Forecasting returns instead of prices exacerbates financial bubbles

Nobuyuki Hanaki[1] · Cars Hommes[2] · Dávid Kopányi[3] ·
Anita Kopányi-Peuker[4] · Jan Tuinstra[5]

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 individual choice 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. In order to understand the impact of these effects found at the individual level on market dynamics, we consider 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 presentation format of past information, we do find that markets are significantly more unstable when subjects have to forecast returns instead of prices. Our results therefore show that the elicitation format may exacerbate, or even create, bubbles and crashes in financial markets.

Keywords Experimental finance · Expectation formation · Asset market · Presentation formats

JEL Classification C72 · D43 · G41

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 the way in which forecasts are elicited, as well as the format in which past data is presented. 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., 2019) find that return expectations are higher when return forecasts, instead of price forecasts, are elicited. Furthermore, they show that return expectations are lower when past returns, instead of past prices, are shown to the subjects. 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., 2019), for an overview).

These findings imply 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 expectations guide investment decisions and thereby have a direct impact on realized market returns. They therefore have an effect on the performance of other traders' investment strategies as well. If the format of the information presented, or the format of the forecasting task, leads to systematic differences 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. Although many important behavioral tendencies have been established in individual

---

[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 transformed 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).

decision-making environments, it is not always clear if and how these findings translate to market dynamics. An important motivation for our paper is to investigate the extent to which aggregate behavior is affected by the choice of variable to be predicted as well as by the presentation of past performance.

In the studies mentioned above 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. However, these feedback effects can be substantial: market prices are determined by the trading decisions of individual traders which are informed by these traders' expectations. The earlier contributions therefore only investigate part of the story: they miss the link between individual decisions and market realizations, and thereby the channel through which the presentation of information and of the forecasting task may have an actual impact on the functioning of financial markets. We go beyond the existing literature by explicitly taking expectations feedback into account. The focus of our study therefore does not lie on individual forecasting behavior, but on the effect on market dynamics instead. In this so-called *learning to forecast* experiment[4] subjects' average expectations of prices/returns are an important determinant of realized prices/returns. This setting allows us to investigate whether price volatility is affected by how information about past performance is presented and by which variable is being forecast. Obviously this cannot be studied when the time series are exogenously given. Moreover it is not clear whether the individual differences observed in previous contributions also result in differences at the market level. As a second contribution, we run an additional treatment where subjects are free to choose the format in which they give their prediction and the format in which information is presented to them. In this way we can study whether the aggregate effects that we identify in our main treatments are robust.

In our main experiment, we use a 2×2 between-subjects design similar to the one used in Glaser et al. (2019). 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 past information are presented. Subjects are paid for their forecasting accuracy. We find that the format of past information (either presented as a return bar chart or as a price line chart) does not have a significant effect on the resulting market dynamics, but we provide evidence that the format in which forecasts are elicited does have an effect. In particular, we find that asking for return forecasts tends to increase price volatility, when compared to asking for price forecasts. Unfortunately, the general picture is somewhat obfuscated by the fact that in a number of markets at least one of the subjects submits a very high prediction for a certain period, possibly as the consequence of a typing error, or because of an inclination to experiment with the forecasting environment. These extreme predictions,

---

which occur much more often in the treatments where subjects' task is to predict the price, have a substantial and persistent effect on the ensuing market dynamics. Correcting for these outlier markets substantially strengthens our main result that forecasting returns leads to more price volatility than forecasting prices. An analysis of subjects' individual forecasts suggests that the difference between predicting returns and predicting prices 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.

In each of the four experimental treatments described above, we fixed both the format in which past information is presented as well as the format of the prediction task. In practice, investors and financial advisors may indeed be prone to using particular formats, due to specific regulation, because of an established habit or default bias, or because the investors they are advising demand that advice in a specific format. Nevertheless, investors and financial advisors may have a choice in how they want information about past performance of assets to be presented, and whether they prefer to predict future prices or future returns as an input to an investment decision (taken by themselves, or by the investor they are advising). Given that predicting returns tends to lead to more volatile prices than predicting prices, the question arises whether, if the prediction format can be chosen freely, one of the formats will emerge as the dominant choice, and how price volatility will be affected. In an attempt to answer these questions, we run an additional treatment of our experiment where in each period subjects can choose whether to submit a price forecast or a return forecast. In addition, in each period subjects can also decide whether past performance of the asset is presented either by past prices or past returns, and they are allowed to switch between those formats as often as they want. We find no clear preference for predicting prices or returns in this treatment with endogenous choice; about half of the predictions are price predictions, and the other half are return predictions. We do find that subjects have a preference to observe information about the variable they will forecast. About 80% of the decisions fall into this category.[5] Combining these two insights we find that two out of our four initial treatments emerge as a natural environment for participants: the treatment where subjects observe and predict prices and the treatment where they observe and predict returns. In terms of market dynamics, we do not find systematic significant differences between this additional treatment and the four initial treatments.

Summarizing, we provide evidence that thinking about returns, instead of thinking about prices, has a substantial impact on the performance of financial markets. These findings are also important, as subjects do not have a clear preference for forecasting prices, which would result in more stable market dynamics. Policy makers and regulators should take the effect of the forecasting format into account when

---

[5] The elicitation format is typically imposed in surveys only, whereas presentation of past performance can be inferred from investor documents, newspapers or from online sources. This dichotomy makes it difficult to find evidence in the field to support the PP or RR configurations directly. Glaser et al. (2019) find that participants in the ZEW Financial Market Survey (a survey among financial market experts) seem to prefer giving price forecasts over return forecasts.

designing policies aimed at stabilizing financial markets, in particular when such policies are in part based upon survey data or laboratory experiments that choose one particular format for forecasts.

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) forecasts are elicited typically document trend extrapolation, whereas mean reverting behavior is found in studies where price forecasts are elicited. For example, 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. Subjects that have to forecast returns exhibit a stronger tendency to extrapolate past trends: 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.[6]

Glaser et al. (2019) 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 a stronger tendency to extrapolate trends when return forecasts are elicited, they find that these return forecasts are higher. 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. (2019) 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 diminish the likelihood of bubbles. However, we do not find evidence for an effect of the format of past information in our experiment. 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. Their behavior is consistent with stronger expected mean reversion when prices are observed than when returns are observed. Diacon and Hasseldine (2007) find that investment decisions are not significantly affected by the presentation format of the different funds. As opposed to Glaser et al. (2007), Stössel and Meier (2015) find that subjects overestimate returns substantially when seeing a return bar chart. Finally, the five-year ahead return forecasts of the subjects of Huber and Huber (2019) 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. (2019).

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

---

[6] For similar findings see e.g., Andreassen and Kraus (1990) and Czaczkes and Ganzach (1996).

on a common prediction strategy. These results are quite robust, for example with respect to information about the underlying model (Sonnemans and Tuinstra, 2010), the number of subjects in a group (Hommes et al., 2021), 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 establishing that the format of the forecasting task 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, 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 market stability in Sect. 3. In addition, in this section we also present the results after correcting for outliers in different ways. In Sect. 4 we discuss our findings from the additional treatment with endogenous task and information formats. Section 5 concludes. Experimental instructions and other supplementary material, such as time series for all markets and the analysis of subjects' individual forecasting strategies, are presented in the online appendices.

2 Experimental design

The experiment, programmed in PhP, was run in February and March 2017, in September and November 2019 and in May 2022 at the CREED experimental laboratory of the University of Amsterdam.[7] In total 504 subjects (students from various fields) participated in the experiment in 26 sessions.[8] Experimental sessions (up to paying out) lasted for approximately 90 min, with payments for each subject typically between €23.5 and €28. Below we will outline the main features of the experimental design.

2.1 Subjects' task and main treatments

Our design is based on the standard 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

---

[7] We ran the additional treatment with endogenous choice in June and September 2022. This treatment and the differences with the treatments described below, are discussed in more detail in Sect. 4.

[8] Originally we collected seven more groups, but due to either a software error, or a subject abruptly leaving the experiment, we did not use those data for the analysis.

consecutive periods, using information about past prices or information about past returns. Subjects' earnings are based upon their forecasting accuracy.

Following (Glaser et al., 2019), 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 ('information'), 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.[9] We have 19 to 23 markets for each treatment.

In contrast to Glaser et al. (2019) 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[10]

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

Here p_t is the price of the risky asset in period t and $\bar{p}_t^f = \frac{1}{6}\sum_{h=1}^{6} p_{h,t}^f$ 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 value 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 value, with most of the weight on the former).[11]

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

---

[9] Recently, Hommes et al. (2021) 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.

[10] See Online Appendix A for more details about the microfoundation of the asset pricing model and detailed derivations for Eq. (1). 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 Eq. (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.

[11] Predicting the fundamental value 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 topic.

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 shown in Eq. (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 the total demand for the risky asset is higher. In addition, they know that the number of subjects in their market is six.[12]

2.2 Information and incentives

Examples of the decision screens, for treatments PR and RP, are shown in Fig. 1. Subjects can submit their forecast at the top of the screen.[13] 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 up to the last period, and their accumulated earnings thus far (lower right part of the decision screen); (ii) a figure with either a time series of past realized prices or a bar chart of past realized returns (lower left part of the decision screen),[14]; and (iii) the most recent price. To be more precise, in the treatments where subjects receive information about prices (PR and PP), we present the previously realized prices in the table, and a figure showing the time series of the realized prices so far. In treatments where subjects receive information about returns (RR and RP), we present the previously realized returns in the table, and a bar chart for the past returns on the graph. Furthermore, subjects see their own forecast in the table in the variable they have to forecast. So in treatments PR and RR the table contains the return forecasts subjects submitted, whereas in treatments PP and RP the table contains their price forecasts. Note that the most recent price is given for all treatments, also 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

---

[12] Online Appendix B contains the instructions for treatment PR.

[13] 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.

[14] Following (Glaser et al., 2019) we always present past prices as time series and past returns as bar charts, respectively, because this is how they are typically represented in financial markets.

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

		Task	
		Price	Return
Information	Price	PP	PR
		(22)	(19)
	Return	RP	RR
		(23)	(20)

and RR, so that incentives are exactly the same in each treatment.[15] Subjects earn between 0 (if the forecast error in that period is 4% or larger) 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 €5.00 show-up fee.

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%.[16] 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.[17] For the first period subjects do not have any information about prices or returns yet (except that $p_0 = 50$), 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.[18] 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.[19]

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

---

[15] 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)). Subjects' earnings in those experiments are based on *absolute* forecast errors. Assuming that 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.

[16] Note that the limit on price forecasts immediately limits the maximum price via the price generating mechanism of Eq. (1). However, the maximum return expectation by itself does not cap the prices. Therefore we set the maximum price in the PR and RR treatments at 1000. If the market reaches that price, subjects are informed about that during the experiment, but not beforehand.

[17] Eleven out of 504 subjects (2.2%) tried to submit a prediction outside of these bounds.

[18] Note that if subjects predict the midpoint of the interval, the prediction is the same independent of the treatments: predicting a return of 0, combined with $p_0 = 50$, is equivalent with a price prediction of 50.

[19] 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 Eq. (1) is then calculated over the subjects who do submit a forecast on time. Such a situation occurred 117 times in total (0.46% of all forecasts).
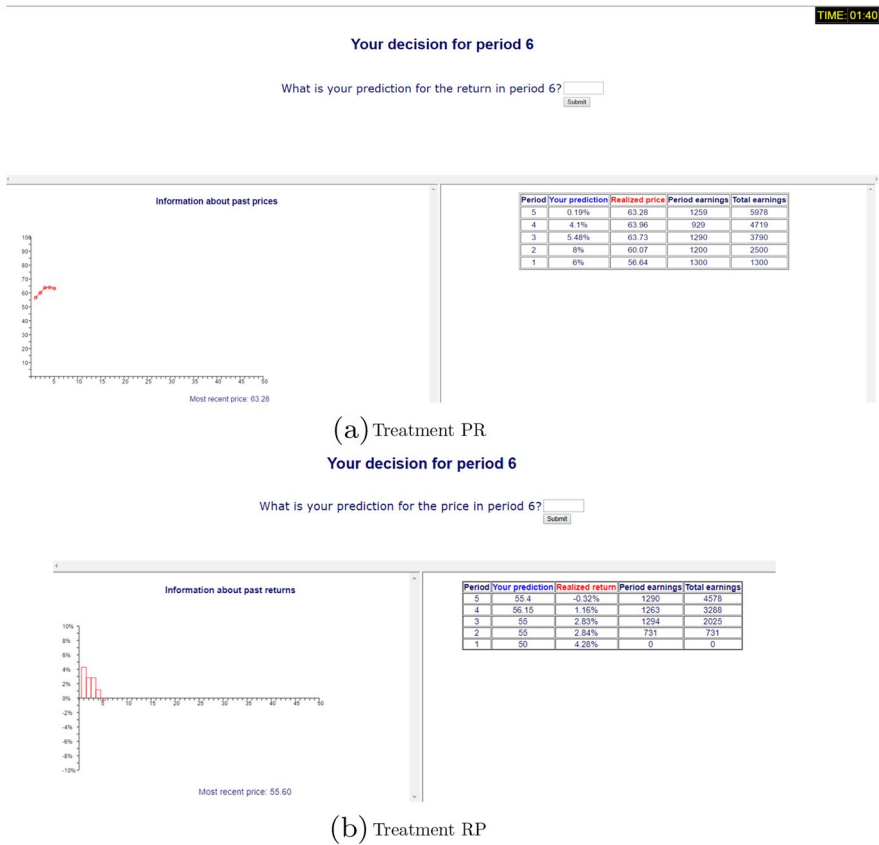
(a) Treatment PR

(b) Treatment RP

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

way. The only difference between treatments is how the task and information are presented to the subjects, 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 individual choice experiments suggest that the format of information and the forecasting task may matter. Indeed, Glaser et al. (2007) and Glaser et al. (2019) provide compelling evidence that expectation formation, for time series of prices/returns that are exogenously given, is affected by how past information and the forecasting task are presented. 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. (2019), on the other hand, find that asking for returns leads to higher forecasts than asking for prices. Both results suggest that at the aggregate market level, 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. (2019) find that showing subjects past returns leads to lower forecasts than when showing them past prices. In earlier learning to forecast experiments price instability typically starts out with a steady increase in asset prices, resulting in a bubble with prices much higher than their underlying fundamental value. Eventually these price increases stop, after which prices start to decline and revert back to the vicinity of the fundamental value, or even overshoot that value. In our setting, lower expectations (due to observing returns) will lead to lower market prices, and this is likely to diminish the extent of price increases and therefore the amplitude of these bubbles. This 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 instability and compare these measures between treatments.

3 The effect of information format and forecasting task on aggregate market outcomes

We present our main findings in this section, starting with providing a first overview of the experimental results in Sect. 3.1. In Sect. 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 market dynamics. Because our main results are slightly diluted by the existence of a number of outlier markets, which are characterised by one or more extreme predictions, we study stability for different subsamples of the data in Sect. 3.3.

3.1 An overview of the experimental results

Figure 2 shows the median market prices in each treatment. That is, for each of the 50 periods we calculate the median price in the given period over the markets in a specific treatment. In order to enhance visibility, we smooth these time series by taking the moving average over 5 periods. The median prices show that there are price oscillations in each treatment but those in treatments PR and RR typically have a larger amplitude (or there are fewer stable markets in these treatments). Figure 2 suggests that the market price in the PP markets (and in the first half of the experiment in the RP markets) stays closer to the fundamental value of 66 whereas markets in the other two treatments exhibit oscillations around the fundamental value. In particular, in 14 of the 45 markets (31%) 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. On the other hand, in treatments PR and RR a much higher fraction

of the markets exhibit prominent fluctuations: only for 4 out of 39 markets (10%) in these two treatments prices are within 10% of the fundamental value for at least 40 consecutive periods. The figure does not provide clear evidence on the effect of observing returns versus observing prices on the volatility of prices and returns.[20]

3.2 Market stability

In this section we formally investigate whether the stability of aggregate market prices is affected by the presentation format of the task and/or information given to the subjects. To that end we use different instability measures, which we calculate for each market on the basis of market prices and returns from period 11 to period 50. We exclude the first ten periods to allow for some learning.

We consider six different measures. Our first three measures are standard measures of price dispersion and volatility: the *standard deviation* of the logreturns (std_r), the *standard deviation* of the market price (std_p) and the *interquartile range* of market prices (*IQR*), which is the difference between the third and the first quartile of the realized price in the given market. That is

$$std_r = \sqrt{\frac{1}{40} \sum_{t=11}^{50} (\ln(p_t/p_{t-1}) - \overline{\ln(1+r)})^2}, \; std_p = \sqrt{\frac{1}{40} \sum_{t=11}^{50} (p_t - \bar{p})^2}, \text{ and}$$

$$IQR = Q_3(p_{11}, \dots, p_{50}) - Q_1(p_{11}, \dots, p_{50}),$$

where $\overline{\ln(1+r)} = \frac{1}{40} \sum_{t=11}^{50} \ln(p_t/p_{t-1})$ is the average realized logreturn over the last 40 periods, $\bar{p} = \frac{1}{40} \sum_{t=11}^{50} p_t$ is the average realized asset price over the last 40 periods and Q_1 and Q_3 denote the first and third quartile, respectively. Our fourth measure, *absolute return* (*AR*) is equal to the median of the absolute returns between periods 11 and 50, that is

$$AR = \text{median}(|r_{11}|, \dots, |r_{50}|).$$

As with the *std* and *IQR* measures, a higher value of *AR* implies higher price volatility.[21]

The four 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 stable but at a level substantially different from the fundamental value these measures will be low, whereas mispricing will be significant. Even though our hypotheses test stability of the market, we investigate two measures of mispricing as well, as they can also be viewed as an indirect measure of stability. These final two measures, *relative absolute deviation* from the fundamental value (*RAD*) and *relative deviation* from the fundamental value (*RD*),

---

[20] Online Appendix C presents a complete set of figures for each market of each treatment. Each figure there includes the realized price and return, and the corresponding forecasts.

[21] 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) or Forsberg and Ghysels (2007)).

Fig. 2 Median price per treatment, smoothed over 5 periods. Treatment ENDO is discussed in Sect. 4
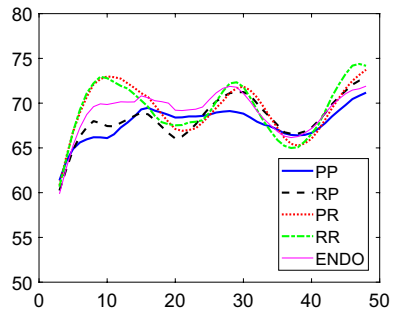
take such deviations from the fundamental value into account (they were introduced in Stöckl et al. (2010), and have become standard bubble measures in the literature on experimental asset markets). These measures are defined as

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

where, in our experiment, $p^* = 66$.[22]

Table 2 summarizes the median values of the instability measures over the markets for each treatment.[23] Furthermore, as our hypotheses test either observing or forecasting prices versus returns, we also report the values for the merged treatments. That is, to test Hypothesis 1 (differences between forecasting price and return) we 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 between observing price and return) we merge PP and PR (into P*) and RP and RR (into R*).[24] The values of the measures AR, RAD and RD are reported in percentages. Although the within-treatment variation in each of the measures is substantial, the measures are to a large extent mutually consistent. The median values show a picture consistent with Fig. 2 the median value of the instability measures is lower in treatments PP and RP than in treatments PR and RR. This is in line with our earlier observation that treatments PP and RP are more stable than the other two treatments.

---

[22] 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 RP13 in Figure C.5 in Online Appendix C, where $p_t > p^*$ for all $t \geq 13$ and consequently $RD \approx RAD$ for this market) or by a low value of RD (if the asset price oscillates around the fundamental value, see e.g., market RR11 in Figure C.12 in Online Appendix C, for which RAD is 6.5 times higher than RD).

[23] Note that these measures will still be nonzero even when every subject predicts the fundamental value, $p_{h,t}^f = p^*$ 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 $std_r = 0.01$, $std_p = 0.50$, $IQR = 0.75$, $AR = 0.66$, $RAD = 0.65$ and $RD = 0.12$ based on the actual noise sequence we used in the experiment.

[24] Values of the instability measures for each of the 84 individual markets, as well as average values and standard deviations per treatment are reported in Tables D.1, D.2 and D.3 in Online Appendix D.

In order to formally test our hypotheses, we use Kolmogorov–Smirnov tests for each measure. The test results are collected in Table 3.[25] Note that we apply one-sided Kolmogorov-Smirnov tests, with the direction given by Hypotheses 1 and 2 – e.g., we test and accept the alternative hypothesis that measure std_r contains lower values for markets in *P than for the markets in *R, which confirms our Hypothesis 1. The tests indicate significant differences ($p < 0.05$) between *P and *R for five of the six instability measures while differences between P* and R* are never significant.[26] Based on these results we do not find support for Hypothesis 2, but we do find support for Hypothesis 1.

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. This is in line with the design chosen in Glaser et al. (2019) but there is an important difference. In Glaser et al. (2019) 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 the prices down and essentially have the possibility to use all past prices for forecasting.

We have established that there is a difference between the distributions of the markets where subjects need to predict prices compared to markets where subjects need to predict returns, with the former tending to be more stable. Now we investigate with a multivariate multiple regression whether these findings also hold true when considering the exact values of the instability measures. Table 4 presents the results of the multivariate multiple linear regression with the six different measures as the dependent variables and two dummy variables as independent variables to test our hypotheses: *P which is one for treatments RP and PP, and P* which is one for treatments PR and PP.[27] The regression results cannot support either hypothesis with almost all dummy variables being insignificant. Based on the above tests our main result is the following:

Result 1 We find some evidence supporting the hypothesis that forecasting prices tends to lead to more stable market dynamics than forecasting returns. The format in which past information is presented has no effect on market stability.

Our result is based upon the observation that, when comparing distributions, forecasting prices clearly leads to more stable market dynamics than forecasting returns. However, this difference is not significant for the regression results. A possible reason behind the difference between the Kolmogorov-Smirnov test and the regression

---

[25] We also make pairwise comparisons between the four individual treatments, the corresponding test results are presented in Table D.4 in Online Appendix D.

[26] Notice that our conclusions remain valid for AR and RAD even if we apply the (very conservative) Bonferroni-Holm correction to account for multiple hypothesis testing.

[27] We decided not to include the interaction term PP in the regression, as that makes our conclusions about the hypotheses more difficult to draw. In the current model the dummy *P tests Hypothesis 1, whereas P* tests Hypothesis 2.

Table 2 Median values of the instability measures over the markets for each treatment, and combined treatments per information or task

	std_r	std_p	IQR	AR	RAD	RD
PP	0.019	4.35	4.52	0.93	6.56	4.89
RP	0.025	5.15	6.33	1.17	7.57	5.49
PR	0.024	5.35	9.54	1.43	8.18	7.40
RR	0.023	5.34	10.26	2.09	9.21	5.74
*P	0.020	4.94	6.12	1.00	7.01	5.35
*R	0.024	5.35	9.78	1.81	8.53	6.23
P*	0.020	4.94	6.37	1.32	7.92	6.50
R*	0.024	5.15	9.45	1.38	8.94	5.62

AR, RAD and RD are reported in percentages

results can be seen when looking at the individual markets. Both the figures in Online Appendix C and the exact values of the six measures per market in Online Appendix D show the same picture. Markets in *P tend to be more extreme than markets in *R. This means that considering the more stable markets (with lower instability measures), we observe more *P markets with smaller values. However, when we look at the other end of the distribution, we again see more *P markets than *R markets. Note however that there are more relatively stable markets than markets with very high price volatility. Furthermore, these markets with extraordinary high price volatility result in very high instability measures which can be seen in Tables D.2 and D.3 in Online Appendix D, as well as in Table D.1 when comparing the medians and the standard deviations. This low number of substantially higher values could lead to positive, but insignificant regression coefficients. Due to this feature of our data, in the next section we investigate different subsamples of the data.

3.3 Outlier markets and sample split

When looking at the data of individual markets, we can see that there are several markets where one out of the six subjects submits a very different forecast than the other five subjects (see individual figures in Online Appendix C). Such outliers have a substantial and long-lasting effect on the market dynamics in the given market (usually, but not necessarily, resulting in larger bubbles) and thereby obfuscate the differences between treatments. These outliers typically come from two sources: either subjects make a typo, or they experiment with very high (or very low) forecasts. While in real life we cannot rule out such a behavior or typos (or "fat-finger errors"[28]) either, it does not happen that often.

In order to investigate the effect these outliers have on our main result, we consider two ways to look at a subsample of the data. First, we remove outlier markets.

---

[28] See e.g., https://www.theguardian.com/business/2022/jun/03/flash-crash-set-off-by-fat-fingered-citigroup-trader-could-cost-50 m, retrieved on 12 September, 2022.

Table 3 Summary of p-values of the Kolmogorov–Smirnov tests for comparing treatments in terms of instability

	std_r	std_p	IQR	AR	RAD	RD
*P vs *R	0.046*	0.053	0.024*	0.002**	0.008**	0.019*
P* vs R*	0.874	0.941	1.000	0.978	0.849	0.326

***Significant at 0.1% level, **Significant at 1% level, *Significant at 5% level. All tests are one-sided. Observations correspond to markets

The number of observations is $n_{*P} = 45$, $n_{*R} = 39$, $n_{P*} = 41$ and $n_{R*} = 43$

Table 4 Multivariate multiple linear regressions for testing Hypotheses 1 and 2

	Dependent variable					
	std_r	std_p	IQR	AR	RAD	RD
*P	0.029* (0.013)	5.890 (6.569)	6.405 (5.803)	0.551 (0.734)	37.62 (25.06)	36.76 (25.19)
P*	−0.002 (0.013)	−2.578 (6.554)	−9.916 (5.790)	−1.051 (0.732)	−32.72 (25.01)	−31.19 (25.13)
constant	0.036** (0.012)	13.76* (5.772)	18.801*** (5.099)	2.989*** (0.645)	30.97 (22.02)	26.80 (22.13)
R^2	0.0529	0.0117	0.0487	0.0313	0.0466	0.0433

***Significant at 0.1% level, **Significant at 1% level, *Significant at 5% level. Standard errors are in brackets

P denotes a dummy being 1 for treatments PP and RP, and P is a dummy being 1 for treatments PP and PR

Observations correspond to markets, the number of observations is 84

Second, as the first method results in an unbalanced distribution of markets across treatments, as an alternative we split the sample evenly per treatment.

We identify a subject's forecast in period $t > 1$ as an outlier if it deviates by more than 50% from the *median* forecast, taken over the six participants in that market in that period. Because typically there is a high degree of coordination between forecasts, a forecast far from this median can indeed be considered an outlier. We do not consider outliers in the first period, $t = 1$, because subjects have no information about past prices in that period, and their forecasts therefore are not necessarily close to each other.[29] Unlike the main analysis, to identify outliers we look at periods before period 11 as well, as having an outlier forecast might have a long-lasting effect on prices. By using this approach, we identify 9 (out of 22) markets in PP, 13 (out of 23) markets in RP, 5 (out of 19) markets in PR and 2 (out of 20) markets in

---

[29] A way around the issue of outliers would have been to let the realized price in Eq. (1) be determined by the median forecast, as in Arifovic and Petersen (2017), or to truncate forecasts that deviate too much from the last price, as in Kopányi-Peuker and Weber (2021). Note, however, that the first approach is not consistent with the underlying asset pricing model, and the second might have made the experiment even more complicated to subjects.

RR as outliers. These markets are classified as outlier markets based on a total of 112 outlier predictions which corresponds to 0.4% of all possible predictions after period 1.[30]

Table 5 confirms the observation that markets that do not feature outliers are much more stable in *P than in *R. When restricting our analysis to markets that do not exhibit such outlier behavior, we see a highly significant difference in the distribution for predicting price vs predicting return, with the former being more stable (see Panel A). Furthermore, the regressions (Panel B) also show that predicting price results in lower instability measures, with a significant negative coefficient for four out of the six measures. Whether subjects observe prices or returns still does not make a difference for stability.

Let us comment on the uneven effect of a typo in the different treatments. Due to the nature of the forecasting task a high forecast may cause larger price fluctuations in the *P treatments than in the *R treatments. For example, when a subject makes a typo by incidentally typing an extra zero, then the forecasted price of that person will be 10 times larger. However, if we consider a relatively stable market in the *R treatment, where the returns are below 10%, such a typo has a much smaller effect. An extra zero would mean a maximum return forecast of 100%, which translates to a price forecast that is only double the original price level. The same holds when submitting the maximum possible return forecast of 300%, which is equivalent with a price prediction that is four times the current price. If we are in a stable market (with prices around 65-70), this price prediction is much lower than the maximum possible price in *P (1000). Outlier markets in *P therefore have the potential to be much more unstable than outlier markets in *R.

One might argue that the subsample we obtain after removing the outliers is selective as we end up with an unbalanced sample with respect to the number of observations per treatment. To circumvent that we consider a sample split based on a median market *per treatment*, which results in a more even split of the total sample as well as a more even split for each treatment. To do so, we combine the six instability measures discussed above to rank the markets per treatment. We implement the ranking in the following way. For each measure we rank the markets in each treatment by giving the market with the lowest value for that measure rank 1, and so on, up to rank n for the market with the highest value of that measure. Subsequently we order all markets in the given treatment by the sum of the ranks they have for the six measures. This gives us an overall ranking for each market in each treatment. We determine the median rank for each treatment, and all markets with a lower rank belong to the stable sample, whereas all markets with a higher ranking belong to the unstable sample.[31]

---

[30] We mark these outlier markets in Online Appendix D. From that appendix it is also clear that not all outlier markets are unstable markets (e.g., RP12), and not all unstable markets are outlier markets (e.g., PP16). Note that the number of outlier markets is consistent with our previous discussion in the sense that these outliers usually result in more instability, and we see more of them in the *P treatments.

[31] It is possible that there are markets exactly with the median rank (e.g., due to ties or an odd number of markets in the treatment). In those cases we include the median market both in the stable and in the unstable sample, but this choice does not influence the results much.

Table 6 shows the results for the Kolmogorov-Smirnov tests (Panels A and C) and the multivariate multiple regressions (Panels B and D). In Panels A and B we present the markets with the better ranking (stable markets) and, for completeness, in Panels C and D we present the results for the more unstable markets. Given that outlier markets often exhibit unstable market dynamics, the results for the stable markets are very similar to the results we obtained after focusing on the non-outlier markets. Predicting price results in a more stable market dynamics in those markets (the regression results are even stronger than in Table 5) compared to the markets where subjects predict returns. There is no significant difference between observing prices and returns. Looking at the unstable markets (see Panels C and D in Table 6), we find no support for our hypotheses. These findings lead us to our second result:

Result 2 After correcting for outliers, whether by only considering non-outlier markets, or by considering only the stable half of the markets, we find that forecasting prices leads to more stable markets than forecasting returns. The differences are strongly significant. The format in which past information is presented still has no effect on market stability.

Even though our main analyses and hypotheses connect with market stability and aggregate market behavior, we have also looked at individual behavior for a potential behavioral explanation behind the results. We investigate how subjects react to past price changes by looking at trend-extrapolation with a regression on decision-level data. Furthermore, we estimate for each individual a forecasting rule to measure the most important factors (past prices or past own forecasts, both with several lags) that enter the decision making process. Both analyses show that subjects forecasting returns extrapolate past price changes stronger than subjects forecasting prices. These results are consistent with the findings of Glaser et al. (2007) who also found stronger trend-extrapolation when predicting returns. The analyses are relegated to Online Appendices E and F.

4 Endogenous choice of task and information

In the main experiment subjects receive information about either past returns or past prices, and they have to submit either price or return forecasts. However, in reality, investors may not be bound to one type of information, but might be free to choose what to look at, and whether to forecast prices or forecast returns. We designed an additional treatment, treatment ENDO, to investigate the effect of the endogenous choice of these two formats on market stability.[32] In Sect. 4.1 we present the design of this additional treatment, and in Sects. 4.2 and 4.3 we discuss the experimental results.

---

[32] We are grateful to the editor for suggesting such a treatment.

Table 5 Kolmogorov-Smirnov tests and multivariate multiple linear regressions after removing outlier markets

Panel A: Kolmogorov–Smirnov test results

	std_r	std_p	IQR	AR	RAD	RD
*P vs *R	0.000***	0.001**	0.000***	0.000***	0.000***	0.000***
P* vs R*	0.991	0.958	0.979	0.963	0.958	0.831

Panel B: Multivariate multiple regression results

Dependent variable

*P	−0.01 (0.01)	−2.56* (1.11)	−4.74* (1.90)	−0.99 (0.52)	−3.74** (1.34)	−2.74*** (0.66)
P*	−0.005 (0.01)	−0.48 (1.09)	−1.36 (1.87)	−0.42 (0.51)	−0.74 (1.32)	0.15 (0.65)
constant	0.03*** (0.004)	6.13*** (0.86)	10.73*** (1.47)	2.53*** (0.40)	9.60*** (1.03)	5.92*** (0.51)
R^2	0.09	0.10	0.12	0.08	0.14	0.25

***Significant at 0.1% level, **Significant at 1% level, *Significant at 5% level. Panel A presents results of one-sided Kolmogorov–Smirnov tests with observations per market: $n_{*P} = 23$, $n_{*R} = 32$, $n_{P*} = 27$, $n_{R*} = 28$. Panel B presents regression coefficients for a multivariate multiple regression. Standard errors are in brackets. *P denotes a dummy being 1 for treatments PP and RP, and P* is a dummy being 1 for treatments PP and PR. Observations correspond to markets, the number of observations is 55

4.1 Experimental design

The design of treatment ENDO is largely the same as the treatments described in Sect. 2. In contrast to these other treatments the ENDO treatment allows for a choice between submitting a price or a return, and for switching between return or price history information. Figure 3 illustrates subjects' screen in this treatment. The change in the table while switching observed information only concerns the market realizations, subjects always see their own forecasts as they have submitted them before.

The incentives, information and time given to subjects are the same as in the main treatments.[33] Important to note is that we give earning examples to subjects in the instructions, so that they can see that there is no monetary advantage of forecasting either variable.[34]

---

[33] Subjects did not make a decision on time in 84 cases (1.65% of all decisions). Furthermore, 13 out of 102 subjects tried to submit prices or returns outside the boundary at least once. The total number of such decisions was 14, out of 5100 decisions.

[34] The instructions can be found in Online Appendix B. Note also that we randomize the order of prices and returns in the instructions and on the decision screen. Half of the subjects always see "prices and returns" in the instructions, and then returns are shown on the top in the decision panel when they make their decisions. For the other half of the subjects, this is the other way around. We choose this randomization in order to minimize the probability that subjects just forecast what they see first. We do not find any significant correlation between the first period choice of prediction task (price versus return) and the variable we show on the top for subjects (p-value of the Spearman correlation test is 0.28), thus there seems to be no particular order effect playing a role in the decisions.

Table 6 Kolmogorov–Smirnov tests and multivariate multiple linear regressions for the split samples

Panel A: Kolmogorov–Smirnov test results—stable markets

	std_r	std_p	IQR	AR	RAD	RD
*P vs *R	0.002**	0.003**	0.003**	0.000***	0.000***	0.001**
P* vs R*	0.823	0.952	0.952	0.929	0.996	0.808

Panel B: Multivariate multiple regression results—stable markets

Dependent variable

	std_r	std_p	IQR	AR	RAD	RD
*P	−0.004* (0.002)	−1.45*** (0.36)	−2.82*** (0.63)	−0.51*** (0.10)	−2.80*** (0.51)	−3.04*** (0.65)
P*	−0.001 (0.002)	−0.51 (0.36)	−1.29* (0.63)	−0.13 (0.10)	−0.79 (0.50)	−0.43 (0.65)
constant	0.02*** (0.002)	3.94*** (0.32)	6.67*** (0.56)	1.33*** (0.09)	7.20*** (0.45)	5.83*** (0.57)
R^2	0.11	0.31	0.37	0.39	0.45	0.36

Panel C: Kolmogorov–Smirnov test results—unstable markets

	std_r	std_p	IQR	AR	RAD	RD
*P vs *R	0.973	0.705	0.138	0.413	0.959	0.847
P* vs R*	0.834	0.956	1.000	0.956	0.834	0.195

Panel D: Multivariate multiple regression results—unstable markets

	Dependent variable					
	std_r	std_p	IQR	AR	RAD	RD
*P	0.06** (0.02)	13.58 (11.44)	15.15 (9.46)	1.58 (1.16)	74.26 (45.76)	72.83 (46.36)
P*	−0.003 (0.02)	−4.92 (11.43)	−18.02 (9.45)	−1.93 (1.16)	−60.82 (45.71)	−58.17 (46.31)
constant	0.05** (0.02)	22.74* (10.20)	30.39** (8.44)	4.57*** (1.04)	54.03 (40.81)	47.17 (41.34)
R^2	0.19	0.04	0.14	0.11	0.10	0.09

***Significant at 0.1% level, **Significant at 1% level, *Significant at 5% level. Panel A and C present results of one-sided Kolmogorov-Smirnov tests with $n_{*P} = 23$, $n_{*R} = 20$, $n_{P*} = 21$, $n_{R*} = 22$ observations for Panel A and $n_{*P} = 23$, $n_{*R} = 21$, $n_{P*} = 22$, $n_{R*} = 22$ observations for Panel C. Panel B and D present regression coefficients for a multivariate multiple regression. Standard errors are in brackets. *P denotes a dummy being 1 for treatments PP and RP, and P* is a dummy being 1 for treatments PP and PR. Observations correspond to markets, the number of observations is 43 in panel B and 44 in panel D

We collected data on 17 markets for this treatment. In total 102 subjects from the University of Amsterdam participated in June and September 2022. The practical procedure was the same as in the main treatments, including incentives, exchange rates and decision time restrictions. The sessions (up to payments) lasted about 100 min and subjects typically earned between €24 and €27.5.
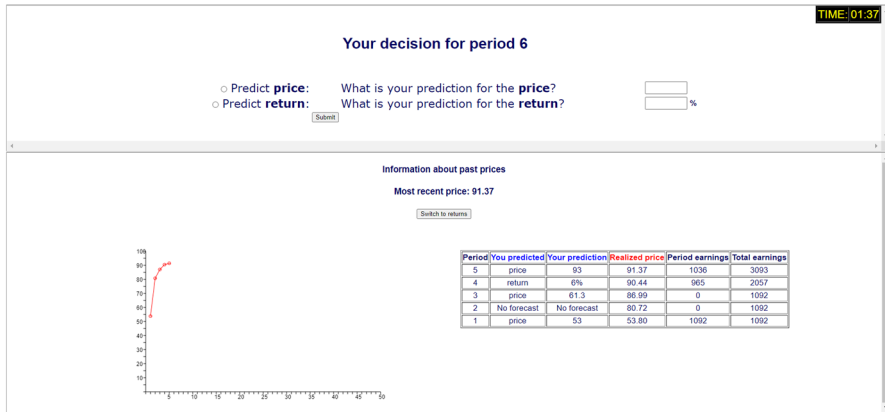
Fig. 3 Example of the decision screen in treatment ENDO

4.2 Stability of the endogenous markets

Figure 2 in Sect. 3.1 not only displays the median prices for the original treatments, but also for treatment ENDO. In the first half of the experiment the median market in this treatment is relatively stable with a price slightly higher than the price in treatments PP and RP. In the second half of the experiment we see similar oscillations as for the median markets in the other treatments.

We use the Kolmogorov-Smirnov test for pairwise comparisons of the six instability measures between treatment ENDO and all original individual and merged treatments.[35] Again, we restrict our analysis to the last 40 periods to account for learning. As we do not have specific hypotheses about the direction of potential differences, all tests are two-sided. The non-parametric tests reveal no systematic differences; most of the test results are not significant at the 5% significance level (see Table D.5 in Online Appendix D). Multivariate regressions confirm these results, we do not find systematic differences between the instability measures of treatment ENDO compared to the other treatments (see Table D.7 in Online Appendix D).

For completeness, we repeat the analysis for the same sample splits that we used for the other treatments. We identify 10 markets as outliers in treatment ENDO.[36] The non-parametric tests and regressions do not reveal systematic differences between treatments for the non-outlier markets or for the stable or unstable markets.

---

[35] The test results, as well as the values per market, the median and mean values, and standard deviations of the instability measures are relegated to Tables D.5 and D.6 in Online Appendix D. Figures for each market (including individual forecasts, realized prices, returns and the number of subjects choosing to forecast a given variable) are shown in Online Appendix C.

[36] Table D.6 in Online Appendix D indicates which markets are outlier markets. Note that our measure is rather strict, and might identify more stable markets as outliers as well, like market ENDO10 where one subject submitted a very high forecast in the very last period (see Figure C.15 in Online Appendix C).

The only significant difference that occurs for multiple instability measures and both for the non-outlier markets as well as for the more stable markets is that the markets from treatment ENDO seem to be a bit more stable than markets in the RR treatment (and also more stable than the *R treatments, see Tables D.8 and D.9 in Online Appendix D). These results hold for the first four measures, but not for RAD and RD. For these two measures the Kolmogorov-Smirnov tests find that treatment ENDO is less stable than treatment PP (and than *P, except for RAD in the stable markets). These results from the sample splits suggest that, consistent with Fig. 2, treatment ENDO lies somewhere between the most and least stable treatments. However, given that we only find these differences for specific treatments and for the sample split, we conclude the following:

Result 3 Giving subjects the freedom to choose the format of the task and the information does not lead to systematically more stable or systematically more unstable market dynamics.

4.3 Choice of task and information in the endogenous markets

In this subsection we investigate the choices subjects make in treatment ENDO. As we already pointed out earlier, subjects do not seem to be influenced by the variable presented to them on the top of the decision screen. In fact, in the first period 57 subjects choose to predict prices, and 38 choose to predict returns. These choices are uncorrelated with the order of the variables on the screen. Looking at the last 40 periods, 58% of subjects (59 out of 102) always predict the same variable, so do not switch prediction task. Throughout these 40 periods, 50.41% of all decisions are price predictions, and 50.49% of the last seen information is price information. Subjects switch prediction task on average 2.94 times throughout the 40 periods (standard deviation of 4.56 and maximum of 21) with a median of zero switches. They switch information while thinking about their prediction (thus within a period) on average 0.65 times with zero median (standard deviation of 1.53) and a maximum of 19 switches within a single period by a single subject. Thus the majority of subjects stick both to the information they saw in the previous period and to the task they chose in the previous period, and are reluctant to change. Importantly, these two variables are the same for the large majority of the decisions. When using the last seen variable in a period as a proxy for the information subjects use to make their decision in that period, we find that in the last 40 periods of the experiment 40.8% of all decisions fall in treatment category PP, 9.6% in RP, 9.6% in PR and 40.0% in RR. The correlation is very strong, subjects are likely to submit a price prediction if they have seen the price chart last, and a return prediction if they have chosen to look at returns (the p-value of the Spearman correlation test is < 0.001). This shows that the PP and RR treatments provide a more natural environment for subjects than the PR and RP treatments.

Finally, one might ask how the group composition in terms of number of subjects choosing to observe or forecast prices affects stability. We used multivariate multiple linear regressions again to investigate this question with markets as observations.

The analysis is relegated to Online Appendix D (see Table D.10). Based on this analysis within treatment ENDO we do not find support for either Hypothesis 1 or Hypothesis 2. This is probably due to the fact that the markets are already relatively close to each other in terms of stability, we have only 17 observations, and that the markets constitute a mix between price forecasters and return forecasters. Also note that this treatment was not designed to test these hypotheses.

To summarize our analysis of treatment ENDO, we find no systematic differences between this treatment and the initial treatments. However, we do find that 80% of the decisions are made such that the information and the decision task are the same, and subjects do not have a clear preference for either predicting price or predicting returns (around 50% of the previously mentioned "same variable" decisions are price decisions, the other half are return decisions). Thus our findings for the initial four treatments are relevant, as all treatment categories, especially PP and RR are observed among the individual decisions. Given our results on the difference between predicting prices and predicting returns, our experiment suggests that regulation that steers investors to think about prices instead of returns facilitates market stability.[37]

5 Concluding remarks

In this paper we fill an important gap in the research on the impact of the presentation format of past information and of the choice of forecasting task on expectation formation. In particular, where previous research focused on individual decision making, we analyze how aggregate behavior is impacted, by acknowledging that expectations have an effect on actual investor behavior, and therefore on realized market dynamics. We go beyond earlier contributions that only consider the effect on forecasting an exogenously given time series, and consider the more realistic environment where market prices are determined endogenously. Our contribution is an example of mapping individual decision biases into aggregate behavior.

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 market 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 observations, see e.g., Sirri and Tufano (1998); Choi et al. (2009) and Greenwood and Shleifer (2014). Our results

---

[37] Also, the pairwise comparison between treatments on the whole sample (Table D.4 in Online Appendix D) supports this policy implication. We find that the tests between PP and RR show a significant difference in the predicted direction for three out of the six instability measures. This is stronger than the differences we find for any other combination of treatments.

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. As a second contribution, we designed an additional treatment where subjects had the opportunity to choose the formats of the forecasting task and the presentation of the information themselves. We find no clear preference for forecasting prices or returns, but we do find a clear preference for forecasting and observing the same variables, a more natural, cognitively less demanding task, corresponding to treatments PP and RR of the main experiment.

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 (adapted from Andreassen (1987)): 60, 62, 65, 68, 70. Under the assumption that the representative price corresponds to the mean or median price of that sequence, the participant's prediction would be around 65. However, the returns corresponding to this series of prices are 3.33%, 4.84%, 4.62% and 2.94%, respectively, with a representative (average) return of around 4%, which translates into a price forecast of about 72.8, instead of the mean reversion resulting from a price forecast of 65. Admittedly, it is not obvious that the average price or return is representative for the sequence of prices or returns.[38] Nevertheless, even if subjects have simple 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. This is consistent with the stronger trend extrapolation we find in treatments PR and RR.

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. We have already discussed our first explanation before Result 1 in Sect. 3.2, namely that we provide subjects in treatments RP and RR with the most recent price, not only with returns. Second, 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. 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

---

[38] In fact, Grosshans and Zeisberger (2018) and Borsboom and Zeisberger (2020), for example, argue that the pattern of past prices, in addition to the level, also plays an important role in participants' risk perception and return expectations. In particular, they find that recent price-developments have a substantial effect on beliefs.

this is – to a certain extent – up to the market participant itself (except professional forecasters and financial analysts who are asked for predictions in a specific format by investors). One may imagine that investors who observe past returns will be more inclined to forecast future returns than investors who 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. Even though our endogenous treatment (ENDO) did not explicitly test this explanation, we do indeed find that subjects have a preference to observe and predict the same variable, although there does not seem to be a preference for either price forecasts or return forecasts. Given that in the first period subjects do not have any past information, it might be that they choose the task first and then the information they want to observe. However, by presenting past performance by means of prices, traders may be nudged into forecasting prices as well, leading to a decrease in price volatility. We leave this question for possible future research.

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 Eq. (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 (2019)).[39] 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 an overview of 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 emerge in 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

---

[39] 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.

that portfolio choices can be explained to a large extent by survey expectations.[40] We therefore believe that our results will translate to an experiment where beliefs of the participants are elicited, either as return forecast or as price forecast, and participants subsequently can trade in the asset. Our conjecture is that in the former case participants show a higher willingness to buy (sell) assets, also against prices higher (lower) than the fundamental value. We leave it to future work to investigate this issue.

Declarations

Conflict of interest Not applicable.

Consent to participate When registering for the experiment, subjects gave their consent to participate.

Consent for publication Publication of the manuscript has been approved by all co-authors.

Ethics approval Ethics approval was given by the University of Amsterdam.

---

[40] Also see (Carlé et al., 2019) and Füllbrunn et al. (2022) who find that forecasting behavior in experimental asset market with trade a la Smith et al. (1988) is generally in line with trading behavior. Relatedly, Hanaki et al. (2018) show that, whether forecasts are explicitly elicited or not does not affect mispricing, again suggesting that forecasts are consistent with trading decisions.

References

Amromin, G., & 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., & Kraus, S. J. (1990). Judgmental extrapolation and the salience of change. *Journal of Forecasting, 9*(4), 347–372.

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

Arifovic, J., & Petersen, L. (2017). Stabilizing expectations at the zero lower bound: Experimental evidence. *Journal of Economic Dynamics and Control, 82*, 21–43.

Bacchetta, P., Mertens, E., & 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., & Hommes, C. (2019). When speculators meet suppliers: Positive versus negative feedback in experimental housing markets. *Journal of Economic Dynamics & Control, 107*, 103730.

Bao, T., Hommes, C., & 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., & 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., & Madrian, B. C. (2013). What goes up must come down? Experimental evidence on intuitive forecasting. *American Economic Review, 103*(3), 570–74.

Borsboom, C., & Zeisberger, S. (2020). What makes an investment risky? An analysis of price path characteristics. *Journal of Economic Behavior & Organization, 169*, 92–125.

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

Carlé, T. A., Lahav, Y., Neugebauer, T., & Noussair, C. N. (2019). Heterogeneity of beliefs and trade in experimental asset markets. *Journal of Financial and Quantitative Analysis, 54*(1), 215–245.

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., & 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., & Metrick, A. (2009). Reinforcement learning and savings behavior. *The Journal of Finance, 64*(6), 2515–2534.

Czaczkes, B., & 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., & 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., & Mason, T. I. (1993). Tests of rational expectations in a stark setting. *The Economic Journal, 103*(418), 586–601.

Forsberg, L., & Ghysels, E. (2007). Why do absolute returns predict volatility so well? *Journal of Financial Econometrics, 5*(1), 31–67.

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

Füllbrunn, S., Huber, C., Eckel, C., & Weitzel, U. (2023). Heterogeneity of beliefs and trading behavior: A reexamination. Forthcoming in the *Journal of Financial and Quantitative Analysis*. https://doi.org/10.1017/S002210902300011X

Ghysels, E., Santa-Clara, P., & 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., & Weber, M. (2019). Thinking about prices versus thinking about returns in financial markets. *Journal of Finance, 74*(6), 2997–3039.

Glaser, M., Langer, T., Reynders, J., & 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., & Shleifer, A. (2014). Expectations of returns and expected returns. *The Review of Financial Studies, 27*(3), 714–746.

Grosshans, D., & Zeisberger, S. (2018). All's well that ends well? On the importance of how returns are achieved. *Journal of Banking & Finance, 87*, 397–410.

Hanaki, N., Akiyama, E., & 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., & 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., & 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., & Sonnemans, J. (2021). Bubbles, crashes and information contagion in large-group asset market experiments. *Experimental Economics, 24*(2), 414–433.

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

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

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

Kirchler, E., Maciejovsky, B., & 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., & Stöckl, T. (2012). Thar she bursts: Reducing confusion reduces bubbles. *American Economic Review, 102*(2), 865–83.

Kopányi-Peuker, A., & Weber, M. (2021). Experience does not eliminate bubbles: Experimental evidence. *Review of Financial Studies, 34*(9), 4450–4485.

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

Marimon, R., Spear, S. E., & 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.

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

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

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

Sonnemans, J., & 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., & Kirchler, M. (2010). Bubble measures in experimental asset markets. *Experimental Economics, 13*(3), 284–298.

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

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Nobuyuki Hanaki[1] · **Cars Hommes[2]** · **Dávid Kopányi[3]** · **Anita Kopányi-Peuker[4]** · **Jan Tuinstra[5]**

✉ Jan Tuinstra
   j.tuinstra@uva.nl

   Nobuyuki Hanaki
   nobuyuki.hanaki@iser.osaka-u.ac.jp

   Cars Hommes
   c.h.hommes@uva.nl

   Dávid Kopányi
   david.kopanyi@acm.nl

   Anita Kopányi-Peuker
   anita.kopanyi-peuker@ru.nl

[1]   Institute of Social and Economic Research, Osaka University, Osaka, Japan

[2]   University of Amsterdam and Tinbergen Institute Amsterdam, The Netherlands, and Bank of Canada, Ottawa, Canada

[3]   The Netherlands Authority for Consumers and Markets, The Hague, The Netherlands

[4]   Radboud University Nijmegen, Institute for Management Research, Nijmegen, The Netherlands

[5]   University of Amsterdam and Tinbergen Institute, Amsterdam, The Netherlands