International Review of Law and Economics 65 (2021) 105973

Contents lists available at ScienceDirect

International Review of Law and Economics

Contracts as reference points: A replication☆

Svenja Hippel[a,*], Sven Hoeppner[b]

[a] University of Würzburg, Department of Economics, Sanderring 2, 97070 Würzburg, Germany
[b] Charles University, Law School, Department for Economics, nám. Curieových 901/7, 116 40 Prague, Czech Republic

ARTICLE INFO

Article history:
Received 13 August 2020
Received in revised form
25 November 2020
Accepted 7 December 2020
Available online 27 December 2020

JEL classification:
C18
C49
C90
D44
D86
D90

Keywords:
Contract theory
Reference points
Self-serving bias
Laboratory experiment
Replication
Bayesian statistics

ABSTRACT

We replicate two treatments of an experimental theory test (Fehr et al., 2011) studying Hart and Moore (2008)'s idea that contracts serve as reference points in trading relationships. In contrast to rigid contracts, flexible contract terms may be perceived in a self-serving manner and, therefore, the contract parties might form subjective entitlements. This reference-dependent perception of flexible contract terms leads to a trade-off of the contractual form. While flexible contracts are, in theory, deemed preferable to rigid contracts, frustrated subjective entitlements may lead to perfunctory performance and shading behavior that is absent in rigid contracts. The results of our replication are mixed. Our findings imply further support for Hart and Moore (2008)'s contracts as reference point hypothesis. However, our replication does not provide reliable evidence for the idea that competition creates objectivity and enhances perceived fairness of the contract terms.

1. Introduction

In a recent influential paper of their academic corpus, Hart and Moore (2008) develop the idea that contracts serve as reference points in trading relationships. An ex ante contract might shape contract parties' subjective entitlements regarding ex post outcomes of a trade: parties might feel entitled to different outcomes within the negotiated contract. Importantly, parties do not form any entitlements regarding outcomes outside of the contract, because ex ante competition during negotiating of the contract supposedly provides objectivity to the contract terms and market participants perceive the initial contract as fair.

Contract parties' subjective entitlements can lead to a trade-off between contractual rigidity and flexibility in a situation where uncertainty about the trading environment resolves over time. A *rigid contract* specifies a price ex ante, bearing the risk of making ex post trade impossible when costs turn out to be high. By contrast, *flexible contracts* only fix an interval of possible trading prices and appear, in principle, advantageous as parties can adjust to the state of nature when uncertainty resolves. However, Hart and Moore (2008) argue that flexible contracts come with the downside that parties suffer from a *self-serving bias*[1] and therefore might form reference points at distant ends of the specified price interval.

☆ We are grateful to the authors of the original paper, especially Christian Zehnder, for sharing their materials with us. We also thank Wladislaw Mill, Daniel Müller, and Patrick Schmitz for helpful comments and suggestions. We are especially indebted to Alexander Streubel who lent us a computer actually powerful enough to run our Bayesian replication analysis with the bootstrapped data. We also owe thanks to Eric Helland who was tremendously patient with us. Finally, we cannot express enough appreciation to the editorial board of the International Review of Law and Economics for providing an outlet for replication studies and, by doing so, for creating an incentive to conduct replications in the first place. We use R version 3.6.1 (2019-07-05) for statistical analysis (R Core Team, 2018). Central parts of our statistical analysis rely on the additional packages *JAGS 4.3.0* (Plummer, 2003), *runjags 2.0.4-6* (Denwood, 2016), and *bridgesampling 1.0.0* (Gronau et al., 2020).

* Corresponding author.
E-mail addresses: svenja.hippel@uni-wuerzburg.de (S. Hippel), hoeppnes@prf.cuni.cz (S. Hoeppner).

[1] See Babcock et al. (1995)'s study, which was successfully replicated in the laboratory by Hippel and Hoeppner (2019), as well as the companion papers Babcock

https://doi.org/10.1016/j.irle.2020.105973
0144-8188/

Subsequently, if a party's perceived entitlement is frustrated, she is aggrieved and *shades* by providing perfunctory rather than consummate performance. As a result, rigid contracts might be more attractive than flexible contracts.

Hart and Moore (2008)'s theory is one of the most important developments in recent years in the field of contract theory. The idea can explain long-term contracts and employment contracts that fix wages in advance and leave discretion to the employer. The paper started a literature stream that tries to shed new light on the theory of the firm with the help of behavioral insights, the investigation of shading behavior through reference points being at its core. This is an instance where laboratory experiments can fruitfully serve as a wind tunnel and empirical and theoretical work naturally go hand in hand.

Fehr et al. (2011) provide the first direct test of the Hart and Moore (2008) model by conducting a laboratory experiment. In their baseline treatment, participants take on either the role of a buyer or the role of a seller. Buyers determine the contract type and the contracts are auctioned off to the sellers. Uncertainty is introduced as the sellers' production cost is initially unknown. Then uncertainty resolves and – given trade is possible – either the transaction is carried out at the ex ante fixed price (rigid contract) or the buyer can choose the final price within the ex ante determined interval (flexible contract). Finally, sellers can decide whether to shade and reduce the product's value for the buyer. Results in Fehr et al. (2011) largely confirm the model's predictions, including the hypothesis of increased shading rates in flexible contracts.

Since then, researchers have modified the experimental setup in order to investigate several research questions regarding the underlying theory and the influence of various aspects of the decision environment (Fehr et al., 2009, 2011, 2015, 2019; Erlei and Reinhold, 2016). In addition to their baseline treatment, Fehr et al. (2011) run a robustness check to study whether competition indeed provides objectivity to the contract terms. Leaving the experimental setup otherwise constant, they remove the competitive auction. After buyers determine the contract type, contracts are randomly allocated to the sellers and the auction outcome is randomly drawn from the results obtained when auctions were present. Fehr et al. (2011) find that when the transaction is governed by rigid contracts, the removal of ex ante competition leads to significantly increased shading behavior.

Also Fehr et al. (2009) study the role of competition in providing objectivity to the negotiated contract terms. Crucially, they implement the removal of competition slightly differently in their experimental setup: the seller is informed about the draw of the auction outcome before determining the contract type. In direct comparison to the baseline data in Fehr et al. (2011)[2], shading rates in rigid contracts spike. Shading rates are about twice as high compared to the baseline data and, in fact, sellers under rigid contracts turn out to shade even more often than sellers under flexible contracts. This finding is important as it suggests that we investigate a contracting situation where details of the experimental setup might have considerable influence on observed shading behavior.

Fehr et al. (2015) investigate the influence of informal agreements. Buyers can make non-binding price announcements and, by doing so, shape seller's expectations. Fehr et al. (2015) find that this opportunity moderately reduces shading rates in flexible contracts. Moreover, in another variation the researchers allow buyers to revise the contract ex post by granting the unilateral right to replace the existing contract with a new one. The results suggest that writing a simple (rigid) contract and revise it ex post if needed, rather than anticipating and including future contingencies in a (flexible) contract from the outset, can under specific circumstances be beneficial for parties. The study also collects additional baseline data with exactly the same setup as in Fehr et al. (2011) and finds similar results, especially very similar shading rates.

While communication in Fehr et al. (2015)'s study happens unilateral from buyer to seller, in a very recent working paper Fehr et al. (2019) extend communication between the parties to free-form. In contrast to nonbinding unilateral price announcements, the evidence from Fehr et al. (2019) indicates that free-form communication increases the potential for aggrievement on the seller-side. The study also replicates the baseline treatment from Fehr et al. (2011) with the difference that they employ a perfect stranger matching, i.e., participants meet each other at most once during the course of the experiment. In this baseline treatment, sellers are almost twice as likely to shade under rigid contracts as in Fehr et al. (2011) (10% compared to 6%).[3]

Finally, another recent paper by Erlei and Reinhold (2016) explores the role of reciprocity during contract choice, i.e., buyers unilaterally selecting the contract type may trigger different responses from buyers. When Erlei and Reinhold (2016) externalize contract choice to a random device, their results indicate that by shading sellers indeed punish buyers for choosing rigid contracts. As the researchers also conduct a baseline treatment, their results also speak to the core idea of the underlying theory. In this regard, the study presents mixed evidence. Erlei and Reinhold (2016) find evidence for reference point effects. Moreover, results regarding many of the experiment's outcomes (e.g., auction outcomes and the relative frequency of contract types) are similar to Fehr et al. (2011)'s results. But Erlei and Reinhold (2016) also observe substantially different shading behavior in the baseline treatment: shading rates under rigid contracts appear to be three times higher (19.5% compared to 6%). In their baseline treatment, Erlei and Reinhold (2016) deviate in two respects from the setup in Fehr et al. (2011), which may influence results. First, the clock auctions used to auction off contracts to the sellers are conducted simultaneously (rather than consecutively). This design difference implies that sellers need to focus on one of the auctions first, which increases the chance of the leaving the other contract to another seller. That is, competitive forces are slightly reduced. Additionally, Erlei and Reinhold (2016) do not provide aggregate information at the end of each period, thereby changing speed and (maybe) direction of learning.

Reference-point driven shading behavior is the most critical aspect for the trade-off between contractual forms. Experimental results so far provide matching evidence on many of the contractual situation's outcomes in support of Fehr et al. (2011)'s original findings. However, baseline shading behavior in rigid contracts, which should not occur at all according to Hart and Moore (2008)'s theory, differs widely between prior studies. In those studies shading rates range from almost negligible 5% to substantial 19%. Many of the prior studies also changed details of the experimental setup, which complicates recognizing the source of such fluctuating shading rates in rigid contracts. Therefore, we replicate Fehr et al. (2011)'s original theory test to shed additional light on the validity of Hart and Moore (2008)'s model. We directly replicate their baseline treatment and the no-competition robustness check, with a strong focus on shading rates. We then use our data and the data from Fehr

and Loewenstein (1997) and Babcock et al. (1997) for evidence on self-serving bias in a legal context.

[2] They do not compare their no-competition data to the robustness check removing competition in Fehr et al. (2011). This might be due to the fact, that Fehr et al. (2009) was actually published earlier than Fehr et al. (2011).

[3] As Fehr et al. (2019) is still an unpublished working paper, the data set is not publicly available at the point of our data analysis.

et al. (2011) and Fehr et al. (2015) to feed our replication analysis and receive a more complete picture.[4]

The results of our replication are mixed. In the replicated baseline treatment, we find that sellers under flexible contracts shade more often than sellers under rigid contracts. With regard to this result, our replication analysis suggests replication success. In the no-competition treatment, we do not find that removing competition during the contracting stage increases shading rates. This result differs from the result of Fehr et al. (2011). Moreover, our replication analysis does not unambiguously suggest replication success regarding this finding. Some auxiliary results suggest (1) that shading behavior is contingent on reference-dependent measures of aggrievement both in the baseline treatment and the no-competition treatment, (2) that flexible contracts are associated with lower auction outcomes after learning effects have been accounted for, and (3) that shading behavior under rigid contracts is both increased and more heterogeneous as compared to Fehr et al. (2011)'s results. With regard to the Hart and Moore (2008)'s theory, our results imply further support for their contracts as reference point hypothesis. However, our replication does not provide supporting evidence for the idea that competition creates objectivity and enhances perceived fairness of the contract terms.

The remainder of this paper is organised as follows. Section 2 summarizes the experimental design of the original study. Section 3 reports the details of the replication study. Section 4 presents our analysis of the replication data and evaluates replication results in light of prior studies. Section 5 discusses our findings and concludes.

2. Original experimental design

2.1. Baseline treatment

The experimental design is an exact replication of the baseline treatment in Fehr et al. (2011). Participants play 15 rounds of the market game displayed in Fig. 1. They are equally split into roles of buyers and sellers. Roles remain fixed throughout. Participants play the market game in groups of four, consisting of two buyers and two sellers. Groups are randomly reshuffled before the beginning of each round.

On the market, buyers and sellers can trade a good. Each seller can sell up to two units of the good, each buyer can buy at most one unit of the good. As there is oversupply, sellers compete for buyers. When a buyer purchases a unit of the good, her payoff is the difference between her valuation for the product and the price of the good. The buyer's valuation for the product depends on the seller's ex post quality choice. Buyers value a good of normal quality with 140 and a good of low quality with 100. When a seller sells a unit of the good, his payoff is the difference between the price of the good and the production cost. The seller's production cost depend on a realized state of nature, which can be good or bad. The good state occurs with a probability of 80%, the bad state with the remaining probability of 20%. For goods of normal quality, production costs are 20 and 80 in the good state and in the bad state, respectively. The production costs for low quality goods are slightly higher, specifically 25 and 85 in the good state and in the bad state, respectively. Except for the state of nature, participants know all parameters of the game at the start of the experiment.

The market game is comprised of two stages, an ex ante *contracting stage* and an ex post *trading stage*. In between of the two stages, the state of nature resolves and the contract parties receive corresponding information.

2.1.1. Contracting stage

First, each buyer decides whether she will offer a rigid or a flexible contract for the purpose of buying the good. The rigid contract fixes the price ex ante. The flexible contract, by contrast, specifies ex ante only a price range. The upper bound of that range is fixed to 140, i.e., the buyer's valuation of the product at normal quality.

Next, the two contracts of the buyers are auctioned off to the sellers in the group. The sellers participate in two consecutive auctions. The sequence of the two auctions is randomized on the group level. Specifically, the sellers can competitively determine the price (rigid contract) or the lower price bound (flexible contract) in an inverse clock auction. The auction starts at a price or lower price bound of 35 and increases by 1 price unit every half second until a maximum of 75. Each seller can accept the current contract at any time during the auction by clicking a button. The first seller who accepts the contract at the displayed price or lower price bound receives the contract. The other seller realizes an outside option of 10. Depending on who accepts the contracts first, one seller might end up with both contracts.

2.1.2. State of nature

After both contracts have been auctioned off, the computer randomly determines the state of nature for each contract independently. The state of nature determines the seller's production costs. Contract parties observe the realized state of nature and are informed whether trade can take place or not. Under rigid contracts, trade can only occur in the good state. As the price in rigid contracts is at most 75, in the bad state the price never covers the seller's production cost of 80. Mutually beneficial transactions are not feasible. If trade does not occur, buyers and sellers realize an outside option of 10. Under flexible contracts, however, trade can always take place because the price range allows the buyer to choose prices that cover the seller's cost in both states of nature.

2.1.3. Trading stage

When the buyer chose a flexible contract in the contracting stage, she now determines the actual trading price. In the good state, the buyer can choose any price between the lower price bound endogenously determined in the auction and the exogenously set upper bound of 140. In the bad state, buyers are required to ensure that sellers are not worse off than her outside option of 10. Therefore, the lower price bound under flexible contracts is set to 95 when the auction result was smaller than that amount.[5] When the buyer chose a rigid contract, she has no decision regarding the trading price which was determined through the auction.

Under both contract types, the seller then observes the trading price. Finally, the seller determines the quality of the product. He can either provide a good of normal quality or a good of low quality (shading). Choosing low quality reduces the buyer's valuation of the product by 40, from 140 to 100. Shading comes at an additional cost of 5 for the seller himself.

2.1.4. Market information

At the end of the round all participants learn about their profit for the round. Buyers also receive some aggregated information about the market outcome. They are informed about profits of buyers under both contract types averaged over all past periods. They also learn how many buyers have opted for each contract type in

[4] Unfortunately, we could not retrieve the data from Erlei and Reinhold (2016) for this purpose.

[5] This amount is the sum of the seller's cost of 80 in the bad state, the outside option of 10, and an additional 5 in order to not disturb the seller's incentives for shading.

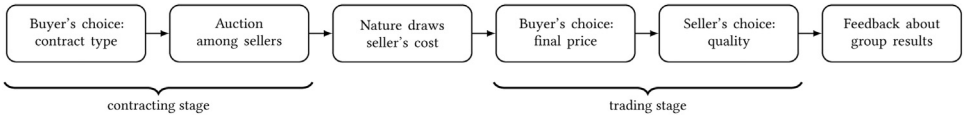

Fig. 1. Market game.

the current round over all groups.[6] This feedback stage is the main difference in the baseline treatment between Fehr et al. (2011) and Erlei and Reinhold (2016). Apart from their profit, buyers in Erlei and Reinhold (2016)'s baseline treatment did not receive any additional market information.

2.2. No-competition treatment

In one of their experimental robustness checks, Fehr et al. (2011) test the argument of the underlying theory that competition will provide objectivity to the contract terms such that parties perceive the initial contract to be fair. This fairness perception serves to create no reference points outside the contract. To investigate this idea, the authors remove the competitive element, i.e., the auctions that determine contract terms from the contracting stage. Instead, Fehr et al. (2011) exogenously determine the price and the lower bound of the price range, respectively, by randomly drawing this variable from the empirical distribution generated in the baseline treatment. After the buyers chose the contract type and the contract terms have been randomly determined for both buyers, each contract is randomly and independently assigned to one of the two sellers.

3. Replication study

3.1. Replication hypotheses

Hart and Moore (2008) provide the theory underlying the experimental design. From that model we derive predictions for the different stages and participant decisions in the market game. First, we start from the original *reference point hypotheses* that Fehr et al. (2011, pp. 503-504) formulate and that the authors use to guide their data analysis. Second, we derive our replication hypotheses based on the central ideas of Hart and Moore (2008)'s model for the baseline as well as the no-competition treatment.

The first original hypothesis concerns the auction outcome. "Market forces imply that the fixed price in rigid contracts and the lower bound of the price range in flexible contracts end up at the competitive level, i.e., 35". This prediction corresponds with standard economic theory, as Hart and Moore (2008)'s model does not speak to behavioral forces in the auction itself.

Central to the model in Hart and Moore (2008) is the shading behavior of sellers in the trading stage. "In rigid contracts sellers never choose low quality irrespective of the price level. In flexible contracts sellers' quality provision is price dependent. Heterogeneity in seller entitlements implies that the frequency of shading is decreasing in the price. Given the price dependence of quality, buyers may not choose the lowest price available in flexible contracts." The predicted difference in seller's shading behavior between contract types creates the trade-off between contractual rigidity and flexibility in the first place.

The hypothesized shading behavior, in turn, influences the buyer's contract choice via the buyer's profit from trade. As the reference point effect causes a lower shading rate in rigid contracts as compared to flexible contracts, "[b]uyers' profits in flexible contracts are lower than predicted by the standard model. If the impact of the reference dependent preferences is strong, buyers may even make higher profits in rigid contracts than in flexible contracts." Ex ante competition through the auction process in the contracting stage is essential for the trade-off to occur. Therefore, "[e]liminating ex ante competition increases shading in rigid contracts." Testing this hypothesis is the central intention of the no-competition treatment.

Shading behavior of sellers is at the core of the dynamics in the market game. As a result of reference points at distant ends of the specified price interval, the model predicts that shading only occurs in flexible contracts, given the buyer's do not adjust their prices appropriately. Therefore, with our first replication hypotheses we focus on the difference in shading rates between the two contract types:

Replication Hypothesis 1 (RH1). In the good state, more shading occurs in flexible contracts than in rigid contracts, i.e., the rate of low quality is higher in flexible contracts than in rigid contracts.

Ex ante competition being necessary to shape contractual reference points constitutes the second crucial element of Hart and Moore (2008)'s theory. When contracts are negotiated in a competitive market, the terms of the contract are perceived as objective and mutually fair. Therefore, competition makes the contract terms salient as a reference point. Without competition, the contractual terms should not serve as a reference point. Low prices in rigid contracts are no longer justified by competitive market forces and, consequently, one would expect to see more shading in rigid contracts. Accordingly, we derive our second replication hypothesis concerning the between-treatment difference in shading behavior:

Replication Hypothesis 2 (RH2). In rigid contracts, the frequency of shading is higher when there is no ex ante competition for prices than when there is ex ante competition for prices.

3.2. Sample size

We determine our required sample size based on power calculations, using session-level frequencies of normal quality provision from the original study.[7] Regarding RH1, we look at the session level frequency of normal quality provision in the good state for each contract type. The original paper found session level frequencies of 0.89, 0.97, 0.95, 0.91, and 0.96 under rigid contracts and 0.78, 0.76, 0.79, 0.67, and 0.75 under flexible contracts (Fehr et al., 2011, fn. 16). We derive the required sample size for a one-sided Wilcoxon signed-rank test (as in the original paper) using the following assumptions: (1) the data points are distributed according to a normal parent distribution, (2) $\alpha = 0.05$, and (3) $1 - \beta = 0.9$. Given the original data, we compute an effect size (Cohen's d) of $dz = 3.627$. We use the Lehman method with continuity correction to compute power. These assumptions and settings result in a required sample size of $N = 5$.

Similarly, we derive the required sample size regarding RH2. With ex ante competition, the session-level frequencies of nor-

[6] To maintain statistical independence, buyers only learned about groups within their matching group, i.e., only other participants they could potentially also be matched with.

[7] We employed G*Power (Faul et al., 2007) for the power calculations and sample size estimations.

mal quality provision in rigid contracts are mentioned above. In the original no-competition treatment, session level frequencies of normal quality provision in rigid contracts were 0.88, 0.85, 0.88, 0.85, and 0.73 (Fehr et al., 2011, fn. 30). We compute an effect size (Cohen's d) of $dz = 1.174$. With the same assumptions and settings as above, we obtain a required sample size of $N = 9$.

Therefore, we used a minimum required sample size of $N = 9$ for each of the two treatments (baseline and no-competition). While the original experiment generated one observation for both rigid and flexible contracts with 28 participants per experimental session, we employed two matching groups of 12 participants per session. This approach corresponds to the procedure used by the original authors in a more recent study (Fehr et al., 2015) and facilitates generating more independent observations. Therefore, we aimed at sampling 10 matching groups (in 5 sessions) for each treatment, satisfying our required sample size with a total of 240 participants.

3.3. Procedure

The experiment was conducted in two separate waves in February and May 2019 at the experimenTUM laboratory of the Technical University of Munich. In each wave, we collected five matching groups of the baseline treatment and used the resulting distribution of auction outcomes to randomly draw prices and lower price bounds in the five matching groups of the no-competition treatment. In total, 240 participants (120 in each treatment) took part in the experiment, almost all of them being students. Participants were recruited from the laboratory's subject pool using ORSEE (Greiner, 2015). Every participant could take part in one session only. Participants sat in visually isolated cubicles during the experiment. Sessions lasted about 1.5 h. All 15 rounds of the market game were payoff-relevant. During the experiment, payoffs where presented in points with an exchange rate of 37 points = 1 Euro. Participants earned 25.46 Euro (about 28.48 US Dollar) on average, including a show-up fee of 4 Euro.

The experiment was programmed with z-Tree (Fischbacher, 2007). Before the start of the experiment, participants received paper instructions, depending on their role as a seller or buyer. We received the original materials, i.e., the paper instructions for both treatments as well as the computer program for the baseline treatment, from the authors of the original study.[8]

Experimenters stood ready to clarify the instructions, if needed. All participants had to pass extensive control questions to ensure full understanding of the market game. As in the original study, the sellers could practice the auctions in two trial auctions before the market game started.

3.4. Differences between the studies

Next to the no-competition treatment, Fehr et al. (2011) also conduct another experimental robustness check that we did not include in our replication as it is particularly concerned with the bad state of nature, which is not in the focus of our replication hypotheses.[9] In this additional treatment, Fehr et al. (2011) reduce the upper bound of the price range in flexible contracts to 95 such that in the bad state only one price is available, and, according to Hart and Moore (2008)'s theory, this should leave no room for aggrievement by buyers. In the good state, the second robustness check only tests a possible reduction of an extreme self-serving bias at an unclear strength and does not eliminate the effect completely. Therefore, we considered it less important than the no-competition treatment, which provides sharp evidence on the necessity of ex ante competition to shape contractual reference points as a behavioral channel in the proposed reference point effect.

We made only minor changes to the materials we received from the original authors and we are certain that these changes did not affect the participants' understanding or perception of the situation. We slightly streamlined the very extensive and repetitive paper instructions in order to decrease an otherwise exhaustive reading time. Additionally, in the no-competition treatment we provided information about the price distribution on separate paper sheets that we handed out together with the paper instructions, using the same wording and description as in the original instructions of the treatment. Regarding the computer screens, we slightly changed the feedback screen that participants saw at the end of each round. Specifically, we changed the information regarding the different contract types from an absolute number to a percentage in order to obscure the presence of multiple matching groups during the sessions. We also had to adjust the conversion rate of points to Euro to adhere to the standard for average per hour payments of the laboratory. While in the original study the exchange rate was 15 points to 1 Swiss Franc, we converted 37 points to 1 Euro. We obviously also used a slightly different subject pool as we collected data in another laboratory in another, albeit neighbouring country. Finally, in contrast to the original study we did not exclude subjects with a background in economics or psychology.

In fact, our subject pool is very comparable to Erlei and Reinhold (2016)'s study. They also collected their data in Germany and did not exclude participants with specific study backgrounds. Similar to the participant pool at the Technical University of Munich, their participants' backgrounds were mainly business administration and industrial engineering.

However, Erlei and Reinhold (2016)'s baseline treatment involves noteworthy changes of the experimental design that may explain at least parts of their different results in shading rates. First, the authors conducted simultaneous, instead of sequentially randomized auctions. That is, both auctions appeared on the seller's screens at the same time. Second, they excluded learning possibilities of buyers by not providing market information at the end of the rounds. They provided neither information regarding the number or frequency of chosen contracts per type in the current round nor information about the aggregate profitability of rigid and flexible contracts over all rounds.

4. Results

In analysing our replication data, we first focus on our replication hypotheses. That is, the analysis centers on shading rates between contract types and the removal of ex ante competition. We then extend the analysis beyond our replication hypotheses by investigating the reference-dependence of shading and the dynamics of outcomes over periods. In doing so, we follow the original analysis in Fehr et al. (2011) but also look for new ways to disentangle behavioral effects. Finally, we conduct a replication analysis using Bayesian model comparison. We report a summary in this section and expound the detailed Bayesian replication analysis in Appendix A.[10]

[8] We thank the original authors for their kind cooperation and, especially, Christian Zehnder for his fast reply to our request.

[9] Their corresponding hypothesis predicts that "[l]owering the upper bound o the price range leads to less shading in flexible contracts, in particular in the bad state of nature." (Fehr et al., 2011, p. 504).

[10] Both data sets are available online. For the original data, please visit: https://doi.org/10.1257/aer.101.2.493. For the replication data, please visit: https://doi.org/10.1016/j.irle.2020.105973.

Table 1
Summary of outcomes of the baseline treatment. Values are the averages of outcome variables.

Contract type	Rigid		Flexible	
State of nature	Good	Bad	Good	Bad
Average price	41.04	–	48.97	97.68
Relative frequency of low quality	0.11	–	0.20	0.22
Average auction outcome	41.04		39.88	
Relative frequency of contract type	0.43		0.57	
Buyers' average profit (per state)	94.76	10.00	83.62	34.22
Sellers' average profit (per state)	20.35	10.00	27.27	16.02
Buyers' average profit (across states)	79.09		73.47	
Sellers' average profit (across states)	18.44		24.96	

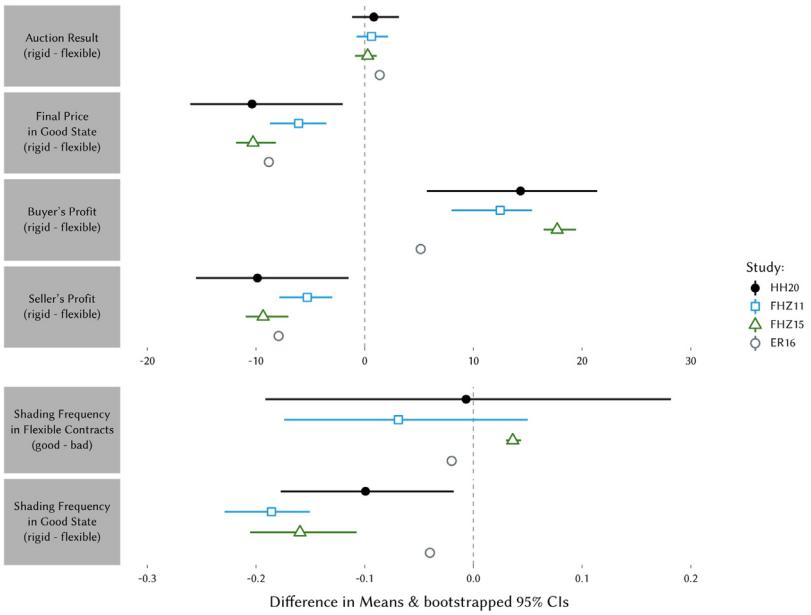

Fig. 2. Summary of aggregate replication findings (HH20) in comparison to the results of the original theory test (FHZ11), of Fehr et al. (2015)'s baseline treatment (FHZ15), and of Erlei and Reinhold (2016)'s baseline treatment (EH16). The figure plots differences in means and bootstrapped 95% CIs. We could not bootstrap 95% CIs for Erlei and Reinhold (2016)'s results because we could not retrieve their data.

4.1. Analysis of replication data

4.1.1. Baseline treatment

In general, most of our point estimates in the baseline treatment appear to be fairly close to the results of Fehr et al. (2011). Our data, however, exhibits much more heterogeneity. Table 1 reports our aggregate outcomes. In 42.67% of all transactions, buyers chose the rigid contract. Flexible contracts were chosen by buyers 57.33% of the time. The average auction outcomes are very similar between contract types, i.e., 41.04 under rigid contracts and 39.88 under flexible contracts. By contrast, final prices look very different. For feasible rigid contracts, the average final price equals the average auction outcome of 41.04. For flexible contracts, in the good state the average final price is 48.97, i.e., an increase of 9.09 over the average auction outcome. Moreover, sellers under feasible rigid contracts shade in 11% of transactions, whereas the average shading rate under flexible contracts is 20% in the good state and 22% in the bad state.

Fig. 2 plots for each relevant outcome the difference in means on the matching group level and bootstrapped 95% confidence intervals, visually comparing our results to prior results.[11] One noteworthy overarching result is that the bootstrapped confidence intervals of our matching group averages are much wider compared to the bootstrapped confidence intervals that are based on the previous data. The reason for this result is that our matching group averages exhibit a much higher heterogeneity (as measured by, e.g., standard deviation) than the matching group or session averages of the prior studies.

In the replication data, the difference in auction results between rigid and flexible contracts is very close to zero. We cannot reject the hypothesis that matching group averages of auction outcomes for rigid contracts and flexible contracts are equal (Wilcoxon signed-rank test: $V = 39$, $p = 0.275$). Visually, the results of the other studies are similar. By contrast, in the good state, buyers under flexible contracts pay substantially higher prices than buyers under rigid contracts. Here, we can reject the hypothesis that the matching group averages of final prices for rigid contracts and flexible contracts in the good state are equal (Wilcoxon signed-rank test: $V = 0$, $p = 0.002$). Also this result appears to be consistent with findings in prior studies. Moreover, buyers' profits appear to be consistently higher under rigid contracts as compared to flexible contracts – at the expense of sellers' profits, of course, which are consistently lower in rigid contracts than under flexible contracts. Most importantly, the difference of shading frequencies in the good

[11] Unfortunately, we could not retrieve the data from Erlei and Reinhold (2016). Therefore, Fig. 2 does not include bootstrapped 95% confidence intervals for their results. Nevertheless, we wanted to include their results in our comparison and do so by using the means reported in their paper.

Table 2
Summary of outcomes of the no-competition treatment. Values are the averages of outcome variables.

Contract type	Rigid		Flexible	
State of nature	Good	Bad	Good	Bad
Average price	40.67	–	47.25	98.07
Relative frequency of low quality	0.17	–	0.25	0.31
Average random price / lower price bound	40.67		39.50	
Relative frequency of contract type	0.39		0.61	

state is consistently negative, a result which we will investigate in detail.

RH1 predicts that, in the good state, more shading occurs under flexible contracts than under rigid contracts, i.e., the rate of low quality is higher when contract terms are flexible rather than rigid. The relative frequency of low quality provision in the good state is 0.109 under rigid contracts and 0.202 under flexible contracts.[12] We can reject the hypothesis that the shading frequencies on the matching-group level are equal for rigid and for flexible contracts (Wilcoxon signed-rank test: $V = 45$, $p = 0.084$). While this result is weakly significant, it aligns with RH1: sellers under flexible contracts on average decrease the quality more often when the good state of nature occurs than sellers under rigid contracts.

Result 1. In the good state, sellers shade more often under flexible contracts than under rigid contracts.

In Section 4.2.1 we investigate the price dependence of shading behavior and which potential reference points drive the results. In this analysis, we find further consistent evidence supporting RH1.

4.1.2. No-competition treatment

The no-competition treatment replicates one of Fehr et al. (2011)'s experimental robustness checks. Table 2 reports the aggregate outcomes in the no-competition treatment. In 39.33% of the transactions, buyers chose the rigid contract. Consequently, buyers opted for flexible contracts 60.67% of the time. The slight shift to flexible contracts as compared to the baseline treatment does not suggest that the no-competition treatment has an effect on contract choice (Chi-squared test of independence: $\chi^2 = 1.931$, $p = 0.165$). The randomly determined contract terms are very similar. The average fixed price for rigid contracts is 40.67 and the lower price bound for flexible contracts is 39.50.[13] However, buyers under flexible contracts in the good state pay 47.25 on average, i.e., 7.75 more than the randomly determined lower price bound. Similar to the baseline treatment, we can reject the hypothesis that the matching group averages of final prices under rigid contracts and under flexible contracts in the good state are equal (Wilcoxon signed-rank test: $V = 0$, $p = 0.002$).

Moreover, Table 2 also reports an increase in shading rates under flexible contracts as compared to rigid contracts. While sellers under feasible rigid contracts provide low quality in 16.67% of transactions, average shading rates under flexible contracts are 24.65% in the good state and 31.04% in the bad state. In contrast to the baseline treatment, however, we cannot reject the hypothesis that the average shading rates on the matching-group level are equal (Wilcoxon signed-rank test: $V = 14$, $p = 0.193$).

RH2 echoes a prediction in line with Hart and Moore (2008)'s reference-point model. Specifically, the theory posits that ex ante competition is necessary to shape contractual reference points, i.e., competition makes the contract terms salient as a reference point. In the absence of competition, low prices in rigid contracts are no longer justified by competitive market forces. Compared to the baseline treatment, in the no-competition treatment more shading should occur under rigid contracts. In fact, 16.67% of sellers provide low quality in rigid contracts when prices are determined by the exogenous random device. When sellers compete for rigid contracts in auctions, by contrast, shading only occurs in 10.86% of the cases. However, we do not find this difference to be statistically significant. In contrast to Fehr et al. (2011), we cannot reject the hypothesis that shading rates under feasible rigid contracts are equal on the matching-group level (Wilcoxon rank sum test: $W = 35.5$, $p = 0.289$).

To delve deeper into the data and control for seller and matching group idiosyncrasies, we regress an indicator for shading on an indicator for the no-competition treatment. In a simple OLS estimation, we initially find a weakly significant effect of the no-competition dummy on seller's shading choices ($\beta_{\text{NoComp}} = 0.058$, $p = 0.039$; $\beta_{\text{Intercept}} = 0.109, p < 0.001$). However, once we cluster standard errors on the matching group level the result goes away ($\beta_{\text{NoComp}} = 0.058$, $p = 0.294$; $\beta_{\text{Intercept}} = 0.109$, $p < 0.001$). When controlling for the variation in sellers and matching groups by estimating random effects in a mixed effects model, we also cannot find a treatment effect ($\beta_{\text{NoComp}} = 0.035$, $p = 0.569$; $\beta_{\text{Intercept}} = 0.156$, $p = 0.002$).

In sum, we do not find evidence supporting RH2. Substituting the auction, where sellers compete for contracts, with a random device that ex ante fixes contract terms does not cause statistically different shading rates.

Result 2. Sellers under rigid contracts do not shade more often without ex ante competition for prices than when sellers compete for contracts by means of auctions.

4.2. Auxiliary analysis

4.2.1. Price dependence of shading

We want to better understand the behavioral mechanisms behind the sellers' shading choices. While buyers under both contract regimes can pay very similar prices when the good state of nature occurs, a core insight from Hart and Moore (2008)'s model is that a buyer under a flexible contract has an incentive to increase her price in order to prevent her seller to be aggrieved, thus also preventing shading. As shading is thought to be price-dependent, the authors of the original paper analyse the relationship between quality choices and final prices. We use their analysis as a starting point.

For the baseline treatment, Fig. 3 illustrates the relationship between quality choices and final prices in the replication data. As we can consider feasible transactions under rigid contracts only, we also focus on the good state under flexible contracts. Similar to Fehr et al. (2011), in the right panel of Fig. 3 we observe a positive correlation between final prices paid by buyers and normal

[12] Prior studies found the following shading rates under rigid and flexible contracts, respectively, in the good state: 0.063 and 0.251 (Fehr et al., 2011); 0.053 and 0.210 (Fehr et al., 2015); and 0.195 and 0.235 (Erlei and Reinhold, 2016). Our observed difference in shading rates between contract types falls right in between what was observed previously.

[13] The randomization device worked. There is no significant difference to fixed prices (Wilcoxon rank sum test: $W = 58$, $p = 0.579$) or to lower price bounds (Wilcoxon rank sum test: $W = 47$, $p = 0.853$) that were determined by auction in the baseline treatment.

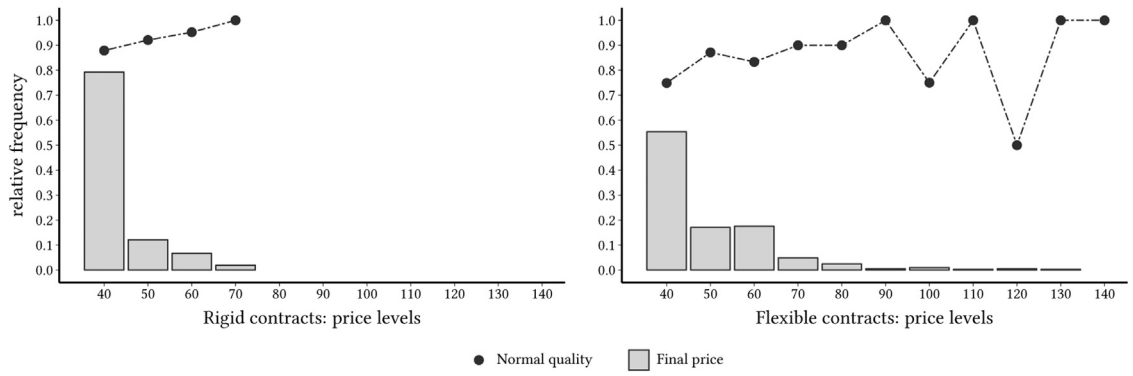

Fig. 3. Relative frequency of normal quality by price levels and distribution of final prices in the good state. Final prices have been rounded to the nearest multiple of ten.

quality provided by sellers under flexible contracts.[14] In contrast to the original study, however, the left panel suggests a similar correlation under rigid contracts. Under both contract types, the frequency of normal quality appears to be increasing in price, i.e., sellers shade less when receiving a higher price. In addition, over the entire price bracket of 40 to 70 the frequency of normal quality is higher under rigid contracts than under flexible contracts, which further supports RH1.

As the extent of shading may be a function of prices paid, we continue investigating the result of the non-parametric test regarding shading between contracts by employing regression analysis. Following Fehr et al. (2011), we regress an indicator for sellers shading on *price increments*, i.e., the actual price less the competitive price of 35, an indicator for flexible contracts, and their interaction term. Using price increments allows us to interpret the OLS constant as the frequency with which sellers shade when buyers offer the competitive price of 35 in rigid contracts. Table 3 reports the results of estimating (1) a linear probability model with clustered standard errors on the matching group level, (2) a linear mixed effects model with random intercepts for individual sellers and matching groups, and (3) a marginal effects estimation from a probit model with clustered standard errors on the seller and matching group level. When considering the difference between final prices and competitive prices, i.e., price increments, the intercept of 0.123 in the simple linear probability model indicates that prices close to the competitive level already trigger some baseline shading under rigid contracts. A change of the contract regime is associated with a noteworthy and significant increase in shading. Across models (1) to (3), the coefficient of the flexible contract dummy suggests that sellers are about 10% to 12% more likely to choose low quality at competitive prices. Note that this is quite different from the original results, where the same estimation yields significant effects of flexible contracts of 33.5% for the OLS model and 29.8% for probit marginal effects.[15] In our data, moreover, sellers' quality choice under rigid contracts does not depend on the price increment and the estimation does not yield an interaction effect between price increments and the flexible contract indicator. The coefficients of the price increment and the interaction of price increment and the flexible contract dummy are close to zero and not significant. All in all, we find only weak evidence regarding the price dependence of shading when employing the same methods as the original authors. However, similar to Fehr et al. (2011) we can reject the hypothesis that the coefficients of price increment and the interaction effect are jointly zero. This result obtains in the simple (F-test: $F = 3.289$, $p = 0.038$) and the mixed effects estimation (Wald-test: $\chi^2 = 17.145$, $p < 0.001$).

We are reluctant to shrug off the weak evidence regarding the influence of price increments. Instead, we explore three other possible price-dependent measures. First, we consider *price differentials*, i.e., the actual price less the auction result.[16] The seller might well compare the final price to the auction result, i.e. to the lowest possible choice the buyer could have made. Price differentials measure by how much the buyer increased his final price above the minimal choice to please the seller and to prevent aggrievement. When sellers reciprocate in kind, one would expect them to shade less. To investigate this idea, we substitute price increments with price differentials but otherwise estimate the same model structures as before. Note that we omit the interaction term because, by design, the price differential only occurs under flexible contracts. In these and the following models, the OLS constant represents the average frequency with which sellers shade under rigid contracts. In models (4) to (6) in Table 3, we capture a similarly strong and significant, positive effect of the flexible contract indicator on shading. Moreover, in both linear probability models we also uncover a negative and (weakly) significant effect of the price differential on shading. In the mixed effects LPM estimation, which performs better than the simple LPM estimation (AIC 421.64 vs. 597.47), an increase of the final price over the auction outcome by one unit is associated with 0.5% decrease in shading. This result supports the idea that shading is contingent on final prices relative to the auction outcome.

The original theory posits that a "party's ex post performance depends on whether the party gets what he is entitled to relative to the outcomes permitted by the contract" (Hart and Moore, 2008, p. 2). Flexible contracts permit any outcome from the feasible price range between the auction result and the upper bound of the possible price range, i.e., [auctionresult, 140]. When assuming that the seller always feels entitled to the most favorable outcome under the contract terms (cf.: Hart and Moore, 2008, p. 8), the probability of shading should increase relative to the distance of the final price to the most favorable outcome of 140. Therefore we also compute the relative aggrievement for each seller under a flexible contract in the good state, i.e., the difference between upper bound of the possible price range and the actual price divided by the difference

[14] We discount observations with a price level of 90 or higher because the number of observations is very small.

[15] We obtain these results by estimating our regression models on the original data. Moreover, our estimation results are analogous to the results reported in Table 3 of Fehr et al. (2011, p. 511), with the exception that their dependent variable is normal quality instead of shading. That means the difference is not driven by the estimation procedure. Rather, the huge effect of flexible contracts in the original data results from both a lower shading rate in feasible rigid contracts (0.06 in the original data vs. 0.11 in our data) and higher shading rates in flexible contracts in the good state (0.25 in the original data vs. 0.20 in our data). Shading behavior under each contract type is more extreme than in our study.

[16] We are indebted to an anonymous referee for suggesting price differentials as explanatory variable.

Table 3
Regression analysis: shading in baseline treatment contingent on contract type and price or reference point measures. "Flexible contract" is an indicator for flexible contracts. "Price increment" is the actual price less the competitive price of 35. "Price differential" is the actual price less the auction outcome. "HM aggrievement" measures the share of the possible price interval (auction outcome to upper price bound) that the buyer withholds from the seller. "Experiential aggrievement" measures the difference between the average experienced aggrievement and the current aggrievement. The latter two measures are standardized. All linear probability models (LPM) cluster standard errors on the matching group level. The linear probability mixed effect models (LPM ME) estimate random intercepts for each seller and each matching group. Probit models (Probit MFX) report marginal effects with clustered standard errors on the seller and on the matching group level.

Dependent variable	Shading (good state)											
	Final price relative to competitive price			Final price relative to auction outcome			HM aggrievement			Experiential aggrievement		
	LPM (1)	LPM ME (2)	Probit MFX (3)	LPM (4)	LPM ME (5)	Probit MFX (6)	LPM (7)	LPM ME (8)	Probit MFX (9)	LPM (10)	LPM ME (11)	Probit MFX (12)
Price increment	−0.003	−0.004	−0.004	–	–	–	–	–	–	–	–	–
(Final price – 35)	(0.003)	0.003	(0.005)	–	–	–	–	–	–	–	–	–
Price differential	–	–	–	−0.003*	−0.005***	−0.003	–	–	–	–	–	–
(Final price – auction outcome)	–	–	–	(0.002)	(0.001)	(0.003)	–	–	–	–	–	–
HM	–	–	–	–	–	–	0.044*	0.071***	0.044	–	–	–
Aggrievement	–	–	–	–	–	–	(0.026)	(0.015)	(0.039)	–	–	–
Experimental	–	–	–	–	–	–	–	–	–	0.052***	0.046***	0.041***
Aggrievement	–	–	–	–	–	–	–	–	–	(0.016)	(0.015)	(0.012)
Flexible contract	0.111**	0.120***	0.100**	0.121***	0.132***	0.116***	0.094***	0.088***	0.090***	0.094***	0.083***	0.090***
	(0.047)	(0.033)	(0.046)	(0.041)	(0.027)	(0.038)	(0.033)	(0.025)	(0.0.36)	(0.032)	(0.025)	(0.039)
Price increment × flexible contract	0.000	0.000	0.002	–	–	–	–	–	–	–	–	–
	(0.004)	(0.003)	(0.005)	–	–	–	–	–	–	–	–	–
Intercept	0.123***	0.149***	–	0.109***	0.126***	–	0.109***	0.125***	–	0.109***	0.128***	–
	(0.041)	(0.040)	–	(0.033)	(0.038)	–	(0.033)	(0.037)	–	(0.033)	(0.036)	–
N	723	723	723	723	723	723	723	723	723	723	723	723
AIC	600.88	439.87	631.29	597.47	421.64	628.18	597.91	417.17	628.75	595.52	429.53	627.70

* $p < 0.10$.
** $p < 0.05$.
*** $p < 0.01$.

between upper bound and the auction outcome. We dub the measure *HM aggrievement* because it is a direct application from Hart and Moore (2008). We estimate the same model structure, now plugging in HM aggrievement. Models (7) to (9) in Table 3 report the results. We obtain very similar effects for baseline shading under rigid contracts, i.e., for the constant. Moreover, the positive and strongly significant effect of flexible contracts on shading is somewhat smaller compared to models containing price differentials as explanatory variable. Importantly, the linear probability models estimate a positive and significant effect of the HM aggrievement measure. Moreover, in absolute terms, the coefficients are much larger than the coefficients of price increments or price differentials. The measure seems to explain the price-dependence of shading much better.

While Hart and Moore (2008) assume for simplicity that the seller always feels entitled to the most favorable outcome under the contract terms, the original theory also recognizes that "when the contract permits more than one outcome, each party may feel entitled to a different outcome" (Hart and Moore, 2008, p. 3). Therefore, in a last step we compute an experiential measure taking into account participants' experienced price choices in the experiment so far. Specifically, we compute the HM aggrievement measure for each participant and for each flexible contract transaction and average this measure. We use this average experienced aggrievement as an individual-specific reference point and compare it to the current HM aggrievement in a period. Table 3 refers to this measure as *experiential aggrievement*. Models (10) to (12) report the estimation results, which are qualitatively similar to the results with HM aggrievement and to the results with price differentials. We observe significant baseline shading under rigid contracts and the flexible contract indicator has a positive and significant effect on shading. Finally, experiential aggrievement has a strong and significant positive association with shading across all models, including the probit model.

To sum up, we hardly find that shading is contingent on price increments alone. However, our results clearly show that shading is contingent on how the final price relates to reference point measures, specifically to how the final price absolutely or relatively divides the possible outcome space permitted by the contract and to how current aggrievement relates to experienced aggrievement. Also note that investigating the price dependence of shading consistently confirmed that flexible contracts in the good state increase the likelihood of shading as compared to rigid contracts. The results reported in Table 3, thus, also complement our earlier results regarding RH1.

Auxiliary Result 1. Although final prices alone have at most weak effects on shading, in the baseline treatment shading behavior is clearly contingent on different reference-dependent measures of aggrievement.

Remember that we did not find differences in shading when removing competition by substituting the auction with a random draw. To investigate shading choices in flexible contracts without competition, we engage in the same regression analysis as in the baseline treatment. Table 4 reports the results. Across almost all models and all reference point measures, a change of the contract regime to flexible contracts is significantly associated with much more shading. For instance, when looking at price increments, shading under flexible contracts at competitive prices is increased by 12% to 16.2%. The only exception occurs for the marginal effects of the probit model with HM aggrievement as reference measure. In this model, the HM aggrievement measure has the strongest effect on shading of all reference measures and possibly explains away the difference between rigid and flexible contracts. Note, in addition, that with the original data we estimate (weakly) significant effects of the flexible contract indicator of 11% for the OLS model and 10.2% for probit marginal effects.[17] The coefficients for the flexible dummy indicator also suggest at least the same effect size as compared to the baseline treatment. This finding is in stark contrast to the original results. In Fehr et al. (2011)'s study, the estimated coefficients were about three times higher in their baseline treatment as compared to both their no-competition treatment and our baseline treatment. In the original study, this result occurs because sellers under rigid contracts shade much less and sellers under flexible contracts shade much more in the baseline treatment as compared to the no-competition treatment. While, as a consequence, Fehr et al. (2011, p. 517) find that the elimination of ex ante competition substantially reduces contract-specific differences in shading choices, we do not find such a distinct effect. Finally, regarding price differentials and HM aggrievement, we find robust evidence for reference-dependent shading also in the no-competition treatment.

Auxiliary Result 2. Shading remains contingent on reference-dependent measures of aggrievement also in the no-competition treatment. In contrast to the original study, the removal of competition does not remove contract-specific differences in shading behavior.

4.2.2. Time trends

Further following Fehr et al. (2011) we investigate how the outcome variables in the baseline treatment develop by period. Regarding auction outcomes in rigid and flexible contracts, the upper panel of Fig. 4 illustrates that both the fixed price in rigid contracts and the lower price bound in flexible contracts decrease over time, converging towards the competitive price of 35. An OLS regression (Table 5) confirms this visual result. We regress the auction outcome on the period variable, an indicator for flexible contracts, and the interaction term of the two. We account for idiosyncrasies on the participant level by estimating seller fixed effects. The estimated coefficient of the period variable is negative, sizeable, and statistically significant. By contrast, the estimated coefficients for the contract type indicator and the interaction term appear to be statistically insignificant. In Fig. 4, the results from the first rounds visually appear to be different from the results from the last ten periods. To account for participants' learning effects during the initial periods of the experiment, we estimate the same regression model for the subset of the data covering the last ten periods.[18] While the estimated coefficient of the period variable is less negative in the last ten periods but remains strongly significant, we find a significant negative coefficient for the flexible contract indicator in the last ten periods. Note that this result differs from our prior finding, where auction results across contract types were not different on the matching group level (compare Section 4.1.1). Moreover, focusing on the last ten periods yields a significant positive interaction between the flexible contract indicator and the period variable. In short, auction outcomes are lower for flexible contracts, but they decline less in time as compared to rigid contracts.

Auxiliary Result 3. In the last ten rounds of the experiment, flexible contracts are associated with lower auction outcomes as compared to rigid contracts.

The summary in Table 1 suggests that buyers under flexible contracts pay more than the lower price bound when the good state of nature occurred. While auction outcomes are very similar across contract types, buyers under flexible contracts pay more (48.97)

[17] Our estimation results are analogous to the results reported in Table 3 of Fehr et al. (2011), with the exception that their dependent variable is normal quality instead of shading.

[18] We thank an anonymous referee for suggesting to also analyze the last ten periods.

Table 4
Regression analysis: shading in no-competition treatment contingent on contract type and price or reference point measures. "Flexible contract" is an indicator for flexible contracts. "Price increment" is the actual price less the competitive price of 35. "Price differential" is the actual price less the auction outcome. "HM aggrievement" measures the share of the possible price interval (auction outcome to upper price bound) that the buyer withholds from the seller. "Experiential aggrievement" measures the difference between the average experienced aggrievement and the current aggrievement. The latter two measures are standardized. All linear probability models (LPM) cluster standard errors on the matching group level. The linear probability mixed effect models (LPM ME) estimate random intercepts for each seller and each matching group. Probit models (Probit MFX) report marginal effects with clustered standard errors on the seller and on the matching group level.

Dependent variable	Shading (good state)											
	Final price relative to competitive price			Final price relative to auction outcome			HM aggrievement			Experiential aggrievement		
	LPM (1)	LPM ME (2)	Probit MFX (3)	LPM (4)	LPM ME (5)	Probit MFX (6)	LPM (7)	LPM ME (8)	Probit MFX (9)	LPM (10)	LPM ME (11)	Probit MFX (12)
Price increment	−0.005**	−0.004	−0.007*	–	–	–	–	–	–	–	–	–
(Final price – 35)	(0.002)	(0.003)	(0.004)	–	–	–	–	–	–	–	–	–
Price differential	–	–	–	−0.007***	−0.008***	−0.009***	–	–	–	–	–	–
(Final price – auction outcome)	–	–	–	(0.002)	(0.001)	(0.003)	–	–	–	–	–	–
HM	–	–	–	–	–	–	0.090***	0.100***	0.119***	–	–	–
Aggrievement	–	–	–	–	–	–	(0.026)	(0.017)	(0.044)	–	–	–
Experimental	–	–	–	–	–	–	–	–	–	0.021	0.017	0.020
Aggrievement	–	–	–	–	–	–	–	–	–	(0.022)	(0.017)	(0.020)
Flexible contract	0.132**	0.162**	0.120***	0.133***	0.152***	0.127***	0.080*	0.092***	0.061	0.080*	0.091***	0.079*
	(0.052)	(0.034)	(0.044)	(0.045)	(0.029)	(0.040)	(0.045)	(0.027)	(0.048)	(0.044)	(0.028)	(0.044)
Price increment × flexible contract	−0.002	−0.004	−0.001	–	–	–	–	–	–	–	–	–
	(0.002)	(0.003)	(0.002)	–	–	–	–	–	–	–	–	–
Intercept	0.194***	0.186***	–	0.167***	0.169***	–	0.167***	0.169***	–	0.167***	0.169***	–
	(0.053)	(0.039)	–	(0.046)	(0.037)	–	(0.046)	(0.037)	–	(0.046)	(0.037)	–
N	712	712	712	712	712	712	712	712	712	712	712	712
AIC	732.37	601.51	715.76	734.38	589.20	716.62	733.82	583.55	715.88	753.93	617.41	739.93

* $p < 0.10$.
** $p < 0.05$.
*** $p < 0.01$.

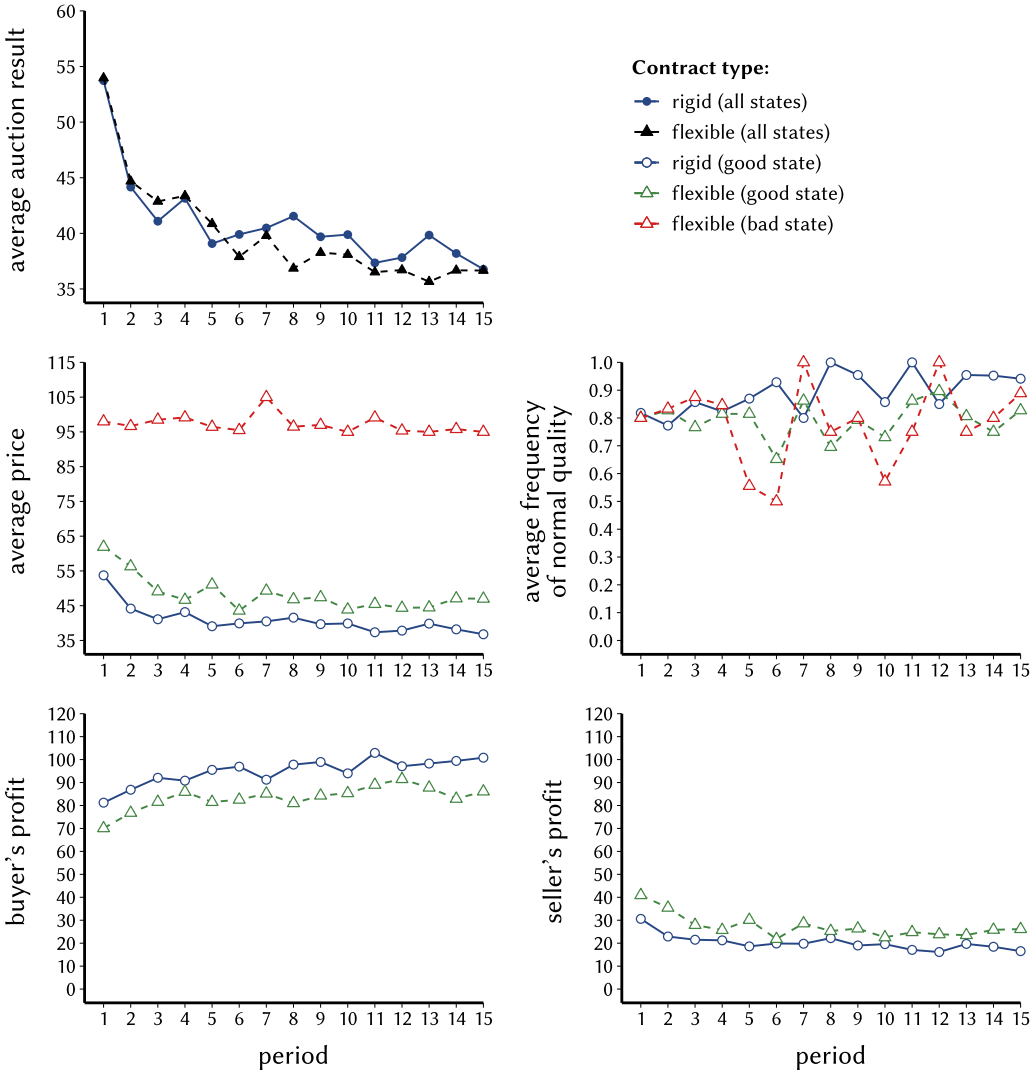

Fig. 4. Average outcomes by period.

Table 5
Regression analysis of outcome variables in the baseline treatment. The models for auction outcomes and normal quality estimate seller fixed effects. The model for final prices estimates buyer fixed effects. The models for buyer's and seller's profit estimate fixed effects for each buyer-seller pair. All models use heteroskedasticity and autocorrelation-consistent (HAC) standard errors. The models for final prices, normal quality, and buyer's and seller's profit only consider the good state.

	Auction outcome		Final prices		Normal quality		Buyer's profit		Sellers's profit	
	(OLS)		(LPM)		(OLS)		(OLS)			
	All periods	Last 10 periods	All periods	Last 10 periods	All periods	Last 10 periods	All periods	Last 10 periods	All periods	Last 10 periods
Period	−0.826***	−0.437***	−0.831***	−0.678***	0.008	0.010	1.069***	1.030***	−0.835***	−0.479***
	(0.085)	(0.069)	(0.109)	(0.158)	(0.005)	(0.007)	(0.195)	(0.321)	(0.156)	(0.187)
Flexible Contract	−1.250	−3.025**	12.067***	11.568*	−0.054	−0.053	−12.586***	−11.336	11.228***	16.347**
	(1.217)	(1.288)	(3.559)	(6.108)	(0.058)	(0.107)	(3.610)	(7.988)	(3.363)	(6.441)
Flexible Contract × Period	0.044	0.199**	0.122	0.488*	−0.005	−0.004	−0.153	−0.616	0.204	0.344
	(0.099)	(0.088)	(0.219)	(0.296)	(0.006)	(0.009)	(0.294)	(0.459)	(0.225)	(0.289)

* $p < 0.10$.
** $p < 0.05$.
*** $p < 0.01$.

on average than buyers under rigid contracts (41.04). The middle-left panel of Fig. 4 indicates that the price difference between flexible contracts (good state) and rigid contracts is robust across periods. We confirm this visual result, again using regression analysis (Table 5). We rely on the same model structure as before, but now include final prices as dependent variable and estimate buyer fixed effects. The estimated coefficient for the flexible contract dummy is positive, large, and statistically significant. Otherwise the results are similar to the case of auction outcomes. The estimated effect of the period variable is negative and significant and there is no discernible interaction between period and contract type. When excluding the first five periods, the coefficients for the

period variable and the flexible contract indicator change slightly in size and the indicator variable for flexible contracts becomes only weakly significant. In addition, the estimation uncovers a weakly significant positive interaction effect between the flexible contract indicator and the period variable. That is, flexible contracts yield higher final prices than rigid contracts and final prices under flexible contracts decrease less over time.

Do the increased prices paid by buyers under flexible contracts in the good state prevent shading by sellers? A core result in the original study, our first replication hypothesis, and our first result was that, in the good state, frequencies of normal quality are lower under flexible contracts than under rigid contracts (compare Section 4.1.1). In our study, sellers under rigid contracts provide normal quality in 89% of the cases when a mutually beneficial transaction is feasible. Sellers under flexible contracts, however, provide normal quality in 80% of the time when the good state of nature occurs. Even though buyers pay higher prices on average, under flexible contracts sellers provide normal quality less often than under rigid contracts. As the middle-right panel of Fig. 4 illustrates, quality choice appears to have no clear overall trend over time for both contract types. Moreover, shading rates fluctuate substantially under both contract types. We again investigate time effects between contract types with regression analysis (Table 5). We plug-in normal quality as dependent variable and otherwise rely again on a linear model structure with seller fixed effects. The estimation yields no effects of either period, contract type, or the interaction thereof. This result does not change when considering the last ten periods only. Note that the effect of the flexible contract indicator becomes significant once the interaction with the period variable is removed.

The effects of contract type and period on final prices naturally affect the profit distribution between buyers and sellers in the good state. Compared to rigid contracts, flexible contracts shift the profit distribution towards sellers. In the good state, buyers' average profit decreases from 94.76 under rigid contracts to 83.62 under flexible contracts. Conversely, sellers' average profit increases from 20.35 under rigid contracts to 27.27 under flexible contracts (see Table 1). We can reject the hypothesis that the matching-group averages of buyer's profit (Wilcoxon signed rank test: $V = 54, \quad p = 0.004$) and seller's profit (Wilcoxon signed rank test: $V = 0, \quad p = 0.002$) are equal in rigid and flexible contracts. The lower panels of Fig. 4 show that these payoff differences between flexible contracts and rigid contracts are persist across rounds. Using buyer's profit and seller's profit as dependent variables, we estimate the same regression model, now including fixed effects for each buyer-seller pair. The estimation results for all periods confirm the visual impression. The flexible contract indicator has a significant and sizeable negative effect on buyer's profits and, conversely, a significant and sizeable positive effect on seller's profits. When considering only the last ten periods, the effect of flexible contracts on buyer's profit, however, is not significant while the effect of flexible contracts on seller's profit even increases. Similarly, every additional period has a significant positive effect on buyer's profit and a significant negative effect on seller's profit. This effect persists after excluding the first five periods. The regression analysis does not reveal interaction effects.

For the no-competition treatment, we exploit our per-period data to shed light on our null result regarding RH2. Remember, our results diverge from Fehr et al. (2011) in that sellers under rigid contracts do not shade more often without ex ante competition as compared to competitive auctions. By depicting the per-period data of relative shading frequencies under rigid contracts both in the baseline treatment and in the no-competition treatment, Fig. 5 suggests a reason for why we may not be able to pick up a treatment difference. In the Fehr et al. (2011) data, the relative frequency of shading under rigid contracts in the baseline treatment lies in a very narrow range from 0.036 to 0.115. In our data, by contrast, this range extends from 0 to 0.227. In fact, the difference between the per-period shading in the two baseline treatments is significant (Wilcoxon rank sum test: $Z = -2.055$, $p = 0.040$), whereas the difference between the per-period shading in the two no-competition treatments is not (Wilcoxon rank sum test: $Z = 0.021$, $p = 0.992$).

In other words, sellers under rigid contracts in baseline treatment shade on a consistent low level in the Fehr et al. (2011) study, whereas the shading behavior that we observe fluctuates similarly to both no-competition treatments.

Auxiliary Result 4. In comparison to Fehr et al. (2011)'s results, seller's shading behavior under rigid contracts in the baseline treatment is elevated and more heterogeneous.

4.3. Replication analysis: summary

In our replication analysis, we evaluate how our results on RH1 and RH2 fit in with the previous results. Bayesian model comparison provides the framework for this evaluation. We assume that the data generating process is the same for all studies. We use the previous results to inform different sets of priors, which get updated given our data. In short, we investigate which previous results are more likely in light of our new data. We provide all technical details in Appendix A.

Our results so far indicate weak non-parametric evidence that sellers shade more often under flexible contracts than under rigid contracts. Our regression analysis found this result to be much more robust. Our Bayesian model comparison allocates next to all prior credibility from a null model (without any effect) to either the results of Fehr et al. (2011) or Fehr et al. (2015), the latter of which receive the major share of that credibility. That is, our data strongly support either of the prior results over a null result. Comparing the two prior studies, our observed data are about 1.66 times more likely given the prior information from Fehr et al. (2015) rather than from Fehr et al. (2011) (compare the upper-left panel of Fig. 6 in the appendix). In any case, our replication supports that shading rates are higher under flexible contracts. A robustness check strengthens this finding. In this regard, we judge our replication as successful.

Result 3. The replication supports the existence of higher shading rates among sellers under flexible contracts compared to sellers under rigid contracts.

As to shading rates under rigid contracts with and without competition, we found no evidence for increased shading under rigid contracts without competition. When modeling a very strict null effect, our Bayesian procedure strongly supports the prior results of Fehr et al. (2011). However, the results are sensitive to how strict we model the null effect. If the constraints on the null effect are lax enough, this result flips such that the null model receives most posterior credibility. Moreover, a robustness check does not overcome this reversal pattern.

Strictly speaking, if we put a high degree of belief into a null effect, the result of the model comparison supports Fehr et al. (2011). However, this finding is at odds with the analysis of our data. We reconcile these findings as follows: The treatment effect that we measure is too small to generate a significant test result, yet it is too large to confirm an idealized null hypothesis. Although we prefer high precision priors as idealized beliefs for the null effect, the fact that we see a reversal of posterior model probabilities also under the conditions of our robustness check suggests to stay very cautious about replication success (compare Fig. 7 in the appendix). For a conclusive result, more data are required.

Result 4. The replication does not unambiguously support the existence of higher shading rates under rigid contracts without ex ante competition for prices compared to when sellers compete for contracts by means of auctions.

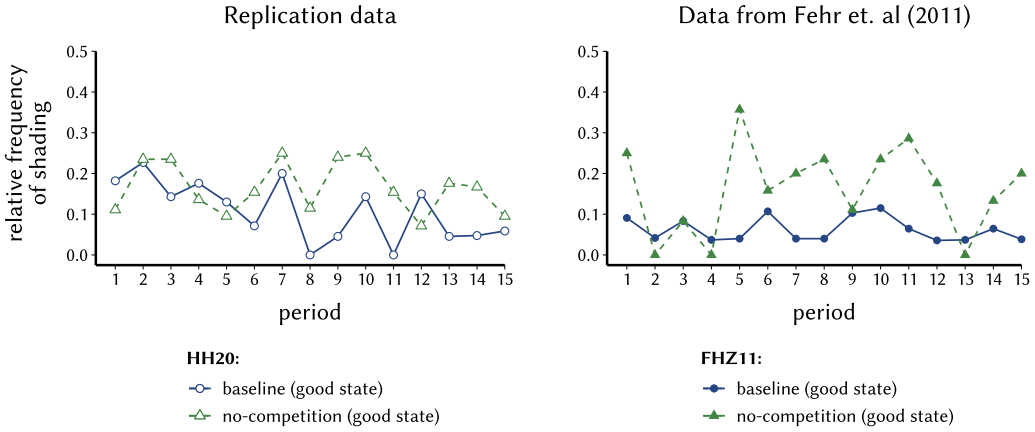

Fig. 5. Frequency of shading under rigid contracts by period with and without ex ante competition. The left panel shows our results. The right panel shows the result of Fehr et al. (2011).

5. Discussion & conclusion

Hart and Moore (2008) develop a comprehensive model explaining how contracts serve as reference points in trading relationships. At the same time, the authors are astonishingly vague in specifying the behavioral channels driving this effect. With respect to the origins of diverging reference points of sellers and buyers, the authors state that they "... do not model why these differences in entitlements arise, but [...] have in mind the kinds of effects described in the self-serving bias literature" (Hart and Moore, 2008, p. 8). The experiment of Fehr et al. (2011) provides evidence that the model's mechanic is indeed empirically relevant as they find significantly more shading in flexible compared to rigid contracts. We find the same pattern in our replication, which confirms our RH1. Moreover, with respect to RH1, our replication analysis suggests replication success.

The importance of reference-dependent measures in the baseline (see Auxiliary Result 1) provide additional support that aggrievement due to a self-servingly biased expectation about the final price plays a key role. If buyers set too low final prices in flexible contracts, sellers are aggrieved and shade. Auxiliary Result 3 adds a piece to this story. From the seller's perspective, a flexible contract is a way around the trade-off caused by competition among buyers, i.e., the trade-off between waiting longer in the clock auction to receive a higher price and the risk of missing the contract. In a flexible contract, the seller can accept the contract right away, trusting – maybe even expecting – that the seller will adjust the price to a seemingly fair level. In line with this reasoning, we find that flexible contracts are associated with lower auction outcomes (Auxiliary Result 3). That is, after some learning period sellers accept the flexible contract much faster than they accept the rigid contract. This expectation of a fair price then sets the ground for potential aggrievement if the seller chooses a sufficiently low final price.

The lack of support for RH2 deserves discussion, however. In contrast to Fehr et al. (2011), under rigid contracts we do not find higher shading rates without competition for prices than when sellers compete for contracts through auctions. This result is not caused by a systematic difference in price. The shading rate under rigid contracts in the baseline treatment is 0.109. By contrast, in the prior studies these shading rates are 0.063 (Fehr et al., 2011) and 0.052 (Fehr et al., 2015), respectively. That is, our shading rates under rigid contracts in the baseline treatment is about twice as high, even in the presence of competition. Additionally, shading behavior under rigid contracts in the baseline treatment fluctuates much more in our study than in Fehr et al. (2011) (Auxiliary Result 4). The effect of removing competition is, therefore, much more difficult to identify statistically. In the no-competition treatment though, shading rates (0.167 and 0.164, respectively) as well as the level of data heterogeneity are very similar between the two studies.

Fehr et al. (2011, p. 496) admit that their low shading rates under rigid contracts in the baseline treatment are somewhat "puzzling". Our average shading rate of 0.109 is very close to the shading behavior of 0.10 reported in Fehr et al. (2019)'s working paper, and Erlei and Reinhold (2016) find (with their slightly different setup) an even higher shading rate of 0.195 in the presence of competition. However, the decision environment used in all these experiments is rather rich in detail, such that there may well be additional behavioral forces present that occurred to a varying degree across the different studies. We would like to emphasize three additional behavioral channels that may impact results: (1) reference points outside the contract; (2) commitment effects; and (3) fairness considerations. Our comparatively higher shading rate under rigid contracts in the baseline treatment suggests that we pick up these effects more than Fehr et al. (2011) did, which obfuscates measuring the effect of removing competition.

First, the theory suggests that under a rigid contract with competitively determined contract terms sellers cannot be aggrieved and, consequently, shading should not occur. However, our results indicate some non-negligible degree of baseline shading under competitive rigid contracts. The occurrence of baseline shading under competitive rigid contracts could be a sign that competition is not fully working as predicted by the theory and that some subjects set reference points outside the contract. In fact, Hart and Moore (2008) discuss the possibility of reference points outside the contract. In such a case, sellers consider reasonable alternative price intervals. In real world settings, such alternatives might be, e.g., market prices of substitute products or services or wages of colleagues. In the artificial environment of the experiment, the reasonable alternative price interval would probably be the possible price interval in the auction, i.e. [35, 75], which applies for both contract types. To the extent that oversupply forces prices to the competitive level, which is observed by Fehr et al. (2011) as well as in our replication, the range of outcomes permitted by the flexible contract includes this alternative interval of outcomes outside the contract. As both outcome spaces greatly overlap, we cannot observe whether shading under flexible contracts in the baseline treatment occurs due to a reference point inside or outside the contract.

Second, in light of our results we wonder whether competition is really that crucial. The experiment of Brandts et al. (2016) does not feature auctions and oversupply, but buyer-seller pairs and take-

it-or-leave-it-offers. Even without competition through auctions, their results also provide supporting evidence for RH1. Moreover, in the regression analysis concerning our no-competition treatment we also find an increase in shading under flexible as compared to rigid contracts. Also, shading is still reference-dependent in line with the reference point proposed by Hart and Moore (2008) (see Auxiliary Result 2). Therefore, the basic mechanism appears somewhat robust even in the absence of competition. This is good news for the reference point hypothesis! However, the finding again emphasizes the question for the underlying behavioral mechanism. We conjecture that there may be a commitment-and-consistency effect at work.[19] Sellers in Fehr et al. (2011)'s experimental paradigm actively accept—one could even say: hunt—the contract through the auction. Similarly, sellers in Brandts et al. (2016)'s design actively accept the take-it-or-leave-it offer. In each case, sellers actively and quite publicly commit themselves to the contract and its terms. Afterwards, committed sellers consistently follow through with their commitment by not shading. The commitment, however, is less clearly defined under flexible contracts because the contract terms, i.e., the price, are not finalized. This lack of concrete terms, weakens sellers' commitment, which leads to more shading. Moreover, a similar commitment is missing in the no-competition treatment. Future research should start exploring this behavioral channel.

Third, by proposing rigid contracts, buyers in fact make unfair offers to sellers because rigid contracts typically lead to an uneven distribution of the gains from trade favouring buyers. Models of inequity aversion (e.g.: Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) suggest that there should be a substantial amount of shading by sellers to counteract the uneven distribution of surplus. On the other hand, the price in a rigid contract is beyond the control of the buyer and this lack of control might make seller's less resentful.[20] Intention-based fairness (e.g.: Rabin, 1993; Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006) would similarly call for negative reciprocation upon ungenerous intentions. Erlei and Reinhold (2016)'s investigation aims at disentangling this behavioral channel from the original reference point hypothesis by introducing exogenously determined contract types. Their findings give a first hint that additional forces might be present.

In fact, the riddle of the different studies obtaining remarkably different baseline shading rates from 0.05 to 0.19 in rigid contracts remains empirically unsolved. Our study clearly supports the idea that baseline shading is a non-negligible phenomenon. With our data, we unfortunately cannot answer the question *why* shading behavior of sellers under rigid contracts differs wildly across the different studies. Thus, as a final lesson from this replication, we call for important future research. If we are to better understand the effects of contract design on shading behavior, future research needs to explore drivers of shading in rigid contracts. Moreover, we encourage researchers to identify and disentangle the different behavioral channels underlying the evidence for Hart and Moore (2008)'s contractual reference points that has been produced to date. Finally, future work should also investigate how these different behavioral forces interact.

Author statement

The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

[19] For an introduction to commitment and consistency as behavioral mechanisms, see Cialdini (2007, Ch. 3).

[20] We thank an anonymous referee for pointing out this lack of choice effect to us.

Appendix A. Replication analysis in detail

To develop an idea about how closely our results resemble previous findings, we employ Bayesian model comparison. We let different hypotheses compete for our data, namely a skeptic null model doubting the presence of a particular effect and a proponent model representing the belief that the previously identified result is true (cf.: Bem et al., 2011; Dienes, 2011; Verhagen and Wagenmakers, 2014).

As our two replication hypotheses RH1 and RH2 speak to shading frequencies, our Bayesian models rely on Ferrari and Cribari-Neto (2004)'s beta regression framework. For each matching group i, let Y_i be the shading rate and let θ_i be an indicator for flexible vs. rigid contracts (with respect to RH1) or the no-competition vs. the baseline auction treatment (with respect to RH2). Formally:

$$Y_i \sim Beta(\phi\,\mu_i, \phi\,(1-\mu_i)),$$

where

$$logit(\mu_i) = \beta_0 + \beta_1\,\theta_i$$

such that the expected value of Y_i is $\mu_i \in [0, 1]$ and the concentration around μ_i is determined by $\phi > 0$. We assign an uninformative prior for the concentration parameter, $\phi \sim Gamma(0.1, 0.1)$. The competing Bayesian models only differ in the remaining priors for β_0 and β_1, i.e., these model-specific priors will represent the skeptic null hypothesis and the proponent hypotheses. We are mainly interested in the posterior distribution of β_1.

Our approach to model the proponent hypotheses is similar in spirit to the Bayes factor test for replication success developed by Verhagen and Wagenmakers (2014). We tightly link the proponent model to the results of the original experiment. In fact, we assume that the posterior distributions of β_0 and β_1 as obtained from the original experiment accurately describe the proponent hypothesis. We also assume that the proponent started out with uninformative priors, $\beta_j \sim Cauchy(0, 2.5), j \in \{0, 1\}$ (Gelman et al., 2008). Then we use original data reported by prior studies to update the vague prior distributions of β_0 and β_1 to the corresponding original posteriors. Finally, we use the information of these posteriors, i.e., mean and standard deviation, to model informed normally distributed priors for our proponent models.

To model the skeptic null hypothesis, generally, we want to use quite precise, i.e., informed priors. General-purpose priors that are centered on zero and have fat tails are agnostic, i.e. they actually imply that any effect is plausible. Our null models, by contrast, describe a sceptic who is convinced that no effects exist for switching from rigid to flexible contracts (RH1) and for removing competition during the price setting stage (RH2). Additionally, the prior is supposed to describe our beliefs about baseline shading under rigid contracts. We observe baseline shading in all studies that employ this experimental paradigm (Fehr et al., 2011, 2015; Erlei and Reinhold, 2016). Therefore, regarding the β_0-prior in our skeptic null model, we proceed similarly to the proponent model. We pool all available baseline shading frequencies and use this data to update a naive Bayesian beta regression model as explained above, with the exception that $\theta_i = 0 \quad \forall\, i$. Then we use information from the resulting β_0-posterior as prior for the skeptic model. For describing the skeptic effect of θ_i, by contrast, we start out with a precise β_1-prior centered on zero, $\beta_1 \sim Normal(0, 0.05)$. Modeling the skeptic null model involves actually an interesting choice. We could have also modeled a rather precise β_0-prior centered on zero. This choice would have described a world without any baseline shading in rigid contracts, much like both the neoclassical and the reference-point model predictions for rigid contracts in Fehr et al. (2011). But our replication test did not aim at finding base-

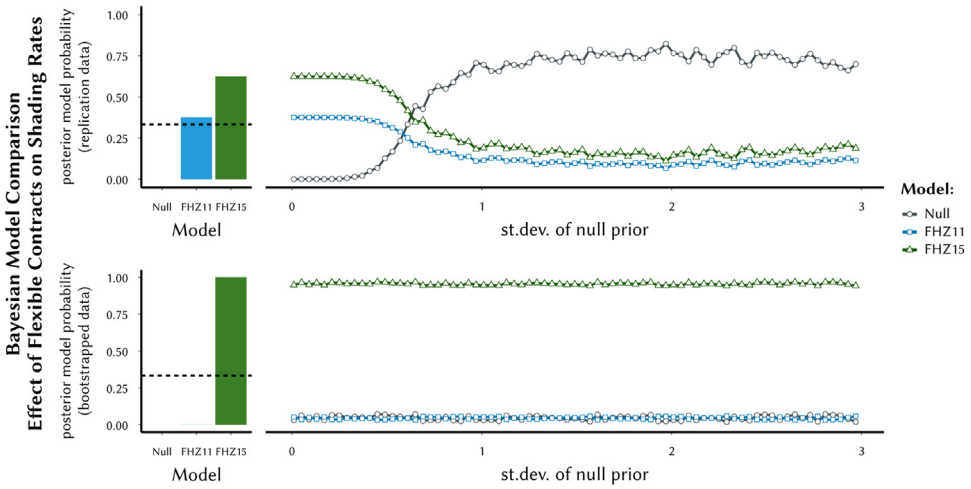

Fig. 6. Bayesian model comparison for the effect of flexible contracts on shading rates. The upper panel displays posterior model probabilities of three competing beta regression models given our replication data. The models only differ in their priors, representing prior knowledge according (1) to a null effect, (2) to the results of Fehr et al. (2011), and (3) to the results of Fehr et al. (2015). The lower panel shows results after bootstrapping the replication data. In each panel, the left-hand side plot displays posterior model probabilities when the relevant prior of the null model for the effect of flexible contracts is very precise (st. dev. $\sigma = 0.05$). The right-hand side plot displays the posterior model probabilities as a function of the precision of the relevant null model prior for contract types. The data in the right-hand side plot of the lower panel are vertically jittered to increase visibility.

line shading in the first place. Therefore, we assume some baseline shading and model β_0-prior accordingly. After all, modeling the β_0-prior for baseline shading makes the model more difficult to reject and, thus, create a harder test, given that baseline shading is present in all studies.

Finally, we confront skeptic and proponent models with our replication data. After obtaining posterior samples for the models, we estimate marginal likelihoods with bridge sampling (Meng and Wong, 1996; Meng and Schilling, 2002) to obtain posterior model probabilities for each competing models.[21]

A.1 Shading under flexible contracts

Regarding the effect of switching from rigid to flexible contracts on shading rates, we inform two proponent models with the previous results of Fehr et al. (2011) and Fehr et al. (2015).[22] The upper panel of Fig. 6 illustrates the results of confronting the models with the data from our replication attempt. The left-hand side plot depicts the posterior model probabilities when the β_1-prior of the null model is very precise (standard deviation $\sigma = 0.05$). From left to right, the posterior model probabilities are 0.000, 0.376, and 0.624 for the skeptic model and the two proponent models, dubbed FHZ11 and FHZ15, respectively. Next to all prior credibility has shifted from the null model to one of the proponent models. Between the two proponent models, FHZ15 receives the major share of that credibility. Given our replication data, each of the proponent models is much more likely than the null model. The log-scaled Bayes factors in favor of the proponent over the null model are 9.113 for FHZ11 and 9.621 for FHZ15, respectively (cf.: Raftery, 1995). That is, our data strongly support either of the proponent models over the null model. Comparing the two proponent models, the observed data are about 1.66 times more likely under the FHZ15 proponent model than under the FHZ11 proponent model. The log-scaled Bayes factor in favor of FHZ15 over FHZ11 is 0.507, implying that our data provide weak support for FHZ15 over FHZ11.

Note, however, that we could only gather ten data points per contract type on the matching-group level. Each updated posterior constitutes a compromise between the prior distributions and the information in the data. With ten data points per contract type, we are concerned whether the data indeed overwhelm the prior. Put differently, we cannot be certain whether our choice regarding the precision of the null model's β_1-prior has a negligibly influence on the parameter posterior and, consequently, also the posterior model probabilities.

To gauge this possibility, we conduct a sensitivity check by running the model comparison and estimating posterior model probabilities for a range of precision values of the null model's β_1-prior. In terms of standard deviation σ, we vary σ from 0.01 to 3.00. The right-hand side plot in the upper panel of Fig. 6 depicts the posterior model probabilities as function of the precision of the null model's β_1-prior. The plot shows that the relation of the proponent models to the null model indeed depends on the precision of its β_1-prior. Our results are fairly robust up until $\sigma \approx 0.4$. With further decreasing precision of the β_1-prior the null model soaks up more and more credibility such that its posterior probability increases, until it is much more likely than any of the proponent models.[23] Although the results are susceptible to the choice of priors, our confidence rests in the results at high levels of precision. The purpose of our model comparison lies in testing the proponent models against a very strict null hypothesis and the posterior probabilities are stable in a high precision range, i.e., when $\sigma \leq 0.4$.

Nevertheless, as a final step we push against the small N problem by resampling our data. Within each matching group and for each contract type, we resample participant's quality choices 100 times, each time computing the mean. With the 2000 bootstrapped

[21] We use *JAGS 4.3.0* (Plummer, 2003) with *runjags 2.0.4-6* (Denwood, 2016) to sample from the models. We use *bridgesampling 1.0.0* (Gronau et al., 2020) for bridge sampling.

[22] Unfortunately, we could not use any of Erlei and Reinhold (2016)'s results as prior information for yet a different proponent model. Erlei and Reinhold (2016) do not report matching-group level data and, therefore, we could not estimate original posteriors that would inform our model comparison priors. Moreover, we could not retrieve the original data. The data is not archived on the article's website and we could not establish contact with the remaining author.

[23] As model priors are not overwhelmed by the data, the β_1-prior of the null model can become vague enough such that the posterior probabilities of the two proponent models will start to dominate again. At $\sigma = 100$, for instance, the posterior model probabilities are 0.092, 0.341, and 0.567 for the null, the FHZ11, and the FHZ15 model, respectively.

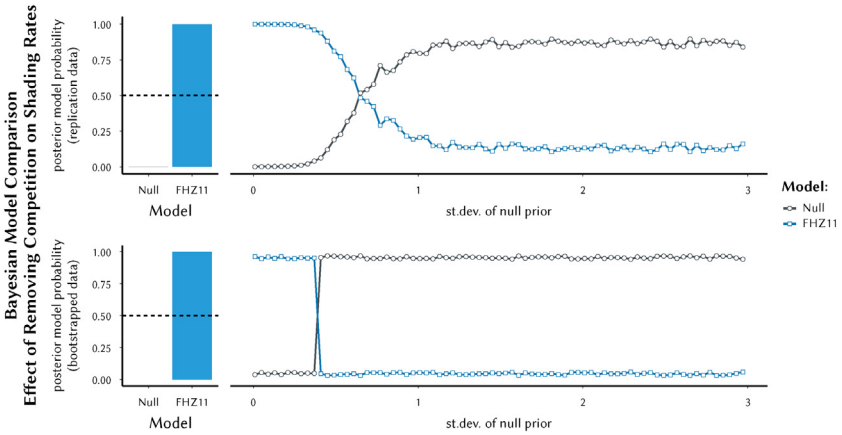

Fig. 7. Bayesian model comparison for the effect of removing competition on shading rates. The upper panel displays posterior model probabilities of two competing beta regression models given our replication data. The models only differ in their priors, representing prior knowledge according (1) to a null effect and (2) to the results of Fehr et al. (2011). The lower panel shows results after bootstrapping the replication data. In each panel, the left-hand side plot displays posterior model probabilities when the relevant prior of the null model for the effect of removing competition is very precise (st. dev. $\sigma = 0.05$). The right-hand side plot displays the posterior model probabilities as a function of the precision of the relevant null model prior for the no-competition treatment. The data in the right-hand side plot of the lower panel are vertically jittered to increase visibility.

shading rates, we conduct our model comparison again.[24] Note that our Baysian beta regression models now slightly change: to account for non-independence of bootstrapped observations within matching groups, we add random intercepts on the matching-group level with uninformative priors, $\nu_k \sim Cauchy(0, 2.5)$, $k \in \{1, \ldots, 10\}$. The lower panel of Fig. 6 illustrates the results of the model comparison with the bootstrapped replication data. The Bayesian algorithm assigns nearly all posterior credibility to the proponent hypothesis modeling the results of Fehr et al. (2015). The estimated log-scaled Bayes factors in favor of the FHZ15 model over both the null and the FHZ11 model are 652.574 and 299.560, respectively.[25] Moreover, the sensitivity analysis indicates that this result is independent of the precision of the null model's β_1-prior.[26] Given the bootstrapped data, the model comparison clearly contradicts the null model and supports a proponent hypothesis, albeit not one that reflects Fehr et al. (2011)'s data but rather Fehr et al. (2015)'s results. We interpret these results as replication success.

A.2 Shading without competition

With respect to the effect of removing competition on shading rates, only Fehr et al. (2011) provide the data to inform a proponent model. The upper panel of Fig. 7 visualizes the results of confronting the null and the proponent model with our replication data. Given a very precise β_1-prior of the null model (standard deviation $\sigma = 0.05$), the model probabilities are very close to 0 for the null model and very close to 1 for the proponent model. The log-scaled Bayes factors in favor of the proponent over the null model is 8.158, i.e., our data strongly support the proponent over the null model.

Our sensitivity analysis reveals that this result depends on the precision of the null model's β_1-prior. The posterior model probabilities are robust until $\sigma \approx 0.3$. At $\sigma \approx 0.7$ posterior model probabilities are about equal. Finally, posterior model probabilities strongly support the null model from $\sigma \approx 1.2$ onwards.

In contrast to the results for contract types, however, the reversal also occurs when conducting the same analysis with the highly concentrated data of the bootstrapped shading means. When β_1-prior of the null model is very precise ($\sigma = 0.05$), the log-scaled Bayes factor in support of the FHZ11 model over the null model is 470.693. However, from $\sigma = 0.37$ to $\sigma = 0.41$ the relation of the posterior model probabilities sharply reverses.[27]

The result of this replication analysis is difficult to interpret. Strictly speaking, if we put a high degree of belief into the skeptic model, our data do not support the null, but rather the FHZ11 model. However, in contrast to Fehr et al. (2011) we did not find a significant effect on shading rates between the baseline and the no-competition treatment, using either a Wilcoxon rank sum test or regression analysis. On first sight, these results seem at odds. We reconcile the findings as follows: The treatment effect that we measure is too small to generate a significant test result, yet it is too large to confirm an idealized null hypothesis. Nevertheless, the interpretation depends on the subjective requirements of what constitutes a skeptic belief. If one is more lenient and considers a null model featuring β_1-prior with $\sigma = 1$ as appropriately skeptic, the results suggest replication failure. Although we prefer high precision priors as idealized beliefs for the skeptic model, the fact that we see a reversal of posterior model probabilities also with the bootstrapped data suggests to stay very cautious about replication success.

References

Babcock, L., Loewenstein, G., 1997. Explaining bargaining impasse: the role of self-serving biases. J. Econ. Perspect. 11 (1), 109–126.
Babcock, L., Loewenstein, G., Issacharoff, S., 1997. Creating convergence: debiasing biased litigants. Law Soc. Inq. 22 (4), 913–925.
Babcock, L., Loewenstein, G., Issacharoff, S., Camerer, C., 1995. Biased judgments of fairness in bargaining. Am. Econ. Rev. 85 (5), 1337–1343.
Bem, D.J., Utts, J., Johnson, W.O., 2011. Must psychologists change the way they analyze their data? J. Person. Soc. Psychol. 101 (4), 716–719.
Bolton, G.E., Ockenfels, A., 2000. ERC: a theory of equity, reciprocity, and competition. Am. Econ. Rev. 90 (1), 166–193.
Brandts, J., Ellman, M., Charness, G., 2016. Let's talk: how communication affects contract design. J. Eur. Econ. Assoc. 14 (4), 943–974.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Q. J. Econ. 117 (3), 817–869.
Cialdini, R.B., 2007. Influence: The Psychology of Persuasion, 1st ed. Collins Business, New York.

[24] We are indebted to Alexander Streubel for lending us a computer that was acutally powerful enough to do the analysis with the bootstrapped data.

[25] In fact, the FHZ11 model also is clearly much more credible than the null model. The log-scaled Bayes factor in support of the FHZ11 model over the null model is 353.014. Given our bootstrapped data the FHZ15 model, however, is so overwhelming that there is no visual difference between the null and the FHZ11 model.

[26] The posterior model probabilities remain robust, also at standard deviation values of 10, 50, and 100.

[27] The posterior model probabilities stay at these levels at standard deviation values of 10, 50, and 100.

Denwood, M.J., 2016. runjags: an R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. J. Stat. Softw. 71 (9), 1–25.
Dienes, Z., 2011. Bayesian versus orthodox statistics: which side are you on? Perspect. Psychol. Sci. 6 (3), 274–290.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games Econ. Behav. 47 (2), 268–298.
Erlei, M., Reinhold, C., 2016. Contracts as reference points – the role of reciprocity effects and signaling effects. J. Econ. Behav. Organ. 127, 133–145.
Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games Econ. Behav. 54 (2), 293–315.
Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. G*power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39 (2), 175–191.
Fehr, E., Hart, O., Zehnder, C., 2011. Contracts as reference points – experimental evidence. Am. Econ. Rev. 101 (2), 493–525.
Fehr, E., Hart, O., Zehnder, C., 2015. How do informal agreements and revision shape contractual reference points? J. Eur. Econ. Assoc. 13 (1), 120–121.
Fehr, E., Hart, O., Zehnder, C., 2019. Contracts, conflicts and communication, Technical Report, Working Paper.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. 114 (3), 817–868.
Fehr, E., Zehnder, C., Hart, O., 2009. Contracts, reference points, and competition-behavioral effects of the fundamental transformation. J. Eur. Econ. Assoc. 7 (2–3), 561–572.
Ferrari, S.L.P., Cribari-Neto, F., 2004. Beta regression for modeling rates and proportions. J. Appl. Stat. 31 (7), 799–815.
Fischbacher, U., 2007. z-tree: Zurich toolbox for ready-made economic experiments. Exp. Econ. 10 (2), 171–178.
Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.-S., 2008. A weakly informative default prior distribution for logistic and other regression models. Ann. Appl. Stat. 2 (4), 1360–1383.
Greiner, B., 2015. Subject pool recruitment procedures: organizing experiments with orsee. J. Econ. Sci. Assoc. 1 (1), 114–125.
Gronau, Q.F., Singmann, H., Wagenmakers, E.-J., 2020. bridgesampling: an R package for estimating normalizing constants. J. Stat. Softw. 92 (10), 1–29.
Hart, O., Moore, J., 2008. Contracts as reference points. Q. J. Econ. 123 (1), 1–48.
Hippel, S., Hoeppner, S., 2019. Biased judgements of fairness in bargaining: a replication in the laboratory. Int. Rev. Law Econ. 58, 63–74.
Meng, X.-L., Schilling, S., 2002. Warp bridge sampling. J. Comput. Graph. Stat. 11, 552–586.
Meng, X.-L., Wong, W.H., 1996. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Stat. Sin. 6 (4), 831–860.
Plummer, M., 2003. JAGS: a program for analysis of Bayesian graphical models using gibbs sampling. In: Hornik, K., Leisch, F., Zeileis, A. (Eds.), Proceedings of the 3rd International Workshop on Distributed Statistical Computing. Technical University Vienna, Vienna.
R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rabin, M., 1993. Incorporating fairness into game theory and economics. Am. Econ. Rev. 83 (5), 1281–1302.
Raftery, A.E., 1995. Bayesian model selection in social research. Sociol. Methodol. 25, 111–163.
Verhagen, J., Wagenmakers, E.-J., 2014. Bayesian tests to quantify the result of a replication attempt. J. Exp. Psychol.: Gen. 143 (4), 1457–1475.