat{h}\right) = \pi_s\left(reject|\hat{h}\right) - \pi_s\left(accept|\hat{h}\right)$ is negative. Thus, there exists a $\theta_B > 0$ who is indifferent between accepting and rejecting the challenge. As $\hat{v}$ decreases (i.e., the lie size increases), $\pi_B\left(accept|\hat{h}\right)$ increases and $\pi_S\left(accept|\hat{h}\right)$ decreases. Thus, the threshold θ_B must increase as $\hat{v}$ decreases. By the intermediate value theorem, it follows that for each announcement $\hat{v}$, there exists a cutoff $\theta_B^*(\hat{v})$ such that B's with $\theta_B \geq \theta_B^*$ will reject the challenge of a S. Further, the cutoff $\theta_B^*(\hat{v})$ is increasing as the size of the lie increases. ■

The above lemma shows that for each buyer type θ_B, there is a cutoff lie size for which B will reject a challenge but accept a challenge for a lower announcement. Let $\hat{v}^*(\theta_B)$ be this cutoff lie. If $\theta_S = 0$, and θ_B is commonly known, an announcement of $\hat{v}^*(\theta_B)$ will not

Table 8: Cutoff lies by B's following high effort as a function of his sensitivity parameter.

$\hat{v}^*(\theta_B)$	Main	High-Benefits	Low-Fine
260	$[0-0.020)$	$[0-0.019)$	$[0-0.067)$
240	$[0.020-0.048)$	$[0.019-0.045)$	$[0.067-0.171)$
220	$[0.048-0.078)$	$[0.045-0.073)$	$[0.171-0.283)$
200	$[0.078-0.109)$	$[0.073-0.103)$	$[0.283-0.409)$
180	$[0.109-0.143)$	$[0.103-0.134)$	$[0.409-0.550)$
160	$[0.143-0.178)$	$[0.134-0.167)$	$[0.550-0.709)$
140	$[0.178-0.216)$	$[0.167-0.203)$	$[0.709-0.891)$
120	$[0.216-0.256)$	$[0.203-0.241)$	$[0.891-1.10)$
100	$[0.256-\infty)$	$[0.241-\infty)$	$[1.10-\infty)$

be challenged. B thus has an incentive to announce $\hat{v}^*(\theta_B)$ since a larger lie will lead to a challenge and a smaller lie will not be challenged but generates less surplus for B.

Table 8 shows the intervals of θ_B that generate each potential $\hat{v}^*(\theta_B)$ under three parameterizations. In order to make the intervals easily comparable, we have used the payments on the truth-telling path as a reference point and scaled θ_B by 320. This is the difference in payments between a rejected challenge after a lie and the truth-telling payment in the Main treatment and can be thought of as normalizing θ_B by λ_{BSB} so that a number corresponds to the psychic utility the buyer gains by destroying \$1 of the sellers profit after a challenge. The difference between the cutoff lie and the truth is increasing in θ_B and the magnitude of θ_B that results in a small lie in the Main Treatment and the High-Benefits treatments is small. By contrast, the parameter must be much higher to generate challenges in the Low-Fine treatment.

Under most constructions of the reference points, announcements equal to this cutoff lie and less than the truth-telling announcement will be considered unkind actions by S. This is due to the fact that in these subgames either (i) S does not challenge or (ii) S challenges and the challenge is rejected. As the payment to S is below the payment along the truth-telling path and S's payment will always be below B's, all feasible announcements in $\hat{v} \in [\hat{v}^*, v)$ are regarded as unkind if the reference point is the payment along the truth-telling path or the equitable split of surplus.

Table 9 shows the intervals of θ_S that will reject announcements at or below each potential announcement threshold if it is known with probability one that B will reject. For comparison, we have again used the truth-telling equilibrium payments as a reference point and scaled θ by 320. The magnitude of S's sensitivity parameter required for S to challenge any lie that will be rejected is very large. This is due to the fact that S's private cost of having a challenge rejected is very high. As we will discuss in the next section, when there is heterogeneity in the sensitivity to reciprocity, some lies in each announcement bracket may be made by B's who will accept a challenge. In this case, the expected loss of challenging in equilibrium will be much lower and S's reciprocity has much more of an expected impact.[15]

[15]For example, if there is a 50 percent chance that a challenge will be accepted, a S in the Main Treatment will challenge an announcement of 200 if $\theta_S \geq .487$ as opposed to 7.696 in the case of 100 percent rejections.

Table 9: Level of S's sensitivity parameter necessary to reject an announcement and all lower announcements.

$\hat{v}^*(\theta_B)$	Main	High-Benefits	Low-Fine
260	∞	∞	∞
240	27.065	24.338	27.099
220	12.495	11.243	11.733
200	7.696	6.926	6.780
180	5.333	4.800	4.406
160	3.943	3.547	3.047
140	3.035	2.729	2.188
120	2.402	2.157	1.607
100	1.939	1.739	1.195

Values shown for the case where S believes his challenge will be rejected with probability 1.

As some of B's lies may be challenge, B's utility is maximized when he makes the largest lie possible that he would reject if challenged and that the S will therefore not challenge. By construction of the retaliation equilibrium, B's announcement will be the one which minimizes

$$\hat{v} = \arg\min_{v} v$$

subject to (1) $v \geq v^*$ and (2) $U_S(e, v, nochallenge|e, v) \geq U_S(e, v, challenge, reject|e, v)$. Condition (1) simply restricts lies to the set of all lies in which B will prefer to reject the counter offer rather than accept it. Condition (2) rules out announcements that B knows will be challenged despite the fact that B will reject the challenge.

As we can see from B's problem, B's announcement is influenced both by his own sensitivity parameter and by the parameter of S. If B has a large sensitivity parameter and $\theta_S = 0$, he will make very large lies and count on S not to challenge them. If on the other hand $\theta_S > 0$ and B fears that he will be challenged if he makes a sufficiently large lie, he will reduce the size of his lie. In the extreme case where the reference point is above the truth-telling equilibrium and θ_S is very large, he may make generous announcements to avoid the cost of inappropriate challenges.

Given the fact that low announcements may be made by B, it is straightforward to see why low effort may be an equilibrium outcome of the game. Faced with a B who has a high sensitivity parameter, it may be the case that S prefers low effort as a way to mitigate the effect of lies. This will be the case, for instance, if B's optimal announcement after high effort is less than 220.

Our retaliation equilibrium can thus rationalize the following empirical regularities:

1. Lie sizes are (weakly) increasing in the negative reciprocity of B.

2. The size of B's lies may be constrained by S's preferences for negative reciprocity.

3. S's who believe B's are reciprocal may choose low effort.

8.3 Retaliation Equilibria with a Distribution of Types

Thus far we have discussed the retaliation equilibrium in the case of a single B and a single S type. In this section, we demonstrate that when there is heterogeneity in the population, a partial-pooling equilibrium may exist in which some non-reciprocal individuals pretend to be reciprocal types. In this equilibrium, challenging a B does not lead to rejection with probability one. As such, the cost of challenges decreases and S's reciprocity plays a much bigger role in shaping the equilibrium.

As our goal is illustrative, we study the simplest environment in which partial-pooling equilibrium emerge. Consider a version of the Main treatment in which there is a fraction $\mu < \frac{1}{2}$ of reciprocal B's with $\theta_B^r = 0.4$ and a fraction $1 - \mu$ with $\theta_B^{nr} = 0$. Further, assume there is a single type of seller, θ_S, who will not challenge an announcement of 240 if B's reject counter offers following an announcement of 240 with probability one.

We now argue that there is no separating equilibrium or pooling equilibrium. To see that there are no separating equilibrium, suppose that all reciprocal types choose to announce 240 and all non-reciprocal types announce 260. In this case, S will never challenge the announcement of 240. As such, a non-reciprocal B prefers to announce 240 instead and receive a payoff of 85 instead of 70. Thus, a set of strategies where the reciprocal types announce 240 and the non-reciprocal types announce 260 is not an equilibrium. To see that no pooling equilibrium exists, consider a candidate pooling equilibrium where all buyers announce 240. In this case, a seller who challenges B will receive -370 with probability μ and 495 with probability $1 - \mu$. Since $\mu < \frac{1}{2}$, the expected value of challenging is higher than not challenging and thus all announcements of 240 are challenged. As non-reciprocal types would then prefer to announce 260, this also cannot be an equilibrium. Finally, the case where all types announce 260 can not be an equilibrium since reciprocal types prefer to announce 240 whether or not they are challenged.

While no separating and pooling equilibrium exist, a strategy in which the non-reciprocal types mix between announcing 240 and 260 can be sustained as part of a partial-pooling equilibrium. In this equilibrium, S's challenge 4.48% of the time making the non-reciprocal buyers indifferent between announcing 240 and 260. The non-reciprocal buyers mix between announcements of 240 and 260 such that S's receive the same utility from challenging and not challenging.[16] Reciprocal buyers always announce 240. Both reciprocal and non-reciprocal buyers will accept counter offers for announcements below 220 and thus we assume for convenience that S's have an out of equilibrium belief that an announcement below 240 is from a non-reciprocal type and that a counter-offer would be accepted.

Remark 1 *When there is heterogeneity in θ_B, a mixed-strategy retaliation equilibrium may exist in which non-reciprocal B's mix between lying and making truthful announcements and S's mix between challenging and not challenging small lies.*

While we have analyzed the mixing strategy with only two groups of B's, we note that the mixed-strategy retaliation equilibrium is a common outcome when there is heterogeneity in preferences for reciprocity, and these preferences are private information. As long as there

[16] In the case of $\theta_S = 0$, the mixing probabilities are such that a proportion $\frac{370\mu}{495(1-\mu)}$ of non-reciprocal individuals announce 240. The mixing proportion decreases as θ_S increases.

are some reciprocal types in the environment, it will always be the case that non-reciprocal B's will mimic the reciprocal types. This leads S's to challenge some small lies and generates destructive disagreement in equilibrium. We thus see mixing and destructive disagreement as a near-ubiquitous feature of the SPI mechanism.

Taking the intuition from both the initial model and the extension to unobserved heterogeneity, our model of retaliation predicts the following:

1. Small lies are predicted to exist in all treatments. With unobserved heterogeneity, some of these small lies will be challenged and rejected.

2. In the High-Benefits treatment, S's have a lower cutoff sensitivity for which they are willing to challenge a small lie. With unobserved B heterogeneity, this results in fewer lies in equilibrium.

3. In the Low-Fine treatment, B's require much higher levels of negative reciprocity to reject challenges. We thus expect fewer small lies in this treatment.

4. In the High-Benefits treatment, if some S's view the "equitable" split of surplus (rather than the truth-telling split) as the reference point, they may challenge a truthful announcement. B's may make generous announcements to avoid this potential outcome.

9 Appendix B: The No-False-Challenge Treatment

In the Buyer-Advantage treatment, we found that a fear of inappropriate challenges was a potential driver for B's to make generous announcements. In this appendix we report on an additional control treatment that eliminates the ability of S's to challenge B when he has made a truthful announcement. Such a mechanism would not be feasible in practice, because it requires that S's 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 B announces the true value of 120, he cannot be challenged, and the game ends. Likewise, after observing high effort, if B 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 8 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% 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 $< .01$, errors clustered by individual). The announcement distributions are also significantly different in clustered version of the chi-squared test developed by Donner & Klar (2000) (p-value $< .01$).

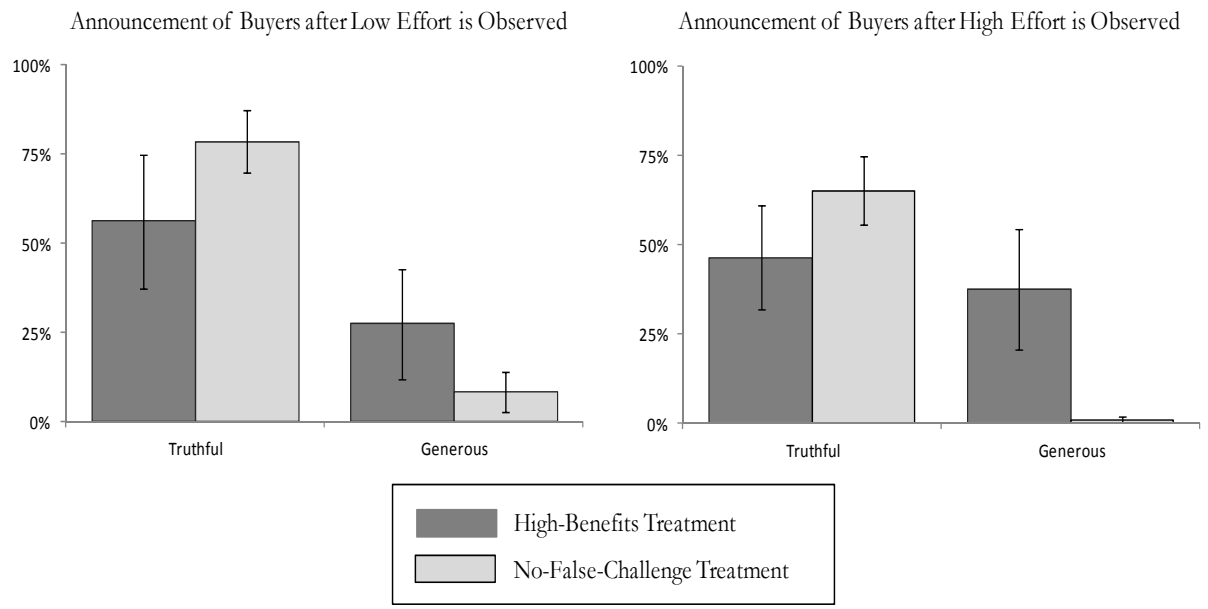


Figure 8: Comparison of Truthful Announcements and Generous Announcements in the High-Benefits and the No-False-Challenge Treatments

S's challenge behavior is similar in the two treatments with 58.5% of small lies being challenged in the High-Benefits treatment and 59.0% of small lies being challenged in the No-False-Challenge treatment. B's willingness to reject the challenges of small lies are also similar with 79.2% of challenges being rejected in the High-Benefits treatment and 87.0% of challenges being rejected in the No-False-Challenge treatment. Neither difference is significant (S's Challenge Behavior: p-value = .962; B's Rejection Behavior: p-value = .544).

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 B's opt-out rate declines from 64.6% in the High-Benefits treatment to 52.1% in the No-False-Challenge treatment, S's opt-out rate increases from 3.5% to 10.2%. Thus, on net, the overall increase in retention rates is small (65.7% vs 58.2%) and not significant (p-value = .394, errors clustered by buyer).

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 Buyer-Advantage treatment. We leave further study of mechanisms such as this one (what Moore (1992) calls "simple sequential mechanisms") to future research.