The Effect of Framing on the Emergence of Bubbles and Crashes in a Learning to Forecast Experiment

Nobuyuki Hanaki*[1], Cars Hommes†[2,3], Dávid Kopányi‡[2,3] and Jan Tuinstra§[2,3]

[1]Université Côte d'Azur, Nice, France

[2]University of Amsterdam, The Netherlands

[3]Tinbergen Institute, The Netherlands

May 1, 2019

Abstract

Expectations of future returns are pivotal for investors' trading decisions, and are therefore an important determinant of the evolution of actual returns. Evidence from laboratory experiments with exogenously given time series of returns suggests that subjects' return forecasts are substantially affected by how they are elicited and by the format in which subjects receive information about past asset performance. We study the effect of framing on asset price dynamics in a learning to forecast experiment where prices and returns are endogenously determined and depend directly upon subjects' forecasts. We vary both the variable (prices or returns) subjects observe and the variable (prices or returns) they have to forecast, with the same underlying data generating process for each treatment. Although there is no significant effect of the format of the presentation of past information, we do find that the incidence and amplitude of bubbles increase significantly when subjects have to forecast returns instead of prices. This is due to subjects coordinating on forecasting strategies that tend to exhibit stronger trend extrapolation when return forecasts are elicited. Our results therefore show that framing may exacerbate, or even create, bubbles and crashes in financial markets.

JEL classification: C63, C72, D43

Keywords: experimental finance, expectation formation, asset market, framing effects

*email: nobuyuki.hanaki@unice.fr

†email: c.h.hommes@uva.nl

‡email: d.kopanyi@uva.nl

§email: j.tuinstra@uva.nl

1 Introduction

Return expectations are one of the main determinants of traders' investment decisions and therefore play a crucial role in financial market dynamics. Consequently, a profound insight in how these expectations are formed will contribute substantially to our understanding of financial market phenomena such as excess volatility and the emergence of bubbles and crashes. It is therefore hardly surprising that, in the last couple of decades, considerable effort has been put in analyzing expectation formation, both by using data from questionnaire studies and by using data from laboratory experiments.[1]

However, the forecasts collected from these survey studies and laboratory experiments may be systematically affected by framing. In particular, it has been demonstrated that the way in which forecasts are elicited, as well as the format in which past data are represented, can have a substantial effect on the forecasts provided by the participants in these studies. Glaser et al. (2007), for example, show that subjects that have to forecast returns exhibit a stronger tendency to extrapolate past trends than subjects that are asked to forecast prices.[2] In addition, in a recent study Glaser et al. (2018) find that return expectations are higher when return forecasts are elicited than when price forecasts are elicited, and that return expectations are lower when past returns are shown to the subjects than when past prices are shown. This is of particular interest since both formats are used in investor documents of mutual funds, on financial websites, and so on, and may change according to changes in regulation.[3] Moreover, some well-known financial market surveys differ in how they elicit forecasts (see Glaser et al., 2018, for an overview).

The effect of framing implies that, at a minimum, results on expectation formation from questionnaire studies and laboratory experiments should be interpreted with care, in particular since these results are used in economic policy debates (see e.g. Glaser et al., 2007 and Hoffmann et al., 2017) and discussed in the popular press, thereby partially shaping the expectations of the general public. Moreover, return

[1] For studies using survey data see e.g. Frankel and Froot (1987) for exchange rate expectations, Bacchetta et al. (2009), Amromin and Sharpe (2014) and Greenwood and Shleifer (2014) for financial market expectations, Case et al. (2012) for expectations on U.S. housing prices and Carroll (2003), Branch (2004) and Malmendier and Nagel (2016) for inflation expectations. For laboratory experiments on expectation formation, see e.g. Schmalensee (1976), Dwyer et al. (1993), Hey (1994), Kelley and Friedman (2002) and Beshears et al. (2013).

[2] Obviously, in order to be able to compare the forecasts between the two treatments, the price forecasts are translated back to return forecasts for the treatment where price forecasts are elicited.

[3] For example, Directive 2009/65/EC of the European Union stipulates that investment funds have to give prospective buyers a so-called Key Investor Information Document (KIID) that includes past information on the fund in the form of a return bar chart. By January 1, 2018, this document was replaced by the Key Investor Document (KID) with no requirements on the presentation of past information, although it has to include different possible future performance scenarios (see Regulation 1286/2014 of the EU).

expectations guide investment decisions and therefore have a direct impact on realized returns. If framing leads to a systematic bias in expectation formation, financial market behavior as measured by, for example, price volatility, is likely to be affected as well. In particular, if eliciting return forecasts leads to stronger trend extrapolation this may increase the incidence of bubbles and crashes. Similarly, if presenting past performance by return bar charts (instead of price line charts) leads to more moderate return forecasts, then bubbles may be less likely to occur.

In the studies mentioned above, and in many others, subjects need to forecast the next realization of a predetermined and exogenously given time series of prices or returns (either simulated or based on historical stock price data). These studies therefore abstract from any effect that return forecasts have on return realizations. In this paper we go beyond the existing literature and study framing in a laboratory experiment where this expectations feedback is explicitly taken into account. In this so-called *learning to forecast* experiment[4] subjects' average expectations of prices/returns are an important determinant of realized prices/returns, which allows us to investigate the effect that framing has on price volatility and the endogenous emergence of bubbles and crashes in asset prices.

We use a 2×2 between-subjects design similar to the one used in Glaser et al. (2018). Depending on the treatment subjects either see past prices or past returns, and either have to forecast the next price or the next return, for 50 consecutive periods. The underlying data generating mechanism is the same for all four treatments, with the only differences between the treatments in how the forecasting task and the information is framed. Subjects are paid for their forecasting accuracy. We find that the format in which past information is represented (either as a return bar chart or as a price line chart) does not have a significant effect on the resulting price dynamics, but the format in which forecasts are elicited does. In particular, we find that asking for return forecasts increases price volatility and the incidence of bubbles and crashes substantially, when compared to asking for price forecasts. Analysis of subjects' individual forecasts reveals that this is due to a tendency of subjects to coordinate on forecasting strategies that extrapolate trends (in prices) more strongly when forecasting returns than when forecasting prices. We therefore provide evidence that thinking about returns, instead of thinking about prices, has a substantial impact on the performance of financial markets. Policy makers and regulators should take this into account when designing polices aimed at stabilizing financial markets.

Our findings are consistent with earlier research that suggests that trend extrapolation may explain differences in forecasting behavior between different elicitation formats. Glaser et al. (2007) review the literature and show that questionnaire studies and laboratory experiments where return (or price change)

[4]This approach to studying expectation formation in self-referential economic models was introduced by Marimon et al. (1993) to analyze price forecasts in an overlapping generations model.

forecasts are elicited typically document trend extrapolation, whereas mean reverting behavior is found in studies where price forecasts are elicited. For example, in one of the studies discussed in Andreassen and Kraus (1990) subjects are presented with five values of an exponential time series and have to predict the next five values. Subjects' forecasts are higher (and more accurate) when they see both the first five values and four changes and have to predict the next five changes than when they are only given the first five values and have to predict the next five values. In Study 1 of Czaczkes and Ganzach (1996) subjects need to predict future stock prices on the basis of past changes in stock earnings. In one treatment they have to predict prices, and get feedback about the price, in the second treatment they have to predict price changes and get feedback about price changes. The predictions in the price change treatment tend to be more extreme, which again suggests that subjects have a stronger tendency to extrapolate trends in that treatment. However, in both of these studies not only the forecast elicitation mode, but also the information or feedback given to the subjects differs between treatments, making it difficult to draw precise conclusions about the effect of the format of the forecasting task. In the questionnaire study presented in Glaser et al. (2007) the effect of the format of the task is isolated: in all treatments subjects observe different series of either increasing, stable or decreasing historical prices of actual stocks from the German stock exchange.[5] In one treatment subjects are asked to forecast prices and in another treatment they are asked to forecast returns. Subjects that have to forecast returns exhibit a stronger tendency to extrapolate past trends than subjects that have to forecast prices: that is, return forecasts are higher (lower) after prices have been increasing (decreasing) when return forecasts are elicited directly than when these forecasts are derived from elicited price forecasts.

Glaser et al. (2018) do not only vary the elicitation task but also, in a 2×2 between-subjects design, the format in which past data are presented, either as a line chart of past prices or as a bar chart of past returns. Instead of finding that return forecasts are more extreme when return forecasts are elicited than when price forecasts are elicited, they find that these return forecasts are higher. A possible explanation for the difference with the studies discussed above may be that the past data used in Glaser et al. (2018) are random sequences from a normal distribution, and therefore less likely to feature the increasing or decreasing trends that are imposed in e.g. Andreassen and Kraus (1990) and Glaser et al. (2007), and that also endogenously emerge in the learning to forecast experiment we discuss in this paper. Nevertheless, higher return forecasts would increase demand for the asset and are therefore likely to increase the incidence of bubbles in asset prices as well. Glaser et al. (2018) also find that return forecasts are lower when past returns are shown than when past prices are shown which, by a similar argument, has the potential to

[5] In one of their treatments subjects are also given past returns, in addition to past prices, but this does not have an effect on the forecasts.

diminish the likelihood of bubbles. However, we do not find evidence for an effect of the format of past information. Indeed, results from other experimental work that focuses on the effect of the presentation of past information are mixed as well. Andreassen (1988) lets subjects trade an artificial stock and investigates to what extent subjects *track* the price, that is, sell when the price is high, and buy when the price is low. He finds that subjects track the price more when they observe past prices than when they observe past price changes. Although subjects' forecasts were not elicited, this behavior is consistent with stronger expected mean reversion when prices are observed than when returns are observed. Diacon and Hasseldine (2007) study the effect of the presentation format of different funds (in terms of the fund value or the % yield), but do not find that investment decisions are significantly effected by this. In the different treatments of the experiment presented by Stössel and Meier (2015) subjects either see a return bar chart or a price line chart. As opposed to Glaser et al. (2007), they find that subjects overestimate returns substantially when seeing a return bar chart. Finally, Huber and Huber (2019) show their subjects either prices or returns, and elicit one- and five-year ahead return forecasts. The five-year ahead return forecasts, in particular, are more extreme when subjects observe returns than when subjects observe prices. This seems to be consistent with the results from Andreassen (1988), but not necessarily with those from Glaser et al. (2018).

Our experiment is also related to previous work on learning to forecast experiments. In applications to financial markets, with subjects forecasting future prices on the basis of past prices, learning to forecast experiments typically exhibit persistent deviations of realized prices from fundamentals and the endogenous emergence of bubbles and crashes (see e.g. Hommes et al., 2005, 2008 and Heemeijer et al., 2009), as well as a remarkably high degree of coordination of individual forecasts on a common prediction strategy. These results are quite robust, for example with respect to information about the underlying model (Sonnemans and Tuinstra, 2010), with respect to the number of subjects in a group (Hommes et al., 2018), and with respect to letting subjects make trading decisions, instead of, or in addition to, letting them predict future prices (Bao et al., 2017). Our results show that when eliciting returns (instead of eliciting prices, which is the case in the previous learning to forecast experiments), these persistent deviations from fundamental values are exacerbated even further.

Finally, by showing that framing may increase mispricing and lead to bubbles and crashes in asset prices, our paper contributes to the literature that shows that framing may have important effects on financial market decisions (see e.g. Kirchler et al., 2005, Kirchler et al., 2012, and Anufriev et al., 2019).

The remainder of this paper is structured as follows. Section 2 presents the experimental design, the underlying asset pricing model and our main hypotheses. We discuss our main results on price stability, and on individual prediction strategies, in Section 3. Section 4 concludes. Experimental instructions and

		task	
		price	*return*
stimulus	*price*	*PP* (8)	*PR* (8)
	return	*RP* (7)	*RR* (8)

Table 1: Treatments in the 2×2 design and the number of markets (in parenthesis)

other supplementary material are presented in the Appendices.

2 Experimental design

The experiment, programmed in PhP, was run in February and March 2017 at the CREED experimental laboratory of the University of Amsterdam. In total 198 subjects (students from various fields) participated in the experiment. Experimental sessions lasted for approximately 90 minutes, with payments for each subject typically between €24 and €28. Below we will outline the main features of the experimental design.

2.1 Subjects' task and treatments

Our design is based upon the typical learning to forecast experimental design (see Hommes et al., 2005, 2008, and Heemeijer et al., 2009, for examples or Hommes, 2011, for an overview). Subjects are told that their role is that of an advisor to a pension fund. This pension fund has to decide how much of its wealth to invest in a risky asset, and bases its decision upon the forecast provided by the subject. The task of the subjects is to forecast the price or the return of the risky asset for 50 sequential periods, using information about past prices or information about past returns. Subjects' earnings are based upon their forecasting accuracy.

Following Glaser et al. (2018), we vary: (i) the manner in which forecasts are elicited ('task'), either by asking for a forecast of the price, p_t, or by asking for a forecast of the return (i.e. the relative price *change*), $r_t = \frac{p_t - p_{t-1}}{p_{t-1}}$; and (ii) the way in which information is provided to the subjects ('stimulus'), again either as a time series of past prices, or as a bar chart of past returns. This gives a 2×2 between-subjects design with four treatments, PP, RP, PR and RR, where, for example, PR means that subjects observe prices (P) and forecast returns (R), and similarly for the other treatments – see Table 1. Six subjects are

active in each market in each treatment.[6] We have seven markets in treatment RP and eight markets in each of the other three treatments.[7]

In contrast to Glaser et al. (2018) and many other experimental studies on expectation formation, in learning to forecast experiments the realization of the variable that subjects need to predict is not exogenously given, but determined by the subjects' predictions. In particular, when subjects predict a higher price/return for the risky asset, the pension funds they advise will demand more of this asset, in order to reap the potential capital gains. Increased aggregate demand for the risky asset will then drive the price/return of this asset up instantaneously. We formalize this 'expectations feedback' by the pricing equation from Bao et al. (2017), which is given by[8]

$$p_t = 66 + \frac{1}{1.05}\left(\overline{p}_t^f - 66\right) + \varepsilon_t. \tag{1}$$

Here p_t is the price of the risky asset in period t and $\overline{p}_t^f = \frac{1}{6}\sum_{h=1}^{6} p_{h,t}^f$ is the average price forecast of the subjects for period t, averaged over the six subjects in the same market. Furthermore, ε_t is a small demand shock with $\varepsilon_t \sim N(0, 0.25)$. Note that the (rational expectations) fundamental price in this market equals 66, and is constant over time. Moreover, expectations feedback is *positive* (an increase in the average forecast increases the price realization) and the *feedback strength*, given by the discount factor 1/1.05, is high (abstracting from the demand shocks, the realized price is a weighted average of the average price forecast and the fundamental price, with most of the weight on the former).[9]

It is straightforward to transform (expected) prices into (expected) returns, and the other way around

[6] Recently, Hommes et al. (2018) ran learning to forecast experiments with large groups of up to 100 subjects. Individual and aggregate behavior in these large groups is very similar to the behavior in groups of six.

[7] For RP we initially had a total of nine markets but we decided to exclude two of these markets from the analysis. In one of the markets one subject decided to stop participating and left the laboratory in period 10. From that point on the effective number of subjects in that market was five instead of six. In the other market we encountered a problem with the experimental software. The decisions of one subject in periods 7, 22, 23, 32, and 34 were treated as 0 by the software and this distorted the dynamics of prices and returns to a large extent. We had a similar issue in one of the markets in treatment RR (market RR4), but we decided to keep that market because the problem occurred only from period 32 onwards and it had only a very minor effect on the dynamics (a return forecast of 0 has a much less dramatic impact on the realized price/return than a price forecast of 0).

[8] Heemeijer et al. (2009) and Sonnemans and Tuinstra (2010) use the same price generating mechanism, but with a fundamental value of 60, instead of 66. Bao et al. (2012) also consider (1), but with a fundamental value that undergoes several permanent shocks and, in the course of their experiment, moves from 56 to 41 and then to 62. See Appendix A for more details about the microfoundation of the asset pricing model and detailed derivations for pricing equation (1).

[9] Let us remark that predicting the fundamental price is the optimal choice when agents are price takers, leading to the highest individual and aggregate payoffs. See Online Appendix C in Bao et al. (2017) for a more detailed discussion on this question.

(using $r_t = (p_t - p_{t-1})/p_{t-1}$ or $p_t = (1 + r_t)\, p_{t-1}$). Equation (1) therefore generates prices (and hence returns) for each of the four treatments. Also note that the realization of demand shocks ε_t is the same for each market and for each treatment.

Subjects do not have full knowledge of the price/return generating mechanism (1). However, they are provided with qualitative information about how the market works. That is, they are explained that: (i) a higher forecast will lead the pension fund they advise to buy more units of the risky asset; and (ii) the market price will be higher if total demand for the risky asset is higher. In addition, they know that the number of subjects in their market is six.

2.2 Information and incentives

Examples of the decision screens, for treatments PR and RP, are shown in Figure 1. Subjects can submit their forecast at the top of the screen.[10] Depending on the treatment, the information subjects have when they need to submit a forecast for period t consists of: (i) a table with the realized prices or returns, their own forecasts, their earnings in the last period, and their accumulated earnings thus far (lower right part of the decision screen); (ii) a figure with (for treatments PP and PR) a time series of past realized prices or (for treatments RP and RR) a bar chart of past realized returns (lower left part of the decision screen)[11]; and (iii) the most recent price. Note that this most recent price is also given for treatments RP and RR, since otherwise the task for subjects in treatment RP, where they observe past returns but have to forecast the price, would become overly complicated. For treatment RR this most recent price is not required, but it is given in order to provide subjects with the same information in treatments RP and RR. For all treatments a price of $p_0 = 50$ is shown on the initial decision screen.

Subjects are paid based on their forecasting accuracy. In particular, the number of points earned by subject h in period t is given by:

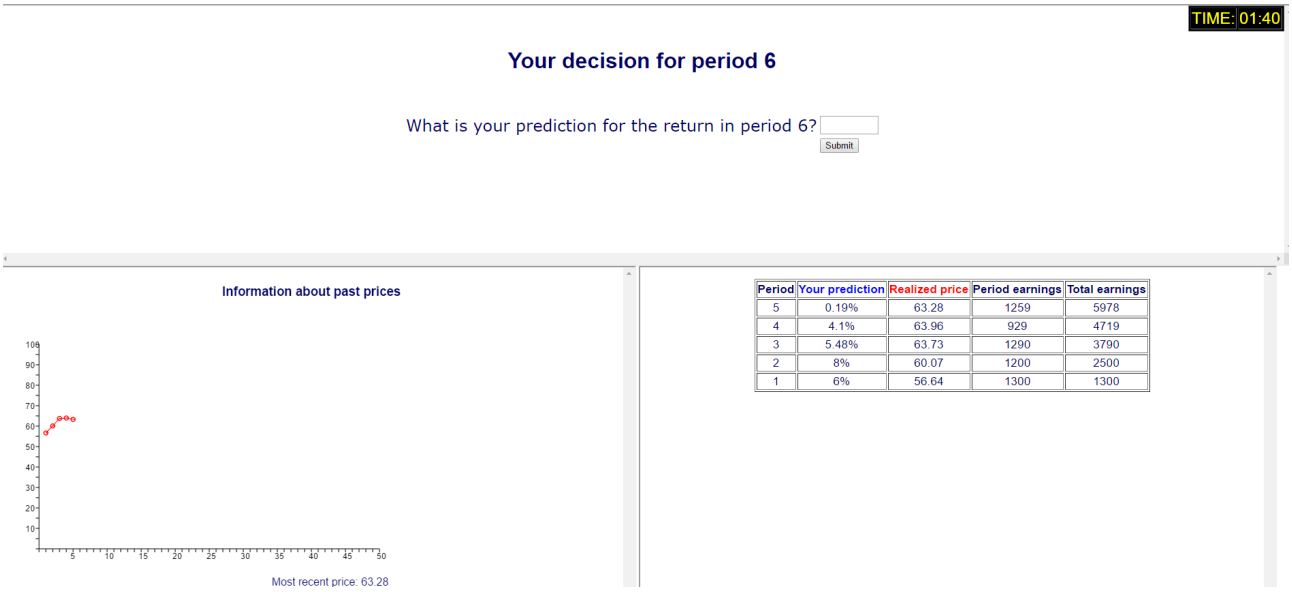
$$\text{payoff}_{h,t} = 1300 \times \max\left\{1 - 625 \times F_{h,t}^2, 0\right\},$$

where $F_{h,t}$ is a measure of the forecast error made by participant h in period t. Here we take $F_{h,t} = \frac{p_{h,t}^e - p_t}{p_{t-1}}$ for treatments PP and RP and $F_{h,t} = r_{h,t}^e - r_t$ for treatments PR and RR, so that incentives are exactly the same in each treatment.[12] Subjects earn between 0 (if the forecast error in that period is 4% or larger)

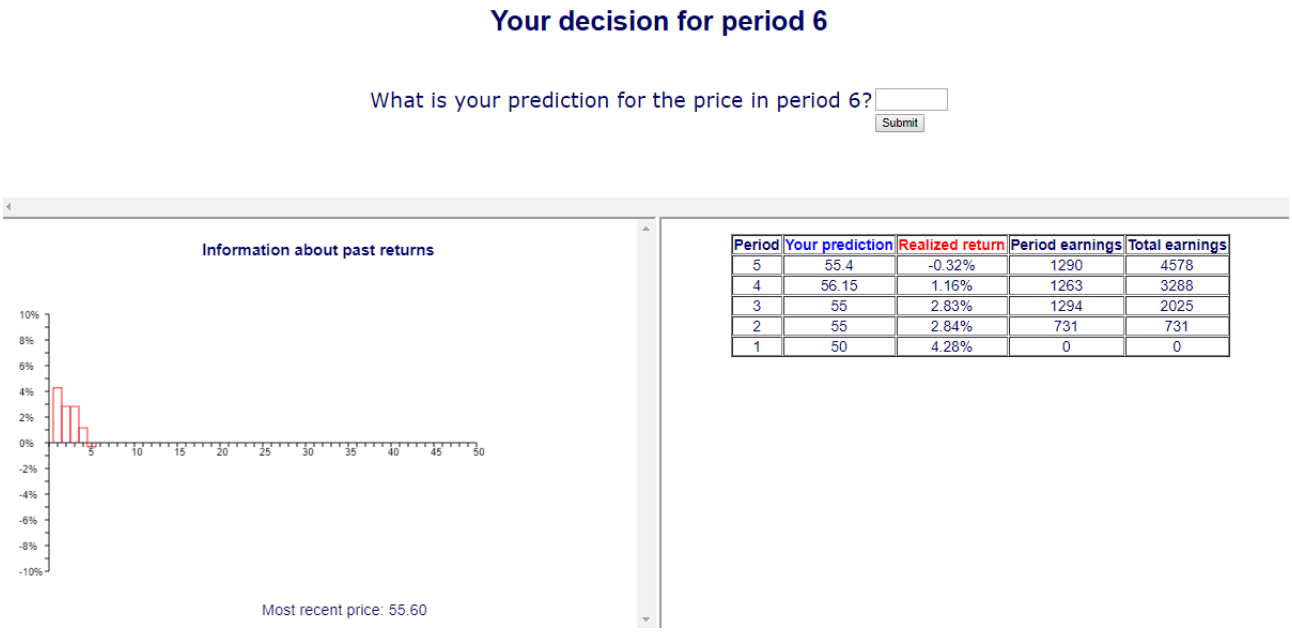
[10] When subjects forecast returns, they have to type the number without the % sign. For example, a forecast of 2.34% has to be submitted as 2.34.

[11] Following Glaser et al. (2018), we present past prices as time series and past returns as bar charts, respectively, because this is how they are typically represented in financial markets.

[12] Note that for treatments PP and RP subjects are rewarded on the basis of *relative* forecast errors. This is different from earlier Learning to Forecast experiments where prices had to be predicted (e.g. Heemeijer et al., 2009 and Bao et al., 2012).



(a) Treatment PR



(b) Treatment RP

Figure 1: Example of the decision screen in treatments PR (panel a) and RP (panel b)

and 1300 (if the forecast is correct) points per period. At the end of the experiment subjects' total points over the 50 periods are transformed into euros (with 2600 points giving €1.00), in addition to a €7.00 show-up fee.

Subjects' earnings in those experiments are based on *absolute* forecast errors. Since forecast errors scale with the price level, our current design penalizes coordination on a price bubble to a lesser extent than those earlier experiments do.

Let us elaborate on some further details of the experiment. We impose upper and lower bounds for price and return forecasts, with price forecasts restricted to be between 0 and 1000 and return forecasts restricted to be between -100% and 300%. In order not to provide focal points, subjects are not informed about these restrictions beforehand, but they receive an (individual) message as soon as they try to submit a forecast that violates a restriction that is relevant for them.[13] For the first period subjects do not have any information about prices or returns yet, but in the instructions we suggest that the price (return) in the first period is likely to be in the interval $[0, 100]$ ($[-10\%, 10\%]$), although subjects are not obliged to choose a forecast from that interval. Subjects have two minutes to make their decision during each of the first 10 periods and one minute for each of the periods 11 to 50.[14,15]

2.3 Hypotheses

Essentially, in each of the four treatments subjects are asked to perform the same task, have the same information for doing the task and are rewarded in the same way. The only difference between treatments is how the task and information are framed, either in terms of prices or in terms of returns. One might therefore conjecture that behavior of subjects is independent of the treatment. However, previous experimental research has shown that framing may matter. Indeed, Glaser et al. (2007) and Glaser et al. (2018) provide compelling evidence that framing has a notable effect on how people form expectations, for time series of prices/returns that are exogenously given. In particular, Glaser et al. (2007) conclude that subjects have a larger tendency to extrapolate trends when they forecast returns than when they forecast prices. Glaser et al. (2018), on the other hand, find that asking for returns leads to higher forecasts than asking for prices. Both results suggest that, in an environment with positive expectations feedback, the incidence of bubbles, as well as price volatility, will be higher when subjects forecast returns than when they forecast prices. This leads to our first hypothesis.

Hypothesis 1 Forecasting *returns instead of prices leads to* more unstable *market dynamics.*

In addition, Glaser et al. (2018) find that showing subjects past returns leads to lower forecasts than

[13] Only five out of 186 subjects (2.7%) tried to submit a prediction outside of these bounds.

[14] If a subject does not submit a forecast on time, then the pension fund advised by this subject is inactive in that period, and the subject will have zero earnings for that period. The average expectation used in (1) is then calculated over the subjects who do submit a forecast on time. This situation occurred 22 times in total (0.24% of all forecasts).

[15] The average decision times were 24.6, 29.7, 28.7 and 25.5 seconds in treatments PP, RP, PR and RR, respectively. Differences in decision times are significant at the 5% level (using 2-sided Kolmogorov-Smirnov tests) between PP vs. RP, PP vs. PR, RP vs. RR and PR vs. RR. Moreover, decision times are significantly lower when subjects observe and forecast the same variable ($p=0.0008$), which makes intuitive sense.

when showing them past prices. In our setting, lower expectations will lead to lower market prices, and this is likely to inhibit the emergence of bubbles, which gives rise to our second hypothesis.

Hypothesis 2 Observing *returns instead of prices leads to* more stable *market dynamics.*

We will test Hypotheses 1 and 2 by considering different measures of stability and compare these measures between treatments.

3 The effect of framing in learning to forecast experiments

We present our main findings in this section, starting with providing a first overview of the experimental results in Section 3.1. In Section 3.2 we investigate whether asking for returns instead of asking for prices, or providing past returns instead of providing past prices, has an effect on price dynamics. Finally, we study how the subjects' prediction strategies vary across treatments in Section 3.3.

3.1 An overview of the experimental results

Previous learning to forecast experiments with positive expectations feedback are characterized by persistent deviations of realized asset prices from their fundamental values, often leading to the endogenous emergence of bubbles and crashes, see e.g. Heemeijer et al. (2009), Bao et al. (2012) and Bao et al. (2017). These three earlier contributions use an experimental design that is essentially the same as our treatment PP.[16] Our other three treatments differ from that benchmark treatment in how the task for the subjects and/or how the information provided to them is framed. Our objective is to understand whether and, if so, how price volatility and the incidence of bubbles and crashes is affected by these differences.

Figures 2–5 show both market prices and returns (red and blue curves, respectively) in each of the 31 markets of the experiment. In addition, these figures show the individual forecasts (black curves) of the six subjects constituting a market (that is, forecasted prices for treatments PP and RP in Figures 2 and 3, and forecasted returns for treatments PR and RR in Figures 4 and 5). These figures illustrate that there is substantial heterogeneity in the development of market prices, both within treatments as well as between treatments. In some markets prices are quite stable, and remain in the vicinity of the fundamental value of 66 throughout the experiment (e.g. market PP4), in other markets there are persistent oscillations around this fundamental value, with no apparent tendency to converge (e.g market RR8). Unfortunately, comparisons between the treatments are somewhat obfuscated by the presence of *outliers*: individual

[16] The only differences are the realization of the demand shocks ε_t, the precise value of the fundamental value and the incentive scheme.

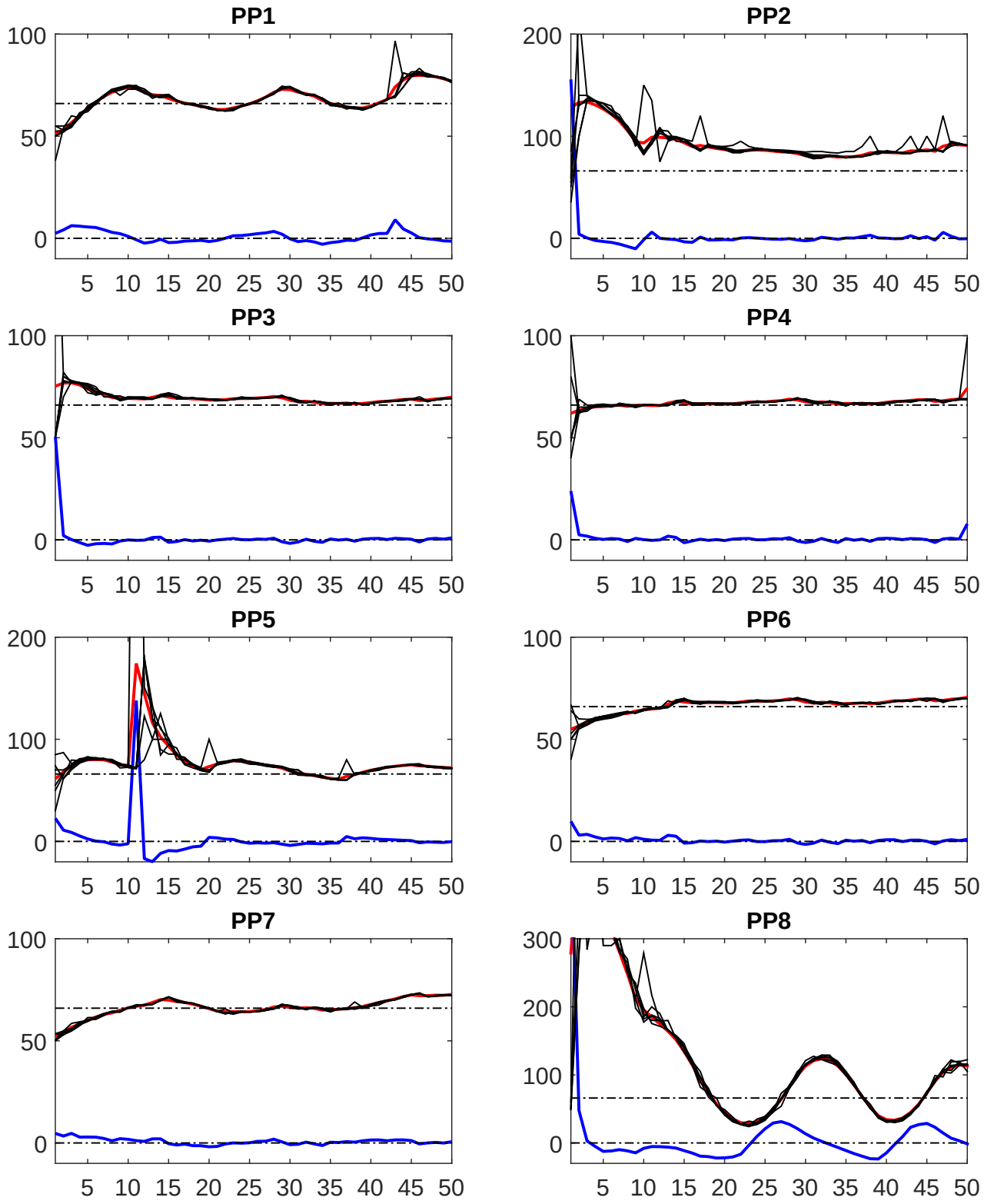


Figure 2: Price forecasts (black), prices (red) and returns (blue) in treatment PP. Note the different vertical scaling for markets PP2, PP5 and PP8.

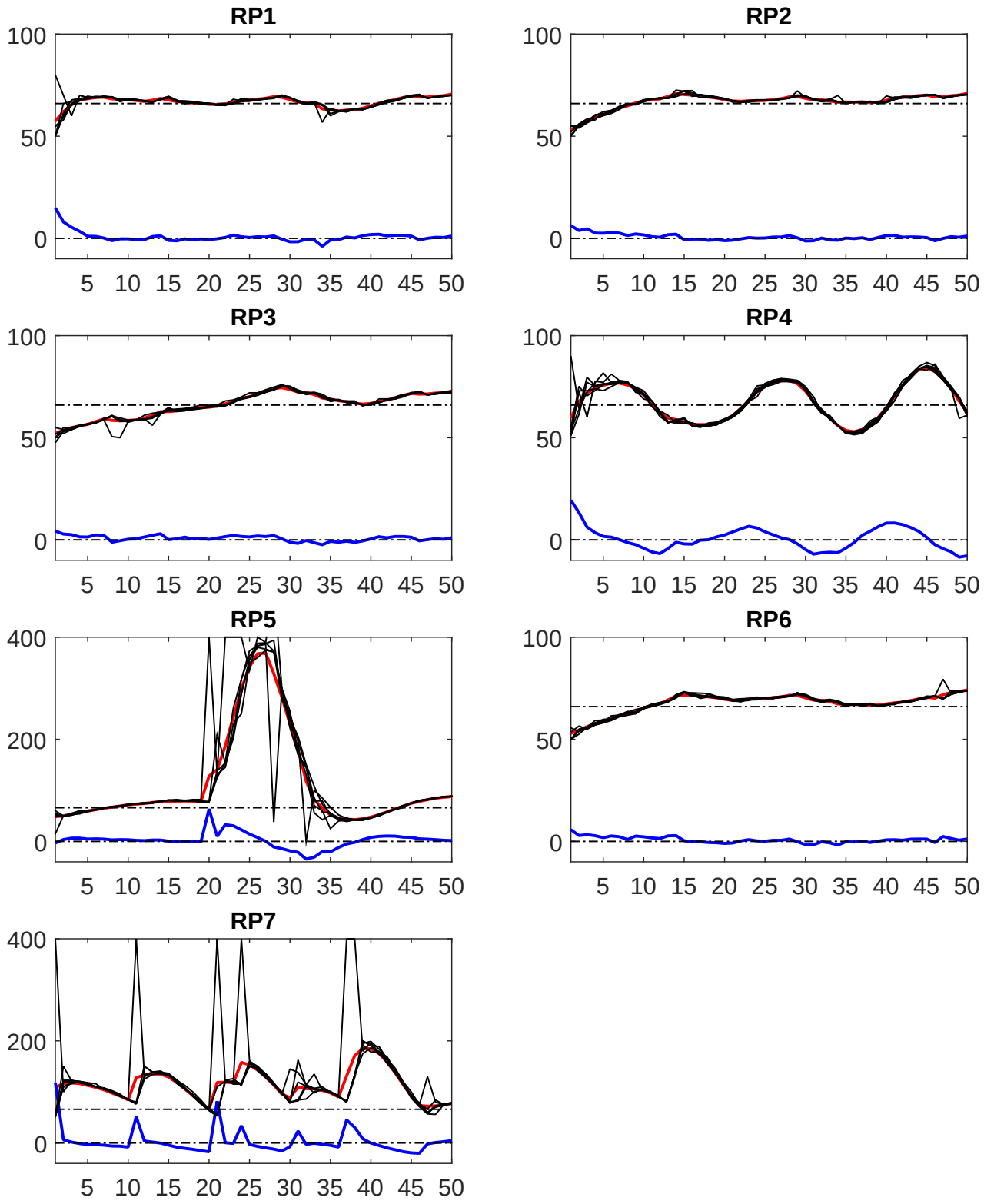


Figure 3: Price forecasts (black), prices (red) and returns (blue) in treatment RP. Note the different vertical scaling for markets RP5 and RP7.

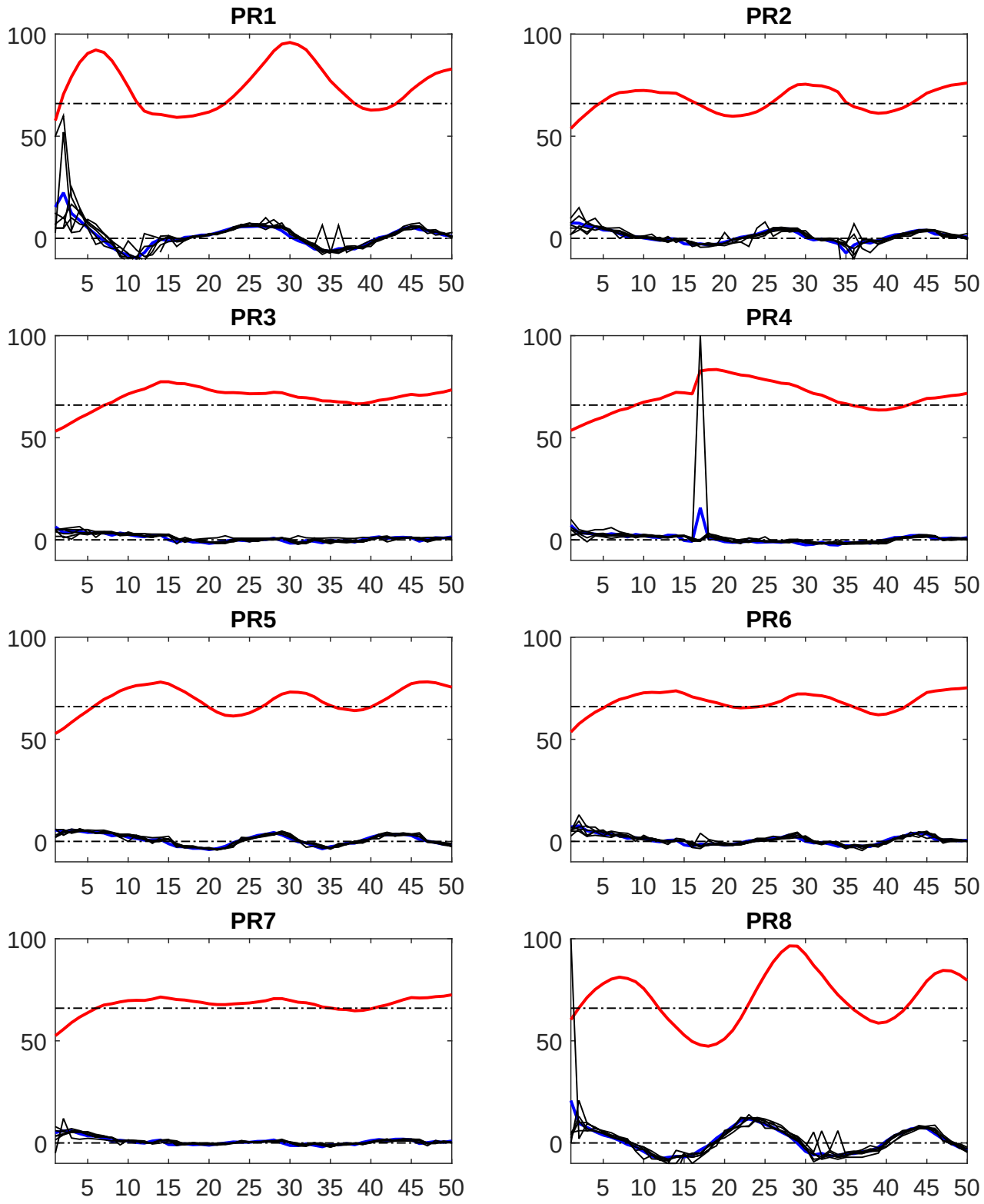


Figure 4: Return forecasts (black), prices (red) and returns (blue) in treatment PR.

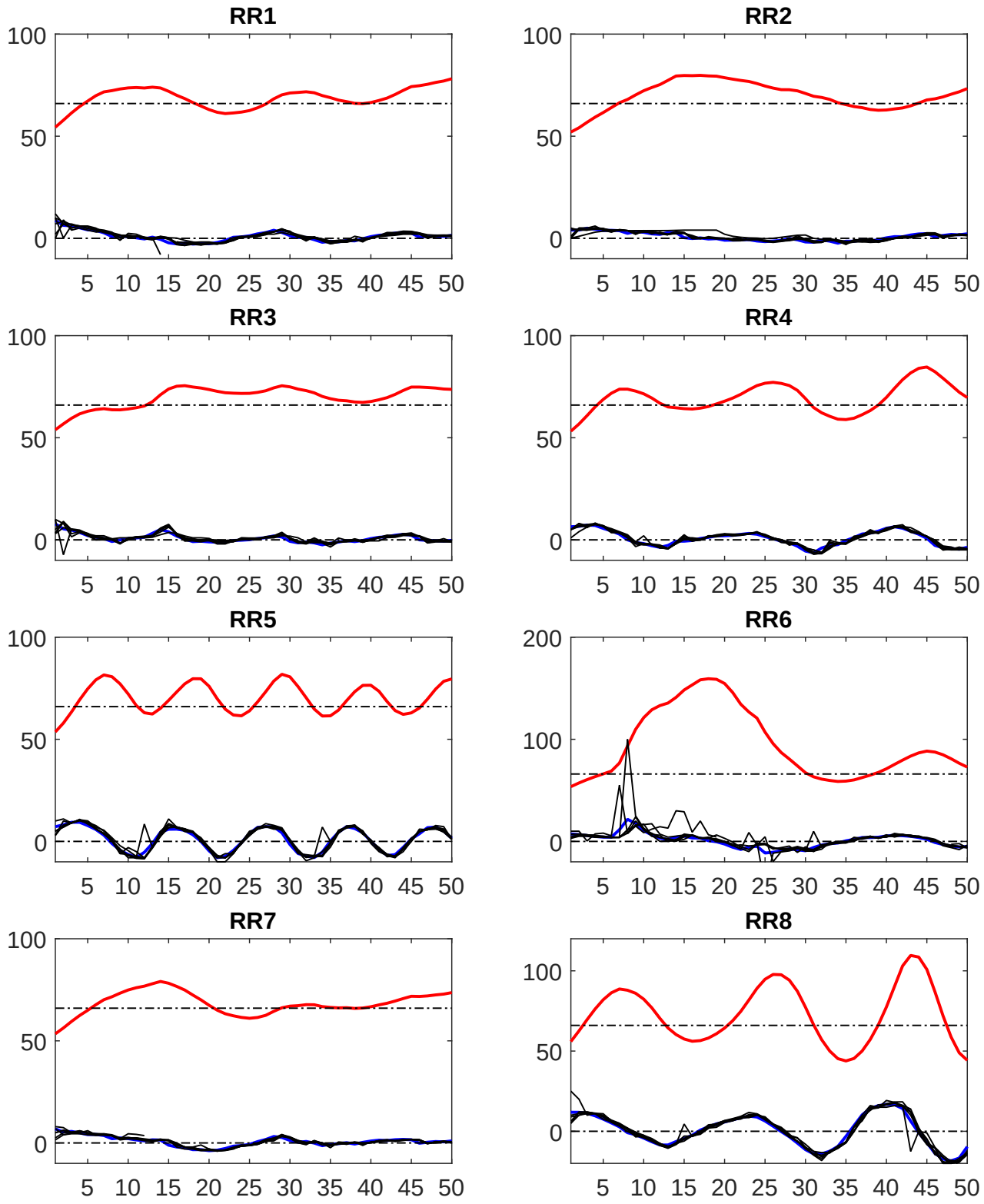


Figure 5: Return forecasts (black), prices (red) and returns (blue) in treatment RR. Note the different vertical scaling for markets RR6 and RR8.

forecasts that deviate substantially from the forecasts of the other subjects in the same period. These outliers may be due to typos by the subjects, or because a subject wants to experiment with the decision environment.[17] Although the incidence of outliers is quite low (comprising less than 0.3% of all submitted forecasts), such an outlier may have a prolonged effect on the dynamics of realized prices and returns. This occurs most often in treatments PP and RP where the price dynamics seem to be affected permanently in five of the fifteen markets (markets PP2, PP5, PP8, RP5 and RP7), whereas there only seems to be a lasting impact in two out of sixteen markets in treatments PR and RR (markets PR4 and RR6).

Notwithstanding the existence of these outliers, visual inspection of Figures 2–5 allows us to draw some preliminary conclusions. First, it seems that deviations from fundamental values and persistent oscillations in prices and returns are more prevalent in the treatments where subjects have to forecast returns than in treatments where they have to forecast prices. For example, in seven of the fifteen markets in treatments PP and RP prices are within 10% of the fundamental value (that is, in the interval $[59.4, 72.6]$) for at least 40 consecutive periods (this happens for markets PP3, PP4, PP6, PP7, RP1, RP2 and RP6). Moreover, for five of the other eight markets outliers seem to play a role in generating the oscillations. On the other hand, in treatments PR and RR most markets exhibit prominent fluctuations, with market PR7 the only market where prices are within 10% of the fundamental value for at least 40 consecutive periods, whereas in only two of the other fifteen markets in these treatments fluctuations are driven by outliers. Second, the figures do not provide clear evidence on the effect of observing returns versus observing prices on volatility of prices and returns. Finally, and consistent with earlier positive feedback learning to forecast experiments (e.g. Hommes et al., 2005, Heemeijer et al., 2009), both in treatments where prices have to be forecasted and in those where returns have to be forecasted there is substantial coordination between individual forecasts in the same market, with most predictions close together in almost all periods. This is remarkable because subjects do not see each others forecasts, but only the aggregate price or return.

3.2 Price stability

In this section we investigate more thoroughly whether the stability of aggregate market prices is affected by the framing of the task and/or information given to the subjects. To that end we use different instability

[17] For example, subject 1 in market PP5 submits a price forecast of 721.05 in period 11, which is likely a typo: The prediction of the five other subjects for that period are between 71.06 and 72.50, and the last observed price is 73.31. Subject 1's intention may well have been to submit 72.05 instead. However, the vast majority of the outliers in the experiment seem to be due to experimentation. Subject 4 in market RP5, for example, has six price predictions of either 400 or 500 (these are much higher than previous prices and the predictions of other subjects) between periods 20 and 28 and a price prediction of 0 in period 32.

measures, which we calculate on the basis of market prices from period 11 to period 50. We exclude the first ten periods to allow for some learning.

We consider five different measures. Our first two measures are standard measures of price dispersion and price volatility: the *standard deviation* of the market price (*std*) and the *price range*, which is the difference between the highest and lowest realized price over the sample (*range*). That is

$$std = \sqrt{\tfrac{1}{40}\sum_{t=11}^{50}(p_t - \bar{p})^2}, \quad \text{and} \quad range = \max\{p_t\}_{t=11}^{50} - \min\{p_t\}_{t=11}^{50},$$

where $\bar{p} = \frac{1}{40}\sum_{t=11}^{50} p_t$ is the average realized asset price over the last 40 periods. Note that our measure *range* is closely related to the (price) *amplitude* measure introduced in Porter and Smith (1995).[18]

Our third measure, AR, is based upon returns instead of prices and is equal to the median of the absolute returns between periods 11 and 50, that is[19]

$$AR = \text{median}_{t\in\{11,\ldots,50\}} |r_t| .$$

As with the *std* and *range* measures, a higher value of AR implies higher price volatility. We take the median of absolute returns in AR in order to restrict the effect of outliers – note that outliers in prices will have a substantial effect on *std* and *range*.

The three measures discussed above measure price volatility, but do not necessarily capture *mispricing* (i.e. deviations of realized prices from the fundamental value) very well. For example, if prices are relatively constant but at a level substantially different from the fundamental value these measures will be low, whereas mispricing will be significant. Our final two measures, *relative absolute deviation* from the fundamental value (RAD) and *relative deviation* from the fundamental price (RD), do take such deviations from the fundamental value into account (they were introduced in Stöckl et al., 2010, and have become standard measures of mispricing and overvaluation, respectively, in the literature on experimental

[18] The price amplitude measure used by Porter and Smith (1995), applied to experimental asset markets with trading, is

$$\max_t \left\{\frac{p_t - p_t^*}{p_1^*}\right\} - \min_t \left\{\frac{p_t - p_t^*}{p_1^*}\right\},$$

where p_t is the mean transaction price in period t, p_t^* is the fundamental value in period t and p_1^* is the initial fundamental value. As opposed to the majority of experimental asset markets, where the fundamental value decreases over time, in our case the fundamental value is constant, meaning that for our experiment the price amplitude measure equals our *range* measure, divided by $p^* = 66$.

[19] In financial market research absolute returns are also used frequently, as they are found to predict future return volatility quite well (see e.g. Ghysels et al., 2006).

asset markets). These measures are defined as

$$RAD = \frac{1}{40}\sum_{t=11}^{50}\frac{|p_t - p^*|}{p^*} \text{ and } RD = \frac{1}{40}\sum_{t=11}^{50}\frac{p_t - p^*}{p^*},$$

where, in our experiment, $p^* = 66$. Note that a high value of RAD means that the asset price deviates persistently from the fundamental value. This can be accompanied by a high value of RD (for example, if the asset is structurally overvalued, see e.g. market PP3, where $p_t > p^*$ for all $t \geq 2$ and consequently $RD = RAD$ for this market) or a low value of RD (if the asset price oscillates around the fundamental value, see e.g. market PR8, for which RAD is almost three times RD).

Tables 7-9 in Appendix C report the values of the five measures discussed above for each of the 31 markets.[20] Note that every market corresponds to one observation, and we therefore have eight observations for treatments PP, PR and RR and seven for treatment RP. Although the variation in each of the measures is substantial, the measures appear to be – to a large extent – mutually consistent.[21] In particular, for each of the five measures seven of the eight markets with the lowest value of that measure are either from treatment PP or from treatment RP, that is, from treatments where subjects have to forecast the price.[22] Not surprisingly, these treatments are also overrepresented under the highest values of the different measures, but this is predominantly due to the fact that most of the markets that are effected by outliers are from those two treatments.[23]

For each measure we use Kolmogorov-Smirnov tests to see whether there are statistically significant differences between the treatments. To test Hypothesis 1 (differences in forecasting price or return) we

[20] Note that these measures will still be nonzero when every subject predicts the fundamental value, $p^f_{h,t} = p^*$ for every h and t, due to the small random shocks in the price generating mechanism. In fact, the measures will then be equal to $std = 0.50$, $range = 2.20$, $AR = 0.01$, $RAD = 0.01$ and $RD = 0.00$ based on the actual noise sequence we used in the experiment.

[21] Correlations between the measures are quite high, in particular between *std* and *range* ($\rho = 0.99$) and between RAD and RD ($\rho = 0.97$), with the correlation between any two of those four measures at least equal to 0.86. The lowest correlations are between AR and the other measures, with those correlations ranging between 0.55 (AR and RD) and 0.73 (AR and RAD). Part of this difference may be due to the markets with outliers. If we exclude those seven markets the correlations between each of the *std*, *range*, AR and RAD measures is at least 0.91. In this case the correlations with RD are lower, and range between 0.37 (RD and AR) and 0.60 (RD and RAD).

[22] For the measures *std*, *range*, AR and RAD the lowest values are obtained by markets PP3, PP4, PP6, PP7, RP1, RP2, RP6 and PR7. The lowest values for measure RD are obtained by markets PP4, PP6, PP7, RP1, RP2, RP3, RP4 and PR2. Note that the three markets with a low value of RD, but relatively higher values for the other measures (markets RP3, RP4 and PR2) are all markets where, although the price does fluctuate considerably, it oscillates around the fundamental value, which reduces the value of RD.

[23] For each of the measures the five markets with the highest value of that measure are markets affected by outliers. The only exceptions are markets RR8 and PR8 which have the second and fifth highest value of AR, respectively.

all data	**std**	**range**	**RAD**	**RD**	**AR**
*P vs *R	0.0592*	0.0592*	0.0215**	0.0215**	0.0231**
P* vs R*	0.9854	0.9973	0.9643	0.7351	0.9346
excl. outlier markets	**std**	**range**	**RAD**	**RD**	**AR**
*P vs *R	0.0052***	0.0052***	0.0008***	0.0008***	0.0015***
P* vs R*	0.9094	0.9094	0.9094	0.6840	0.9094

Notes: ***: significant at 1% level, **: significant at 5% level, *: significant at 10% level. All test are one-sided. Observations correspond to markets, the number of observations is $n_{*P} = 15$, $n_{*R} = 16$, $n_{P*} = 16$ and $n_{R*} = 15$ in the upper panel and $n_{*P} = 10$, $n_{*R} = 14$, $n_{P*} = 12$ and $n_{R*} = 12$ in the lower panel.

Table 2: Summary of p values in the Kolmogorov-Smirnov tests for comparing treatments in terms of instability.

merge treatments PP and RP (into *P) on the one hand and PR and RR (into *R) on the other hand, and to test Hypothesis 2 (differences in observing price or return) we merge PP and PR (into P*) and RP and RR (into R*). The test results are collected in Table 2.[24] Note that we apply one-sided Kolmogorov-Smirnov tests, with the direction given by Hypotheses 1 and 2 – e.g. we test whether the value of measure *std* is significantly smaller for the markets in *P than for the markets in *R, since this is what Hypothesis 1 predicts.

From Table 2 we see that differences between *P and *R are significant at the 5% level for three measures (*AR*, *RAD* and *RD*) and for the other two they are significant at the 10% level. On the other hand we do not find significant differences between observing prices and observing returns. The picture is somewhat distorted by the fact that most markets with outliers (which lead to less stable price dynamics) happen to be in the treatments where prices are forecasted. Excluding the markets with outliers (that is, excluding markets PP2, PP5, PP8, RP5, RP7, PR4 and RR6) makes the differences between forecasting prices and forecasting returns much more apparent: The difference between *P and *R is now significant at the 1% level for all five measures.

Hypothesis 1 therefore is supported by the data, whereas Hypothesis 2 is not. To corroborate this finding we use the five measures discussed above to characterize the number of stable markets in each treatment and test whether this number is different for different merged treatments. For each measure we rank the 31 markets by giving the market with the lowest value for that measure rank 1, and so on, up to

[24] We also make pairwise comparisons between the four individual treatments, the corresponding test results are presented in Appendix D.

	all data	**excl. outlier markets**
*P vs *R	0.0062***	0.0009***
P* vs R*	0.7672	0.8114

Notes: ***: significant at 1% level, **: significant at 5% level, *: significant at 10% level. All test are one-sided. Observations correspond to markets, the number of observations is $n_{*P} = 15$, $n_{*R} = 16$, $n_{P*} = 16$ and $n_{R*} = 15$ in the left column and $n_{*P} = 10$, $n_{*R} = 14$, $n_{P*} = 12$ and $n_{R*} = 12$ in the right column.

Table 3: Summary of p values in the Wilcoxon rank-sum test for comparing treatments in terms of the number of stable markets.

rank 31 for the market with the highest value of that measure. Subsequently we order all 31 markets by the average rank they have for the five measures, which gives the following stability ranking: PP3, PP4, PP6, RP2, RP1, PP7, PR7, RP6, RP3, PR6, PR3, PR2, RR7, RR3, RR1, PP1, PR5, RP4, RR2, PR4, PP2, RR4, RR5, PR1, PR8, PP5, RR8, RR6, RP7, PP8 and RP5, with PP3 the most stable, and RP5 the most unstable market. Based upon the time series in Figures 2-5 we classify the first eight of these markets as *stable*, and the remaining 23 markets as *unstable*.[25] Note that seven of the eight stable markets are from treatments PP and RP again (with market PR7 the only exception). Table 3 collects p values of Wilcoxon rank-sum tests on the differences in the number of stable markets between treatments. The results are consistent with our earlier findings: The differences between predicting prices and predicting returns (*P vs *R) is significant at the 1% level and there is no difference between observing prices and observing returns (P* vs R*). Excluding the seven markets with outliers gives more significant results again.

Based on the above our main result is the following:

Result 1 *Forecasting prices leads to more stable price dynamics than forecasting returns. It does not matter for price stability whether subjects observe prices or returns.*

A possible reason for the absence of an effect of observing prices versus observing returns might be that in each treatment of our experiment we show the most recent price to the subjects in each period.

[25]Markets PP3, PP4, PP6 and RP2 are clearly stable. We also classify markets RP1, PP7, PR7 and RP6 as stable as they exhibit only minor oscillations around the fundamental value. Market RP3 features steadily increasing prices until period 30 and market PR6 exhibits clear oscillations, which is why we classify these two markets as unstable. The other markets, starting with PR3, clearly show an unstable pattern. However, our results are qualitatively the same when we also classify RP3 and PR6 as stable markets.

This is in line with the design chosen in Glaser et al. (2018) but there is an important difference. In Glaser et al. (2018) subjects have to make a one-time forecast: In their treatments RP and RR subjects could observe all past returns and only the most recent price. In our experiment, however, subjects have to make a forecast for 50 consecutive periods so even though we show them only the most recent price in treatments RP and RR, they can write them down and essentially have the possibility to use all past prices for forecasting.

Having established that framing of the forecasting task does affect subject behavior and aggregate market dynamics, we will try to explain this effect in the next subsection.

3.3 Forecasting task and trend extrapolation

We start by investigating how subjects respond to price changes in the different treatments. Figure 6 shows a scatter plot of $p^f_{h,t+1} - p_t$ against $p_t - p_{t-1}$ for the four treatments.[26] Obviously, there is a strong positive relation between the expected change in the price and the last observed price change. That is, in each of the four treatments subjects have a tendency to extrapolate trends: if they observe a price increase (decrease) in the previous period they expect that the price will again increase (decrease) in the current period.

The slopes of the corresponding linear regressions, together with their 95% confidence intervals, are reported in Table 4. The differences in slopes are significant, with the slopes higher in treatments RR and PR – where they are relatively close to 1 – than in treatment PP and, in particular, treatment RP.[27] We therefore find that, although trend extrapolation plays a role in each treatment, it is clearly stronger in treatments where returns need to be forecasted. In those treatments subjects believe, on average, that a price change will continue into the next period almost one-for-one. In the treatments where prices need to be forecasted, in particular in treatment RP, there is a stronger tendency for subjects to believe

[26] Return forecasts in treatments PR and RR are transformed to price forecasts by $p^f_{h,t} = \left(1 + r^f_{h,t}\right) p_{t-1}$. As before, we use the data of the last 40 periods only. Moreover, we remove the outliers from the data set to get a clearer picture about the relation between the variables. A forecast is classified as an outlier according to the following rule. If the most recent price change was positive, the forecast $p^f_{h,t}$ is an outlier if it exceeds the most recent price by at least 25% or if it is lower than the most recent price by at least 5%: $p^f_{h,t} \notin (0.95p_{t-1}, 1.25p_{t-1})$. Similarly, if the most recent price change was negative, the forecast is considered an outlier if it exceeds the most recent price p_{t-1} by at least 5% or if it is lower than the most recent price by at least 25%: $p^f_{h,t} \notin (0.75p_{t-1}, 1.05p_{t-1})$. Using this rule the following number of forecasts per treatment are excluded: PP: 24 (1. 25%); RP: 41 (2.53%); PR: 8 (0.42%) and RR: 12 (0.63%).

[27] The confidence intervals already show that the slopes are significantly different but we also performed regressions with treatment interaction terms to compare slopes between pairs of treatments. Most p values of the interaction terms are 0.0000 with the exception of PP vs. PR (0.0088) and PR vs. RR (0.0052). Thus, differences in reactions to price changes are highly significant.

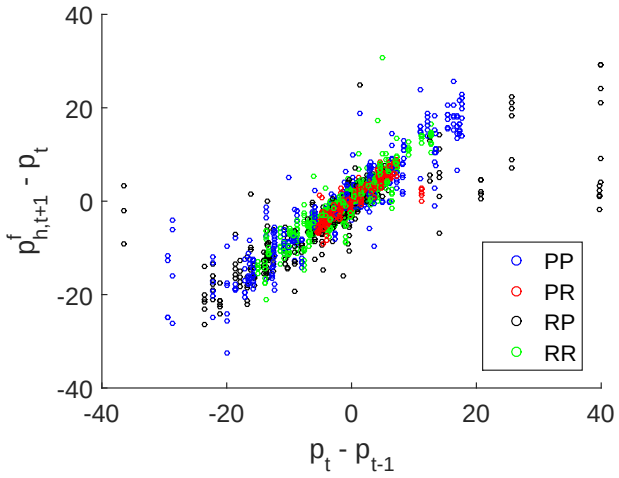


Figure 6: A scatter plot of $p^f_{h,t+1} - p_t$ vs. $p_t - p_{t-1}$.

	slope	confidence interval
PP	0.8765	(0.8584, 0.8947)
RP	0.5839	(0.5624, 0.6053)
PR	0.9243	(0.9057, 0.9430)
RR	0.9670	(0.9477, 0.9862)

Table 4: Reactions to price changes: estimated slopes and confidence intervals.

that, although the change in price will continue, the price change will decrease in size. This is consistent with Glaser et al. (2007), who also explain their results by stronger trend extrapolation (in prices) when subjects have to forecast returns.

We therefore have the following result:

Result 2 *Subjects tend to extrapolate trends in past price changes more strongly when they need to forecast returns than when they need to forecast prices.*

Figure 6 and Table 4 present aggregate forecasting behavior. We can also investigate forecasting behavior at the individual level. To that end we estimate the following forecasting rule for each individual subject

$$p^f_{h,t+1} = C_h + \sum_{l=0}^{3} \beta_{hl} p_{t-l} + \sum_{l=0}^{3} \gamma_{hl} p^f_{h,t-1} + \varepsilon_{h,t+1}, \tag{2}$$

on data from the last 40 periods of the experiment. Note that we have 48 subjects in treatments PP, PR and RR and 42 subjects in treatment RP. For each of these 186 subjects we estimate the forecasting model (2). Tables 5 and 6 summarize the results (individual estimation results are available upon request).

	C	p_t	p_{t-1}	p_{t-2}	p_{t-3}	p^f_t	p^f_{t-1}	p^f_{t-2}	p^f_{t-3}
PP	11	47	34	14	10	16	11	6	13
RP	12	42	30	13	4	17	12	9	7
PR	9	48	43	18	8	15	4	0	7
RR	16	48	46	28	16	12	8	8	11

Table 5: Number of subjects with significant coefficient

	C	p_t	p_{t-1}	p_{t-2}	p_{t-3}	p^f_t	p^f_{t-1}	p^f_{t-2}	p^f_{t-3}
PP	1.83	1.64	-0.64	0.01	0.02	0	-0.03	0	-0.02
RP	1.58	1.65	-0.59	-0.02	0	0	-0.05	0	-0.02
PR	0.41	1.94	-1.11	0.05	0.04	0.1	-0.02	0	0
RR	0.51	2.28	-1.65	0.39	0.01	0.02	-0.05	-0.02	0

Table 6: Average coefficients over all subjects

Table 5 presents, for each variable in (2), the number of subjects in each treatment for which the coefficient on that variable is significantly different from zero at the 5% level. Variables p_t and p_{t-1} appear most often: For all but one subject the coefficient on p_t is significantly different from zero, and for more than 80% of the subjects the coefficient on p_{t-1} is significantly different from zero. In addition, variables p_{t-2} (in particular for treatment RR) and $p^f_{h,t}$ feature regularly, but the coefficient of none of the other variables is significantly different from zero for more than a quarter of the subjects.

To better understand the impact that p_t and p_{t-1} (as well as the other variables) have on the forecasts of the subjects, Table 6 presents for each variable the average value of the estimated coefficients on that variable for the different treatments.[28] Some features stand out from this table. First, with the exception of the average coefficient of p_{t-2} in treatment RR, the average estimated values of the coefficients of p_t and p_{t-1} (and of the constant) are substantially larger (in absolute value) than those of the other variables. Second, the average estimated forecasting rule in treatments PP, RP and PR is close to the trend extrapolation rule

$$p^f_{t+1} = p_t + \theta_0 \left(p_t - p_{t-1}\right),$$

with values of θ_0 of around 0.64, 0.59 and 1.11 for treatments PP, RP and PR, respectively. For treatment

[28] The average is calculated over all subjects in the given treatment. If a variable is insignificant in the regression for a given subject, then its coefficient is considered as 0 when calculating the average over all subjects.

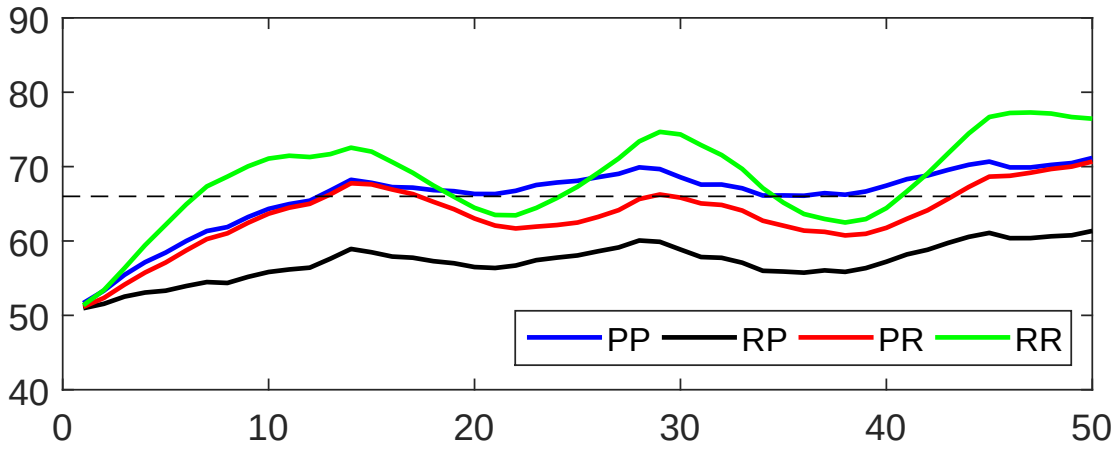


Figure 7: Simulated prices, for each treatment based upon the "average" forecasting rule from Table 6. The dashed line show the fundamental price.

RR the average estimated forecasting rule is close to the more general trend extrapolation rule

$$p^f_{t+1} = p_t + \theta_0 (p_t - p_{t-1}) + \theta_1 (p_{t-1} - p_{t-2}),$$

with $\theta_0 = 1.26$ and $\theta_1 = -0.39$. Note that the main trend extrapolation parameter θ_0 is much higher (and larger than one) for treatments PR and RR than for treatments PP and RP. This is consistent with our finding that the tendency to extrapolate trends is stronger when subjects have to forecast returns.

To illustrate this further, for each treatment we run a simulation of model (2), where we assume that all participants use the same forecasting rule, namely the one given in Table 6 for that treatment. Figure 7 shows the resulting hypothetical dynamics for each of the four treatments – where the same realization of shocks is used as in the experiment.[29] From the figure we see that these hypothetical dynamics are quite stable for the treatments where the price is forecasted, but the simulated prices oscillate more for the treatments where returns have to be forecasted, in particular for treatment RR.[30] What is also apparent is that for the simulated prices there is typically overvaluation for treatment RR, constant undervaluation for treatment RP whereas the simulated price is close to or oscillates around the fundamental value for the other two treatments.

The analysis thus far suggests that subjects in return forecasting treatments tend to use forecasting

[29]For each simulation we use the same initial conditions, with $p_t = 50$ for $t \leq 0$ for prices, and $p^f_{h,t} = 50$ for $t \leq 0$ for predictions.

[30]When computed for (the last 40 periods of) the simulated time series, the values of *std*, *range* and *AR* are lower in PP and RP than in PR and RR. Mispricing, measured by *RAD* and the absolute value of *RD*, is the lowest in PP and PR and the highest in RP.

rules that extrapolate trends stronger. We next investigate to what extent subjects succeed in coordinating on the same forecasting rule. To that end we use the following decomposition of average individual errors (introduced in Hommes et al., 2005) for each experimental market:

$$\frac{1}{240}\sum_{h,t}\left(\frac{p^f_{h,t}-p_t}{p_{t-1}}\right)^2 = \frac{1}{240}\sum_{h,t}\left(\frac{p^f_{h,t}-\overline{p}^f_t}{p_{t-1}}\right)^2 + \frac{1}{40}\sum_{t}\left(\frac{\overline{p}^f_t-p_t}{p_{t-1}}\right)^2 \tag{3}$$

The left hand side of (3) gives, for that particular market, the *average individual relative quadratic forecast error* in prices (recall that subjects are incentivized to minimize this relative quadratic error), averaged both over the six subjects in that market and the last 40 periods. The right hand side expresses this number as the sum of the *average dispersion error*, which measures the variation between individual forecasts, and the *average common error*, which measures how far the average forecast lies from the predicted price. If the average common error explains most of the individual forecast errors, then these forecast errors are positively correlated, and subjects deviate from the correct forecast in similar ways. However, if the average common error is small relative to the average dispersion error, then subjects are approximately correct on average and their forecast errors tend to cancel each other out.[31]

Tables 22-25 in Appendix F summarize the results. Obviously, the largest average individual forecast errors can be found in markets where outliers played an important role and the smallest average individual forecast errors among the markets that we classified as stable, although also some of the unstable markets have surprisingly small average individual forecast errors (e.g. markets RR1 and RR3). About two-thirds (68%) of the average individual forecast error can be explained by the average common error, and for only five of the 31 markets the average dispersion error is larger than the average common error. Moreover, there are no significant differences between (combinations of) treatments in the fraction of the error explained by the common error. From this we conclude that which variable is being observed or forecasted has little impact on the level of coordination. Moreover, in the same treatment subjects tend to coordinate on the same forecasting rules, although these rules extrapolate trends stronger in treatments where returns have to be forecasted.

4 Concluding remarks

In this paper we investigated the effect of the format of the forecasting task and of the format of past information on aggregate market dynamics in a laboratory experiment where realized prices/returns are

[31] When someone did not submit a forecast on time, then his/her 'forecast' is not considered in the average individual and dispersion errors (i.e. the error is considered 0 and we divide by less than 240). The definition of the dispersion error is also slightly modified. In addition, we correct for outliers, as discussed before.

determined by price/return expectations. Although we do not find evidence that the format of past information, which is either presented as a return bar chart or as a price line chart, has a notable impact on aggregate price dynamics, the format of the task (either forecasting prices, or forecasting returns) does have a significant effect. In particular, when subjects are asked to forecast returns, they tend to coordinate on expectation rules that exhibit stronger extrapolation of past trends than when they are asked to forecast prices. This leads to larger price volatility and a higher incidence of bubbles and crashes in those treatments. Earlier empirical research has already shown that financial market participants have a tendency to extrapolate trends in past performance, see e.g. Sirri and Tufano (1998), Choi et al. (2009) and Greenwood and Shleifer (2014). Our results suggest that this tendency increases when investors think in terms of returns instead of prices, and that this may have a substantial adverse impact on financial market stability.

Andreassen (1987, 1988) and Glaser et al. (2007) refer to the representativeness heuristic (see Tversky and Kahneman, 1982) to argue that subjects that think in prices are more likely to predict mean reversion (in prices) than subjects that think in price changes or returns. To illustrate their point, consider the following sequence of monotonically increasing prices: 60, 62, 65, 68, 70. The representative (e.g. mean or median) price is around 65. However, the representative return is around 4% (the returns corresponding to this series of prices are 3.33%, 4.84%, 4.62% and 2.94%, respectively), which translates into a price forecast of about 72.8. That implied price forecast gives rise to trend extrapolation, instead of the mean reversion resulting from a price forecast of 65. Similarly, if we believe subjects have naïve or adaptive expectations (i.e. their forecast is equal to the last realized value of that variable, or it is equal to a weighted average of that last observation and their previous forecast), clearly trends in prices will be extrapolated when returns are forecasted, but not when prices are forecasted. Both explanations are consistent with the stronger trend extrapolation we find in treatments PR and RR.

However, the argument based upon the representativeness heuristic also suggests that the format of past information should have an effect on forecasting behavior: returns are much more salient in treatments RP and RR than in treatments PP and PR. However, we do not find evidence for a significant difference between those treatments. We can think of two possible reasons for this. First, independent of the format of past information, subjects may focus on the variable that they need to forecast. If required, they translate the variable that they observe into the variable that they need to forecast. This would diminish the effect of the format of past information, and would be consistent with the mixed results on the effect of the chart format in the existing experimental literature that we discussed in the Introduction. Second, for treatments RP and RR we provided the subjects, in addition to the bar chart of past returns, also with the most recent price. Following Glaser et al. (2018), we did this in order not to make the forecasting task

in treatment RP overly complicated: if subjects only know the initial price value of 50 they would have to use all past returns to compute the current price level. The effect of the format of past information might be diminished because subjects saw the last price in each of the treatments. Notwithstanding our results, for actual financial markets the format of the presentation of past information may still have a real effect. In our experiment we impose the variable that subjects need to forecast, but in actual financial markets this is – to a certain extent – up to the market participant itself (except professional forecasters and financial analysts that are asked for specific predictions by investors). One may imagine that investors that observe past returns will be more inclined to forecast future returns than investors that observe past prices – making treatments PP and RR the more relevant treatments in our experiment. In this way, the format of the presentation of past information may still have an impact, albeit indirect, upon actual financial market dynamics.

Many learning to forecast experiments with positive expectations feedback feature persistent deviations from fundamentals and the emergence of bubbles and crashes. In those experiments subjects typically observe past prices and have to forecast future prices, as in our treatment PP. Our results show that if return forecasts are elicited instead of price forecasts, these persistent deviations from the fundamental value are exacerbated, although the underlying price/return generating mechanism remains exactly the same.

Two final remarks about the model we chose for investigating forecasting behavior are in order here. First, as in previous learning to forecast experiments the feedback strength we choose is relatively high: the relevant coefficient in equation (1) is 1/1.05, which is approximately equal to 0.95, implying that the realized price (return) will be quite close to the average price (return) forecast. Earlier work on learning to forecast experiments with positive expectations feedback has shown that a smaller value of this feedback strength will mitigate the endogenous emergence of bubbles and crashes in this framework. In fact, deviations from fundamentals quickly vanish and prices converge to fundamentals if the feedback coefficient is about 0.70 or less (see Sonnemans and Tuinstra, 2010, and Bao and Hommes, 2015).[32] From our results we conjecture that, in a learning to forecast experiment where return forecasts are elicited, the feedback strength would have to be even lower to induce convergence to the fundamental value. Second, in our learning to forecast experiment we focus exclusively on expectation formation, whereas in other experimental studies on the endogenous emergence of bubbles and crashes, subjects can buy and sell the asset (see Palan, 2013 for the sizable literature on bubbles in experimental asset markets pioneered by Smith et al., 1988). However, in a recent study, Bao et al. (2017) show that the bubbles and crashes that

[32]Note that for the underlying financial market model such a low feedback strength would coincide with an interest rate of about 43% or higher, which seems quite substantial.

emerge in learning to forecast experiments with positive feedback are robust when subjects can also trade in that asset. In addition, Amromin and Sharpe (2014) and Greenwood and Shleifer (2014) show that portfolio choices can be explained to a large extent by survey expectations.[33] We therefore believe that our results will translate to a situation where subjects also have the possibility to trade. We leave it to future work to investigate this issue.

Acknowledgments

We would like to thank Cees Diks, John Duffy, Martin Weber and participants at the conference Experimental Finance 2017, Nice, France, the 22nd Annual Workshop on Economic Science with Heterogeneous Interacting Agents, Milan, Italy, the 23rd International Conference on Computing in Economics and Finance, New York, U.S.A., the 2017 BEAM Project Workshop in Kyoto, Japan, the 11th Maastricht Behavioral and Experimental Economics Symposium, Maastricht, the Netherlands, the 24th International Conference on Computing in Economics and Finance, Milan, Italy and the 2019 Chapman University Workshop on Experimental & Behavioral Aspects of Financial Markets, Orange, U.S.A. for useful suggestions and comments. This research was funded by the ANR/NWO ORA-Plus project "BEAM" (Behavioral and Experimental Analysis in Macro-finance, ANR-15-ORAR-0004, NWO 464-15-143) as well as by the French government-managed l'Agence Nationale de la Recherche under Investissements d'Avenir UCA^{JEDI} (ANR-15-IDEX-01). In particular, we thank the UCAinACTION project.

References

Amromin, G. and Sharpe, S. A. (2014). From the horse's mouth: Economic conditions and investor expectations of risk and return. *Management Science*, 60(4):845–866.

Andreassen, P. B. (1987). On the social psychology of the stock market: Aggregate attributional effects and the regressiveness of prediction. *Journal of Personality and Social Psychology*, 53(3):490.

Andreassen, P. B. (1988). Explaining the price-volume relationship: The difference between price changes and changing prices. *Organizational Behavior and Human Decision Processes*, 41(3):371–389.

Andreassen, P. B. and Kraus, S. J. (1990). Judgmental extrapolation and the salience of change. *Journal of forecasting*, 9(4):347–372.

[33] Also see Hanaki et al. (2018) who show that, in an experimental asset market with trade a la Smith et al. (1988), whether forecasts are explicitly elicited or not does not affect mispricing (as long as either the forecasting task or the trading task is rewarded). This suggests that, also in these experiments, forecasts are consistent with trading decisions.

Anufriev, M., Bao, T., Sutan, A., and Tuinstra, J. (2019). Fee structure and mutual fund choice: An experiment. *Journal of Economic Behavior & Organization*, 158:449–474.

Bacchetta, P., Mertens, E., and Van Wincoop, E. (2009). Predictability in financial markets: What do survey expectations tell us? *Journal of International Money and Finance*, 28(3):406–426.

Bao, T. and Hommes, C. (2015). When speculators meet constructors: Positive and negative feedback in experimental housing markets. Technical report, CeNDEF Working paper 15-10, University of Amsterdam, Netherlands.

Bao, T., Hommes, C., and Makarewicz, T. (2017). Bubble formation and (in) efficient markets in learning-to-forecast and optimise experiments. *The Economic Journal*, 127(605):F581–F609.

Bao, T., Hommes, C., Sonnemans, J., and Tuinstra, J. (2012). Individual expectations, limited rationality and aggregate outcomes. *Journal of Economic Dynamics and Control*, 36(8):1101–1120.

Beshears, J., Choi, J. J., Fuster, A., Laibson, D., and Madrian, B. C. (2013). What goes up must come down? Experimental evidence on intuitive forecasting. *American Economic Review*, 103(3):570–74.

Branch, W. (2004). The theory of rationally heterogeneous expectations: Evidence from survey data on inflation expectations. *Economic Journal*, 114:592–621.

Brock, W. A. and Hommes, C. H. (1998). Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *Journal of Economic Dynamics and Control*, 22(8-9):1235–1274.

Carroll, C. D. (2003). Macroeconomic expectations of households and professional forecasters. *Quarterly Journal of Economics*, 118(1):269–298.

Case, K. E., Shiller, R. J., and Thompson, A. (2012). What have they been thinking? Home buyer behavior in hot and cold markets. In *Brooking Papers on Economic Activity*, pages 265–315.

Choi, J. J., Laibson, D., Madrian, B. C., and Metrick, A. (2009). Reinforcement learning and savings behavior. *The Journal of Finance*, 64(6):2515–2534.

Czaczkes, B. and Ganzach, Y. (1996). The natural selection of prediction heuristics: Anchoring and adjustment versus representativeness. *Journal of Behavioral Decision Making*, 9(2):125–139.

Diacon, S. and Hasseldine, J. (2007). Framing effects and risk perception: The effect of prior performance presentation format on investment fund choice. *Journal of Economic Psychology*, 28(1):31–52.

Dwyer, G. P., Williams, A. W., Battalio, R. C., and Mason, T. I. (1993). Tests of rational expectations in a stark setting. *The Economic Journal*, 103(418):586–601.

Frankel, J. and Froot, K. (1987). Using survey data to test standard propositions regarding exchange-rate expectations. *American Economic Review*, 77:133–153.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2006). Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, 131(1-2):59–95.

Glaser, M., Iliewa, Z., and Weber, M. (2018). Thinking about prices versus thinking about returns in financial markets. *Journal of Finance (forthcoming)*.

Glaser, M., Langer, T., Reynders, J., and Weber, M. (2007). Framing effects in stock market forecasts: The difference between asking for prices and asking for returns. *Review of Finance*, 11(2):325–357.

Greenwood, R. and Shleifer, A. (2014). Expectations of returns and expected returns. *The Review of Financial Studies*, 27(3):714–746.

Hanaki, N., Akiyama, E., and Ishikawa, R. (2018). Effects of different ways of incentivizing price forecasts on market dynamics and individual decisions in asset market experiments. *Journal of Economic Dynamics and Control*, 88:51–69.

Heemeijer, P., Hommes, C., Sonnemans, J., and Tuinstra, J. (2009). Price stability and volatility in markets with positive and negative expectations feedback: An experimental investigation. *Journal of Economic Dynamics and Control*, 33(5):1052–1072.

Hey, J. D. (1994). Expectations formation: Rational or adaptive or ...? *Journal of Economic Behavior & Organization*, 25(3):329–349.

Hoffmann, A. O., Iliewa, Z., and Jaroszek, L. (2017). Wall street crosses memory lane: How witnessed returns affect professionals' expected returns. SSRN working paper, http://dx.doi.org/10.2139/ssrn.2877366.

Hommes, C. (2011). The heterogeneous expectations hypothesis: Some evidence from the lab. *Journal of Economic Dynamics and Control*, 35(1):1–24.

Hommes, C., Kopányi-Peuker, A., and Sonnemans, J. (2018). Bubbles, crashes and information contagion in large-group asset market experiments. CeNDEF Working paper 18-05.

Hommes, C., Sonnemans, J., Tuinstra, J., and Van de Velden, H. (2005). Coordination of expectations in asset pricing experiments. *The Review of Financial Studies*, 18(3):955–980.

Hommes, C., Sonnemans, J., Tuinstra, J., and Van de Velden, H. (2008). Expectations and bubbles in asset pricing experiments. *Journal of Economic Behavior & Organization*, 67(1):116–133.

Huber, C. and Huber, J. (2019). Scale matters: risk perception, return expectations, and investment propensity under different scalings. *Experimental Economics*, 22(1):76–100.

Kelley, H. and Friedman, D. (2002). Learning to forecast price. *Economic Inquiry*, 40(4):556–573.

Kirchler, E., Maciejovsky, B., and Weber, M. (2005). Framing effects, selective information, and market behavior: An experimental analysis. *The Journal of Behavioral Finance*, 6(2):90–100.

Kirchler, M., Huber, J., and Stöckl, T. (2012). Thar she bursts: Reducing confusion reduces bubbles. *American Economic Review*, 102(2):865–83.

Malmendier, U. and Nagel, S. (2016). Learning from inflation experiences. *The Quarterly Journal of Economics*, 131(1):53–87.

Marimon, R., Spear, S. E., and Sunder, S. (1993). Expectationally driven market volatility: An experimental study. *Journal of Economic Theory*, 61(1):74–103.

Palan, S. (2013). A review of bubbles and crashes in experimental asset markets. *Journal of Economic Surveys*, 27(3):570–588.

Porter, D. P. and Smith, V. L. (1995). Futures contracting and dividend uncertainty in experimental asset markets. *Journal of Business*, 68(4):509–541.

Schmalensee, R. (1976). An experimental study of expectation formation. *Econometrica*, 44(1):17–41.

Sirri, E. R. and Tufano, P. (1998). Costly search and mutual fund flows. *The Journal of Finance*, 53(5):1589–1622.

Smith, V. L., Suchanek, G. L., and Williams, A. W. (1988). Bubbles, crashes, and endogenous expectations in experimental spot asset markets. *Econometrica*, 56:1119–1151.

Sonnemans, J. and Tuinstra, J. (2010). Positive expectations feedback experiments and number guessing games as models of financial markets. *Journal of Economic Psychology*, 31(6):964–984.

Stöckl, T., Huber, J., and Kirchler, M. (2010). Bubble measures in experimental asset markets. *Experimental Economics*, 13(3):284–298.

Stössel, R. and Meier, A. (2015). Framing effects and risk perception: Testing graphical representations of risk for the kiid. Available at SSRN: https://ssrn.com/abstract=2606615 or http://dx.doi.org/10.2139/ssrn.2606615.

Tversky, A. and Kahneman, D. (1982). Judgement of and by representativeness. In Kahneman, D., Slovic, P., and Tversky, A. (editors), *Judgement under uncertainty: Heuristics and Biases*. Cambridge University Press.

APPENDIX

A The asset pricing model

Let us first briefly summarize the market framework used in our learning to forecast experiment, following Heemeijer et al. (2009) and Bao et al. (2017). There are I agents in the market and they can invest in a risky asset and in a risk-free bond. The risky asset pays an uncertain dividend y_t in each period whereas the risk-free bond pays a gross return of $1 + r$.

Agent i's wealth W_i evolves according to

$$W_{i,t+1} = (1 + r)(W_{i,t} - p_t z_{i,t}) + z_{i,t}(p_{t+1} + y_t) = (1 + r)W_{i,t} + z_{i,t}(p_{t+1} + y_t - (1 + r)p_t), \tag{A.1}$$

where p_t is the price of the risky asset in period t (before the dividend is paid) and $z_{i,t}$ is the amount of risky asset agent i buys in period t.

Agents are assumed to have mean-variance preferences, that is they choose the amount of the risky asset in order to maximize

$$E_{i,t}(W_{i,t+1}) - \frac{1}{2} a Var_{i,t}(W_{i,t+1}),$$

where a is a parameter for risk aversion.

This optimization problem leads to the following optimal demand for the risky asset:

$$z^*_{i,t} = \frac{p^e_{i,t+1} + y - (1 + r)p_t}{a Var_{i,t}(p_{t+1} + y_t - (1 + r)p_t)} = \frac{p^e_{i,t+1} + y - (1 + r)p_t}{a\sigma^2}, \tag{A.2}$$

where $p^e_{i,t+1}$ is the price expectation of agent i for the next period and y is the (constant) expected dividend. Notice that we make the assumption that $Var_{i,t}(p_{t+1} + y_t - (1 + r)p_t) = \sigma^2$ for each agent i. That is, we assume that agents can have heterogeneous price expectations but they all believe that the variance in question is equal to σ^2.

The price of the risky asset is governed by the aggregate demand (Z^D_t) and the exogenous aggregate supply (Z^S_t) of the asset according to the following price adjustment mechanism:

$$p_{t+1} = p_t + \lambda\left(Z^D_t - Z^S_t\right) + \varepsilon_t, \tag{A.3}$$

with $\varepsilon_t \sim N(0, 0.5^2)$ and λ is the speed of adjustment.

Assuming that the aggregate supply of the asset is 0 and combining (A.2) and (A.3), we get the following law of motion for prices:

$$p_{t+1} = p_t + \lambda \sum_{i=1}^{I} \frac{p^e_{i,t+1} + y - (1 + r)p_t}{a\sigma^2} + \varepsilon_t \tag{A.4}$$

To further simplify the law of motion, we use the following assumptions about the parameters: $a\sigma^2 = I$ and $\lambda = \frac{1}{1+r}$. This yields

$$p_{t+1} = \frac{1}{1+r}\left(\bar{p}^e_{t+1} + y\right) + \varepsilon_t, \tag{A.5}$$

where $\bar{p}^e_{t+1}$ denotes the agents' average price expectation. An equivalent form of (A.5) is

$$p_{t+1} = p^f + \frac{1}{1+r}\left(\bar{p}^e_{t+1} - p^f\right) + \varepsilon_t, \tag{A.6}$$

where $p^f = \frac{y}{r}$ is the fundamental price of the risky asset.

Thus, in this asset market framework price dynamics is driven by the agents' average price expectations. Notice that agents need to form one-period-ahead forecasts as p_t depends on forecasts for the same period ($\bar{p}^e_t$). Under naive expectations ($p^e_{i,t+1} = p_t$) and $r > 0$ the price converges to the fundamental price p^f.

B Instruction for treatment PR

Welcome to this experiment on decision-making. Please read the following instructions carefully. If you have any questions, please raise your hand, and we will come to your table to answer your question in private.

General information

You are a financial advisor to a pension fund that wants to optimally invest a large amount of money. The pension fund has two investment options: a risk free investment (on a savings account) and a risky investment (on the stock market). As its financial advisor, you have to forecast the stock return for 50 subsequent time periods. The more accurate your forecasts are, the higher your total earnings are.

Your forecasting task

Your only task is to forecast the stock return in each time period as accurately as possible. The stock return is the relative price change compared to the previous period:

$$return_t = (price_t - price_{t-1})/price_{t-1}.$$

The return therefore measures how fast prices are increasing or decreasing. For example, if the price in period t-1 is 50 and the price in period t is 53, then the return in period t is (53-50)/50=0.06, or 6%. The stock return has to be forecasted one period ahead, that is at the beginning of each period you need to forecast what the return will be in that period. It is very likely that the stock return will be between -10% and 10% in the first period. After all participants have given their forecasts for the first period, the stock price for the first period will be revealed and, based upon your forecasting error, your earnings for period 1 will be given. After that you have to give your forecast for the stock return in the second period. After all participants have given their forecasts for period 2, the stock price in the second period will be revealed and, based upon your forecasting error, your earnings for period 2 will be given. This process continues for 50 time periods in total.

The available information for forecasting the stock return in period t consists of all past prices up to period t-1, your total earnings up to period t-1, and your past return forecasts up to period t-1. **Notice that the variable you need to forecast differs from the variable you receive information about: You need to forecast returns but you receive information about prices.**

In each period you have limited time to make your forecasting decision. If you do not submit a forecast during this time frame, your pension fund will be inactive, and you will not earn any points in that given

period. A timer will show you the remaining time for each period (2 minutes for each of the first 10 periods, 1 minute for each of the later periods).

Information about the stock market

The stock price in period t depends on the aggregate demand for the stock and on the supply of stocks. The supply of stocks is fixed during the experiment. The demand for stocks is mainly determined by the aggregate demand of the large pension funds active in the market. In addition, there are some small investors that are active on the stock market. The higher the aggregate demand for stocks is, the higher the realized market price will be. There are 6 large pension funds in the stock market. Each pension fund is advised by a participant of the experiment.

Earnings

Your earnings depend on the accuracy of your forecasts. Your payoff for your forecast in period t is given by

$$1300 * (1 - 625 * e_t^2),$$

where e_t is the forecast error, that is the absolute difference between your forecast of the return in period t and the realized return in that period. The maximum possible points you can earn in each period (if you make no forecast error) is 1300, and the larger your forecast error is, the fewer points you will make. Note, however, that you will never earn negative payoffs in a single period: If your forecast error in a particular period is very large, your payoffs for that period will be zero. There is a Payoff Table on your desk, which shows the points you can earn for different forecast errors.

We will pay you in cash at the end of the experiment based on the points you earned. You earn 0.5 euro for each 1300 points you make plus an additional 5 euros of participation fee.

Background information about the investment strategies of the funds

The precise investment strategy of the pension fund that you are advising and the investment strategies of the other pension funds and of the small investors are unknown. The savings account that pension funds can use for their risk free investment pays a fixed interest rate of 5% per time period. The stock pays an uncertain dividend in each time period. Economic experts have computed that the average dividend is 3.3 euros per period. The realized stock return per period is uncertain and depends upon the (unknown) dividend and upon stock price changes.

Based upon your stock return forecast, your pension fund will make an optimal investment decision. The higher your return forecast is, the more money will be invested in the stock market by the fund, so the larger will be the demand for stocks.

On the next screens you are asked to answer some questions in order to check if the experiment is clear to you.

C Measures of instability and mispricing

In this section we report for each market the values of the three measures of instability (standard deviation, price range and median absolute returns) and of the two measures of mispricing (relative absolute deviation and relative deviation). These measures were introduced in Section 3.2 and we used these values in the statistical tests for comparing treatments.

	markets							
	1	2	3	4	5	6	7	8
PP	5.19	5.28	1.07	1.36	21.39	1.11	2.68	42.55
RP	2.11	1.29	4.02	9.36	98.17	1.89	31.75	
PR	11.41	5.35	2.92	6.03	5.42	3.77	2.03	14.19
RR	4.58	5.66	2.92	7.08	6.46	34.56	4.92	19.1

Table 7: Standard deviation of prices (*std*) over the last 40 periods

	markets							
	1	2	3	4	5	6	7	8
PP	16.74	19.45	4.18	8.53	113.42	5.72	8.52	156.33
RP	8.22	4.54	15.73	31.46	326.34	7.7	119.71	
PR	36.69	16.29	10.82	19.85	16.73	13.23	7.84	49.12
RR	17.06	17.09	10.8	25.74	20.44	100.62	18.03	65.76

Table 8: Range of prices (*range*) over the last 40 periods

	markets							
	1	2	3	4	5	6	7	8
PP	1.61	1	0.56	0.58	2.18	0.61	0.79	14.77
RP	0.78	0.64	1.17	4.24	7.92	0.69	8.15	
PR	3.55	1.94	0.94	1.12	2.04	1.32	0.82	5.24
RR	1.37	1.33	1	2.75	5.19	4.35	1.25	8.49

Table 9: Median absolute returns (AR), over the last 40 periods (in %).

	markets							
	1	2	3	4	5	6	7	8
PP	7.01	31.38	3.89	2.35	20.43	3.67	3.58	57.5
RP	2.91	3.78	6	12.48	103.09	5.49	77.76	
PR	15.23	7.55	8.37	10.25	8.83	6.19	4.7	18.82
RR	7.01	9.88	8.94	9.75	9.48	52.73	6.63	24.6

Table 10: Relative absolute deviation (*RAD*) over the last 40 periods (in percentage points)

	markets							
	1	2	3	4	5	6	7	8
PP	5.35	31.38	3.89	2.32	19.12	3.52	2.61	28.86
RP	1.38	3.78	3.39	0.82	92.12	5.49	77.69	
PR	10.47	2.93	8.37	9.44	6.84	4.86	4.4	6.65
RR	4.88	8.41	8.81	6.23	6.38	49.8	4.62	8.74

Table 11: Relative deviation (*RD*) over the last 40 periods (in percentage points)

D Test results

In this section we report detailed test results for comparing treatments in terms of instability. Table 12 summarizes the p values of the Kolmogorov-Smirnov tests for comparing treatments based on the five instability measures we use in Section 3.2 and Table 13 reports the p values of the Wilcoxon rank-sum tests for comparing treatments in terms of the number of stable markets.

all data	**std**	**range**	**RAD**	**RD**	**AR**
PP vs RP	0.9556	0.753	0.9315	0.4837	0.8702
PP vs PR	0.2643	0.2643	0.0939*	0.2643	0.0939*
PP vs RR	0.1877	0.1877	0.1877	0.1877	0.0497**
RP vs PR	0.4405	0.4405	0.1698	0.1698	0.1951
RP vs RR	0.1698	0.1698	0.0547*	0.0547*	0.1698
PR vs RR	1	1	0.8626	0.8626	1
*P vs *R	0.0592*	0.0592*	0.0215**	0.0215**	0.0231**
P* vs R*	0.9854	0.9973	0.9643	0.7351	0.9346
same variable	0.9863	0.4717	0.9439	0.9623	0.9212
excl. outlier markets	**std**	**range**	**RAD**	**RD**	**AR**
PP vs RP	1	0.7782	1	0.3667	1
PP vs PR	0.0454**	0.0454**	0.0102**	0.0454**	0.0102**
PP vs RR	0.0204**	0.0104**	0.0204**	0.0204**	0.0204**
RP vs PR	0.2239	0.2239	0.0454**	0.0454**	0.0759*
RP vs RR	0.0454**	0.0454**	0.0102**	0.0102**	0.0454**
PR vs RR	0.8424	0.8424	0.8424	0.8424	0.8424
*P vs *R	0.0052***	0.0052***	0.0008***	0.0008***	0.0015***
P* vs R*	0.9094	0.9094	0.9094	0.684	0.9094
same variable	0.9913	0.4333	0.9913	0.9913	0.7864

Notes: ***: significant at 1% level, **: significant at 5% level, *: significant at 10% level. All test are one-sided except for *PP vs RR* and *same variable*. Observations correspond to markets, the number of observations is $n_{PP} = 8$, $n_{RP} = 7$, $n_{PR} = 8$ and $n_{RR} = 8$ in the upper panel and $n_{PP} = 5$, $n_{RP} = 5$, $n_{PR} = 7$ and $n_{RR} = 7$ in the lower panel.

Table 12: Summary of p values in the Kolmogorov-Smirnov tests for comparing treatments in terms of instability.

	all data	**excl. outlier markets**
PP vs RP	0.7855	0.9167
PP vs PR	0.141	0.0455**
PP vs RR	0.0769*	0.0202**
RP vs PR	0.9872	0.9987
RP vs RR	0.141	0.0455**
PR vs RR	0.9872	0.9987
*P vs *R	0.0062***	0.0009***
P* vs R*	0.7672	0.8114
same variable	0.9377	1

Notes: ***: significant at 1% level, **: significant at 5% level, *: significant at 10% level. All test are one-sided except for *PP vs RR* and *same variable*. Observations correspond to markets, the number of observations is $n_{PP} = 8$, $n_{RP} = 7$, $n_{PR} = 8$ and $n_{RR} = 8$ in the left column and $n_{PP} = 5$, $n_{RP} = 5$, $n_{PR} = 7$ and $n_{RR} = 7$ in the right column.

Table 13: Summary of p values in the Wilcoxon rank-sum test for comparing treatments in terms of the number of stable markets.

In the tables *same variable* corresponds to testing differences between observing and predicting the same variable (merging PP and RR) versus observing and predicting different variables (merging RP and PR).

Table 12 shows that most differences between individual treatments are not significant: Out of the 30 comparisons, only the difference for measure AR between PP and RR is significant at the 5% level, four more differences – between PP and PR, and between RP and RR – are significant at the 10% level. However, after excluding the markets with outliers, differences between individual treatments are significant at the 5% level, except for PR vs RR.[34] Finally, there is no difference between observing and predicting the same vs different variables.

We have similar findings when comparing the number of stable markets in Table 13.

[34]These results should be treated with some prudence, since (when excluding markets with outliers) treatments PP and RP now only contain five markets each, with treatments PR and RR each containing seven markets.

E Regression results

In this section we report for each market the estimation output for the individual forecasting rule we used in Section 3.3:

$$p^f_{h,t+1} = C_h + \sum_{l=0}^{3} \beta_{hl} p_{t-l} + \sum_{l=0}^{3} \gamma_{hl} p^f_{h,t-1} + \varepsilon_{h,t+1}.$$

For each treatment we report two tables. The first one shows for each market the number of subjects for which a given variable is significantly different from 0 at the 5% level. The second one shows for each market the average of the parameter estimates for a given variable, where the average is taken over the six subjects in the market. (The coefficients of insignificant variables are set to 0 when calculating the average.)

We also report for each treatment the number of good models, i.e. models where there is no serial correlation in the residuals (Ljung-Box Q test), no heteroskedasticity in residuals (Engle's ARCH test) and there is no model specification error (Ramsey's RESET) at the 5% level. In treatment PP 35 models are good out of 48, in RP 33 out of 42, in PR 40 out of 48 and finally 35 models are good out of 48 in treatment RR.

	C	p_t	p_{t-1}	p_{t-2}	p_{t-3}	p_t^f	p_{t-1}^f	p_{t-2}^f	p_{t-3}^f
1	3	6	5	3	1	1	1	0	0
2	2	5	5	2	0	3	2	1	3
3	0	6	4	0	0	1	1	1	0
4	2	6	2	0	2	4	1	0	1
5	0	6	5	2	2	3	1	1	5
6	3	6	3	2	2	3	2	1	2
7	0	6	6	0	0	0	1	1	0
8	1	6	4	5	3	1	2	1	2
all	11	47	34	14	10	16	11	6	13

Table 14: Number of subjects with a significant coefficient - treatment PP

	C	p_t	p_{t-1}	p_{t-2}	p_{t-3}	p_t^f	p_{t-1}^f	p_{t-2}^f	p_{t-3}^f
1	2.35	1.78	-0.88	0.08	-0.05	0.06	-0.03	0	0
2	1.84	1.29	-0.6	0.12	0	0.24	-0.07	0.02	0
3	0	1.41	-0.46	0	0	0.11	-0.02	-0.02	0
4	1.39	1.23	-0.08	0	-0.05	-0.13	-0.02	0	-0.01
5	0	1.85	-0.66	0	0.04	-0.16	-0.01	-0.06	0
6	8.09	1.31	-0.27	-0.14	0.03	-0.06	0.02	0.02	-0.05
7	0	1.87	-0.89	0	0	0	0.03	-0.02	0
8	1	2.33	-1.33	0	0.23	-0.06	-0.13	0.03	-0.08
all	1.83	1.64	-0.64	0.01	0.02	0	-0.03	0	-0.02

Table 15: Average coefficients over all subjects - treatment PP

	C	p_t	p_{t-1}	p_{t-2}	p_{t-3}	p_t^f	p_{t-1}^f	p_{t-2}^f	p_{t-3}^f
1	2	6	4	2	0	1	2	1	0
2	0	6	5	3	0	1	2	0	1
3	1	6	4	1	0	3	0	2	1
4	2	6	5	2	1	2	1	0	1
5	4	6	4	2	0	4	3	4	3
6	1	6	4	0	1	3	2	1	0
7	2	6	4	3	2	3	2	1	1
all	12	42	30	13	4	17	12	9	7

Table 16: Number of subjects with a significant coefficient - treatment RP

	C	p_t	p_{t-1}	p_{t-2}	p_{t-3}	p_t^f	p_{t-1}^f	p_{t-2}^f	p_{t-3}^f
1	1.07	1.6	-0.49	0.03	0	-0.05	-0.08	-0.02	0
2	0	1.66	-0.69	0.09	0	0.05	-0.09	0	-0.03
3	-0.8	1.82	-0.52	-0.04	0	-0.18	0	-0.06	-0.01
4	2.31	2.16	-1.33	0.19	0.03	-0.03	-0.05	0	-0.02
5	5.09	1.31	-0.38	-0.16	0	0.17	-0.1	0.15	-0.06
6	0.69	1.7	-0.55	0	-0.03	-0.05	-0.07	-0.02	0
7	2.69	1.27	-0.16	-0.21	0.02	0.07	0	-0.05	-0.02
all	1.58	1.65	-0.59	-0.02	0	0	-0.05	0	-0.02

Table 17: Average coefficients over all subjects - treatment RP

	C	p_t	p_{t-1}	p_{t-2}	p_{t-3}	p_t^f	p_{t-1}^f	p_{t-2}^f	p_{t-3}^f
1	1	6	6	3	1	3	1	0	1
2	3	6	6	2	1	0	0	0	1
3	2	6	5	1	0	4	0	0	2
4	0	6	5	2	1	5	0	0	0
5	0	6	6	3	2	3	1	0	1
6	1	6	5	3	0	0	1	0	1
7	1	6	5	2	0	0	1	0	1
8	1	6	5	2	3	0	0	0	0
all	9	48	43	18	8	15	4	0	7

Table 18: Number of subjects with a significant coefficient - treatment PR

	C	p_t	p_{t-1}	p_{t-2}	p_{t-3}	p_t^f	p_{t-1}^f	p_{t-2}^f	p_{t-3}^f
1	0.25	2.28	-1.83	0.32	0.03	0.21	-0.05	0	0.04
2	2.68	1.89	-0.98	0.17	-0.1	0	0	0	-0.02
3	-1.12	1.78	-0.65	-0.08	0	-0.02	0	0	-0.01
4	0	1.3	-0.72	-0.15	0.06	0.49	0	0	0
5	0	2.17	-1.42	0.07	0.1	0.1	-0.05	0	0.03
6	0.42	2.06	-1.19	0.19	0	0	-0.05	0	-0.03
7	0.5	1.78	-0.73	0.01	0	0	-0.04	0	-0.02
8	0.55	2.29	-1.4	-0.11	0.21	0	0	0	0
all	0.41	1.94	-1.11	0.05	0.04	0.1	-0.02	0	0

Table 19: Average coefficients over all subjects - treatment PR

	C	p_t	p_{t-1}	p_{t-2}	p_{t-3}	p_t^f	p_{t-1}^f	p_{t-2}^f	p_{t-3}^f
1	0	6	6	2	2	1	1	2	2
2	4	6	6	3	2	2	1	1	2
3	2	6	6	5	2	0	2	0	0
4	3	6	6	4	1	2	0	0	3
5	2	6	6	3	2	1	0	1	0
6	1	6	5	2	4	3	1	3	2
7	3	6	5	4	1	2	2	1	1
8	1	6	6	5	2	1	1	0	1
all	16	48	46	28	16	12	8	8	11

Table 20: Number of subjects with a significant coefficient - treatment RR

	C	p_t	p_{t-1}	p_{t-2}	p_{t-3}	p_t^f	p_{t-1}^f	p_{t-2}^f	p_{t-3}^f
1	0	2.12	-1.32	0.23	0.06	0.09	-0.08	-0.07	-0.02
2	-0.96	1.94	-1.01	0.11	-0.02	0.05	-0.04	0.02	-0.03
3	-0.07	2.38	-1.85	0.66	-0.1	0	-0.09	0	0
4	1.63	2.39	-1.92	0.49	-0.09	0.02	0	0	0.08
5	4.1	2.18	-1.84	0.32	0.09	0.09	0	0.05	0
6	-0.36	2.18	-1.36	0.08	0.24	0.09	-0.05	-0.11	-0.04
7	-0.62	2.12	-1.1	0.1	0.05	-0.05	-0.06	-0.02	-0.02
8	0.37	2.95	-2.8	1.14	-0.17	-0.1	-0.05	0	0.04
all	0.51	2.28	-1.65	0.39	0.01	0.02	-0.05	-0.02	0

Table 21: Average coefficients over all subjects - treatment RR

F Coordination results

In this section we report for each market the forecasting error decomposition as specified by equation (3) in Section 3.3:

$$\frac{1}{240}\sum_{h,t}\left(\frac{p^f_{h,t}-p_t}{p_{t-1}}\right)^2 = \frac{1}{240}\sum_{h,t}\left(\frac{p^f_{h,t}-\overline{p}^f_t}{p_{t-1}}\right)^2 + \frac{1}{40}\sum_{t}\left(\frac{\overline{p}^f_t-p_t}{p_{t-1}}\right)^2.$$

According to this equation, the average individual error is decomposed into two parts: the average dispersion error (first term on the right hand side) and average common error (second term on the right hand side).

market	avg. individual error	avg. dispersion error	avg. common error
1	2.33	0.68 (29%)	1.65 (71%)
2	8.74	3.97 (45%)	4.77 (55%)
3	0.69	0.13 (19%)	0.56 (81%)
4	1.97	0.14 (7%)	1.83 (93%)
5	439.05	3.55 (1%)	435.5 (99%)
6	0.78	0.22 (29%)	0.55 (71%)
7	0.79	0.21 (27%)	0.58 (73%)
8	31.68	21.64 (68%)	10.04 (32%)

Table 22: Forecast error decomposition in treatment PP. The reported values are multiplied by 10000.

market	avg. individual error	avg. dispersion error	avg. common error
1	1.43	0.84 (59%)	0.59 (41%)
2	0.95	0.38 (40%)	0.57 (60%)
3	0.99	0.43 (43%)	0.56 (57%)
4	3.15	2.16 (69%)	0.99 (31%)
5	148.35	9.59 (6%)	138.75 (94%)
6	0.99	0.26 (26%)	0.73 (74%)
7	404.59	8.15 (2%)	396.43 (98%)

Table 23: Forecast error decomposition in treatment RP. The reported values are multiplied by 10000.

market	avg. individual error	avg. dispersion error	avg. common error
1	2.57	1.54 (60%)	1.04 (40%)
2	2.43	1.27 (52%)	1.16 (48%)
3	1.02	0.39 (38%)	0.64 (62%)
4	6.36	0.23 (4%)	6.13 (96%)
5	0.96	0.25 (26%)	0.72 (74%)
6	1.15	0.5 (43%)	0.65 (57%)
7	0.68	0.11 (16%)	0.58 (84%)
8	2.93	1.46 (50%)	1.47 (50%)

Table 24: Forecast error decomposition in treatment PR. The reported values are multiplied by 10000.

market	avg. individual error	avg. dispersion error	avg. common error
1	0.88	0.21 (24%)	0.67 (76%)
2	1.23	0.45 (36%)	0.78 (64%)
3	0.96	0.28 (29%)	0.68 (71%)
4	1.02	0.24 (23%)	0.78 (77%)
5	1.48	0.41 (27%)	1.07 (73%)
6	9.34	4.51 (48%)	4.83 (52%)
7	0.84	0.13 (16%)	0.7 (84%)
8	4.11	1.29 (31%)	2.83 (69%)

Table 25: Forecast error decomposition in treatment RR. The reported values are multiplied by 10000.