Journal of Economic Behavior and Organization 219 (2024) 510–527

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Reading the market? Expectation coordination and theory of mind☆

Te Bao [a], Sascha Füllbrunn [b], Jiaoying Pei [c], Jichuan Zong [d,*]

[a] School of Social Sciences and NTU-WeBank Joint Research Centre on FinTech, Nanyang Technological University, Singapore
[b] Radboud University, Institute for Management Research, Department of Economics and Business Economics, Heyendaalseweg 141, 6525 AJ Nijmegen, the Netherlands
[c] School of Social Sciences, Nanyang Environment & Water Research Institute, Environmental Process Modelling Centre, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore
[d] School of Finance and Laboratory of Experimental Ecoomics, Dongbei University of Finance and Economics, Dalian, China

ARTICLE INFO

JEL Classification:
C92
G40
G17

Keywords:
Theory of mind
Strategic uncertainty
Asset bubbles
Experimental Finance

ABSTRACT

Suppose that all asset market traders are proficient at *reading the market*. Would markets become more stable, resulting in lower volatility and fewer price bubbles? To answer this question, we test whether Theory of Mind (ToM) capabilities enhance expectation coordination and reduce expectation heterogeneity and price bubbles in learning-to-forecast experiments. We compare the price and expectation dynamics between markets composed of participants with either high or low ToM capabilities as measured by the eye gaze test. Despite an economically substantial difference between the two groups, we find no statistically significant differences in the measures of expectation coordination, price bubbles, market stability, and expectation heterogeneity.

1. Introduction

According to Nobel prize winner Richard Thaler, one primary difficulty of financial forecasting is that: "*this game is identical to Keynes's beauty contest: you have to guess what other people are thinking that other people are thinking*" (Thaler, 2015). The guessing ability here is supposed to be related to the *Theory of Mind* (henceforth ToM), a psychological trait Frith and Happé (1999) raised to describe the capacity to detect intentionality in the market, i.e., "read" the intentions and act successfully upon them.

In this study, we investigate whether a market in which bubbles are inherently likely to emerge and expectations have difficulty converging to the fundamental level would form smaller price bubbles and become less volatile when populated with traders possessing *all* high ToM skills.

We applied Learning-to-forecast experiments (henceforth LtFE, Marimon et al., 1993, Hommes, 2011, 2021) to investigate this question. The participants in our experiment assumed the role of professional forecasters who periodically submitted their expectations of future market prices. After collecting individual expectations, a computer algorithm determines the conditional optimal realization of the price given the forecasters' expectations. Forecasters learn about realized prices and submit new expectations. This

☆ This study is a pre-registered study at AEA RCT Registry (RCT ID: AEARCTR-0007836). It is approved by IRB of Dongbei University of Finance and Economics (F20210503). Financial support from a Tier 1 Grant from the MOE of Singapore (RG 69/19), NTU-WeBank Joint Research Centre on FinTech (NWJ-2020-003) and National Natural Science Foundation of China (72173017) is gratefully acknowledged.

* Corresponding author.
E-mail address: zong@dufe.edu.cn (J. Zong).

https://doi.org/10.1016/j.jebo.2024.01.018
Received 1 December 2022; Received in revised form 20 December 2023; Accepted 19 January 2024

design is beneficial for our research question, as the subjects need to read market signals to make decisions. As we are interested in the correlation between price bubbles and ToM skills, we used positive-feedback LtFE, where subjects usually fail to learn the rational expectation equilibrium, and the market price exhibits long and persistent bubble and crash patterns[1] (Bao et al., 2021).

Before the LtFE, we obtained the subjects' ToM scores using the eye-gaze test (Baron-Cohen et al., 1997). The computer ranked the subjects in each session based on their ToM scores and composed the markets based on the similarity of the ToM scores, which were unknown to the subjects. While we are interested in whether markets perform better when they show high ToM skill, recent experiments have instead considered correlational studies of ToM scores and trading performance. Corgnet et al. (2018) and Hefti et al. (2018) treated the ToM score as an independent trait for each subject; they tested whether high ToM subjects performed better in experimental asset markets than low ToM subjects. By contrast, our design allows us to examine whether markets populated with the highest ToM scores (High-ToM group) are more stable, with fewer bubbles or crashes, compared to markets with the lowest ToM scores (Low-ToM group). Our design aligns with other quasi-experiments that comprise markets based on subject characteristics. For example, Eckel and Füllbrunn (2015) composed markets by subjects' sex (male vs. female markets), Bosch-Rosa et al. (2018) by cognitive ability, Janssen et al. (2019) by the propensity to speculate, and Füllbrunn et al. (2019) by risk attitudes and subjects' gender.

Our results showed no significant differences between the high- and low-ToM groups using several performance measures. However, on average, the low-ToM group showed larger price bubbles, higher price variability, and worse coordination in price forecasts. After estimating the learning heuristics, we found an equal number of subjects using adaptive expectations, trend-extrapolation rules, naïve expectations, and fundamental forecasts in the Low- and High-ToM groups. Hence, ToM skills do not seem to affect aggregated market performance, and there are still heterogeneities in the belief formation of price forecasts.

Our study closely relates to that of Corgnet et al. (2018). They examined the collective impact of individual characteristics, including fluid intelligence, cognitive reflection, and ToM, on subjects' ability to aggregate information in double auction markets. In a static environment, subjects have private signals about the value of an asset that yields a true value in aggregation. Subjects with high ToM skills are supposed to 'read' the private signals from the other subjects' bids and asks, better estimate the true value, and, thus, increase trading profits. They report correlations in market performance at the individual level, but not at the market level, via systematic market composition. In our LtFE, there is an expectation feedback loop between the subjects' expectations and the realized market price; that is, the realized market price is a function of the average expectations. We believe that ToM is more relevant for our study. Subjects in an LtFE need to infer others' thinking and decisions in a dynamic environment in real time, and the prices will endogenously depend on their real-time expectation formation.

Our study is related to the literature that investigates the correlation between forecasters' ToM skills and forecasting performance in *other* financial market experiments. Specifically, those who forecast market prices are not involved in trading. For example, Bruguier et al. (2010) find that subjects with high ToM skills better predict price changes in markets with insiders. Later, Bossaerts et al. (2019) replayed the markets in Bruguier et al. (2010) and De Martino et al. (2013) and studied the neurobiological foundation of the role of ToM. Similarly, Corgnet et al. (2021a) jointly assessed the predictive power of individual characteristics (including ToM) on the forecasting performance of the true value of the asset being traded using market data taken from Corgnet et al. (2018, 2020). Their results showed that high ToM skills benefit the market when there is low mispricing because subjects with higher ToM skills are more attentive to market orders, which, in turn, are better at extracting valuable insights from observing market price dynamics. In contrast, when market dynamics are driven mainly by noise and high mispricing, forecasters with high ToM skills can become destabilized because they fail to recognize that market dynamics are meaningless.[2] Their results resemble Hefti et al. (2018) experimental results, concluding that subjects with high ToM but low analytical skills incur high trading losses due to failure to detect the fundamental value. Note that this strand of literature is different from our study, as we focus on the relationship between traders' *ToM skills* and *the market dynamics* of the same group of traders.

We are also closely related to the literature on the relationship between ToM level and level-k thinking, with the latter being hypothesized to explain strategic behavior, since Keynes (1936) discussed the beauty contest. Although beauty contests differ substantially from LtFE (i.e., mainly incentive structure, information structure, and feedback strength; see Sonnemans and Tuinstra, 2010), subjects in both games have an incentive to coordinate their expectations around others' expectations (Colasante et al., 2020). Georganas et al. (2015) reported the results of two-person games with a unique Nash equilibrium. They found that cognitive ability, measured by the Cognitive Reflection Test (Frederick, 2005) – and ToM skills, measured by the Eye Gaze Test (Baron-Cohen, 1997) – failed to predict players' levels of thinking. In contrast, Fe et al. (2022) conducted a modified Money Request game (designed to trigger level-k thinking: Arad and Rubinstein, 2012) to study how childhood cognitive skills affect strategic sophistication. The game has no pure strategy Nash equilibrium. They found that children's ToM skills – measured by the Imposing Memory Task, Kinderman et al. (1998), and cognitive ability, measured by Raven's Progressive Matrices (Raven et al., 1998) – both predict level-1 behavior.

Our main contribution is threefold.

First, we use a quasi-experiment to examine how market-level ToM influences the formation of individual expectations and aggregate market outcomes. While correlational studies have been conducted on the role of ToM in other experimental market settings,

[1] By contrast, a LtFE is called to display negative feedback when the realized asset price is low if the average price expectation is low. For example, in a production market, a higher expected price leads to increased production and thus lower the realized market price. The typical result from negative-feedback LtFE is that subject usually learn the rational expectation equilibrium quite well, which leads to a rapid convergence of market price towards the rational expectation equilibrium.

[2] Their result resembles Hefti et al.'s (2018) experimental result that in a call market, those investors with only high ToM but low analytical skills may ride on financial bubbles and incur high trading losses due to failure to detect the fundamental value.

our results help determine how ToM matters in expectation formation. Besides, different from previous studies that usually focus on individual trading behavior and performance (see references above), we conduct a more comprehensive study on the impact of ToM on a series of measures on the aggregate quality and stability of the market. In particular, an intuitive channel for ToM to facilitate price stability is that higher ToM market participants are not only better at reading the information from the market price but also at achieving higher expectations coordination with other participants. Compared to previous studies, where there is a lack of a measure for expectation coordination, our experimental setup allows us to directly examine the causal link from the ToM score to expectations coordination and price efficiency. Our results clearly show that a higher ToM does not seem to enhance expectation coordination or reduce price bubbles.

Second, our findings provide additional evidence on how strategic uncertainty matters for market behavior and asset bubble formation (e.g., Akiyama et al., 2017; Zhang and Zheng, 2017; Rholes and Petersen, 2021) and whether ToM fuels or mitigates strategic uncertainty-induced market instability. Our results suggest that although having traders/professionals with better ToM may help enhance market stability at an aggregate level, when categorizing subjects based solely on their ToM skills, there are still heterogeneities in the belief formation of price forecasts.

Finally, we add to the growing experimental literature on experimental asset markets (e.g., Palan, 2013; Baghestanian et al., 2015; Giusti et al., 2016; Holt et al., 2017; Ding et al., 2018; Nuzzo and Morone, 2017; Fenig et al., 2018; Duffy et al., 2019; Sunder, 2020 and Jiang et al., 2021) and, particularly, to the literature on LtFE (e.g., Hommes, 2011,2021; Assenza et al., 2014; Petersen, 2014; Bao et al., 2021).

The remainder of this paper is organized as follows. Section 2 presents the experimental design. Section 3 presents the experimental results, and Section 4 concludes the paper.

2. Experimental setup

2.1. *Experimental design*

The laboratory experiments were performed with 24 participants in each session. Each subject (1) took the eye gaze test (Baron-Cohen et al., 1997), (2) participated in a learning-to-forecast experiment (LtFE), and (3) completed the abbreviated numeracy test (Weller et al., 2013) that measures the numeracy scale and the Cognitive Reflection Test (CRT; Frederick, 2005)[3]; and reported their gender information. Subjects who participated in the experiment conducted in 2022 also completed (4) self-monitoring (a form of attentiveness to social cues) test (Snyder, 1974) and reported their age.

We administered an eye-gaze test to assess the participants' ToM capabilities (Baron-Cohen et al., 1997). This test measures participants' attributes to detect the true mental state of another person by viewing photos of their eyes. This test has been used in most studies dealing with ToM, such as Corgnet et al. (2018), Bruguier et al. (2010), and Hefti et al. (2018), and is sometimes incentivized[4] (each correct question gives a bonus). Other studies have used two tests that employed weighted ToM scores. For example, Bruguier et al. (2010) and Corgnet et al. (2021a) used Heider's test (Heider and Simmel, 1944) and weighed both tests equally into a score. At least 23 different tests claim to measure ToM in different forms (Quesque and Rosetti, 2020), and a combination would probably better measure particular aspects of ToM. However, we decided to follow Corgnet et al. (2018) as a seminal study to implement the unincentivzed ToM test in market experiments, showing that their version already correlated quite well with traders' behavior.

During the test, the participants viewed 36 separate images showing people's eyes, and for each image, they chose one of the four emotions that best described the person's mental state. The number of correct answers determined their 'ToM score.' Higher scores indicated a higher ToM capability, meaning that the subject was better at reading other people's intentions and, consequently, was better at reading 'the market.' Within each session, we ranked ToM scores from 1 to 24 (1 being the highest and 24 being the lowest), randomly breaking ties. We grouped the subjects such that those ranked 19–24 were in the *Low-ToM* group*,* and those ranked 1–6 were in the *High-ToM* group. The remaining two groups were subjects ranked 7–12 (*Middle-High*) and 13–18 (*Middle-Low*). For the analysis, we concentrated on comparing the high and low ToM groups to determine whether the effect unravels for these extreme groups (see, e.g., Bosch-Rosa et al. (2018) and Janssen et al. (2019) for a similar analysis strategy).[5] We included the middle two groups as a robustness check using 25 % as the threshold for grouping the subjects. As shown in Appendix C, we fail to find any difference in market dynamics when including intermediate markets and, hence, all our subjects in the analysis. We did not tell the participants that

[3] The experiment is programmed using zTree (Fischbacher, 2007). The instruction and the screenshot of user interface can be found in Appendix A.

[4] See Ridinger and McBride (2015) for a discussion on how a monetary incentive would affect the theory of mind performance.

[5] In the Appendix C, we repeat our analysis where we include intermediate groups. They provide similar results as when only comparing the high and low groups. Some may wonder if, by chance, a session scores higher than another one on the ToM skills. We also tried creating the experiment-wide "Low-ToM" and "High-ToM" groups by coding the top sixteen groups with the highest ToM score in the total 64 markets as "High-ToM" groups and coding the bottom sixteen groups with the lowest ToM score in the total 48 markets as "Low-ToM" group. Overall, there is only two exception markets (out of 32) where its session-wide grouping is different from experiment-wide grouping.

the assignment of groups depended on their ToM scores.[6]

In part two, we conduct a learning-to-forecast experiment (LtFE) with six subjects in each market, following Hommes et al. (2005, 2008). Each of the six subjects played the role of a financial advisor advising investment funds to buy or sell risky assets. For 50 consecutive periods, each subject had to predict the one-period-ahead market price of financial assets. Therefore, the price of a risky asset is an increasing function of the average price expectations in the market.[7] The determination function is defined in Eq. (1).

$$p(t) = \frac{1}{1+r}\left(\overline{p}^e(t) + d\right) + e_t \tag{1}$$

$p(t)$ is the realized market price in period t, d is the dividend of the risky asset, r is the risk-free interest rate, $\overline{p}^e(t)$ is the average price forecast for period t, and $e_t \sim N(0,1)$ is a small i.i.d. shock to price.[8] In our experimental setting, the interest rate of the risk-free asset r, is 5 %, and the dividend value is 3.3 points.

We incentivize the submission of the price closest to the realized market price. The earnings are $\max\left\{100 - \frac{100}{49}(\textit{prediction error})^2, 0\right\}$ with the preduction error beding$|\, p_t - p_t^e|$. In other words, subjects earn higher earnings when their price forecasts are closest to the realized price; that is, $\min|p^e(t) - p(t)|$. The realized price is a function of the average of all price forecasts; it is equivalent to saying that subjects would earn a higher price when they predict what all others predict in the market.

Imposing the rational expectation condition, where all subjects maximize their payoff through $\min|p^e(t) - p(t)|$, so that $\overline{p}^e(t) = p^e(t) = E(p(t))$ into Eq. (1), a simple computation shows that $p^* = 66$ is the unique REE of the system. In other words, if all agents have rational expectations (i.e., $\overline{p}^e(t) = p^e(t) = p^*$), the realized price then becomes $p_t = p^* + \epsilon = 66 + \epsilon$, i.e., the fundamental price plus a slight white noise. Thus, price forecasts are self-fulfilling and all subjects maximize earnings by minimizing the forecasting error.

Fig. 1 illustrates the computer interface used by the participants during the experiment. The screen displays the four boxes. The top box plots the past realized market price (in red, calculated using Eq. (1)) and the past price forecast history (in blue). In the middle box, the interface provides subjects with information on the current period t, risk-free interest rate, dividend of the risky asset, and total earnings subjects have earned. The bottom box shows the price forecast of the risky asset for the current period t. Finally, the right box shows a table reporting their own price forecasts and the history of the realized price, both up to period $t-1$.

At the end of each period t, the subject received information on the price forecasts submitted by the subject (p_t^e), the actual market price p_t generated by Eq. (1), the prediction error that is calculated based on the difference between price forecasts and market price, that is, $p_t - p_t^e$; their earnings from the period that decrease quadratically with the prediction error, that is, $\textit{earning} = \max\left\{100 - \frac{100}{49}(\textit{prediction error})^2, 0\right\}$; and their total earnings up to period t.

In Part 3, we perform a numeracy test, including a CRT (Weller et al., 2013; Frederick, 2005). Lambrecht et al. (2021) found that both cognitive ability and Theory of Mind (ToM) are associated with higher earnings in a cryptocurrency market. But The two attributes act as substitutes for each other, consistent with the findings of Corgnet et al. (2018). Fe et al. (2022) provide a detailed conceptual framework, arguing that ToM and cognitive ability respond to different cognitive processes in strategic situations involving understanding intentions. Their experimental data showed a low correlation between children's ToM skills and cognitive ability; only ToM skills, but not cognitive ability, predicted whether children direct reciprocity appropriately according to the intentions of the allocators in a gift-exchange game (Fehr et al., 1998). Meanwhile, their empirical results show that the positive effect of primary school spending improves ToM skills but not cognitive ability, highlighting the difference between the two cognitive skills.

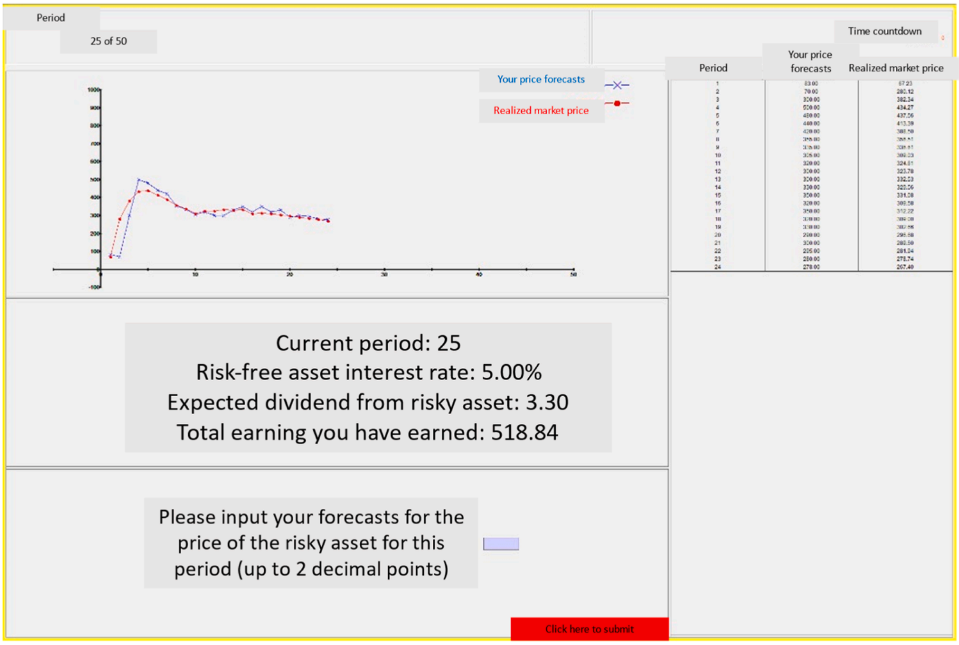
In Part 4, we performed a self-monitoring test (Snyder, 1974) measuring attentiveness to social cues and collecting information on gender and age. Only subjects who participated in our experiment in 2022 completed the study. A total of 144 subjects participated in the Season 2021 experiment, and 240 subjects participated in the Season 2022 experiment.

We collected information on the CRT, numeracy test, self-monitoring test, gender, and age, and conducted a balance check between

[6] We do not inform subjects that they are grouped with subjects of similar ToM skills to avoid the potential common information effect found in Corgnet et al. (2021b) — that information aggregation is enhanced when a high level of CRT is common information for all participants — may weaken our manipulation of ToM skills. We also do not tell subjects their performance in the eye gaze tests to eliminate the effect of the performance feedback on their subsequent behavior in the prediction task.

[7] As can be seen from (1), the asset price will increase when the average price prediction $\overline{p}^e(t)$ made by the subjects goes up and decrease when the average price prediction $\overline{p}^e(t)$ made by the subjects goes down. Therefore, the market is a positive-feedback market.

[8] One alternative way to aggregate the forecasts in a market is to use median forecasts (e.g., Rholes and Petersen, 2021; Petersen and Rholes, 2022), which has the advantage to reduce the impact of one extreme forecasts on the realized price in the experimental economies. In our study, we choose to let the realized price be a function of the average forecasts to make the best comparison with the existing LtFE that employ mean forecasts and use students as subjects. Meanwhile, we acknowledge that some may wonder if aggregating the forecasts using the average forecasts of only 6 subjects may increase the price bubbles, as one subject's extreme forecast would possibly have large impact on the realized price. However, it is found that the bubbles and crashes would also occur in large experimental LtFE asset markets, e.g., when the market size is 21-32 in Bao et al. (2020), and when the market size is between 92 and 104 in Hommes et al. (2021).



Period	Your price forecasts	Realized market price
1	83.00	67.23
2	70.00	280.12
3	300.00	382.34
4	500.00	434.27
5	480.00	437.56
6	440.00	413.39
7	420.00	388.50
8	355.00	356.51
9	335.00	335.61
10	305.00	309.03
11	320.00	324.51
12	300.00	323.78
13	300.00	332.03
14	330.00	320.56
15	350.00	331.08
16	320.00	309.58
17	350.00	312.22
18	320.00	309.00
19	330.00	302.66
20	290.00	295.58
21	300.00	288.50
22	295.00	281.24
23	280.00	278.74
24	278.00	267.40

Current period: 25
Risk-free asset interest rate: 5.00%
Expected dividend from risky asset: 3.30
Total earning you have earned: 518.84

Please input your forecasts for the price of the risky asset for this period (up to 2 decimal points)

Fig. 1. The computer interface is the translated version of the screenshot shown to the subjects, originally written in Chinese. The computer interface we presented to the subjects in Chinese can is in Appendix D (Fig. D.1).

the groups using these variables because they were found to affect individual expectations. For example, Hefti et al. (2018) suggested that analytical and mentalizing capabilities are critical for explaining individual trading behavior. Similarly, Gill and Prowse (2016) found that cognitive ability predicted how quickly adults learned to play at equilibrium in a repeated beauty contest game. The existing literature has also found that CRT and self-monitoring of the subjects may affect market dynamics: In both LtOE (Bosch-Rosa et al., 2018) and LtFE settings (Zong et al., 2017),[9] the price deviation from the REE in markets populated by low-CRT subjects is more significant than in those populated by high-CRT subjects; In Biais et al. (2005), trading performance among high self-monitors are better in a call market. Additionally, we collected information on gender because Baron-Cohen et al. (1997) suggested that females have higher empathy than males and score higher in Eye Tests that measure ToM skills. In turn, the balance check served as a randomization check to ensure that our results were free from the influence of confounding factors.

We preregistered the experiment at the AEA RCT Registry under the RCT ID AEARCTR-0,007,836 on June 30, 2021.[10]

2.2. Conjectures

We measure price and expectation performance following Hommes et al. (2005) and Stöckl et al. (2010), respectively. Table 1 presents an overview of the measures considered.

First, we consider two ‘bubble’ measures as introduced in Stöckl et al. (2010): *Relative Deviation* (or overpricing), the percentage deviation of the period price from the REE (*RD*) and *Relative Absolute Deviation* (or mispricing), the percentage absolute deviation of the price ($p_{k,t}$) from REE (RAD) – Equations (2) and (3). A smaller RD and/or RAD indicates a smaller deviation in the price from the REE, resulting in a smaller price bubble in the market.

As illustrated, we hypothesize that all members of the high-ToM group are more capable of inferring the intentions of other players and the market. Given that all subjects in the market have the same goal of minimizing forecasting error and maximizing payoff achieved when forecasting the price around the REE, we hypothesize that subjects in the high-ToM group will submit a price forecast closer to the REE, leading to smaller price bubbles in the market (lower RD and RAD).

Hypothesis 1. *Low-ToM markets show a higher deviation from REE than High-ToM markets.*

[9] The key difference between LtFE and LtOE is that LtFE requires subjects to submit their price forecasts, and the market price is determined by the average price forecasts. In contrast, LtOE requires subjects to submit their trading/production quantity, and the market price is determined by the total supply and demand. We use LtFE setting in the experiment to directly elicit the price belief and avoid the difficulty of “testing joint hypotheses”. In asset market experiments, the subjects might fail to converge to the equilibrium either when failing to form rational expectations or being unable to calculate the optimal trading quantity given their expectations.

[10] We pre-registered our experiment only for the first wave (Sessions 1-6). The first version of our paper got referee reports suggesting a higher sample size. Hence, we added ten further sessions in the second wave (Sessions 7-16).

Table 1
Overview of measures used for the analysis.

Market level		
Relative Deviation (%)	$RD_k = \frac{1}{50}\sum_{t=1}^{50}\frac{p_{k,t} - REE}{REE} \times 100$	(2)
Relative absolute Deviation (%)	$RAD_k = \frac{1}{50}\sum_{t=1}^{50}\frac{\lvert p_{k,t} - REE\rvert}{REE} \times 100$	(3)
Price dispersion	$PD_k = \sqrt{\frac{1}{50}\sum_{t=1}^{50}(p_{k,t} - \overline{p}_k)^2}$	(4)
Amplitude	$AMP_k = \max p_k - \min p_k$	(5)
Forecast dispersion	$FD_k = \frac{1}{50}\sum_{t=1}^{50}\left(\frac{1}{6}\sum_{i=1}^{6}(p^e_{k,t,i} - \overline{p}^e{}_{k,t})^2\right)$	(6)

Notes. k indicates the market, t indicates the period, i indicates the subject, p^e is the price forecast, p is the price, and REE is the equilibrium price.

The positive feedback LtFE frequently shows persistent bubbles and crashes, that is, a persistent variation in price visualized as a substantial price oscillation (e.g., Fig. 3 in Hommes et al., 2005). We hypothesize that low-ToM markets are more volatile than are high-ToM markets. To measure these differences, we consider Price Dispersion (PD), that is, the standard deviation of prices (Equation (4) in Table 1), and Price Amplitude (AMP), that is, the difference between the highest and lowest prices (Equation (5) in Table 1), as the dependent variable when studying the effect of ToM on market volatility. Hence, we formulate Hypothesis 2.

Hypothesis 2. *Low-ToM markets show higher price variability than High-ToM markets.*

Now, we move from the market-level price dynamics to the forecasting behavior of individuals. As discussed, subjects with better ToM skills are hypothesized to infer the market and, hence, all other opponents' intentions. Therefore, we conjecture that there is faster and stronger coordination of expectations in high-ToM markets.

We measure the coordination of expectations using forecast dispersion, that is, the standard deviation of price forecasts by all group members in each period (Equation (6) in Table 1). By treating forecast dispersion as the dependent variable, we test whether there is better coordination of price expectations in the high-ToM group.

Hypothesis 3. *Low-ToM markets show worse coordination of price expectations than High-ToM markets.*

Finally, we ask whether the subjects in the low- and high-ToM groups use different heuristics when forecasting prices. We consider the following forecasting strategies taken directly from Anufriev et al. (2019): adaptive expectations, trend-extrapolation rules, naïve expectations, and fundamental forecasts (Makarewicz, 2021). Among these, adaptive expectations, trend-extrapolation rules, and naïve expectations rely on past price or past price forecasts. Adopting fundamental forecasts requires understanding the market's intention, coordinating with all players in the market, and submitting a price around the fundamental value or REE of the asset, so that the payoff for everyone is maximized. Therefore, in line with Hypothesis 1, in which we hypothesized that only subjects in the High-ToM group would be able to coordinate expectations around the REE, we hypothesized that there are more subjects in the High-ToM group using fundamental forecasts. Moreover, because subjects with low ToM cannot do so, they will need to rely more on previously disclosed information, resulting in more subjects adopting all the other expectations.

Hypothesis 4. *More subjects adopt adaptive expectation, trend-extrapolation rule, and naïve expectation in Low-ToM groups. In contrast, more subjects in High-ToM group would adopt fundamental forecasts.*

So far, we have considered the hypothesis that we pre-registered, where we only compared markets with the highest (High-ToM) and lowest ToM scores (Low-ToM) in each session.

2.3. Procedure

We conducted six sessions in October and November 2021 (Season 2021) and another ten sessions in November 2022 (Season 2022) at the *Laboratory of Experimental Economics, Dongbei University of Finance and Economics* (DUFE).[11] A total of 384 participants participated in 16 sessions, with 24 participants per session.[12] The subjects who participated in our experiment were undergraduate

[11] We pre-registered our experiment only for the first wave (Sessions 1-6). The first version of our paper got referee reports suggesting a higher sample size. Hence, we added ten further sessions in the second wave (Sessions 7-16). Hence, our decision was not covid related. In fact, both our experiments were in the Late COVID phase, as described in Petersen and Rholes (2022), i.e., their results do not apply to our experiments. According to Table 1, they compared subjects' behavior in three timing treatments: Pre-COVID (between October to November 2019), Early COVID (between April to June 2020), and Late COVID (between October to December 2021). We conducted our experiment from October 12 to November 6, 2021, and November 5 to November 20, 2022.

[12] Note that even though the number of observations is higher than in other studies using learning to forecast experiments. For example, some earlier papers sometimes have only 4, 6, or 10 markets in their consideration, and in compassion our sample size with 16 observations in each group is on the high side. The power is still not fully sufficient. For a power of 80%, we would need more than 1500 participants.

students at DUFE, and each could participate in our experiment only once. The average duration of the experimental sessions was approximately one hour. The average payoff for the subjects was 37.5 RMB (about 5.2 dollars, substantially higher than the hourly wage for part-time work for students in Dalian, standard deviation:17.27RMB). The experiment was programmed using zTree (Fischbacher, 2007). Participants were recruited using the WeChat platform in a laboratory.

3. Results

Table 2 reports the average values of test scores and demographic information across the four groups.

The randomization check showed that females performed better in the eye-gaze test, which is consistent with the findings of Baron-Cohen et al. (1997). Individuals in the low-ToM group typically scored better on the numeracy test. Furthermore, because we grouped participants with similar performances on the eye-gaze test into one market within a session, the ToM score was significantly higher in the high-ToM group than in the low-ToM group ($p < 0.01$).

Our analysis runs OLS regression with only the high- and low-ToM groups. In the analysis to test Hypotheses 1–3, the dependent variable is a market-level measure with standard errors clustered at the session level in the following form: In the analysis testing Hypothesis 4, the dependent variable is an individual-level measure with standard errors clustered at the market level.

$$Y_k = \beta_0 + \beta_L L_k + \beta_C C_k + \varepsilon_k \tag{7}$$

The independent variable L_k is binary; it equals one if market k is in the Low ToM group and zero if it is in the High ToM group.

C_k denotes a vector of cohort-level control variables unique to market k, averaged from the individual-level control variables. Because all the other variables are not significantly different between the two groups at the 5 % significance level, we only control for gender and the numeracy test, but not other balanced individual-level variables when conducting market-level analysis (i.e., when testing Hypotheses 1–3).[13] In contrast, we control for all demographic variables (as listed in Table 2) when conducting individual-level analyses (i.e., when testing Hypothesis 4). We discuss the regression equation and econometric approach in detail in Section 3.4.

3.1. Deviation from equilibrium

Fig. 2 depicts the average and individual realized market prices against REE in both groups across all 50 periods. The large price bubbles observed in our experiment resemble the pattern observed in settings with an upper bound of 1000 (e.g., in Hommes et al., 2008). This pattern differs from the persistent oscillation around the REE observed in other studies, such as Hommes et al. (2005). This difference was caused by the difference in the experimental design. While the asset price depends only on the average expectation of all market participants in Hommes et al. (2008), there is a robot fundamentals trader who always predicts the fundamental value in Hommes et al. (2005), whose influence on the market price is larger when the price deviates more from the fundamental value. Thus, market price is more stable in studies such as Hommes et al. (2005) than in Hommes et al. (2008) and our study. The graphs show similar patterns in the two groups (black line), although some markets show bubbles in the low-ToM group (grey lines).

Table 3 presents the average RD and RAD values for the high- and low-ToM groups. At the aggregate level, the low-ToM group formed a bubble 65 % larger than the high-ToM group in terms of RD and RAD. At the individual market level, all markets, except Market 4 in the Low-ToM group, exhibit a persistent bubble pattern, suggesting that overpricing is common in both groups. However, there is some heterogeneity among the markets within the same treatment group. In almost half of the sessions (i.e., sessions 3, 4, 6, 7, 11, 12, and 13), the Low-ToM group exhibited a smaller bubble size than the High-ToM group in terms of either RD or RAD, which was contrary to the result from the aggregate level.[14]

Both the rank-sum test and the OLS regression, which cluster standard errors at the market level, find no significant difference at the 5 % level in the magnitude of price bubbles between the two groups. Additionally, we performed Cohen's d test on the size of the difference and found that the difference was not larger than the medium level (0.5).

Table 4 reports outcomes from the regressions on RD and RAD using the cohort averages of the subjects' characteristics as controls: CRT, Numeracy test, self-monitoring test, and demographic information. We separate the full and the 2022 samples. Testing hypothesis one, we see that the Low-ToM dummy coefficient suggests a 30 % higher overpricing (columns 1 and 4) and a 50 % higher mispricing (columns 7 and 10), in line with our hypothesis. However, given its insignificant difference, some extreme bubble markets seem to drive this average difference.

Meanwhile, according to columns (2) and (8), markets in the low-ToM groups with higher coort-level CRT scores create smaller bubbles, while those with higher cohort-level numeracy scores create larger bubbles. In contrast, neither test has a significant effect on price bubbles among the high-ToM groups, as shown in columns (3) and (9). Furthermore, the positive effect of the cohort-level numeracy test on the magnitude of price bubbles in the Low-ToM groups is only significant at the 10 % level when we look only at the Season 2022 sample, which additionally controls the self-monitoring test score and age, as shown in columns (5) and (11). The

[13] To make the analysis complete, in the analysis part of Hypothesis 1-3, we would also attach and discuss the regression result controlling all variables that is listed in Table 2.

[14] In fact, the economically large but statistically insignificant price bubbles generated from Low ToM group can be driven by two very bad markets in the Low ToM group (i.e., Market 9 and 10). More specifically, As showed in Table 3, the RAD/RD is less than 200% for most markets in both high ToM and low ToM groups. But the average price deviation is much higher in the low ToM group mainly because there are two markets in the group with a price deviation larger than 500%.

Table 2
Summary statistics on the test performance and demographic information .

	Low-ToM	Middle-low	Middle-high	High-ToM	Low ToM – high ToM	
					Rank-sum (z)	OLS coefficient
Season 2021 and 2022, N = 384						
ToM score	19.46	23.29	25.34	28.11	−4.83***	−8.66***
CRT score	1.96	1.90	2.01	1.92	0.29	0.04
Numeracy test	4.32	4.35	4.18	3.91	1.97**	0.42**
% Female	0.53	0.57	0.70	0.72	−2.35**	−0.19***
Season 2022, N = 240						
Self-monitoring test	13.55	13.28	13.92	14.05	−0.42	−0.50
Age	21.65	21.63	21.38	21.65	−0.15	0.00

Notes: Column "Low ToM – High ToM" calculates the average difference between Low -ToM and High -ToM markets. The null hypothesis is that the measurement in the Low-ToM market has no difference from that in the High-ToM market. For the rank-sum test sub-column, the asterisks report the exact p value for Wilcoxon rank-sum test, as the sample size is smaller than 200 in our experiment. For the OLS sub-column, the asterisks indicate the significance level comparing the Low-ToM markets with the High-ToM market using a treatment dummy in an OLS regression clustered at session-level (Panel Season 2021 and 2022: in total 6; Panel Season 2022: in total 10). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

negative coefficient of the CRT score on the price bubble is no longer statistically significant when we consider only the 2022 sample with additional controls.

Observation 1. *We cannot support Hypothesis 1. Although the Low-ToM group forms a price bubble around 65 % larger on the aggregate level, there is no significant difference in price deviation from REE between Low- and High-ToM markets.*

3.2. *Price variability*

In this section, we consider the price variability between the two groups. According to Fig. 2, prices in both groups seem to fluctuate irregularly without any apparent pattern.

Table 5 calculates price variability using the measurements of price dispersion and amplitude in Table 1. At the aggregate level, the Low-ToM group reveals a higher price variability, about 73 %–93 % larger than that in the High-ToM group, a medium effect size according to Cohen's d. Similar to before, we observe the heterogeneity of the markets within the same group. For example, in sessions 3, 7, 12, and 13, the price variability and amplitude were lower in the low ToM group than in the high ToM group. We found no significant differences in price dispersion and amplitude using the rank-sum test and OLS regressions. This result is consistent when we repeat the analysis for the last 10 and 25 periods, that is, when the subjects are experienced.

Table 6 reports outcomes from the regressions on Price Dispersion and Amplitude analog to Table 4, showing the same pattern. We find a sizable but insignificant coefficients (at 5 % level) for the Low-ToM dummy.

Meanwhile, among the Low-ToM groups, the market with a higher cohort-level numeracy test is more volatile in terms of both the measure of price dispersion and price amplitude, which is robust when we only look at the Season 2022 sample that additionally controls for self-monitoring test performance and age. However, this effect was not observed in the high-ToM group.

Observation 2. *We cannot support Hypothesis 2. Even though average price variability and amplitude are substantially higher in Low-ToM groups, the difference is not significant.*

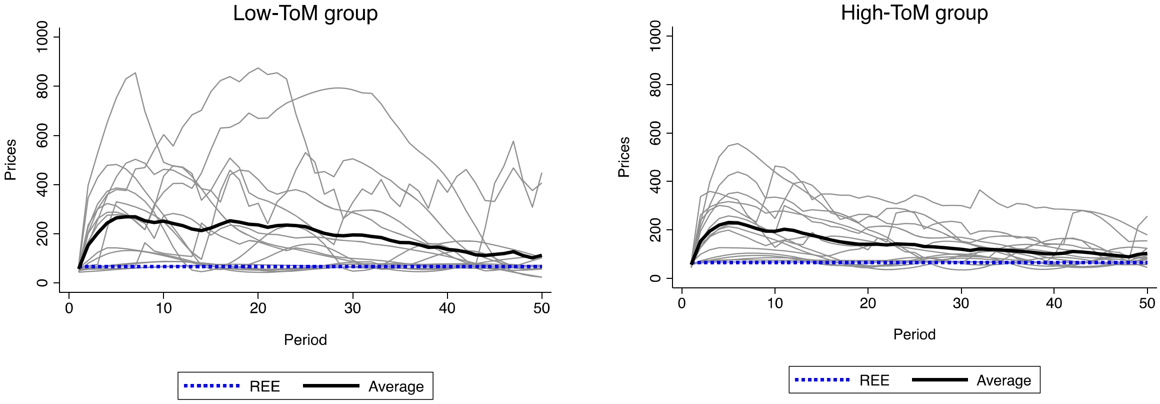


Fig. 2. Realized market price in each of the individual markets in sixteen sessions (grey lines), fundamental value (REE, in blue dotted line), an average value of realized market price (black line) for each period in the low- and high- ToM group. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Relative deviation and relative absolue deviations.

	(RD in%)		(RAD in%)	
Session	Low-ToM	High-ToM	Low-ToM	High-ToM
1	395	138	395	142
2	154	117	164	117
3	60	56	60	75
4	−2	7	9	8
5	14	1	23	2
6	21	63	28	63
7	4	225	13	226
8	68	14	86	26
9	595	116	604	116
10	568	14	568	30
11	133	379	139	380
12	4	9	14	19
13	186	212	194	224
14	397	187	397	197
15	291	232	296	232
16	125	39	135	53
Average	188	113	195	119
Low ToM – high ToM [p-value]				
Rank sum	0.72 [0.491]		0.75 [0.468]	
OLS	34.37 [0.592]		36.96 [0.559]	
Cohen's d	0.46		0.48	

Notes: Row "Low ToM – High ToM" calculates the average difference between Low -ToM and High -ToM markets. The null hypothesis is that the measurement in the Low-ToM market has no difference compared with that in the High-ToM market. p value in the bracket. For the rank-sum test row, we report the exact p value for Wilcoxon rank-sum test, as the sample size is smaller than 200 in our experiment. For the OLS coefficient row, p- value is calculated by comparing the Low-ToM markets with the High-ToM market using a simple OLS regression, where the standard error is clustered at the session level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

3.3. Coordination of price expectation

We use forecast dispersion, as defined in Equation (4) in Table 1, to measure price forecast coordination. Previous studies on LtFE find a high level of coordination in the price expectations of subjects within the same market (e.g., Hommes et al., 2005; Bao and Zong, 2019), where there is a decreasing trend in the standard deviations of price expectations over time.

Fig. 3 depicts the expectation dynamics in both Low- and High-ToM groups. At the aggregate level, the downward trend in the standard deviation of the price forecasts is stronger in the high-ToM group than in the low-ToM market. The average standard deviation of prediction in the High-ToM group was 14.42 after a peak at 154.73 in the 2nd period. It is only half of that in the Low-ToM group, where the average standard deviation of the prediction in the Low-ToM group is 35.74 after the peak at 142.50 in the 2nd period. However, at the individual level, the poor coordination of price expectations in the low-ToM group seems to result from the two low-ToM markets. When the two markets are excluded, there is no clear difference in the coordination of price expectations between the two groups.

Table 7 formally calculates and lists the standard deviations of the price forecasts for each session. The average forecast dispersion doubled in the Low-ToM group compared to the High-ToM group. However, we find no significant difference at the 5 % level using both the rank-sum test and OLS regression because 6 out of 16 sessions (i.e., sessions 4, 6, 7, 11, 12, and 13) reveal an opposite result compared to the aggregate comparison, where there is a smaller forecast dispersion in the Low-ToM group. The size effect was medium according to Cohen's d.

We further conducted a regression analysis to determine whether there was a discrepancy in the coordination speed between the two groups while controlling for demographic characteristics and test scores. We analyze the coordination speed with a random effects model allowing for heterogeneity in price formation between the groups (Breusch and Lagan LM test for random effects: $p < 0.01$ for both models), but not within each group.

Table 8 reports the results. Both groups reveal a convergence pattern to the REE as the "stage" categorical variable has a 1 % significance level with a negative sign (Row 2). The forecasts coordinate about 3.5 units for every five periods. Consistent with the mean comparison, we find no statistically significant difference in the coordination of price forecasts between Low- and High-ToM groups. Overall, the insignificant results align with our findings in the mean comparison, where there is heterogeneity in the expectation dynamics within the same treatment groups. In sum, we conclude that there is no significant difference in coordination when comparing the two groups.

Table 9 reports the regression table, including controls cohort-level performance on CRT, Numeracy test, self-monitoring test, and demographic information, which is averaged from individual-level data. Among the Low-ToM groups, the market with a higher cohort-level numeracy test coordinates worse on the price expectation, which is robust compared to only looking at the Season 2022 sample that additionally controls for self-monitoring test scores and age. The pattern, however, fails to extend to the high-ToM group.

Table 4
Regression table testing the effect of ToM on price bubbles, comparing only low- and high-ToM groups.

	RD						RAD					
Sample	Full			Season 2022 Only			Full			Season 2022 Only		
ToM Group		Low	High		Low	High		Low	High		Low	High
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Low ToM Group	31.61			52.86			34.72			54.87		
	(62.19)			(120.1)			(61.38)			(119.5)		
CRT score	−19.99	−258.0**	66.73	−19.04	−368.9	3.802	−16.20	−246.0**	70.95*	−17.79	−357.9	3.795
	(56.44)	(88.68)	(39.28)	(114.2)	(245.3)	(167.3)	(55.53)	(87.16)	(38.11)	(111.7)	(244.4)	(162.8)
Numeracy test	104.8**	197.3***	35.66	95.48	320.2*	−7.665	101.9**	196.3***	34.11	91.64	313.9*	−8.597
	(43.02)	(61.31)	(52.46)	(75.35)	(166.1)	(109.2)	(42.93)	(60.40)	(51.42)	(75.18)	(165.0)	(106.5)
Female Dummy	−4.498	−162.4	93.11	26.69	−30.03	−22.97	2.254	−170.8	112.8	28.19	−42.70	−12.49
	(132.0)	(216.1)	(185.1)	(211.1)	(307.6)	(279.9)	(127.2)	(214.4)	(178.1)	(203.2)	(304.8)	(273.2)
Self-Monitoring Test				8.022	−43.71	18.61				7.163	−43.96	18.36
				(21.75)	(72.06)	(54.23)				(21.44)	(71.59)	(52.39)
Age				−35.01	−148.5	3.940				−36.32	−147.2	1.746
				(46.36)	(115.5)	(65.48)				(45.58)	(114.7)	(63.80)
Constant	−254.8	−73.09	−221.0	431.8	3361	−165.6	−249.2	−80.90	−231.0	491.1	3357	−110.8
	(224.3)	(312.6)	(319.7)	(1179)	(4044)	(1467)	(223.9)	(308.8)	(312.8)	(1158)	(4015)	(1432)
Number of Markets	32	16	16	20	10	10	32	16	16	20	10	10
R-squared	0.182	0.348	0.143	0.194	0.625	0.070	0.180	0.345	0.154	0.196	0.619	0.067

Robust standard errors are in parentheses (cluster at the session level); *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 5
Price variability of market price in the two groups across sixteen sessions.

	Price Dispersion		Amplitude	
Session	Low-ToM	High-ToM	Low-ToM	High-ToM
1	111	67	465	247
2	74	25	259	144
3	44	63	182	222
4	8	4	32	22
5	22	1	86	4
6	25	17	90	75
7	10	74	33	275
8	127	35	534	207
9	252	79	768	296
10	257	20	817	67
11	124	73	427	411
12	11	13	41	41
13	149	158	458	523
14	192	113	783	357
15	135	60	425	255
16	99	48	341	169
Average	102	53	359	207
Low ToM – high ToM [p-value]				
Rank sum	1.62 [0.110]		1.66 [0.102]	
OLS	37.04* [0.090]		92.73 [0.164]	
Cohen's d	0.76		0.69	

Notes: Row "Low ToM – High ToM" calculates the average difference between Low -ToM and High -ToM markets. The null hypothesis is that the measurement in the Low-ToM market has no difference compared with that in the High-ToM market. p value in the bracket. For the rank-sum test row, we report the exact p value for Wilcoxon rank-sum test, as the sample size is smaller than 200 in our experiment. For the OLS coefficient row, p-value is calculated by comparing the Low-ToM markets with the High-ToM market using a simple OLS regression, where the standard error is clustered at the session level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Still, even when controlling for cohort-level test performance and demographic information, we do not find significant differences in price coordination between the Low- and High-ToM groups, as shown in row 1 of Table 9.

Observation 3. *We cannot support Hypothesis 3. Although Low-ToM group coordinates worse at about 117 % on the price forecasts on the aggregate level, there is no significant difference in the coordination of price expectation between the Low- and High-ToM markets.*

3.4. *Individual forecasting strategies*

To determine whether heterogeneity exists in the forecasting strategies used by markets with different ToM levels, we estimate the following four simple forecasting strategies for each individual *i*. We follow Anufriev et al. (2019) by running time-series regressions for each individual *i*:

- Adaptive expectations: $p^e_{i,t} = p^e_{i,t-1} + \theta_i(p_{t-1} - p^e_{i,t-1}) + \epsilon_{i,t}$
- Trend-following expectations: $p^e_{i,t} = p_{t-1} + \gamma_i(p_{t-1} - p_{t-2}) + \epsilon_{i,t}$
- Naïve expectations: $p^e_{i,t} = p_{t-1} + \epsilon_{i,t}$
- Fundamental forecasts: $p^e_{i,t} = p^f = 66$

For the first two heuristics, we follow Bao et al. (2013), who claim an estimation to be successful if the coefficient estimates. θ_i and γ_i are statistically significant at the 5 % level, with no autocorrelation detected in the errors using the Breusch-Godfrey test used to detect the higher-order serial correlation.

For the last two heuristics, we perform a two-sided Wald test against the null hypothesis that the coefficient is equal to one. We claim an estimation to be successful if we fail to reject the null hypothesis at the 5 % level and detect no autocorrelation in errors using the Breusch-Godfrey test.

For 31.25 % of subjects, for whom more than one expectation rule could be categorized, we characterized the individual into the estimated learning strategy with a smaller mean squared error (MSE).

Table 10 (Parametric Approach Panel) summarizes the categorization of forecasting strategies in the two groups after categorizing subjects who fit both learning rules into only one rule that has a smaller MSE. For adaptive expectations, we found that the estimation was successful for 11 of the 96 subjects in the Low-ToM treatment group and 13 of the 96 subjects in the High-ToM treatment group. For trend-following expectations, we found that the estimation was successful for 52 of the 96 subjects in the Low-ToM treatment group and 44 of the 96 subjects in the High-ToM treatment group. Fourteen of the 96 subjects in the Low-ToM group and 13 of the 96 subjects in the High-ToM group were categorized as naïve forecasters. Finally, only four out of 96 subjects were fundamental forecasters in the High-ToM group, while none were in the Low-ToM group.

Table 6
Regression table testing the effect of ToM on market volatility, comparing only low- and high-ToM groups.

	Price Dispersion						Amplitude					
Sample	Full			Season 2022 Only			Full			Season 2022 Only		
ToM Group		Low	High		Low	High		Low	High		Low	High
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Low ToM Group	36.19*			48.40			88.52			101.4		
	(20.23)			(41.32)			(61.55)			(140.3)		
CRT score	−6.168	−36.98	18.56	−8.025	−65.71	−0.781	−30.51	−256.4	82.48	−106.1	−319.5	−9.841
	(24.18)	(47.26)	(19.18)	(43.72)	(89.94)	(54.81)	(96.80)	(180.4)	(60.31)	(163.4)	(292.0)	(210.1)
Numeracy test	37.25*	90.53***	0.607	37.85	153.7**	−11.66	141.6*	304.1***	34.79	146.2	541.8**	9.503
	(20.61)	(28.25)	(26.09)	(34.83)	(66.12)	(44.02)	(72.62)	(99.55)	(89.66)	(133.8)	(216.2)	(169.4)
Female Dummy	11.67	−87.42	43.99	16.59	−96.43	−21.06	−28.72	−359.6	149.4	−70.08	−513.0	−36.91
	(46.62)	(82.64)	(48.54)	(75.86)	(125.1)	(106.9)	(148.5)	(283.4)	(196.7)	(243.2)	(428.4)	(417.5)
Self-Monitoring Test				−5.548	−7.164	1.242				−8.115	−9.921	−2.454
				(9.384)	(27.70)	(14.29)				(30.18)	(81.08)	(57.84)
Age				−25.73	−37.52	−22.10				−51.93	−84.69	−39.46
				(16.48)	(48.16)	(19.32)				(57.27)	(148.2)	(80.28)
Constant	−88.89	−170.0	−16.38	558.0	541.6	590.8	−266.9	−262.7	−194.2	1191	921.7	1158
	(97.88)	(145.2)	(127.3)	(396.1)	(1615)	(416.0)	(343.7)	(503.6)	(466.4)	(1379)	(4788)	(1803)
Number of Markets	32	16	16	20	10	10	32	16	16	20	10	10
R-squared	0.223	0.343	0.058	0.403	0.582	0.321	0.232	0.390	0.103	0.332	0.591	0.098

Robust standard errors are in parentheses (cluster at the session level); *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

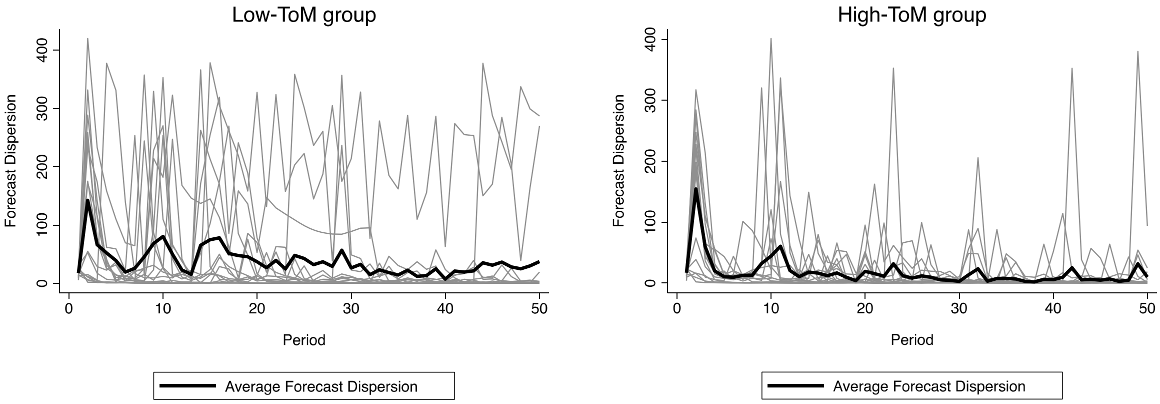


Fig. 3. Standard deviation of price forecasts in each of the individual markets in sixteen sessions (grey lines) and the average value of the standard deviation of price forecasts (black line) for each period in the low- and high-ToM group.

Table 7
Forecast dispersion in the markets of the groups across sixteen sessions.

	Forecast Dispersion	
Session	Low-ToM	High-ToM
1	38	17
2	174	22
3	25	16
4	20	9
5	1	2
6	3	1
7	4	8
8	1	32
9	43	11
10	70	49
11	86	2
12	16	38
13	1	2
14	18	18
15	78	26
16	45	33
Average	188	113
Low ToM – high ToM [p-value]		
Rank sum	0.90 [0.381]	
OLS	8.96 [0.447]	
Cohen’s d	0.59	

Notes: Row “Low ToM – High ToM” indicates the average difference between Low -ToM and High -ToM markets. The null hypothesis is that the measurement in the Low-ToM market has no difference compared with that in the High-ToM market. p value in the bracket. For the rank-sum test row, we report the exact p value for Wilcoxon rank-sum test, as the sample size is smaller than 200 in our experiment. For the OLS coefficient row, p- value is calculated by comparing the Low-ToM markets with the High-ToM market using a simple OLS regression, controlling for gender, with the standard error clustered at the session level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Imposing the null hypothesis of a random assignment of learning strategies across the treatment group, Fisher’s exact test has a p-value of 0.828, 0.312, 1.000, and 0.121 for adaptive, trend-extrapolation, naïve, and fundamental forecasters, respectively. Thus, we conclude that there is no difference in the number of subjects adopting the four learning strategies between the Low- and High-ToM groups at a statistical significance level of 5 %.

Next, we rerun Eq. (7) with Y_k, the average estimated coefficient in market k of adaptive/trend-following expectations from all the subjects using this heuristic. By doing so, it allows us to investigate whether there is a difference θ and γ between Low- and High-ToM groups, or in other words, whether they adopt adaptive/trend-following expectations to the same extent.

The average estimated coefficient for adaptive expectations is $\overline{\theta} = 0.95$ in the low-ToM group, and $\overline{\theta} = 0.78$ in the High-ToM group. The average estimated coefficient for adaptive expectations is $\overline{\gamma} = 0.78$ in the low-ToM group, and $\overline{\gamma} = 0.61$ in the High-ToM group. When comparing the extent to which the two groups use heuristics, we find evidence that for those categorized as users of adaptive and trend-following expectations, the Low-ToM group tends to use the two heuristics to a larger extent at the 5 % significance level.

Table 8
Regression result price forecasts standard deviation.

	(1)	(2)	(3)	(4)
	The standard deviation of price forecasts			
Low ToM	6.63	−1.13	10.57	10.33
	(10.90)	(15.43)	(13.58)	(22.52)
Stage			−3.45***	−3.93***
			(0.94)	(1.42)
Low ToM × Stage			−0.72	−2.08
			(2.29)	(3.08)
Numeracy Test	20.75**	19.95	20.75***	19.95**
	(7.60)	(10.93)	(7.16)	(9.23)
CRT	−16.86	−20.15	−16.86	−20.15
	(14.86)	(15.87)	(14.00)	(13.40)
Female Dummy	−30.22	−31.34	−30.22	−31.34
	(19.69)	(24.56)	(18.55)	(20.73)
Age		−2.17		−2.17
		(6.28)		(5.31)
Self-Monitoring Test		1.96		1.96
		(2.63)		(2.22)
Constant	−9.73	26.49	9.26	48.09
	(39.35)	(134.73)	(39.86)	(114.34)
Observations	32	20	320	200
R-squared	0.23	0.35		
Number of markets (N)			32	20

Note: Column (2) reports the regression comparing the convergence speed of expectations between the Low- and High-ToM markets. The dependent variable was the five-period average of the standard deviation of expectations. "Stage" is a categorical variable where stage=1 if the period is between 1 and 5; stage = 2 if the period is between 6 and 10; …; and stage = 10 if the period is between 46 and 50. Robust standard errors are in parentheses (standard errors are clustered at the session level). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 9
Regression Table testing the Effect of ToM on Forecast Dispersion, comparing only Low- and High- ToM groups.

	The standard deviation of price forecasts					
Sample	Full			Season 2022 Only		
ToM Group		Low	High		Low	High
	(1)	(2)	(3)	(4)	(5)	(6)
Low ToM Group	6.633			−1.127		
	(10.90)			(15.43)		
CRT score	−16.86	−90.17*	5.683	−20.14	−58.49	−7.493
	(14.86)	(49.67)	(5.722)	(15.87)	(37.91)	(18.62)
Numeracy test	20.75**	35.01**	9.668	19.95	56.64**	9.205
	(7.604)	(15.00)	(6.077)	(10.93)	(22.81)	(13.78)
Female Dummy	−30.22	−54.72	−3.119	−31.34	−56.79	−17.93
	(19.69)	(31.41)	(20.84)	(24.56)	(44.43)	(30.78)
Self-Monitoring Test				1.957	−1.443	0.184
				(2.630)	(7.986)	(5.644)
Age				−2.170	−12.69	3.102
				(6.284)	(12.98)	(5.879)
Constant	−9.736	91.84	−29.14	26.47	224.8	−56.06
	(39.35)	(95.01)	(32.04)	(134.7)	(434.6)	(115.7)
Number of Markets	32	16	16	20	10	10
R-squared	0.228	0.382	0.292	0.347	0.663	0.267

Robust standard errors in parentheses (cluster at session level).
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

However, the difference was sensitive to the control and sample examined.

As a robustness check,[15] we adopted the non-parametric approach commonly used in the LtFE literature (e.g., Mokhtarzadeh and Petersen, 2021). More specifically,

More specifically, we first estimate the coefficient and get $\widehat{\theta}_i$ and $\widehat{\gamma}_i$ for each of the subjects on the following adaptive and trend-following expectations. We then calculated the counterfactual predictions for each subject for each of the following expectations:

[15] We thank the anonymous referee for suggesting the non-parametric approach.

Table 10
Categorization of forecasting strategies .

	Treatment Group	Adaptive	Trend Extrapolation	Naïve	Fundamental
Parametric Approach					
Number of Subjects					
Count	Low ToM	11	52	14	0
	High ToM	13	44	13	4
Low ToM – High ToM	[2-sided Fisher's exact p value]	[0.828]	[0.312]	[1.000]	[0.121]
Coefficient					
Average	Low ToM	0.95	0.78	–	–
	High ToM	0.78	0.61	–	–
Low ToM – High ToM	OLS Coefficient, control CRT and numeracy test, female, all sample [*p* value]	0.208 [0.156]	0.192** [0.014]	–	–
	OLS Coefficient, control CRT, numeracy test, gender, age, and self-monitoring, Season 2 sample only [*p* value]	0.355** [0.046]	0.120 [0.341]	–	–
Non-Parametric Approach					
Number of subjects					
Count	Low ToM	10	73	13	0
	High ToM	9	78	9	0
Low ToM – High ToM	[2-sided Fisher's exact p value]	[1.000]	[0.481]	[0.497]	–
Coefficient					
Average	Low ToM	0.76	0.72	–	–
	High ToM	0.67	0.55	–	–
Low ToM – High ToM	OLS Coefficient, control CRT and numeracy test, female, all sample [*p* value]	0.136 [0.281]	0.188* [0.054]	–	–
	OLS Coefficient, control CRT, numeracy test, gender, age, and self-monitoring, Season 2 sample only [*p* value]	0.374 [0.182]	0.090 [0.462]	–	–

Notes: Row "Low ToM – High ToM" in the "Coefficient" section compares the average difference between Low -ToM and High -ToM markets, and the standard error clustered at the market level. The null hypothesis is that the estimated coefficient of adaptive- or trend-following expectation is similar for subjects in the Low- and High-ToM markets. *p* value in brackets. For the OLS coefficient with controls rows, *p*- the value is calculated by comparing subjects in the Low-ToM markets with the High-ToM market using a simple OLS. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

- Counterfactual adaptive prediction: $\widehat{p}^e_{i,t} = p^e_{i,t-1} + \widehat{\theta}_i(p_{t-1} - p^e_{i,t-1})$
- Counterfactual trend-following prediction: $\widehat{p}^e_{i,t} = p_{t-1} + \widehat{\gamma}_i(p_{t-1} - p_{t-2})$
- Counterfactual naïve prediction: $\widehat{p}^e_{i,t} = \widehat{p}_{t-1}$
- Counterfactual fundamental prediction: $\widehat{p}^e_{i,t} = p^f = 66$

We then compute the mean square error between each period's real prediction and the counterfactual predictions, that is, $MSE_{i,t} = (p^e_{i,t} - \widehat{p}^e_{i,t})^2$. Finally, we categorize a subject as a specific user of a heuristic when the average mean square error during all periods of the heuristic is the smallest.

Compared with the parametric approach, we do not count a subject as a non-user of learning heuristics when the estimated coefficient is insignificant at the 5 % level. The non-parametric approach also alleviates the potential autocorrelation problem caused by biased standard errors in the panel structure of the estimation Equations. Another advantage of the nonparametric approach is that every subject can be categorized using a single-learning approach. In contrast, the parametric approach concluded that 21.4 % of the subjects in the Low- and High-ToM groups could be categorized into none of the four heuristics.

Table 10 reports the results in the bottom panel. By the non-parametric approach, we have no fundamental forecasters anymore. There is still no significant difference in the number of subjects using adaptive, trend following, and naïve- expectations between the Low- and High-ToM group. Furthermore, we fail to find any significant difference in the extent of adopting adaptive or trend-following expectations at the 5 % level, when categorizing subjects using a non-parametric approach.

Table 11 shows the regressions that include control cohort-level performance on the CRT, Numeracy test, self-monitoring test, and demographic information averaged from individual-level data. We found some evidence that female subjects with lower CRT scores, numeracy test scores, or self-monitoring tests tended to use adaptive expectations to a larger extent. However, we are cautious about this pattern, as it is sensitive to the sample we are looking at and/or the approach we use to categorize the subjects, probably because there are relatively fewer users with adaptive expectations. However, the pattern does not extend to the use of trend-following expectations, as most subjects in our experiment could be categorized as using trend-following expectations.

Observation 4. *We cannot support Hypothesis 4. An equal number of subjects use adaptive expectation, trend-extrapolation rule, naïve expectation, and fundamental forecast in Low- and High-ToM group. Meanwhile, among the subjects who adopt each rule, there is no significant difference in the extent of adopting the rules between the two groups.*

Table 11
Regression table testing the effect of ToM on learning heuristics, comparing only low- and high-ToM groups.

Approach	Parametric				Non-Parametric			
Heuristics	Adaptive		Trend-Following		Adaptive		Trend-Following	
Sample	Full	Season 2022	Full	Season 2022	Full	Season 2022	Full	Season 2022
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Low ToM Group	0.208	0.355**	0.192**	0.120	0.136	0.374	0.188*	0.0896
	(0.139)	(0.150)	(0.0729)	(0.123)	(0.121)	(0.259)	(0.0942)	(0.119)
CRT score	−0.0879**	−0.0557	−0.0151	−0.00479	−0.117**	−0.0404	0.0355	−0.00917
	(0.0339)	(0.159)	(0.0325)	(0.0430)	(0.0481)	(0.0948)	(0.0334)	(0.0371)
Numeracy test	−0.0164	−0.172***	−0.0200	−0.0266	−0.0480	−0.172***	−0.0178	−0.0209
	(0.0301)	(0.0370)	(0.0193)	(0.0281)	(0.0523)	(0.0454)	(0.0150)	(0.0229)
Female Dummy	0.0626	0.636***	0.0761	−0.00671	−0.0461	0.465**	0.0719	−0.0102
	(0.0821)	(0.125)	(0.0797)	(0.144)	(0.129)	(0.182)	(0.0736)	(0.104)
Self-Monitoring Test		−0.0602***		0.00103		−0.0445		−0.00742
		(0.0113)		(0.0104)		(0.0279)		(0.0100)
Age		0.0160		0.0184		0.00760		−0.00290
		(0.0267)		(0.0272)		(0.0296)		(0.0181)
Constant	0.940***	1.738***	0.658***	0.362	1.097***	1.605*	0.493***	0.900*
	(0.141)	(0.386)	(0.112)	(0.701)	(0.247)	(0.750)	(0.140)	(0.449)
Observations	24	14	96	62	19	12	151	99
R-squared	0.247	0.806	0.099	0.066	0.247	0.659	0.066	0.029

Robust standard errors are in parentheses (cluster at the market level); *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

4. Conclusion

Despite extensive research on how ToM capacity affects individual trading behavior (Bossaerts et al., 2019; Corgnet et al., 2018, 2021a; De Martino et al., 2013; Frith and Frith, 2005; Hefti et al., 2018), little is known about how ToM affects expectation formation. In this study, we explore whether ToM affects expectation formation in the LtFE, which is widely used in experimental economics to study expectation formation in financial markets and macroeconomics experiments (Marimon et al., 1993; Hommes, 2011, 2021; Assenza et al., 2014; Bao et al., 2021). Specifically, when all players in the asset market are proficient at "reading the market," would there be a more stable market with fewer price bubbles or crashes?

We failed to find a statistically significant difference between the high- and low-ToM groups when comparing price and expectation dynamics, despite the large difference between the two groups. Our results suggest that when categorizing participants based solely on their ToM skills, there are still heterogeneities in the belief formation of price forecasts.

Our study also relates to that of Eckel and Füllbrunn (2015). Their study confirmed an inverse relationship between the magnitude of price bubbles and the proportion of female traders in the market. When we grouped the markets using ToM skills, we found that females scored higher on the eye gaze test. However, there is no significant correlation between gender composition and market dynamics in our study.

One limitation of our study is that we used student subjects. Since recent studies find some differences between student and financial professional samples (Holzmeister et al., 2020; Weitzel et al., 2020; Bao et al., 2022; Füllbrunn et al., 2022), future work may consider how ToM or CRT scores matter for financial decision-making by financial professionals.

In our study, we implemented the unincentivized eye gaze test to capture Theory of Mind capacity, following the seminal work of Corgnet et al. (2018). However, since Theory of Mind is a broad concept that refers to one's ability to understand what others are thinking, there are multiple measurements that may be relevant to performance in the financial market (Corgnet et al., 2021). An interesting avenue for future research might be to study whether our results remain robust when measuring Theory of Mind capacity using a different test. Meanwhile, our findings suggest that markets filled with subjects who score higher in the eye gaze test may sometimes "collude" with each other and drive more ups and downs in the asset price. Future studies may provide more comprehensive investigation on the possible double-edge-sword effect of Theory of Mind.

Declaration of competing interest

None.

Data availability

Data will be made available on request.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jebo.2024.01.018.

References

Akiyama, Eizo, Hanaki, Nobuyuki, Ishikawa, Ryuichiro, 2017. It is not just confusion! Strategic uncertainty in an experimental asset market. Econ. J. 127 (605), F563–F580.

Arad, A., Rubinstein, A., 2012. The 11-20 money request game: a level-k reasoning study. Am. Econ. Rev. 102 (7), 3561–3573.

Anufriev, M., Hommes, C., Makarewicz, T., 2019. Simple forecasting heuristics that make us smart: evidence from different market experiments. J. Eur. Econ. Assoc. 17 (5), 1538–1584.

Assenza, T., Bao, T., Hommes, C., Massaro, D., 2014. Experiments on expectations in macroeconomics and finance. Experiments in Macroeconomics. Emerald Group Publishing Limited.

Baghestanian, S., Lugovskyy, V., Puzzello, D., 2015. Traders' heterogeneity and bubble-crash patterns in experimental asset markets. J. Econ. Behav. Organ. 117, 82–101.

Bao, T., Duffy, J., Hommes, C., 2013. Learning, forecasting and optimizing: an experimental study. Eur. Econ. Rev. 61, 186–204.

Bao, T., Zong, J., 2019. The impact of interest rate policy on individual expectations and asset bubbles in experimental markets. J. Econ. Dyn. Control 107, 103735.

Bao, T., Hennequin, M., Hommes, C., Massaro, D., 2020. Coordination on bubbles in large-group asset pricing experiments. J. Econ. Dyn. Control 110, 103702.

Bao, T., Hommes, C., Pei, J., 2021. Expectation formation in finance and macroeconomics: a review of new experimental evidence. J. Behav. Exp. Finance 32, 100591.

Bao, T., Corgnet, B., Hanaki, N., Okada, K., Riyanto, Y.E., & Zhu, J. (2022). Financial forecasting in the lab and the field: qualified professionals vs. smart students. *iser discussion paper*, (1156).

Baron-Cohen, S., Jolliffe, T., Mortimore, C., Robertson, M., 1997. Another advanced test of theory of mind: evidence from very high functioning adults with autism or asperger syndrome. J. Child psychol. Psychiatr 38 (7), 813–822.

Biais, B., Hilton, D., Mazurier, K., Pouget, S., 2005. Judgemental overconfidence, self-monitoring, and trading performance in an experimental financial market. Rev. Econ. Stud. 72 (2), 287–312.

Bosch-Rosa, C., Meissner, T., Bosch-Domènech, A., 2018. Cognitive bubbles. Exp. Econ. 21 (1), 132–153.

Bossaerts, P., Suzuki, S., O'Doherty, J.P, 2019. Perception of intentionality in investor attitudes towards financial risks. J. Behav. Exp. Finance 23, 189–197.

Bruguier, A.J., Quartz, S.R., Bossaerts, P., 2010. Exploring the nature of "trader intuition. J. Finance 65 (5), 1703–1723.

Colasante, A., Alfarano, S., Camacho-Cuena, E., Gallegati, M., 2020. Long-run expectations in a learning-to-forecast experiment: a simulation approach. J. Evolut. Econ. 30, 75–116.

Corgnet, B., Desantis, M., Porter, D., 2018. What makes a good trader? On the role of intuition and reflection on trader performance. J. Finance 73 (3), 1113–1137.

Corgnet, B., DeSantis, M., & Porter, D. (2020). *Let's chat… when communication promotes efficiency in experimental asset markets.*

Corgnet, B., Deck, C., Desantis, M., Porter, D., 2021a. Forecasting skills in experimental markets: illusion or reality? Manage Sci.

Corgnet, B., DeSantis, M., Porter, D., 2021b. Information aggregation and the cognitive make-up of market participants. Eur. Econ. Rev. 133, 103667.

Ding, S., Lugovskyy, V., Puzzello, D., Tucker, S., Williams, A., 2018. Cash versus extra-credit incentives in experimental asset markets. J. Econ. Behav. Organ. 150, 19–27.

De Martino, B., O'Doherty, J.P., Ray, D., Bossaerts, P., Camerer, C, 2013. In the mind of the market: theory of mind biases value computation during financial bubbles. Neuron 79 (6), 1222–1231.

Duffy, J., Jiang, J., & Xie, H. (2019). Experimental asset markets with an indefinite horizon. Available at SSRN 3420184.

Eckel, C.C., Füllbrunn, S.C., 2015. Thar she blows? Gender, competition, and bubbles in experimental asset markets. Am. Econ. Rev. 105 (2), 906–920.

Fe, E., Gill, D., Prowse, V., 2022. Cognitive skills, strategic sophistication, and life outcomes. J. Polit. Econ. 130 (10), 2643–2704.

Fehr, E., Kirchsteiger, G., Riedl, A., 1998. Gift exchange and reciprocity in competitive experimental markets. Eur. Econ. Rev. 42 (1), 1–34.

Fenig, G., Mileva, M., Petersen, L., 2018. Deflating asset price bubbles with leverage constraints and monetary policy. J. Econ. Behav. Organ. 155, 1–27.

Fischbacher, U., 2007. z-Tree: zurich toolbox for ready-made economic experiments. Exp. Econ. 10 (2), 171–178.

Füllbrunn, S., Janssen, D.J., Weitzel, U., 2019. Risk aversion and overbidding in first price sealed bid auctions: new experimental evidence. Econ. Inq. 57 (1), 631–647.

Frith, U., Happé, F., 1999. Theory of mind and self-consciousness: what is it like to be autistic? Mind lang. 14 (1), 82–89.

Füllbrunn, S., Huber, C., König-Kersting, C., 2022. Experimental finance and financial professionals. Handbook of Experimental Finance. Edward Elgar Publishing, pp. 64–72.

Frederick, S., 2005. Cognitive reflection and decision making. J. Econ. Perspect. 19 (4), 25–42.

Frith, C., Frith, U., 2005. Theory of mind. Curr. Biol. 15 (17), R644–R645.

Georganas, S., Healy, P.J., Weber, R.A., 2015. On the persistence of strategic sophistication. J. Econ. Theory 159, 369–400.

Gill, D., Prowse, V., 2016. Cognitive ability, character skills, and learning to play equilibrium: a level-k analysis. J. Polit. Econ. 124 (6), 1619–1676.

Giusti, G., Jiang, J.H., Xu, Y., 2016. Interest on cash, fundamental value process and bubble formation: an experimental study. J. Behav. Exp. Finance 11, 44–51.

Heider, F., Simmel, M., 1944. An experimental study of apparent behavior. Am. J. Psychol. 57 (2), 243–259.

Hefti, A., Heinke, S., Schneider, F., 2018. Mental capabilities, Heterogeneous Trading Patterns and Performance in an Experimental Asset Market. University of Zurich. Department of Economics Working Paper, No. 234.

Holt, C.A., Porzio, M., Song, M.Y., 2017. Price bubbles, gender, and expectations in experimental asset markets. Eur. Econ. Rev. 100, 72–94.

Holzmeister, F., Huber, J., Kirchler, M., Lindner, F., Weitzel, U., Zeisberger, S., 2020. What drives risk perception? A global survey with financial professionals and laypeople. Manage Sci. 66 (9), 3977–4002.

Hommes, C., Sonnemans, J., Tuinstra, J., Van de Velden, H., 2005. Coordination of expectations in asset pricing experiments. Rev. Financ. Stud. 18 (3), 955–980.

Hommes, C., Sonnemans, J., Tuinstra, J., Van de Velden, H., 2008. Expectations and bubbles in asset pricing experiments. J. Econ. Behav. Organ. 67 (1), 116–133.

Hommes, C., 2011. The heterogeneous expectations hypothesis: some evidence from the lab. J. Econ. Dyn. Control 35 (1), 1–24.

Hommes, C., 2021. Behavioral and experimental macroeconomics and policy analysis: a complex systems approach. J. Econ. Lit. 59 (1), 149–219.

Hommes, C., Kopányi-Peuker, A., Sonnemans, J., 2021. Bubbles, crashes and information contagion in large-group asset market experiments. Exp. Econ. 24 (2), 414–433.

Janssen, D.J., Füllbrunn, S., Weitzel, U., 2019. Individual speculative behavior and overpricing in experimental asset markets. Exp. Econ. 22 (3), 653–675.

Jiang, J.H., Puzzello, D., Zhang, C., 2021. How long is forever in the laboratory? Three implementations of an infinite-horizon monetary economy. J. Econ. Behav. Organ. 184, 278–301.

Keynes, J.M. (1936). The general theory of interest, employment, and money.

Kinderman, P., Dunbar, R., Bentall, R.P., 1998. Theory-of-mind deficits and causal attributions. Br. J. Psychol. 89 (2), 191–204.

Lambrecht, M., Sofianos, A., & Xu, Y. (2021). *Does mining fuel bubbles? An experimental study on cryptocurrency markets(No. 703).* AWI Discussion Paper Series.

Makarewicz, T., 2021. Traders, forecasters and financial instability: a model of individual learning of anchor-and-adjustment heuristics. J. Econ. Behav. Organ. 190, 626–673.

Marimon, R., Spear, S.E., Sunder, S., 1993. Expectationally driven market volatility: an experimental study. J. Econ. Theory 61 (1), 74–103.

Mokhtarzadeh, F., Petersen, L., 2021. Coordinating expectations through central bank projections. Exp. Econ. 24, 883–918.

Nuzzo, S., Morone, A., 2017. Asset markets in the lab: a literature review. J. Behav. Exp. Finance 13, 42–50.

Palan, S., 2013. A review of bubbles and crashes in experimental asset markets. J. Econ. Surv. 27 (3), 570–588.

Petersen, L., 2014. Forecast error information and heterogeneous expectations in learning-to-forecast macroeconomic experiments. Experiments in Macroeconomics. Emerald Group Publishing Limited.

Petersen, L., Rholes, R., 2022. Macroeconomic expectations, central bank communication, and background uncertainty: a COVID-19 laboratory experiment. J. Econ. Dyn. Control 143, 104460.

Quesque, F., Rossetti, Y., 2020. What do theory-of-mind tasks actually measure? Theory and practice. Perspect. Psycho. Sci. 15 (2), 384–396.

Raven, J.C., Court, John Hugh, 1998. Raven's Progressive Matrices and Vocabulary Scales, 759. Oxford Pyschologists Press, Oxford.

Rholes, R., Petersen, L., 2021. Should central banks communicate uncertainty in their projections? J. Econ. Behav. Organ. 183, 320–341.

Ridinger, G., McBride, M., 2015. Money affects theory of mind differently by gender. PLoS ONE 10 (12), e0143973.

Snyder, M., 1974. Self-monitoring of expressive behavior. J. Pers. Soc. Psychol. 30 (4), 526.

Sonnemans, J., Tuinstra, J., 2010. Positive expectations feedback experiments and number guessing games as models of financial markets. J. Econ. Psychol. 31 (6), 964–984.

Stöckl, T., Huber, J., Kirchler, M., 2010. Bubble measures in experimental asset markets. Exp. Econ. 13 (3), 284–298.

Sunder, S., 2020. Experimental Asset Markets: a Survey. In: Kagel, J., Roth, A. (Eds.), The *Handbook of Experimental Economics*. Princeton University Press, Princeton, pp. 445–500.

Thaler, R., 2015. Keynes's 'Beauty Contest'. Chicago Booth Review. https://www.chicagobooth.edu/review/keyness-beauty-contest.

Weitzel, U., Huber, C., Huber, J., Kirchler, M., Lindner, F., Rose, J., 2020. Bubbles and financial professionals. Rev. Financ. Stud. 33 (6), 2659–2696.

Weller, J.A., Dieckmann, N.F., Tusler, M., Mertz, C.K., Burns, W.J., Peters, E., 2013. Development and testing of an abbreviated numeracy scale: a Rasch analysis approach. J. Behav. Decis. Mak. 26 (2), 198–212.

Zhang, M., Zheng, J., 2017. A robust reference-dependent model for speculative bubbles. J. Econ. Behav. Organ. 137, 232–258.

Zong, J., Fu, J., Bao, T., 2017. Cognitive ability of traders and financial bubbles: an experimental study. J. World Econ. 40, 167–192.