Effects of different ways of incentivizing price forecasts on market dynamics and individual decisions in asset market experiments

Nobuyuki Hanaki, Eizo Akiyama, Ryuichiro Ishikawa

Effects of different ways of incentivizing price forecasts on market dynamics and individual decisions in asset market experiments*

Nobuyuki Hanaki† Eizo Akiyama‡ Ryuichiro Ishikawa§

January 29, 2018

Abstract

In this study, we investigate (a) whether eliciting future price forecasts influences market outcomes and (b) whether differences in the way in which subjects are incentivized to submit "accurate" price forecasts influence market outcomes as well as the forecasts in an experimental asset market. We consider four treatments: one without forecast elicitation and three with forecast elicitation. In two of the treatments with forecast elicitation, subjects are paid based on their performance in both forecasting and trading, while in the other treatment with forecast elicitations, they are paid based on only one of those factors, which is chosen randomly at the end of the experiment. We found no significant effect of forecast elicitation on market outcomes in the latter case. Thus, to avoid influencing the behavior of subjects and market outcomes by eliciting price forecasts, paying subjects based on either forecasting or trading performance chosen randomly at the end of the experiment is better than paying them based on both. In addition, we consider forecast-only experiments: one in which subjects are rewarded based on the number of accurate forecasts and the other in which they are rewarded based on a quadratic scoring rule. We found no significant difference in terms of forecasting performance between the two.

Keywords: Price forecast elicitation, Experimental asset markets

JEL Code: C90, D84

*We thank anonymous referees and an associate editor for constructive comments. Makoto Soga provided invaluable help in organizing the experiments. This project is partly financed by L'Institute Universitaire de France, the ANR ORA-Plus project "BEAM" (ANR-15-ORAR-0004), and JSPS KAKENHI Grant Numbers 26285043, 26350415, 26245026, and 26590026. It also benefited from financial support from the French government managed by l'Agence Nationale de la Recherche under Investissements d'Avenir UCA^{JEDI} (ANR-15-IDEX-01). In particular, we thank the UCAinACTION project. We also thank Edanz Group (www.edanzediting.com) for editing a draft of this manuscript. The experiments reported in this paper have been approved by the Institutional Review Board of the Faculty of Engineering, Information and Systems, University of Tsukuba (No. 2012R25).

†Université Côte d'Azur, CNRS, GREDEG. Corresponding author. GREDEG, 250 rue Albert Einstein, 06560 Valbonne, FRANCE. E-mail: Nobuyuki.HANAKI@unice.fr

‡Faculty of Engineering, information and Systems, University of Tsukuba. E-mail: eizo@sk.tsukuba.ac.jp

§School of International Liberal Studies, Waseda University. E-mail: r.ishikawa@waseda.jp

1 Introduction

Expectations play a central role in modern macroeconomics and financial economics. Further, various policies, both monetary and fiscal, have increasingly come to be viewed as influencing the expectations of people, in particular market participants (Honkapohja, 2015; Mertens and Ravn, 2014). However, there is still considerable room for research into the dynamics of expectation formations to better understand how such policies influence the expectations and behavior of market participants.

In studying the dynamics of expectation formations, laboratory experiments provide a powerful tool. Unlike field data, whereby researchers rarely have full knowledge of the information on which people base their expectations, laboratory experiments enable researchers to determine what kind of information subjects have access to. For example, recent studies on "learning-to-forecast experiments" (Hommes et al., 2005; Heemeijer et al., 2009; Sonnemans and Tuinstra, 2010; Bao et al., 2012; Bao and Hommes, 2014) have been very successful in demonstrating the kinds of market environments under which the price expectations of subjects quickly converge to the rational expectations equilibrium (REE). These experimental studies also demonstrate the complex dynamics of expectations that do not converge to the REE and contribute to the construction of new type of models of expectation formation/dynamics by providing researchers with data against which proposed models can be tested (see, e.g., Anufriev and Hommes, 2012; Anufriev et al., 2015).[1]

The importance of studying expectation dynamics in other experimental paradigms is increasingly acknowledged among researchers. For example, in a recent survey of a large body of experimental literature that emerged after the seminal study by Smith et al. (1988), Palan (2013) discusses the scarcity of existing studies, as well as the need for more research, investigating the dynamics of expectations regarding future prices in multi-period asset markets.[2]

A potential obstacle to future developments in this direction is that as yet, there is no consensus regarding the most appropriate methodology regarding forecast elicitations. For example, it is not yet understood how eliciting forecasts influences market outcomes such as the degree of mispricing or trade volumes.[3] One recent study investigating this issue is Bao et al. (2013) using the framework

[1] Beshears et al. (2013) conduct forecasting-only experiment to study whether subjects correctly perceive the dynamics of a mean reverting time series. They show that if the mean-reverting process is fast, most of the subjects recognize it. But if it is slow, none of the subjects does.

[2] See also Powell and Shestakova (2016) and Nuzzo and Morone (2017) for a more recent survey of this rapidly growing body of literature.

[3] This lack of methodological consensus was evident in a discussion held among prominent scholars in the field,

of the "learning-to-forecast experiment." These authors examine cobweb markets in which subjects undertake (1) only forecasting (with computers implementing optimal trading behavior based on the submitted forecasts), (2) only trading (i.e., there is no elicitation of price forecasts), or (3) both forecasting and trading. Bao et al. (2013) find that market prices converge to the REE, but at significantly different speeds. The convergence is fastest when subjects only undertake forecasting and slowest when subjects undertake both forecasting and trading, suggesting that when subjects need to engage in multiple tasks, they take longer to learn "optimal" trading behavior.

Furthermore, researchers use different means of incentivizing subjects for their performance in forecasting and trading when they engage in both activities. In some studies, subjects are rewarded for both their forecasting and trading performances. Typically, in these studies, forecasting performance is rewarded by a bonus payment in addition to the reward from trading performance (Haruvy et al., 2007; Akiyama et al., 2014, 2017; Bosch-Rosa et al., 2017).[4] In other studies, subjects are rewarded based on their performance in relation to one of the two activities, but not both, and the activity that is used for this purpose is randomly determined at the end of the experiment (Bao et al., 2015). In comparing the two incentive schemes, researchers often discuss the possibility of subjects hedging between forecasting and trading when they are incentivized in relation to both activities (see, for example, Bao et al., 2015). While the observed trading behaviors may well differ between the two incentive schemes discussed above, to the best of our knowledge, there is no systematic experimental study investigating precisely how this difference in the way in which subjects are rewarded influences their behavior and forecasts.

This study aims to fill this gap in the literature. Using the experimental asset market paradigm pioneered by Smith et al. (1988), we investigate (a) whether (and how) eliciting future price forecasts influences market outcomes and (b) whether (and how) differences in the way in which subjects are incentivized in relation to their forecasting and trading performance, i.e., whether they are rewarded for either forecasting or trading (randomly determined at the end of the experiment), or for both, influence market outcomes, as well as the forecasts they submit. Further, to investigate (c) how two different ways of measuring forecast performance, a quadratic scoring function and a step function,

such as Vernon Smith, Charles Noussair, and Bruno Biais, during the annual meeting of the Society for Experimental Finance held in Nijmegen, in the Netherlands in June 2015. Vernon Smith noted that in analyzing the data from the initial experiments related to Smith et al. (1988) with one-period-ahead forecast elicitation, they felt that eliciting a forecast was influencing the trading behavior of subjects, and thus market outcomes, and therefore ceased to elicit forecasts in the subsequent sessions.

[4]Marimon et al. (1993) was the first study to offer a small bonus for the subject with the best forecast in addition to the rewards provided for the main experimental task.

and (d) how two different levels of reward for "accurate" forecasts impact subjects' forecasting performance, we conduct additional forecasting-only experiments in which subjects do not trade at all, but merely forecast the prices observed in our treatment without forecast elicitation.

We find that eliciting forecasts can significantly increase the magnitude of mispricing when subjects are rewarded for both trading and forecasting performance. If they are only rewarded for either forecasting or trading performance, chosen randomly at the end of the experiment, the market outcomes are not significantly different from those in the treatment without forecast elicitation. Thus, to avoid influencing the behavior of subjects and market outcomes by eliciting price forecasts compared with the benchmark case without forecast elicitation, subjects should be rewarded based on either forecasting or trading performance (to be determined randomly at the end of the experiment) rather than on both forecasting and trading, as has often been done in previous studies. In our forecasting-only experiments, we did not observe any statistically significant differences in terms of subjects' forecasting performance between the two scoring methods, nor between the two reward levels. We also observed poorer forecasting performance in the second and third rounds of the forecasting-only experiments compared with the experiments in which subjects undertook both trading and forecasting.

The rest of the paper is organized as follows. Section 2 presents the experimental design. The results are reported in Section 3. Section 4 presents the design of the additional forecasting-only experiments and the results. Section 5 presents a summary and concluding remarks.

2 Experiment

We consider asset market experiments a la Smith et al. (1988) with and without forecast elicitation. In total, we consider four treatments: one without forecast elicitation, T (trade-only treatment), and three with forecast elicitation. The three treatments with forecast elicitation, which we label BONUS, ForT (forecast or trade), and F&T (forecast and trade), differ in terms of the way in which subjects are incentivized.[5]

In the BONUS treatment, subjects receive a bonus payment for the accuracy of their forecasts, calculated as a fraction of the payment they receive based on their trading performance. The amount of the bonus is relatively small compared with the payment based on the trading performance. This

[5]Please refer to Appendix for English translation of the script used for the instruction videos. We have distributed handouts based on the instruction videos that are shown to our subjects as well.

type of incentive scheme has been employed by Haruvy et al. (2007), Akiyama et al. (2014, 2017), and Bosch-Rosa et al. (2017).

In the ForT treatment, subjects are paid based on *either* the accuracy of their forecasts or their trading performance, but not both. Which one is used for payment is randomly determined at the end of experiment. To the best of our knowledge, this type of incentive scheme has not been employed in the experimental framework a la Smith et al. (1988), but has been employed in other asset market experiments called "learning-to-forecast and learning-to-optimize" experiments by Bao et al. (2015).

Finally, in the F&T treatment, subjects are paid based on *both* the accuracy of their forecasts and their trading performance. To set the expected rewards from forecasting and trading performance at the same level as that used in the ForT treatment, the exchange rate between the experimental currency unit (ECU) and the real currency (Japanese yen) was set at 50% of that used in the ForT treatment.

Next, we describe the experimental market, which was the same for all four treatments. Then, we illustrate the way in which forecasts were elicited and how subjects were incentivized in the three treatments with forecast elicitation.

2.1 Markets

In all the treatments, groups of six subjects trade an asset with a life of 10 periods. Each trader receives four units of the asset and 520 ECUs as their initial endowment. Each unit of the asset pays a dividend of 12 ECUs at the end of each period, and after the final dividend payment at the end of period 10, the asset loses its value. Thus, the fundamental value (FV) of the asset at the beginning of period t, FV_t, is $12(11-t)$ ECUs. We employ a call market structure for trading, as in van Boening et al. (1993), Haruvy et al. (2007), Akiyama et al. (2014, 2017), and Bosch-Rosa et al. (2017). In call markets, unlike in continuous double auctions, there is one market clearing price for the asset in each period. Having only one price per period is an advantage for experiments with future price forecasts because the forecasted future prices are very clearly defined.[6]

In our call market experiment, subjects can submit a buy order and a sell order in each period by separately specifying a price and a quantity for each type of order. Therefore, if subject i decides to submit a buy order in period t, he or she has to specify the maximum price at which he or she

[6]Our experimental setup is based on that used by Akiyama et al. (2014, 2017).

is willing to buy a unit of asset (b_t^i for bid), and the maximum number of units he or she is willing to buy (d_t^i) in that period. Similarly, to submit a sell order in a period, subject i has to specify the minimum price at which he or she is willing to sell a unit of asset (a_t^i for ask), and the maximum number of units he or she is willing to sell (s_t^i) in that period. Of course, subjects can decide not to submit either a buy or a sell order by setting the quantities in both types of order to zero. We impose three constraints on the orders subjects can submit: an admissible price range, a budget constraint, and a relationship between b_t^i and a_t^i in the case where the subject submits both buy and sell orders. The admissible price range is set such that when $d_t^i \geq 1$ ($s_t^i \geq 1$), b_t^i (a_t^i) must be an integer between 1 and 2000, i.e., $b_t^i \in \{1, 2, ..., 2000\}$ ($a_t^i \in \{1, 2, ..., 2000\}$). The budget constraint simply means that neither borrowing of cash nor short selling is allowed.[7] The final constraint is such that when a trader submits both a buy order and a sell order, the ask must be no less than the bid, i.e., $a_t^i \geq b_t^i$. We imposed a 60-second nonbinding time limit for submitting orders. When the time limit is reached, the subjects are told via a flashing message in the upper right corner of their screen to submit their orders as soon as possible.

Once all of the traders have submitted their orders, the price that clears the market is calculated,[8] and all transactions are processed at that price, p_t, among traders who submitted buy orders such that $b_t \geq p_t$ and sell orders such that $a_t \leq p_t$.[9]

2.2 Forecast elicitations

In addition to trading assets in the call market, in the BONUS, F&T, and ForT treatments, at the beginning of each period (i.e., before submitting their orders), subjects were asked to forecast the prices in each of the remaining periods. This allowed us to study the evolution of long-run as well as short-run price forecasts of subjects. This elicitation method was first introduced by Haruvy

[7]Thus, the budget constraint implies (i) $d_t^i \times b_t^i \leq$ cash holdings, and (ii) $s_t^i \leq$ units of asset on hand, at the beginning of period t.

[8]When there are several such prices, the lowest one is chosen as the market clearing price. The choice of the lowest price is made using the same principle used in the design of Haruvy et al. (2007). We considered using the mid-price in the case of multiple prices because it is more realistic in the sense that it is used in several markets, such as the Tokyo Stock Exchange or the NASDAQ. However, we chose not to employ the mid-price, partly to avoid deviating from the approach used in previous studies (Haruvy et al., 2007; Akiyama et al., 2014, 2017), and also because our algorithm treats $a_t^i = 2000$ in the case of $s_t^i = 0$, and $b_t^i = 1$ in the case of $d_t^i = 0$. As a result, when all of the traders submit $s_t^i = d_t^i = 0$, the market clearing price is going to be 1000. Given that the maximum FV in our experiment was 120, we considered that this was too high, and opted for returning the lowest price, i.e., 1 in this case. This can bias prices downward, but because this feature is the same in all of the treatments we consider, it should not influence the effect of the treatment that we are investigating. In our data consisting of 1440 periods (30 periods × 12 markets × 4 treatments), however, there was no period in which $\sum_i s_t^i = \sum_i d_t^i = 0$. There were 7 periods with $\sum_i d_t^i = 0$ (but with $\sum_i s_t^i > 0$), and 4 periods with $\sum_i s_t^i = 0$ (but with $\sum_i d_t^i > 0$).

[9]Any ties among the last accepted buy or sell orders are resolved randomly. It is possible that no transaction will take place given the computed market clearing price.

et al. (2007) and has since been used in other studies (Akiyama et al., 2014, 2017; Bosch-Rosa et al., 2017). In period t, each subject submitted $11 - t$ forecasts, thus subjects submitted a total of 55 price forecasts over the 10 periods. We imposed a 120-second nonbinding time limit for submitting price forecasts. When the time limit was reached, the subjects were told, via a message flashing in the upper right corner of their screen, to submit their forecasts as soon as possible.

2.2.1 BONUS treatment

In the BONUS treatment, subjects were told that they would receive the following bonus payment based on the accuracy of their forecasts:

$$\begin{aligned}\text{Bonus (in ECUs)} = 0.5\% \times \text{number of ``accurate'' forecasts} \\ \times \text{final cash holding in period 10.}\end{aligned}$$

An "accurate" forecast was defined as a forecast that was within 10% of the realized price.[10] Therefore, if all 55 forecasts were "accurate," the subject would receive 27.5% of his or her final cash holding (resulting from trading and dividends) as a bonus payment. This is the incentive scheme used in Akiyama et al. (2014, 2017).

2.2.2 ForT treatment

In the ForT treatment, subjects were told that they would be rewarded based *either* on their trading performance (i.e., based on their final cash holding resulting from trading and dividends) *or* on the number of "accurate" forecasts they provided, with the method used for payment to be determined randomly at the end of the experiment.

Subjects were informed that in cases where their reward was based on the accuracy of their forecasts, they would be paid according to the following formula:

$$\text{Payment based on forecast performance (in ECUs)} = 40 \times \text{number of ``accurate'' forecasts.}$$

An "accurate" forecast was defined in the same way as in the BONUS treatment, that is, a forecast that was within 10% of the realized price. Thus, if all 55 of a subject's forecasts proved to be

[10] Let $f^i_{t,p}$ be the forecast for period p price submitted by subject i in the beginning of period t. $f^i_{t,p}$ is considered as an accurate forecast if $0.9P_p \leq f^i_{t,p} \leq 1.1P_p$ where P_p is the realized price in period p.

"accurate," the subject would receive 2200 ECUs.

A reward of 40 ECUs per "accurate" forecast in the ForT treatment may seem generous given that the value of the initial endowment (four units of asset and 520 ECUs) is 1000 ECUs. However, we chose this value based on the average number of "accurate" forecasts by subjects reported in Akiyama et al. (2014). As noted above, Akiyama et al. (2014) used the same experimental setting as that used in our BONUS treatment, and also repeated the same market condition three times using the same group of subjects, as we have done in this study. The average number (and standard deviation) of "accurate" forecasts by the 168 subjects was 11.7 (11.08), 25.71 (15.99), and 39.02 (16.47) for the first, second, and third market rounds, respectively. Thus, the average number of "accurate" forecasts per subject per 10-period market round was approximately 25. By setting the reward for an "accurate" forecast at 40 ECUs, we tried to equate the expected reward from trading performance (i.e., the final cash holding of 1000 ECUs) with that from forecasting performance (i.e., the number of "accurate" forecasts). However, we did not explain our reasoning to the subjects.

2.2.3 F&T treatment

In the F&T treatment, subjects were told that they would be rewarded based on *both* their trading performance (that is, based on their final cash holding resulting from trading and dividends) *and* on the number of "accurate" forecasts they provided. The forecast performance was measured in exactly the same way as in the ForT treatment. However, to set the expected payment in the F&T treatment at the same level as that in the two remaining treatments with forecast elicitation, the exchange rate between ECUs and JPY was halved in the F&T treatment compared with that used in the other treatments.

In all four treatments, the same group of traders, with identical initial endowments of cash and assets, participated in the same 10-period market three times. Each 10-period market is termed a "round." Thus, the experiment consisted of three rounds of a 10-period market with identical initial endowments and an identical group of subjects.[11] In the ForT treatment, the type of performance on which the reward for each round was based (i.e., either forecasting performance or trading performance) was determined independently (i.e., independent across rounds but the same across subjects

[11] Before commencing Round 1, there was a practice period to allow subjects to familiarize themselves with the user interface. Subjects were given their initial endowment of cash and assets, and asked to enter their price forecasts (in the treatments with forecast elicitation) and their orders for period 1. Information regarding the resulting market clearing price was not provided to the subjects.

in a session) at the end of the experiment. In addition to a participation fee of 600 JPY, subjects were paid based on an exchange rate of 1 ECU = 1 JPY for the BONUS and ForT treatments and 1 ECU = 0.5 JPY for the F&T treatment.

3 Results

The experiment was conducted at the University of Tsukuba in Japan between July 2015 and July 2017.[12] A total of 288 subjects (three sessions involving 24 subjects for each of the four treatments) participated in the experiment. Thus, there were 12 markets, each involving six subjects, for each of the four treatments. The subjects had never participated in a similar experiment before, and each subject only participated in one experimental session.

The experiment lasted for two to three hours (longer with forecast elicitation), including the provision of instructions and the post-experiment questionnaire. In addition to their participation fee, subjects were paid an average of 3000 JPY in the T (trade only) treatment, 3400 JPY in the BONUS treatment, 3230 JPY in the ForT treatment, and 2670 JPY in the F&T treatment.

Next, we summarize the market-level outcomes, prices, and volumes relating to each of the four treatments. We then compare the orders submitted by subjects in the four treatments and their price forecasts in the BONUS, ForT, and F&T treatments.

3.1 Market outcomes

3.1.1 Prices

Figure 1 shows the price dynamics over 10 periods for three rounds of the four treatments. The results for the four treatments are shown in separate rows: T (top row), BONUS (2nd row), ForT (3rd row), and F&T (4th row).

It can be seen that there is a greater tendency toward overpricing in the F&T and BONUS treatments than in the T and ForT treatments. To better compare the magnitude of the mispricing observed in the four treatments, we compute the relative absolute deviation (RAD) and relative deviation (RD) proposed by Stöckl et al. (2010). For each market m, the RAD^m and RD^m are

[12] The experiment was computerized using z-Tree Fischbacher (2007).

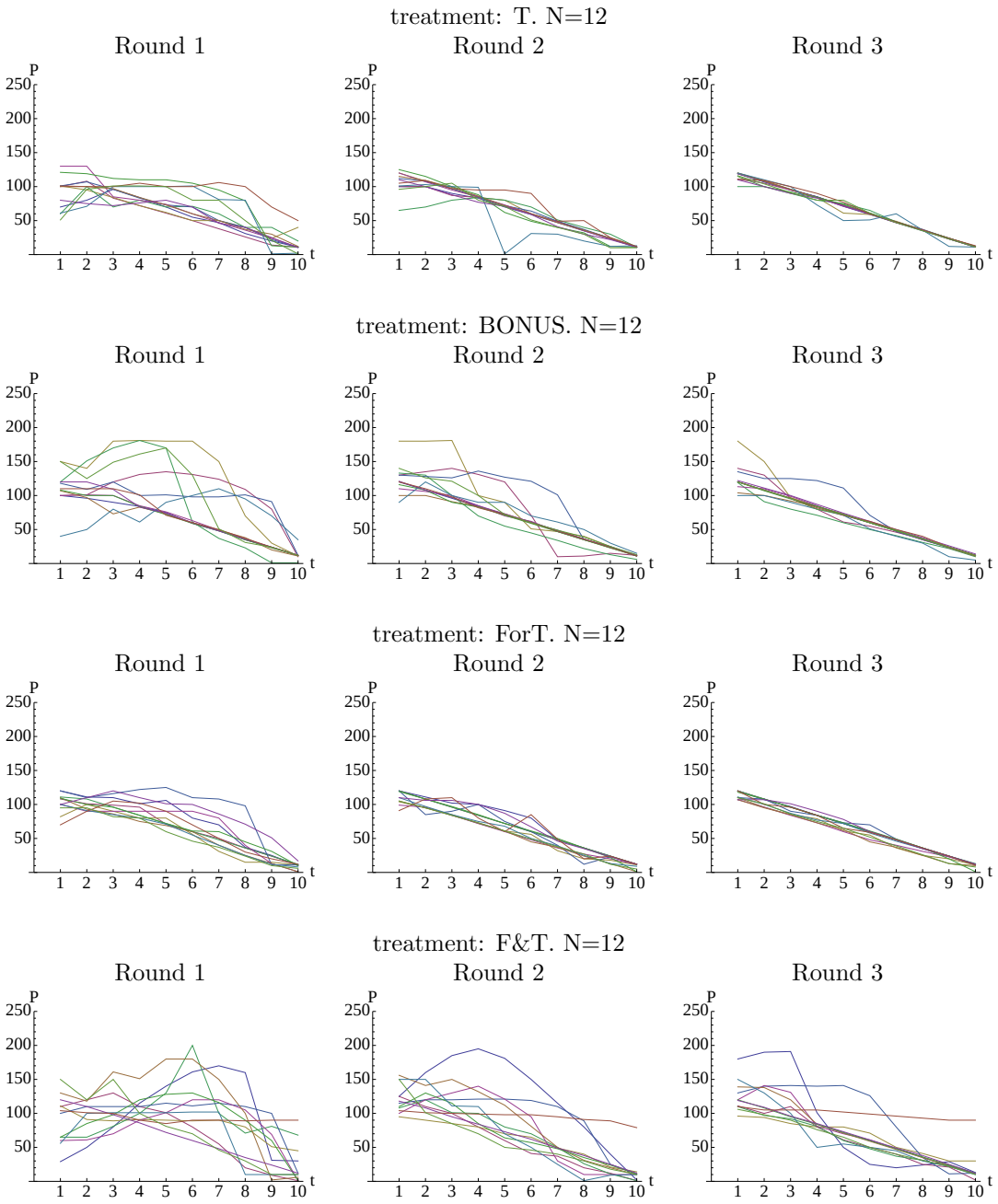
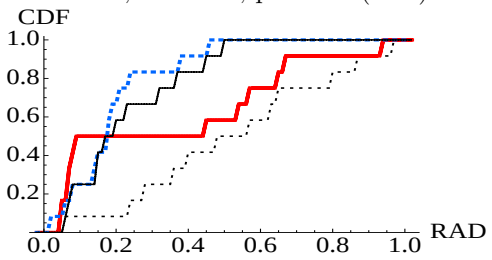


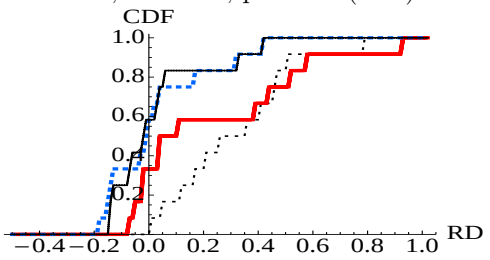
Figure 1: Price dynamics for three rounds of the four treatments: T (1st row), BONUS (2nd row), ForT (3rd row), and F&T (4th row).

RAD, Round 1, p=0.030 (KW)



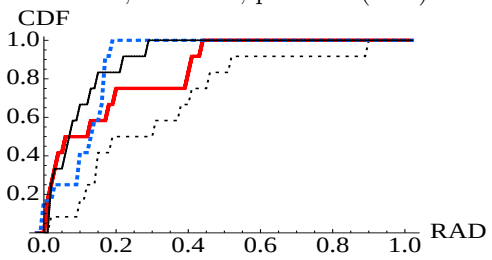
	Mean	Std. Dev	Min	Max
T	0.226	0.148	0.059	0.494
BONUS	0.346	0.318	0.041	0.930
ForT	0.186	0.125	0.014	0.453
F&T	0.517	0.282	0.011	0.964

RD, Round 1, p=0.003 (KW)



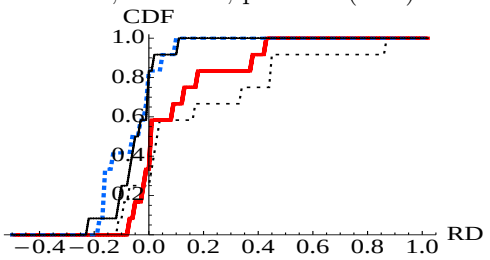
	Mean	Std. Dev	Min	Max
T	0.021	0.177	-0.147	0.412
BONUS	0.238	0.323	-0.074	0.927
ForT	0.023	0.188	-0.189	0.411
F&T	0.314	0.224	0.008	0.785

RAD, Round 2, p=0.059 (KW)



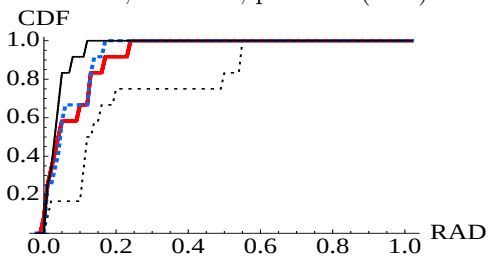
	Mean	Std. Dev.	Min	Max
T	0.096	0.086	0.106	0.288
BONUS	0.156	0.166	0.002	0.430
ForT	0.109	0.066	0	0.185
F&T	0.305	0.245	0.015	0.898

RD, Round 2, p=0.009 (KW)



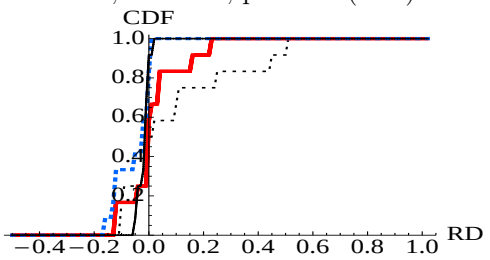
	Mean	Std. Dev.	Min	Max
T	-0.048	0.082	-0.227	0.102
BONUS	0.085	0.164	-0.079	0.424
ForT	-0.065	0.095	-0.185	0.097
F&T	0.172	0.296	-0.112	0.868

RAD, Round 3, p=0.036 (KW)



	Mean	Std. Dev.	Min	Max
T	0.035	0.0308	0.005	0.111
BONUS	0.073	0.0748	0	0.230
ForT	0.063	0.0575	0.002	0.162
F&T	0.211	0.1979	0.002	0.547

RD, Round 3, p=0.096 (KW)



	Mean	Std. Dev.	Min	Max
T	-0.018	0.020	-0.056	0.012
BONUS	0.012	0.099	-0.129	0.221
ForT	-0.055	0.063	-0.162	0.003
F&T	0.089	0.208	-0.109	0.508

Figure 2: Distribution of RAD^m (left) and RD^m (right) in T (thin solid black), BONUS (thick red), ForT (thick dashed blue), and F&T (thin dashed black) treatments over three rounds. Summary statistics and p-values relating to the Kruskal–Wallis (KW) test for multiple comparisons are also presented.

defined as

$$RAD^m = \frac{1}{T}\sum_{p=1}^{T}\frac{|P_p^m - FV_p|}{|\overline{FV}|} \tag{1}$$

$$RD^m = \frac{1}{T}\sum_{p=1}^{T}\frac{P_p^m - FV_p}{|\overline{FV}|}, \tag{2}$$

where $T = 10$ and P_p^m is the realized price in period $p \in \{1, 2, ..., 10\}$ in market m. FV_p is the FV of the asset in period p. $|\overline{FV}| = |\frac{1}{T}\sum_{p=1}^{T} FV_p|$.

Figure 2 shows the empirical cumulative distribution function (CDF) of the RAD^m (top) and RD^m (bottom) for the T (thin solid black), BONUS (thick red), ForT (thick dashed blue), and F&T (thin dashed black) treatments. It also shows the summary statistics (mean, standard deviation, minimum, and maximum) as well as p-values from the Kruskal–Wallis (KW) tests for multiple comparisons.

The top panel of Figure 2 shows that in Round 1, the *RAD* and *RD* are greatest for the F&T treatment, followed by the BONUS treatment.[13] The *RAD* and *RD* for the T and ForT treatments are both smaller than those for the other two treatments, and the KW tests show that the differences are statistically significant. Statistically significant differences between the *RAD* and *RD* across the four treatments continue to be observed in Rounds 2 and 3.[14] However, a different pattern emerges in Rounds 2 and 3, where the *RAD* in the BONUS treatment moves much closer to that in the T and ForT treatments. In fact, the *RAD* becomes statistically significantly different in the BONUS and F&T treatments in Rounds 2 and 3, although the difference is only marginal in the former (p=0.093 and p=0.033, respectively, based on a two-sided two-sample permutation test, PT).

Observation 1 *Eliciting long-run price forecasts does not have a significant impact on mispricing compared with the market without forecast elicitation if subjects are rewarded based on either trading performance or forecasting performance chosen randomly at the end of the experiment, but not both.*

3.1.2 Trading volumes

We now turn our attention to trading volumes and compute the *turnover* (TO), which is the proportion of outstanding assets that are traded over 10 periods. For market m, turnover, TO^m, is

[13] The difference between the F&T and BONUS treatments is not statistically significant; p=0.172 for *RAD* and p=0.509 for *RD* based on a two-sided two-sample permutation test.

[14] The differences are only marginal in relation to the *RAD* in Round 2 and the *RD* in Round 3.

defined as follows:

$$TO^m = \sum_{p=1}^{10} \frac{Q_p^m}{24}, \tag{3}$$

where Q_p^m is the number of units of asset traded in period p in market m, and the total number of outstanding assets is 24 (recall that each of the six traders commenced with four units of the asset).

Figure 3 shows the CDF of TO^m across three rounds in the four treatments: T (thin solid black), BONUS (thick red), ForT (thick dashed blue), and F&T (thin dashed black). Summary statistics and p-values from the KW test are also reported. It can be seen from the p-values that TO is not statistically significantly different across the four treatments in any of the three rounds. Thus, we make the following observation.

Observation 2 *Eliciting long-run price forecasts does not have a significant effect on trading volumes compared with markets without forecast elicitation.*

3.2 Orders

We now analyze the orders submitted by subjects to investigate how individual trading behavior is influenced by different incentive schemes regarding the forecast elicitation. To summarize the orders submitted by subjects across 10 periods in a market, we employ a measure of the potential loss that can be generated by these orders. Akiyama et al. (2014) define the potential loss by subject i in market m, $PL^{i,m}$, as:

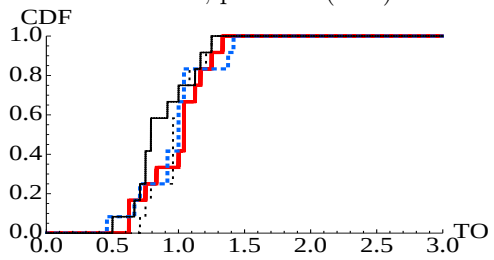
$$PL^{i,m} \equiv \frac{1}{1000} \sum_t \left(d_t^{i,m} \max(b_t^{i,m} - FV_t, 0) + s_t^{i,m} \max(FV_t - a_t^{i,m}, 0) \right), \tag{4}$$

where $b_t^{i,m}$ and $a_t^{i,m}$ are the maximum price at which i is willing to buy and the minimum price at which i is willing to sell a unit of asset, respectively, as specified in subject i's orders submitted in period t. $d_t^{i,m}$ and $s_t^{i,m}$ are the maximum quantities demanded and supplied in association with $b_t^{i,m}$ and $a_t^{i.m}$, respectively. The potential loss is normalized by the value of the initial endowment (1000) so that $PL^{i,m}$ denotes the share of the initial endowment that subject i would potentially lose if his or her orders were executed at the prices submitted.[15]

Based on the individually computed $PL^{i,m}$, we define the average potential loss of the traders

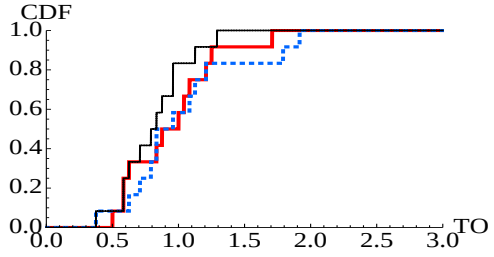
[15]It should be noted that submitting such orders may not result in any losses in our experiment because the actual trading prices can differ from those submitted by the subjects.

Round 1, p=0.558 (KW)



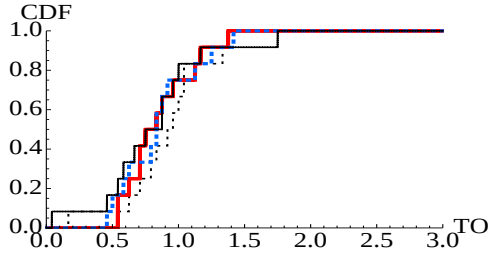
	Mean	Std. Dev	Min	Max
T	0.868	0.226	0.500	1.250
BONUS	0.986	0.232	0.625	1.333
ForT	0.962	0.270	0.458	1.417
F&T	0.972	0.166	0.708	1.250

Round 2, p=0.412 (KW)



	Mean	Std. Dev.	Min	Max
T	0.809	0.255	0.375	1.292
BONUS	0.941	0.351	0.500	1.708
ForT	1.021	0.452	0.375	1.917
F&T	1.007	0.265	0.667	1.542

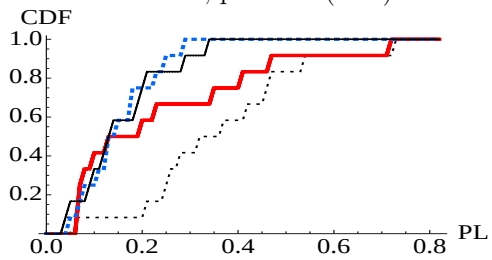
Round 3, p=0.746 (KW)



	Mean	Std. Dev.	Min	Max
T	0.806	0.420	0.042	1.750
BONUS	0.851	0.262	0.542	1.375
ForT	0.851	0.296	0.458	1.417
F&T	0.903	0.327	0.167	1.417

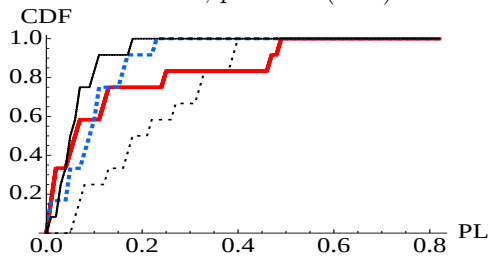
Figure 3: Distribution of turnover TO^m for the T (thin solid black), BONUS (thick red), ForT (thick dashed blue), and F&T (thin dashed black) treatments over three rounds. The p-values from the KW test are also presented.

Round 1, p=0.012 (KW)



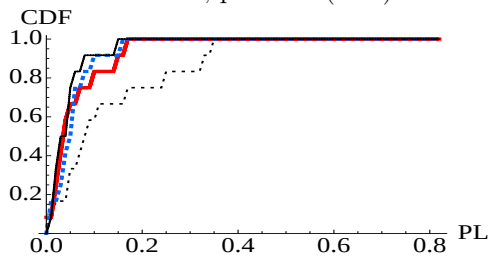
	Mean	Std. Dev	Min	Max
T	0.153	0.091	0.039	0.338
BONUS	0.237	0.208	0.063	0.719
ForT	0.150	0.074	0.048	0.285
F&T	0.356	0.177	0.034	0.722

Round 2, p=0.009 (KW)



	Mean	Std. Dev.	Min	Max
T	0.061	0.449	0.008	0.174
BONUS	0.138	0.171	0.004	0.480
ForT	0.093	0.065	0.001	0.222
F&T	0.215	0.122	0.057	0.397

Round 3, p=0.098 (KW)



	Mean	Std. Dev.	Min	Max
T	0.042	0.040	0.001	0.148
BONUS	0.056	0.054	0.000	0.166
ForT	0.053	0.042	0.002	0.160
F&T	0.127	0.117	0.001	0.341

Figure 4: Distribution of the within-market average potential loss, PL^m, in the T (thin solid black), BONUS (thick red), ForT (thick dashed blue), and F&T (thin dashed black) treatments over three rounds. Summary statistics and p-values from the KW test are also presented.

in each market, PL^m, and use these as independent observations.

Figure 4 shows the CDF of the PL^m across three rounds in the four treatments: T (thin solid black), BONUS (thick red), ForT (thick dashed blue), and F&T (thin dashed black)). We observe that in all three rounds the PL^m distribution from the F&T treatment lies to the right of those from the other three treatments. In Rounds 1 and 2, the PL^m from the BONUS treatment lies to the right of the two remaining treatments, ForT and T. In Round 3, the PL^m from the BONUS treatment is similar to those from the ForT and T treatments. In fact, the PL^m from the BONUS and F&T treatments in Round 3 are marginally statistically significantly different (p= 0.068, PT). This is very similar to what was observed in relation to RAD and RD. These results suggest that the greater mispricing that is observed in the F&T and BONUS treatments compared with the ForT and T treatments is the result of subjects in the former two treatments submitting orders that deviate further from the FV compared with the latter two treatments.

Subjects in the F&T treatment submit orders that deviate significantly more from the FV than do those in the BONUS treatment in Round 3. This difference between F&T and BONUS treatment can be due to the fact that the payoff from the forecasting task is more positively correlated with the payoff from trading task in BONUS treatment. This may make subjects to be more careful in submitting their orders because lower payoffs from trading result in lower payoff from forecasting.

To better understand the differences in the deviations of orders from the FV across treatments, we now look at buy orders and sell orders separately. We define the potential loss from buying and the potential loss from selling, $PBL^{i,m}$ and $PSL^{i,m}$, respectively, for subject i in market m as:

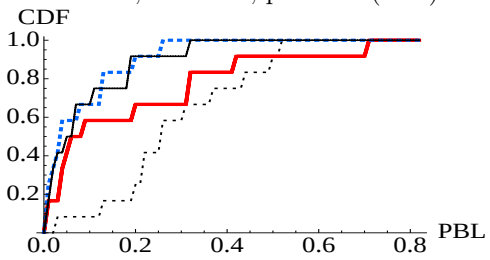
$$PBL^{i,m} \equiv \frac{1}{1000} \sum_t d_t^{i,m} \max(b_t^{i,m} - FV_t, 0) \tag{5}$$

$$PSL^{i,m} \equiv \frac{1}{1000} \sum_t s_t^{i,m} \max(FV_t - a_t^{i,m}, 0). \tag{6}$$

$PBL^{i,m}$ and $PSL^{i,m}$ are normalized by the value of the initial endowment to enable comparison with $PL^{i,m}$. Similar to $PL^{i,m}$, based on the individually computed $PBL^{i,m}$ and $PSL^{i,m}$, we define the average potential loss from buying and average potential loss from selling for traders in each market, PBL^m and PSL^m, respectively, and use them as independent observations.

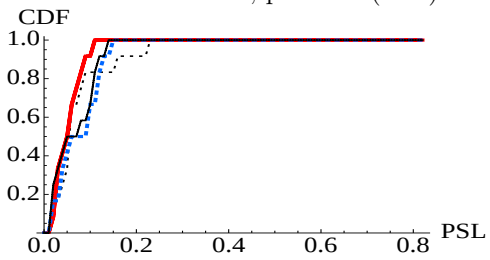
Figure 5 shows the CDF of the PBL^m (left) and PSL^m (right) across three rounds in the four treatments. The four treatments are represented by the same colored lines as used in Figure 4. It

PBL^m, Round 1, p=0.003 (KW)



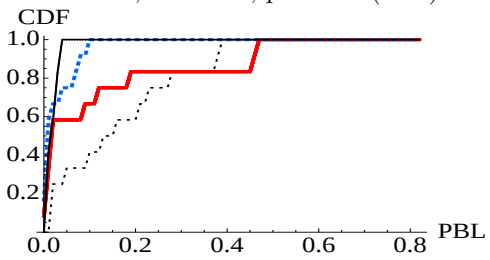
	Mean	Std. Dev	Min	Max
T	0.086	0.096	0.006	0.315
BONUS	0.185	0.218	0.004	0.708
ForT	0.075	0.082	0.004	0.252
F&T	0.283	0.148	0.021	0.515

PSL^m. Round 1, p=0.733 (KW)



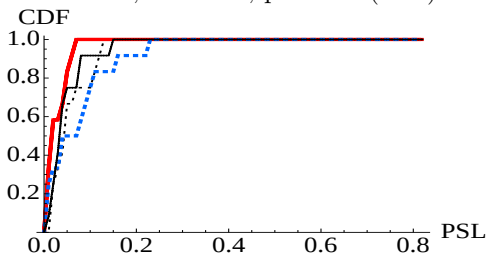
	Mean	Std. Dev	Min	Max
T	0.067	0.044	0.017	0.133
BONUS	0.052	0.029	0.011	0.109
ForT	0.075	0.046	0.014	0.141
F&T	0.073	0.061	0.013	0.226

PBL^m, Round 2, p=0.004 (KW)



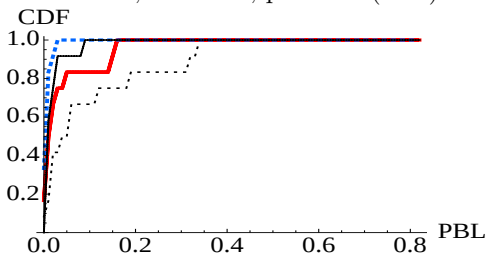
	Mean	Std. Dev.	Min	Max
T	0.017	0.013	0.000	0.039
BONUS	0.113	0.172	0.000	0.461
ForT	0.026	0.034	0.000	0.091
F&T	0.161	0.134	0.012	0.386

PSL^m, Round 2, p=0.257 (KW)



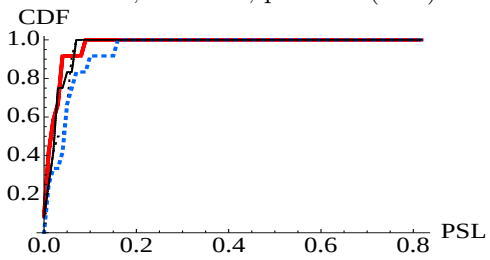
	Mean	Std. Dev.	Min	Max
T	0.045	0.038	0.003	0.144
BONUS	0.025	0.020	0.004	0.064
ForT	0.069	0.068	0.001	0.222
F&T	0.053	0.041	0.011	0.125

PBL^m, Round 3, p=0.020 (KW)



	Mean	Std. Dev.	Min	Max
T	0.016	0.024	0.001	0.086
BONUS	0.035	0.055	0.000	0.156
ForT	0.006	0.007	0.000	0.027
F&T	0.096	0.121	0.000	0.340

PSL^m, Round 3, p=0.309 (KW)



	Mean	Std. Dev.	Min	Max
T	0.026	0.021	0.000	0.065
BONUS	0.021	0.023	0.000	0.081
ForT	0.047	0.042	0.001	0.151
F&T	0.032	0.024	0.001	0.061

Figure 5: Distribution of the within-market average potential loss from buying, PBL^m (left), and average potential loss from selling, PSL^m (right), in the T (thin solid black), BONUS (thick red), ForT (thick dashed blue), and F&T (thin dashed black) treatments over three rounds. Summary statistics and the p-values from the KW test are also presented.

can be seen that there is a similar ordering of PBL^m to that found in relation to PL^m above, i.e., the PBL^m is greatest in the F&T treatment, followed by the BONUS, T, and ForT treatments. The KW test shows that the PBL^m is statistically significantly different across the four treatments. However, this is not the case in relation to the PSL^m. The CDFs for the PSL^m shown in the right-hand panels of Figure 5 are very close to each other, and the KW test indicates that the differences across the four treatments, if any, are not statistically significant. This absence of the significant difference in PSL^m across treatments is probably due to much smaller range of possible downward deviations of asks from FV (used in defining PSL^m) compared to the possible upward deviations of bids form FV (used in defining PBS^m). Thus, in later analyses wherein we investigate the relationship between forecasts and trading behavior across the three treatments with forecast elicitation, we only focus on buy orders.

3.3 Forecasting performance

We now look at the three treatments in which price forecasts were elicited, BONUS, ForT, and F&T, with a view to determining whether differences in the way in which subjects are incentivized to submit "accurate" forecasts influences their forecasting performance.

Table 1 shows the average numbers of "accurate" forecasts and their standard deviations over three rounds of the three treatments. The average numbers of "accurate" forecasts do not differ across the three treatments in Round 1 (p=0.702, KW test).[16] However, in Rounds 2 and 3, the average number of accurate forecasts in the F&T treatment is significantly smaller than that in the ForT and BONUS treatments.[17]

The reason for the poorer forecasting performance in the F&T treatment compared with the other two treatments is not only the differences in incentives across the treatments, but also the greater mispricing observed in the F&T treatment compared with the other two treatments. This can be seen from Table 2, which shows the results of regressing the number of "accurate" forecasts on treatment dummies (D_{BONUS} and $D_{F\&T}$, which take a value of 1 in the relevant treatment and 0 otherwise) and RAD. The estimated coefficient for $D_{F\&T}$ is negative in Rounds 2 and 3, but not statistically significant. Conversely, the estimated coefficient for RAD is negative and statistically

[16] We are using individual as an independent observation for statistical tests. The result does not change even if we use within-group average as an independent observation instead (p=0.809, KW test).

[17] If we use the within-group average as an independent observation, the difference is only marginally statistically significant in Round 3. p=0.103 in Round 2 and p=0.089 in Round 3 (KW test).

Table 1: The average number (standard deviation) of "accurate" forecasts in the three treatments with forecast elicitation.

Treatment	number of subjects	Number of "accurate" forecasts		
		Round 1	Round 2	Round 3
ForT	72	12.65	31.10	39.69
		(11.55)	(16.37)	(11.73)
BONUS	72	12.74	28.39	38.29
		(10.82)	(19.03)	(17.37)
F&T	72	11.53	18.92	27.38
		(10.85)	(15.18)	(16.51)
p-values (Kruskal–Wallis)		0.702	0.001	0.001

Table 2: Relationship between subjects' forecasting performance (number of "accurate" forecasts) and mispricing (RAD) in the market

	(1) Round 1	(2) Round 2	(3) Round 3
Constant	15.94***	34.60***	41.97***
	(2.310)	(4.153)	(3.033)
D_{BONUS}	2.917	-1.209	-1.045
	(2.758)	(5.552)	(4.815)
$D_{F\&T}$	4.748	-5.888	-7.022
	(2.896)	(6.542)	(4.803)
RAD	-17.70***	-32.17**	-35.89
	(4.238)	(13.90)	(25.19)
adj. R^2	0.145	0.170	0.175
N	216	216	216

Standard errors corrected for within group correlations in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

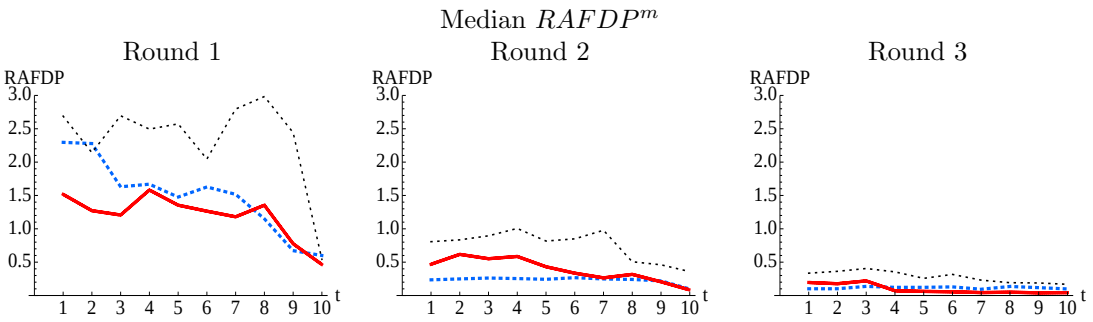
significant except for Round 3.

We further investigate the dynamics of forecasting performance by computing the deviations of the forecast prices from the realized prices. We define the relative absolute forecast deviation from prices (RAFDP) and the relative forecast deviation from prices (RFDP) as follows:

$$\text{RAFDP}_t^{i,m} = \frac{1}{T-t+1}\sum_{p=t}^{T}\frac{|f_{t,p}^{i,m} - P_p^m|}{P_p^m} \tag{7}$$

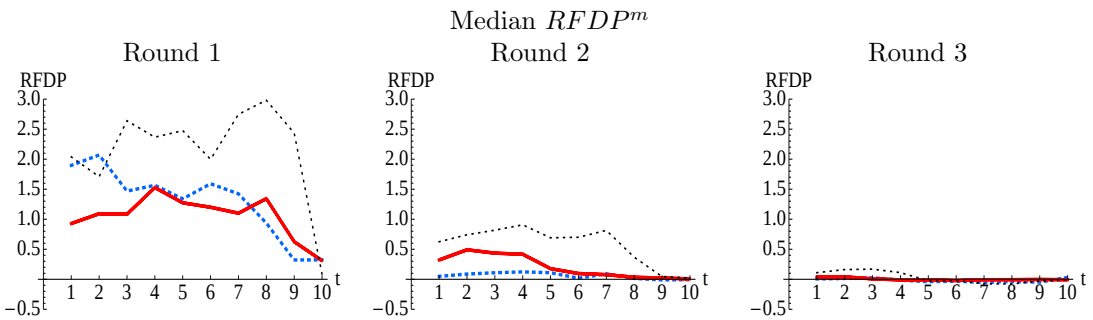
$$\text{RFDP}_t^{i,m} = \frac{1}{T-t+1}\sum_{p=t}^{T}\frac{f_{t,p}^{i,m} - P_p^m}{P_p^m}, \tag{8}$$

where P_p^m is the realized price in market m in period p. We take an average across traders in the same market in each period and use this as an independent observation.



p-values of KW test

t	1	2	3	4	5	6	7	8	9	10
Round 1	0.273	0.495	0.600	0.926	0.752	0.572	0.354	0.344	0.400	0.819
Round 2	0.070	0.148	0.092	0.120	0.191	0.151	0.203	0.209	0.109	**0.027**
Round 3	0.169	0.182	0.239	0.193	0.311	0.142	**0.026**	**0.022**	**0.016**	0.081



p-values of KW test

t	1	2	3	4	5	6	7	8	9	10
Round 1	0.362	0.461	0.627	0.918	0.758	0.580	0.383	0.304	0.489	0.954
Round 2	0.116	0.148	0.101	0.125	0.303	0.185	0.298	0.228	0.820	0.961
Round 3	0.264	0.399	0.526	0.215	0.789	0.900	0.486	0.191	0.655	0.346

Figure 6: Dynamics of the median $RAFDP_t^m$ (top) and $RFDP_t^m$ (bottom) in the BONUS (thick red), ForT treatment (thick dashed blue), and F&T (thin dashed black) treatments over three rounds. P-values from KW tests are also presented for each period.

Figure 6 shows the dynamics of the median $RAFDP_t^m$ and $RFDP_t^m$ for the BONUS (thick red), ForT treatment (thick dashed blue), and F&T (thin dashed black) treatments over three rounds. P-values from the KW test are also shown for each period. The top panel shows that the median $RAFDP_t^m$ in the F&T treatment is much larger than that in the BONUS and ForT treatments during Round 1. However, these differences are not statistically significant across the three treatments at the 5% significance level. This is in line with the absence of statistically significant differences in terms of the number of accurate forecasts among the three treatments in Round 1. The same is true for most of the periods during the three rounds, except for a few periods toward the end of Round 2 and Round 3, for which the p-values are shown in bold. Taking the direction of the deviations into account, as we do in $RFDP_t^m$, does not change the result. In fact, $RFDP_t^m$ is not statistically significantly different across the three treatments in any of the 30 periods over the three rounds. Thus, we make the following observation.

Observation 3 *The differences in the incentive schemes across the three treatments did not result in statistically significant differences in forecasting performance in terms of deviation of the forecast prices from the realized prices.*

3.4 Forecasts and orders

Why do subjects in the F&T and BONUS treatments tend to submit orders that deviate further from the FV compared with those in the ForT treatment? An intuitive reason is that subjects in the F&T and BONUS treatments take more trading risks because they have an additional source of payment compared with those in the ForT treatment. This implies that for a given level of price forecasts, those in the F&T and BONUS treatments submit orders at higher prices than those in the ForT treatment. Further, it is possible that subjects in the ForT treatment try to hedge between earnings from forecasting and trading to ensure their earnings by, for example, forecasting higher current period prices while submitting orders at lower prices, such that a negative correlation is observed between their forecasts and bids. To address this question, we investigate the relationship between subjects' forecasts and bids (buy orders). We focus on buy orders because of the significant across-treatment variations in buy orders noted above.

Because we elicit not only short-term forecasts but also long-term forecasts, we study the relationships between both types of forecasts and trading behavior. Note that in their analysis of the

relationship between forecasts and trading behavior based on data from Haruvy et al. (2007), Carle et al. (2017) report that while short-term forecasts (i.e., those of the price in the current period) have a statistically significant relationship with trading behavior in the current period, long-term forecasts (summarized as the relative deviation of forecasts from the FV for all of the remaining periods) do not. The way Carle et al. (2017) characterize long-term forecasts, however, fails to capture what type of price paths subjects are expecting. For example, subjects may expect prices to remain constant in the future, or to go up first and then go down. Such differences in dynamics will be hidden if one averages over these future forecasts as done by Carle et al. (2017). Thus, in our analyses, we characterize them differently by focusing on the first peak or the first valley that appears in subjects' long-term forecasts.

Let us consider a series of forecasts that subject i (in market m) submits at the beginning of period t. We define i's maximum long-term forecast in period t, $h_t^{i,m}$, and its timing, $\tau h_t^{i,m}$, as well as i's long-term minimum forecast in period t, $l_t^{i,m}$, and its timing, $\tau l_t^{i,m}$, as follows:

$$
\begin{aligned}
h_t^{i,m} &= \max_{k>t} f_{t,k}^{i,m} \\
\tau h_t^{i,m} &= \min \operatorname*{argmax}_{k>t} f_{t,k}^{i,m} \\
l_t^{i,m} &= \min_{k>t} f_{t,k}^{i,m} \\
\tau l_t^{i,m} &= \min \operatorname*{argmin}_{k>t} f_{t,k}^{i,m}.
\end{aligned}
$$

We then consider the following five types of forecast paths:

(1) Peak then valley: if $l_t^{i,m} < f_{t,t}^{i,m} < h_t^{i,m}$ and $\tau h_t^{i,m} < \tau l_t^{i,m}$

(2) Valley then peak: if $l_t^{i,m} < f_{t,t}^{i,m} < h_t^{i,m}$ and $\tau l_t^{i,m} < \tau h_t^{i,m}$

(3) Up: if $f_{t,t}^{i,m} \leq l_t^{i,m} < h_t^{i,m}$

(4) Down: if $l_t^{i,m} < h_t^{i,m} \leq f_{t,t}^{i,m}$

(5) Flat: if $l_t^{i,m} = h_t^{i,m} = f_{t,t}^{i,m}$.

Finally, we define the maximum price rise (or fall) that subject i expects between period t and the price peak (or price valley), $\Delta f_t^{i,m}$, and the number of periods before the peak (or valley), $\Delta \tau_t^{i,m}$,

as follows:

$$\Delta f_t^{i,m} = \begin{cases} h_t^{i,m} - f_{t,t}^{i,m} & \text{if (1) Peak then Valley or (3) Up} \\ l_t^{i,m} - f_{t,t}^{i,m} & \text{otherwise} \end{cases}$$

$$\Delta \tau_t^{i,m} = \begin{cases} \tau h_t^{i,m} - t & \text{if (1) Peak then Valley or (3) Up} \\ \tau l_t^{i,m} - t & \text{otherwise} \end{cases}.$$

We are interested in determining whether the relationship between subjects' short-term forecasts $f_{t,t}^{i,m}$, as well as the characteristics of their long-term forecasts as measured by $\Delta f_t^{i,m}$ and $\Delta \tau_t^{i,m}$, and their buy orders differ across the three treatments, and if so, how.

Table 3 shows the results of the OLS regressions. The standard errors are corrected for the within-market clustering effect. Note that the estimated coefficients of $f_{t,t}^{i,m}$ are positive and statistically significant in Rounds 2 and 3, and positive but not statistically significant in Round 1. Thus, there is no clear evidence of subjects hedging their earnings from forecasting and trading because as noted above, such hedging would imply a negative correlation between bids and forecasts.

We now consider the treatment effects. In Round 1, we do not observe any statistically significant treatment effects. None of the estimated coefficients of the dummy variables, D_{BONUS} and $D_{F\&T}$, their interaction terms with the time trend, t, short-term forecasts, $f_{t,t}^{i,m}$, or the characteristics of long-term forecasts, $\Delta f_t^{i,m}$ and $\Delta \tau_t^{i,m}$, is statistically significant.

However, in Rounds 2 and 3, we observe some statistically significant differences across treatments. For the relevant range of the short-term forecasts (up to three times the FV), the bids are highest in the BONUS treatment, followed by the F&T treatment, compared with the ForT treatment in the early periods.[18] This is an indication that subjects in the F&T and BONUS treatments take more risks and submit higher bids in relation to the same level of short-term forecasts compared with those in the ForT treatment. These significant differences between the BONUS and ForT treatments persist in Round 3. Furthermore, the estimated coefficient of $\Delta f_t^{i,m} \times D_{F\&T}$ is positive and statistically significant. Thus, in the F&T treatment, if subjects forecast a price increase between the current period and the expected peak period, they raise the bids more than they do in

[18] This is basically the result of the very high and statistically significant estimated coefficients of D_{BONUS} and $D_{F\&T}$. Although the estimated coefficient of their interaction terms with $f_{t,t}^{i,m}$ is negative and statistically significant, the predicted bids will be higher in the BONUS treatment, followed by the F&T treatment, compared with the ForT treatment unless $f_{t,t}^{i,m}$ is very large relative to the FV. These treatment differences become smaller in later periods as one can observe from the negative and statistically significant estimates of $t \times D_{BONUS}$ and $t \times D_{F\&T}$.

Table 3: Dependent variable $b_t^{i,m}$

	(1) Round 1	(2) Round 2	(3) Round 3
Constant	67.87***	-20.88	6.092
	(18.10)	(12.84)	(32.70)
D_{BONUS}	28.07	122.4***	95.61***
	(21.47)	(28.72)	(34.94)
$D_{F\&T}$	-21.22	77.51***	51.64
	(23.56)	(17.43)	(35.27)
$f_{t,t}^{i,m}$	0.148	0.757***	0.852***
	(0.111)	(0.0538)	(0.227)
$f_{t,t}^{i,m} \times D_{BONUS}$	-0.0148	-0.518***	-0.743***
	(0.138)	(0.159)	(0.249)
$f_{t,t}^{i,m} \times D_{F\&T}$	0.257	-0.120*	-0.258
	(0.169)	(0.0692)	(0.240)
t	-4.060***	2.032	-0.632
	(1.445)	(1.259)	(3.194)
$t \times D_{BONUS}$	-1.463	-11.16***	-8.989**
	(1.734)	(2.788)	(3.382)
$t \times D_{F\&T}$	2.899	-7.029***	-4.665
	(2.344)	(1.980)	(3.508)
$\Delta f_t^{i,m}$	0.0259	-0.0235	-0.0113
	(0.0211)	(0.0329)	(0.0461)
$\Delta f_t^{i,m} \times D_{BONUS}$	0.0249	0.0669	0.0703
	(0.0512)	(0.0416)	(0.0811)
$\Delta f_t^{i,m} \times D_{F\&T}$	-0.00225	0.144**	0.136**
	(0.0272)	(0.0619)	(0.0501)
$\Delta \tau_t^{i,m}$	1.658	4.455***	0.503
	(1.466)	(0.952)	(2.056)
$\Delta \tau_t^{i,m} \times D_{BONUS}$	-2.780	-4.867***	1.337
	(1.761)	(1.754)	(2.320)
$\Delta \tau_t^{i,m} \times D_{F\&T}$	-1.471	-6.182***	-0.980
	(1.997)	(1.238)	(2.131)
adj. R^2	0.235	0.638	0.736
N	1248	1204	1257

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

ForT treatment. This is also the case in Round 3. Based on these results, we make the following observation.

Observation 4 *When subjects are rewarded based on both forecasting and trading performance, they tend to submit higher bids in relation to the same forecasted prices compared with the case in which they are rewarded based on either forecasting or trading performance.*

4 Forecast-only experiments

Thus far, we have investigated the effect of eliciting long-run price forecasts, as well as the way in which forecasting performance is incentivized, on market outcomes, forecasting performance, and trading behavior. We have seen that eliciting long-run forecasts, per se, does not have a significant effect on price dynamics and trading behavior if subjects are rewarded based on either their trading performance or their forecasting performance. However, mispricing is greater when there is forecast elicitation if subjects are rewarded based on both forecasting and trading performance. Despite the significant differences between the three treatments with forecast elicitation in terms of market outcomes, there were no statistically significant differences in terms of forecasting performance among the three treatments with forecast elicitation once the magnitude of mispricing was controlled for.

In designing experiments with an incentive scheme similar to that used in the ForT treatment, it is necessary to know the average forecasting performance to set an appropriate piece rate for an "accurate" forecast so that the expected rewards from forecasting and trading are comparable. We determined the piece rate for an "accurate" forecast based on the average forecasting performance reported by Akiyama et al. (2014) in their BONUS treatment. When there is no data set available, it is necessary to run sufficient pilot experiments to enable the computation of an appropriate piece rate. One possible way to find an appropriate piece rate without engaging numerous subjects in pilot experiments is to conduct "forecast-only" experiments in which subjects are asked to forecast the prices observed in the market under the T treatment without engaging in trading (see, for example, Section 4 in Akiyama et al., 2014). The question remains, of course, as to whether the forecasting performance of subjects in forecast-only experiments is comparable to that of subjects who engage in both trading and forecasting. Here, we report the results of our forecast-only experiments to address this question. In addition, we investigate how the introduction of a quadratic scoring rule instead

of a piece rate for an "accurate" forecast influences performance in forecast-only experiments.[19]

4.1 Procedure

In our forecast-only experiments, subjects were told that they would be participating in a "price forecasting game" in which they would be asked to forecast the prices observed in an "asset trading game" experiment that had been conducted in the past. Subjects received identical instructions to those provided in the T treatment. Following the instructions, subjects were reminded that they were not going to trade assets, but rather to forecast the prices that were observed over three rounds of 10 randomly selected trading periods. They then received instructions in relation to forecasting that were similar to the instructions provided to subjects in the ForT treatment. That is, they were told that at the beginning of each period they would be asked to forecast the prices for all of the remaining periods. They were also told that once they had submitted their forecasts, the realized price for the current period would be displayed on the screen. Finally, subjects were told that they would be rewarded based on their forecasting performance as follows:

$$\text{Payment based on forecast performance (in ECUs)} = F \times \text{number of "accurate" forecasts.}$$

As in the other treatments with forecast elicitation, we defined "accurate" forecasts as those that fell within 10% of the realized price. We tested two values of F: 40 and 50. $F = 40$ is identical to the ForT treatment, while $F = 50$ was used to check the robustness of our results against changes in the piece rate. We called the forecast experiment with $F = 40$ the FO40 treatment, and that with $F = 50$ the FO50 treatment.

We also conducted a forecast-only experiment with a quadratic scoring rule. We called this treatment FOQS (forecasting only, quadratic scoring). In this treatment, subjects were told that they would be rewarded based on the following formula:

$$\text{Payment based on forecast performance (in ECUs)} = \sum_{t=1}^{10}\sum_{p=t}^{10} \max\left(50 - 0.1\left(100 \times \frac{P_p - f_{t,p}}{P_p}\right)^2, 0\right),$$

[19] Beshears et al. (2013) use a piece-rate incentive scheme in their forecasting only experiment.

where P_p is the realized price in period p and $f_{t,p}$ is the forecast price for period p elicited at the beginning of period t. Note that $\left(100 \times \frac{P_p - f_{t,p}}{P_p}\right)$ is the forecast error (in percentage terms). Thus, if the forecast price is equal to the realized price, i.e., the forecast error is zero, the subject will receive 50 ECUs. If the forecast error is within 10%, the subject will receive 40 ECUs. If the forecast error is greater than 22.36%, the subject will not receive a reward. We selected this formula because a 10% forecast error provided the same reward as that under the FO40 treatment.

Similar to the other treatments, subjects in the forecast-only experiment participated in three rounds of a 10-period market. We randomly selected a series of prices observed in one of the 12 sessions in the T treatment for each subject. At the end of the experiment, subjects were rewarded based on their total earnings over the three rounds of the experiment in addition to their participation fee of 600 JPY.

4.2 Results

The experiment was conducted at the University of Tsukuba (Japan) between July and December 2016.[20] A total of 73 subjects (26 in FO40, 24 in FO50, and 23 in FOQS) participated in the forecast-only experiment. These subjects had never participated in an asset market experiment before. On average, subjects in the FO40, FO50, and FOQS treatments earned 2400, 2800, and 3810 JPY, respectively, in addition to their participation fee. The experiment lasted for about two hours, including the provision of instructions and completion of the post-experiment questionnaire. The higher rewards obtained in the FOQS treatment compared with the FO40 and FO50 treatments were the result of the additional rewards that subjects were able to obtain when their forecast errors were greater than 10%.

The top half of Table 4 shows the average number (and standard deviation) of "accurate" forecasts in each of the three rounds in the FO40, FO50, and FOQS treatments. Note that in the FOQS treatment, we simply counted the number of forecasts with an error of no more than 10%. The results for the ForT, F&T, and BONUS treatments are also shown for comparison.

Of interest here is the comparison between the outcomes of the the forecast-only treatments and the three other treatments involving trading. In Round 1, the average number of "accurate" forecasts is not significantly different across the five treatments ($p = 0.825$ based on the KW test using an individual as a unit of observation.). In Rounds 2 and 3, the number of accurate forecasts

[20] The experiment was computerized using z-Tree Fischbacher (2007).

Table 4: Average number (standard deviation) of "accurate" forecasts.

Treatment	number of subjects	Number of "accurate" forecasts		
		Round 1	Round 2	Round 3
FO40	26	10.08	19.85	30.12
		(5.73)	(11.12)	(16.52)
FO50	24	10.17	17.25	28.58
		(6.61)	(10.11)	(15.73)
FOQS	23	13.30	22.13	32.91
		(10.09)	(12.68)	(18.28)
p-values (KW), FO40, FO50, FOQS		0.345	0.274	0.774
ForT	72	12.65	31.10	39.69
		(11.55)	(16.37)	(11.73)
BONUS	72	12.74	28.39	38.29
		(10.82)	(19.03)	(17.37)
F&T	72	11.53	18.92	27.38
		(10.85)	(15.18)	(16.51)
p-values (KW), All 6 treatments		0.843	0.002	0.006

increases in all treatments. However, the number of accurate forecasts is significantly lower in the forecast-only treatments compared with the ForT and BONUS treatments. There is no statistically significant difference between the forecast-only treatments and the F&T treatment.

This result suggests that the forecast-only experiment is a good guide to setting an appropriate piece rate for forecasting performance in the first round of an asset market experiment with forecast elicitation for inexperienced subjects. However, unfortunately it does not provide a reliable guide to piece rates for later rounds or experiments involving experienced subjects. This is in sharp contrast with Bao et al. (2013) that report, prices converge much quicker to the fundamental value when subjects are only forecasting, compared to the case where subjects are trading and forecasting. In our view, the difference between our results and those of Bao et al. (2013) is as follows. In Bao et al. (2013), trades are done optimally and automatically by computer program given submitted forecasts. As a result, the price series to be forecasted are less noisy, which certainly help subjects to better forecast. In our forecast only experiment, on the contrary, subjects are asked to forecast quite noisy price series that result from other subjects trading. It is quite likely that the substantial behavioral uncertainty that exist in our forecast-only experiments prevented subjects to outperform those subjects who trade and forecast.

We now compare the three forecast-only treatments. The results of the KW test presented in

Table 4 show that the number of accurate forecasts is not statistically significantly different among the three treatments. Thus, we can conclude that the 25% difference in the piece rate for accurate forecasts between the FO40 and FO50 treatments does not have a significant influence on the results. Furthermore, the introduction of a quadratic scoring scheme does not have a statistically significant effect in terms of the number of accurate forecasts compared with the piece rate for accuracy within 10%.

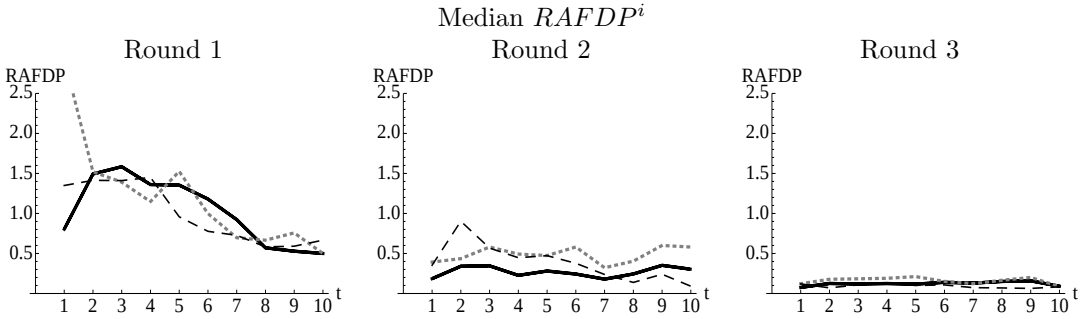
How do forecast deviations from the realized prices, RAFDP and RFDP, differ across the three forecast-only treatments? Figure 7 shows the dynamics of the median RAFDP (top) and median RFDP (bottom) for each round of the three forecast-only treatments: FO40 (thick solid black), FO50 (thick dashed gray), and FOQS (thin dashed black). The p-values from the KW test are also presented. Except for the first period in Round 1 and the last period in Round 2, the median RAFDP is very similar in each of the three treatments. In fact, except for these periods, they are not statistically significantly different across the three treatments at the 5% significance level. The RFDP comparisons across the three treatments are basically the same. Thus, we summarize these results with the following observation.

Observation 5 *Differences in the piece rate (FO40 vs FO50) or the scoring method (FOQS vs FO40 or FO50) do not result in statistically significant differences in forecasting performance.*

5 Summary and conclusion

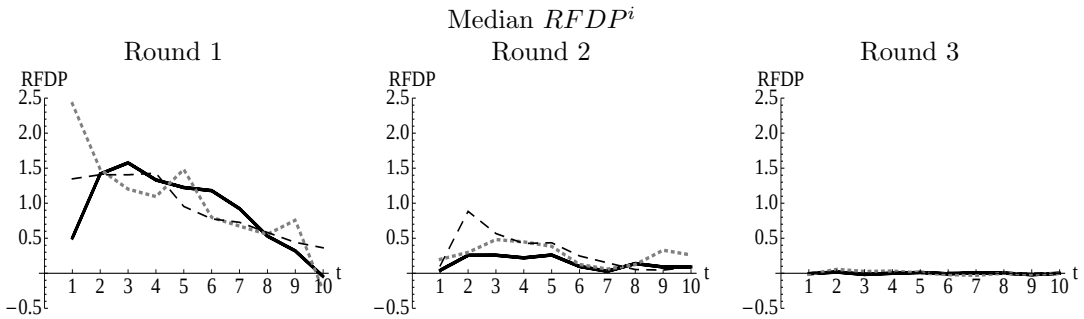
In this paper, we investigated (a) whether eliciting future price forecasts influences market outcomes and (b) whether differences in the way in which subjects are incentivized for their performance in forecasting and trading, i.e., whether they are rewarded based on both forecasting and trading or on either forecasting or trading, influence market outcomes as well as the forecasts they submit in the experimental asset market paradigm pioneered by Smith et al. (1988).

We examined four treatments: one without forecast elicitation (T treatment) and three with forecast elicitation. In one of the latter three treatments, subjects were paid based on either their forecasting performance or their trading performance, but not both, with the payment method determined randomly at the end of the experiment (ForT treatment). In the two other forecast elicitation treatments, subjects were rewarded based on both their forecasting and trading perfor-



p-values of Kruskal-Wallis test

t	1	2	3	4	5	6	7	8	9	10
Round 1	**0.045**	0.663	0.739	0.743	0.597	0.517	0.760	0.867	0.576	0.507
Round 2	0.249	0.729	0.702	0.533	0.712	0.296	0.359	0.095	0.070	**0.025**
Round 3	0.784	0.563	0.891	0.588	0.405	0.631	0.730	0.358	0.159	0.333



p-values of Kruskal-Wallis test

t	1	2	3	4	5	6	7	8	9	10
Round 1	0.352	0.906	0.933	0.987	0.965	0.723	0.988	0.959	0.828	**0.042**
Round 2	0.828	0.546	0.412	0.623	0.901	0.277	0.391	0.825	0.548	0.864
Round 3	0.493	0.674	0.483	0.269	0.902	0.304	0.303	0.952	0.475	0.751

Figure 7: Dynamics of the median $RAFDP_t^i$ (top) and $RFDP_t^i$ (bottom) in the FO40 (thick solid black), FO50 (thick dashed gray), and FOQS (thin dashed black) treatments over three rounds. P-values from the KW test are also presented for each period.

mances (BONUS and F&T treatments). In the BONUS treatment, subjects were rewarded for their forecasting performance in the form of bonus payments that were a proportion of the payment based on their trading performance. In the F&T treatment, subjects were rewarded equally for both their forecasting and trading performances.

While we found no statistically significant differences between the T and ForT treatments in terms of mispricing, trading volumes, and the potential loss that can be generated by the orders subjects submit, significantly greater mispricing was observed in the BONUS and F&T treatments compared with the T treatment. Thus, to avoid influencing the behavior of subjects and market outcomes by eliciting price forecasts compared to the benchmark case without forecast elicitation, subjects should be rewarded based on either their forecasting or trading performance (to be determined randomly at the end of the experiment) instead of being rewarded based on both forecasting and trading performance, as has been done in previous studies that employed variants of the experimental setting proposed by Smith et al. (1988).

We conjecture that the reason why the BONUS and F&T treatments result in greater mispricing than the T and ForT treatments is because subjects take greater risks in relation to their trading activity in the BONUS and F&T treatments than in the T and ForT treatments. We find some support for this conjecture. For a given level of price forecasts, subjects in the BONUS and F&T treatments tended to submit higher bids than those in the ForT treatment.

We also conducted forecast-only experiments in which subjects were asked to forecast prices observed in the T treatment without engaging in trading. The results suggest that this type of experiment is useful for understanding the forecasting performance of inexperienced subjects (i.e., in the first round of a multi-period market), but not for experienced subjects (such as in the second or third rounds of the same markets). In terms of the number of accurate forecasts, the performance of experienced subjects in the forecast-only experiments was significantly worse compared with those who undertook both trading and forecasting. In our forecast-only experiments, we varied the piece rate for an "accurate" forecast (40 vs 50 ECUs for forecasts falling within 10% of the realized price) and introduced a quadratic scoring rule to measure subjects' forecasting performance. These variations did not produce any statistically significant differences in forecasting performance.

References

Akiyama, E., N. Hanaki, and R. Ishikawa (2014): "How do experienced traders respond to inflows of inexperienced traders? An experimental analysis," *Journal of Economic Dynamics and Control*, 45, 1–18.

——— (2017): "It is not just confusion! Strategic uncertainty in an experimental asset market," *Economic Journal*, 127, F563–F580.

Anufriev, M. and C. Hommes (2012): "Evolutionary selection of individual expectations and aggregate outcomes in asset pricing experiments," *American Economics Journal, Microeconomics*, 4, 35–64.

Anufriev, M., C. H. Hommes, and T. Makarewicz (2015): "Simple forecasting hueristics that make us smart: Evidence from different market experiments," CENDEF Working paper 15-07, University of Amsterdam.

Bao, T., J. Duffy, and C. Hommes (2013): "Learning, Forecasting, and Optimizing: An Experimental Study," *European Economic Review*, 61, 186–204.

Bao, T. and C. Hommes (2014): ""When Constructors Meet Speculators: an Experiment on Housing Supply Elasticity and Price Stability," Mimeo, University of Groningen.

Bao, T., C. Hommes, and T. Makarewicz (2015): "Bubble foremation and (in)efficient markets in learning-to-forecast and -optimize experiments," *Economic Journal*, forthcoming.

Bao, T., C. Hommes, J. Sonnemans, and J. Tuinstra (2012): "Individual expectations, limited rationality and aggregate outcomes," *Journal of Economic Dynamics and Control*, 36, 1101–1120.

Beshears, J., J. J. Choi, A. Fuster, D. Laibson, and B. C. Madrian (2013): "What Goes Up Must Come Down? Experimental Evidence on Intuitive Forecasting," *The American economic review*, 103, 570–574.

Bosch-Rosa, C., T. Meissner, and A. Bosch-Domènech (2017): "Cognitive Bubbles," *Experimental Economics*, forthcoming, doi:10.1007/s10683-017-9529-0.

Carle, T. A., Y. Lahav, T. Neugebauer, and C. N. Noussair (2017): "Heterogeneity of beliefs and trade in experimental asset markets," *Journal of Financial and Quantitative Analysis*, forthcoming.

Fischbacher, U. (2007): "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10, 171–178.

Haruvy, E., Y. Lahav, and C. N. Noussair (2007): "Traders' Expectations in Asset Markets: Experimental Evidence," *American Economics Review*, 97, 1901–1920.

Heemeijer, P., C. Hommes, J. Sonnemans, and J. Tuinstra (2009): "Price stability and volatility in markets with positive and negative expectations feedback: An experimental investigation," *Journal of Economic Dynamics and Control*, 33, 1052–1072.

Hommes, C., J. Sonnemans, J. Tuinstra, and H. van de Velden (2005): "Coordination of expectations in asset pricing experiments," *Review of Financial Studies*, 18, 955–980.

Honkapohja, S. (2015): "Monetary policies to counter the zero interest rate: An overview of research," Research Discussion Paper 18/2015, Bank of Finland.

Marimon, R., S. E. Spear, and S. Sunder (1993): "Expectationally driven market volatility: An experimental study," *Journal of Economic Theory*, 61, 74–103.

Mertens, K. R. S. M. and M. O. Ravn (2014): "Fiscal policy in an expectations-driven liquidity trap," *Review of Economic Studies*, 81, 1637–1667.

Nuzzo, S. and A. Morone (2017): "Asset markets in the lab: A literature review," *Journal of Behavioral and Experimental Finance*, 13, 42–50.

Palan, S. (2013): "A Review of bubbles and crashes in experimental asset markets," *Journal of Economic Surveys*, 27, 570–588.

Powell, O. and N. Shestakova (2016): "Experimental asset markets: A survey of recent developments," *Journal of Behavioral and Experimental Finance*, 12, 14–22.

Smith, V. L., G. L. Suchanek, and A. W. Williams (1988): "Bubbles, Crashes, and Endogenous Expectations in Experimental Spot Asset Markets," *Econometrica*, 56, 1119–1151.

Sonnemans, J. and J. Tuinstra (2010): "Positive expectations feedback experiments and number guessing games as models of financial markets," *Journal of Economic Psychology*, 31, 964–984.

Stöckl, T., J. Huber, and M. Kirchler (2010): "Bubble measures in Experimental Asset Markets," *Experimental Economics*, 13, 284–298.

van Boening, M. V., A. W. Williams, and S. LaMaster (1993): "Price bubbles and crashes in experimental call markets," *Economics Letters*, 41, 179–185.