Does Paying for Data Change Investment Decisions?*

Philipp Chapkovski[1], Arzu Işık[2], Mariana Khapko[3], and Marius Zoican[2]

[1]University of Duisburg-Essen
[2]Haskayne School of Business, University of Calgary
[3]Rotman School of Management, University of Toronto

Abstract

Do investors trade differently when paying for market data versus getting it for free? In an investment experiment, we randomize whether subjects pay for or freely receive identical order flow data. Payment inflates perceived data informativeness by 11%, increases responsiveness of return forecasts to data by 31%, and leads to 14% larger forecast errors. We further document a sunk-cost mechanism concentrated among low financial literacy investors: payment raises investment levels while reducing the response of investment to own return forecasts. Offering data as a paid add-on distorts beliefs and behavior, with the burden falling disproportionately on the least sophisticated investors.

Keywords: market data, endowment effect, financial literacy, investment behavior, experiment
JEL Codes: G11, G12, G14

*Philipp Chapkovski (filipp.chapkovskii@uni-due.de) is a postdoctoral researcher at the University of Duisburg-Essen. Arzu Işık (arzu.isiktopbas@ucalgary.ca) is a doctoral student at Haskayne School of Business, University of Calgary. Mariana Khapko (mariana.khapko@rotman.utoronto.ca) is affiliated with University of Toronto Scarborough and the Rotman School of Management. Marius Zoican (contact author, marius.zoican@haskayne.ucalgary.ca) is the Canada Research Chair in Financial Technology at Haskayne School of Business, University of Calgary. Marius Zoican acknowledges funding from the Canada Research Chairs Program and the Social Sciences and Humanities Research Council of Canada (SSHRC Insight Grant 435-2025-0100). A pre-registration describing the experimental protocol was filed on January 20, 2026 and is available at Open Science Foundation/OSF.

1 Introduction

> "No matter what sort of data you're after — Forex, money market, fixed income, or equity — you need to acquire it. Even if it is a one-off thing for your trade, you need to acquire it."
> — Interactive Brokers Campus, "What Every Trader Should Know About Market Data"

Does paying for market data change how individual investors interpret and use it? Retail brokerages such as Robinhood, Interactive Brokers, or TD Direct Investing increasingly un-bundle market data from their core trading services, selling near-real-time order flow and market depth as an add-on for prices ranging between $5 and $50 per month.[1] Order flow is valuable for hedge funds with the speed and infrastructure to extract short-term signals (Bartlett et al., 2023; Kwan et al., 2024), but retail traders lack these capabilities. Having paid for data, traders may overtrade on it to justify the cost: The act of paying may itself alter how traders value the information.

The question matters because retail investors now account for a substantial, and rapidly growing, share of market activity. In April 2025, individual traders represented 36% of U.S. equity and options order flow, up from under 10% before the pandemic.[2] The share is higher still in short-term derivatives: retail accounts for roughly 60% of same-day (0DTE) options volume (Bryzgalova et al., 2023), now more than half of all S&P 500 options activity. As retail participation grows, so does the importance of understanding how these traders respond to platform design — including add-on data products sold separately from core brokerage services.

We conjecture that the act of *paying* for market data shapes the beliefs and strategies of retail traders. Specifically, we hypothesize an endowment effect driven by endogenous information acquisition. By purchasing access to market feeds, investors come to treat the data as an asset they own rather than as a neutral input. This sense of ownership increases the perceived signal value of complex market data, leading investors to rely on it more heavily when forming beliefs. As a result, beliefs respond more strongly to noisy data and pass through more aggressively into portfolio choices — even when the data itself has no predictive value given investors' available tools.

[1] See Chapkovski et al. (2025b) for a detailed snapshot of the retail market data landscape.

[2] See J.P. Morgan: Who is buying U.S. equities (June 4, 2025); Financial Times: Are retail traders the captains now? (December 9, 2025)

These effects should be heterogeneous across investors, generating selection into data acquisition. Overconfident investors, in particular, should be more likely both to purchase data ex ante and to overestimate their ability to extract value from it ex post, amplifying belief distortions and the pass-through from beliefs to portfolio decisions.

This paper provides the first controlled evidence on how actively paying for market data affects belief formation and investment behavior. We build on recent work by Andries et al. (2025) and develop an online randomized controlled experiment in oTree (Chen et al., 2016), in which participants complete a sequence of independent investment rounds. In each round, participants observe a graphical display of 16 historical realizations of stock returns and order-flow imbalance, mapped to half-hour intervals within a trading day. Using this information, participants form beliefs about next-period returns and choose how much of their endowment to invest in the risky asset.

Crucially, the informational content of order flow varies across rounds. In predictable rounds, lagged order imbalance forecasts next-period returns with a correlation of approximately 0.75; in baseline rounds, the imbalance is pure noise and carries no predictive content. Participants are informed that the data may or may not be predictive but are never told which rounds are predictable, forcing them to infer signal quality from the data itself.

To isolate the causal effect of paying for information from selection effects, we introduce a two-stage design. Treatment participants choose each round between paying E$5 for accessing order-flow data or receiving no data. After they choose, we randomly draw the round type: either the choice are binding (paid data condition), or all participants receive data for free regardless of their choice (free data condition). Control participants always receive free data without being asked. This design generates random variation in payment among participants who chose data, cleanly identifying the sunk cost effect. The free data condition also allows us to observe beliefs for participants who declined to pay, separating selection from treatment effects.

We document three main findings. First, paying for data distorts beliefs. Participants who pay for order-flow data are 8 percentage points more likely to perceive uninformative data as informative: that is, an 11% increase relative to baseline beliefs. This causal effect of payment amounts to 61% of the selection effect, indicating that both payment for data and selection into payment substantially inflate perceived informativeness. Notably, this endowment effect in beliefs is equally strong among participants with high and low financial literacy. However, the unconditional ability to discriminate between informative and uninformative data varies sharply: high-literacy participants are three times better at adjusting beliefs when data is actually informative.

Second, the distortion propagates from beliefs to forecasts and investment behavior. Among participants who perceive data as informative, payment increases the sensitivity of return forecasts to order-flow imbalance by 31%. Paying for data leads to "noise overfitting": when data is objectively uninformative, return forecast errors are 13.9% larger in paid rounds than in free rounds (corresponding to an 1.08 percentage points increase).

Third, to understand the pass-through from forecasts to investments, we develop a simple model of sunk costs where payment for data introduces psychological negative utility from under-utilization. Investors feel pressure to justify the expenditure through action, and the most tangible action is to trade. Paying for data leads investors to unconditionally take more risk and, as a consequence, to respond less to their own forecasts. Our experiment provides supporting evidence for these mechanisms. Among participants who perceive data as informative, payment for data increases the sensitivity of return forecasts to order-flow imbalance by 31%. In addition to increasing forecast responsiveness, paying for data also worsens forecast accuracy. Payment also raises investment levels by 7 percentage points while simultaneously reducing the elasticity of investment to beliefs by 15%.

Importantly, the sunk cost investment effect is concentrated among participants with low financial literacy, who invest nearly 15 percentage points more of their endowment when they pay for data and exhibit sharply attenuated forecast pass-through. High-literacy participants show no such effects, suggesting that financial sophistication offers some protection against sunk-cost distortions in how investments respond to beliefs and forecasts; even though it offers no protection against distortions in beliefs and forecasts themselves.

Finally, we characterize selection into data purchase. Both overconfidence and financial literacy predict willingness to pay for data: a one-standard-deviation increase in either raises purchase probability by 5–6 percentage points.

2 Literature review

We contribute to the experimental literature on individual investors in modern financial markets. Closest to our paper, Andries et al. (2025) show that when predictive signals are transparent and freely available, investors update forecasts in the right direction but underreact in their portfolio choices. At the same time, Chapkovski et al. (2025b) document that retail investors are willing to pay more for data than the value they can extract from it. We instead isolate the behavioral consequences of actively paying for information. In our design, participants in the costly and free data conditions observe identical order-flow histories; the only difference is whether access requires an active purchase decision. This allows us to cleanly identify how *investment* in data, rather than the data itself, shapes beliefs and investment behavior.

Beyond identifying the causal effect of paying for information, our design also sheds light on who pays for data, contributing to the literature on endogenous information acquisition (Huber et al., 2011; Fuster et al., 2022). By observing voluntary purchase decisions, we characterize selection into data acquisition and show how overconfidence, financial literacy, and risk preferences shape willingness to pay for complex market data.

Our paper is also related to work studying how platform architecture influences retail trading behavior. Chapkovski et al. (2025a, 2026) show that gamification increases trading volume and risk-taking and can reinforce trading mistakes, while Yelagin (2024) finds that gamification raises return volatility but makes retail order flow less toxic. Other studies document that engagement tools such as price notifications, alerts, and smartphone interfaces increase attention and risk-taking (Arnold et al., 2022; Kalda et al., 2021), and that attention-steering features on Robinhood contribute to poor performance (Barber et al., 2022). We complement this literature by studying a distinct and increasingly important platform feature: the sale of granular market data to retail traders.

We further contribute to the literature on retail trading and investor heterogeneity. Earlier studies show that retail trades can predict returns, with aggressive trades anticipating news and passive orders providing contrarian liquidity (Kaniel et al., 2008; Kelley and Tetlock, 2013). More recent evidence highlights substantial heterogeneity across retail investors, with order flow on platforms such as Robinhood differing in sophistication and predictive content from broader retail flow (Eaton et al., 2022). Our design speaks directly to this heterogeneity by measuring both selection into data acquisition and belief updating conditional on acquisition, using financial literacy and overconfidence as proxies for investor sophistication.

Third, we contribute to the literature on the value of financial data and endogenous information acquisition. Recent work treats data as an economic asset whose value depends on how it improves decision-making and equilibrium outcomes (Veldkamp, 2023; Farboodi et al., 2024). By holding the information environment fixed and varying only whether investors actively pay for access, our experiment allows us to compare willingness to pay with the value investors can extract from identical data. We show that paying for information can distort belief formation and investment responses, generating systematic overvaluation even when processing capacity is unchanged.

Finally, our hypotheses relate to the endowment effect, which shows that ownership can distort valuation even when it does not affect underlying payoffs. Classic experiments document that individuals assign higher value to goods they own, creating gaps between willingness to pay and willingness to accept (Kahneman et al., 1990), with effects strongest among inexperienced agents and attenuating with experience (List, 2011). While this literature focuses on physical or financial assets, we study an endowment effect in the use of information. In our setting, paying for data creates a sense of ownership over an informational input, increasing its perceived precision and its influence on beliefs and portfolio choices. Our design directly tests this mechanism by holding information fixed and varying only whether access requires an active purchase decision.

3 Experiment design

3.1 Round investment game

Our experimental design builds on the framework of Andries et al. (2025). In each round, participants observe past realizations of two variables: (i) a "stock return" r_t and (ii) an order flow imbalance ω_t. The display consists of 16 past observations — visually mapped to half-hour intervals over a trading day — to mimic an intraday trading environment.

Across rounds, the order imbalance may or may not predict returns. Participants know this but are not told which specific rounds contain predictive information, nor the fraction of such rounds (set to 50%). We refer to rounds where the imbalance is uninformative as *baseline* rounds, and to those where it predicts returns as *predictable* rounds.

In baseline rounds, stock returns are drawn from a random walk process

$$r^b_{t+1} = \mu + \epsilon^b_{t+1}, \tag{1}$$

where $\mu = 15\%$ and $\epsilon^b_t \sim \mathcal{N}\left(0, \sigma^2_b\right)$ with $\sigma = 12\%$. Conversely, in predictable rounds, stock returns load on the lagged order imbalance:

$$r^p_{t+1} = \mu + \omega_t + \epsilon^p_{t+1}, \tag{2}$$

where the order imbalance is identically and independently distributed as $\omega_t \sim \mathcal{N}\left(0, \sigma^2_\omega\right)$ and $\epsilon^p_t \sim \mathcal{N}\left(0, \sigma^2_p\right)$. We impose $\sigma^2_p + \sigma^2_\omega = \sigma^2_b$ to ensure that r^b and r^p share the same unconditional distribution. The only difference between baseline and predictable rounds is thus the presence of an informative signal that is correlated with returns.

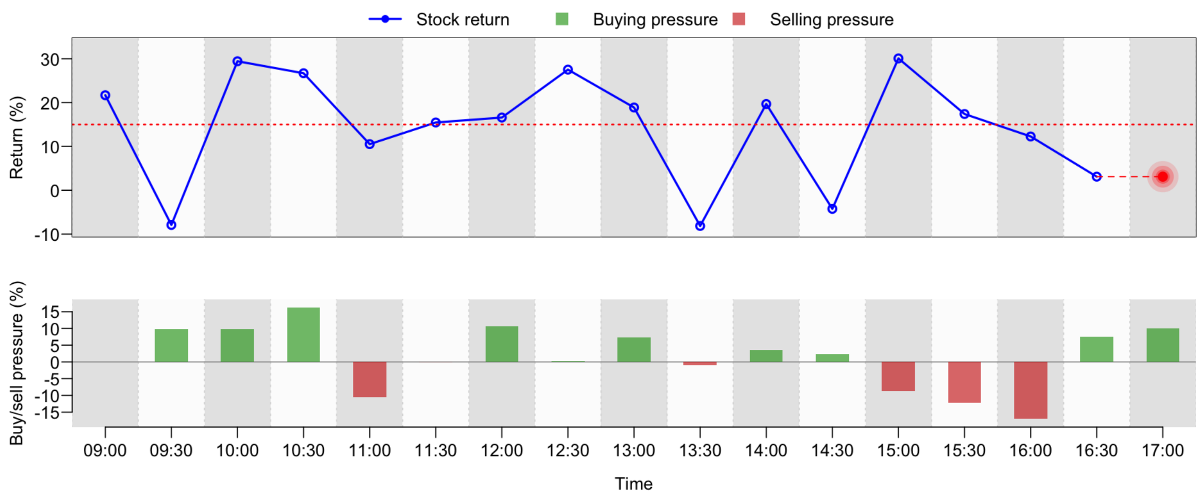
In the experiment, we set $\sigma_\omega = 9\%$ which implies $\sigma_p = 7.9\%$. This parametrization yields a correlation of $\text{Corr}\left(r^p_{t+1}, \omega_t\right) = \frac{0.09^2}{0.12 \times 0.09} = 0.75$, ensuring that the imbalance is sufficiently informative to be visually detectable in the time-series display, while still leaving substantial residual noise. This mirrors realistic intraday settings, where order-flow imbalance carries predictive content but is far from a perfect signal. Figure 1 illustrates the information displayed to participants in each round.

At the start of each round, participants receive an endowment of E$100 (experimental dollars). They choose how much to invest in the risky stock; the remainder is held in cash at a zero return. Leverage and short selling are not permitted. In addition to choosing their portfolio, participants report (i) whether they believe the order-flow imbalance is informative in that round and (ii) their forecast of the next-period return.

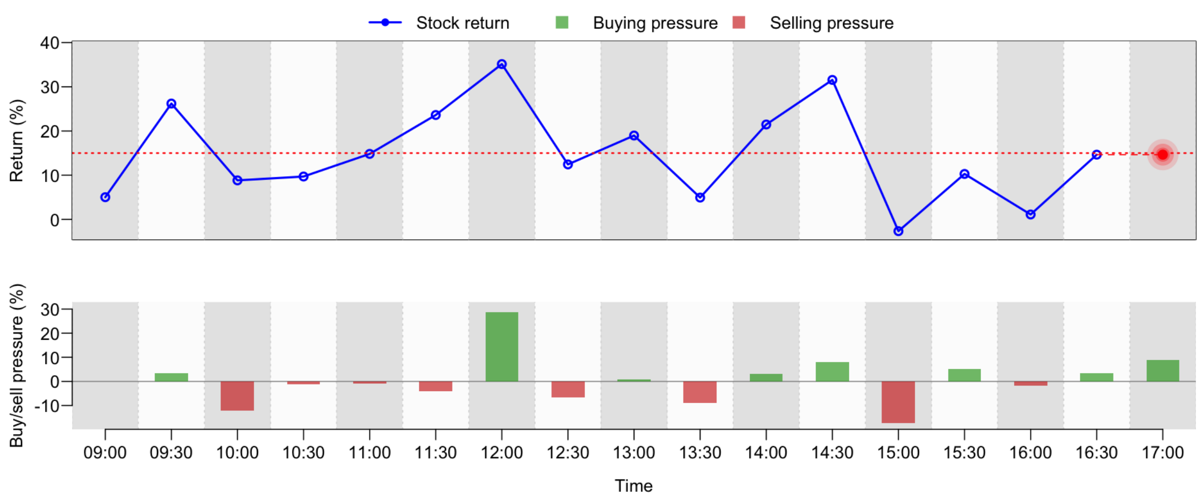
At the end of a round, participants receive feedback on the realized return and on whether the imbalance was in fact informative. They earn E$1 for correctly identifying the round type and an additional E$1 for providing a forecast within one percentage point of the realized return.

Figure 1: **Example screens shown to participants.**

Each panel displays 16 past half-hour returns and order-flow imbalances (buy/sell pressure). The return to be predicted (at 17:00) is marked by a red fuzzy circle.



(a) Baseline round: order imbalance is uninformative.



(b) Predictable round: order imbalance is informative.

3.2 Estimating the value of order imbalance data

To quantify the ex ante value of data, we impose structure on investor preferences. We assume investors have constant relative risk aversion (CRRA) over wealth, $U(w) = \frac{w^{1-\gamma}}{1-\gamma}$, where γ is the risk aversion coefficient.

Under CRRA preferences and normally distributed returns, the classical result in Merton (1969) implies that investors allocate a constant share of wealth to the risky asset equal to $\alpha = \frac{\mathbb{E}r}{\gamma\sigma^2(r)}$. Using our experimental parameters and a risk-aversion level of $\gamma = 24$ (calibrated to estimates in Andries et al., 2025), the optimal portfolio share in baseline rounds is $\alpha_b^\star = 43.4\%$. This allocation yields a certainty equivalent of E\$103.31 for an initial endowment of E\$100.

To estimate optimal investment in predictable rounds, we simulate 20,000 realizations of the order-flow imbalance. For each simulated draw ω_i, the optimal risky share is

$$\alpha_{r,i}^\star = \max\left\{0, \min\left\{\frac{\mu + \omega_i}{\gamma\sigma_p^2}, 1\right\}\right\}, \tag{3}$$

that is, the Merton share projected onto the admissible interval $[0, 1]$ reflecting the no-leverage and no-shorting constraints.

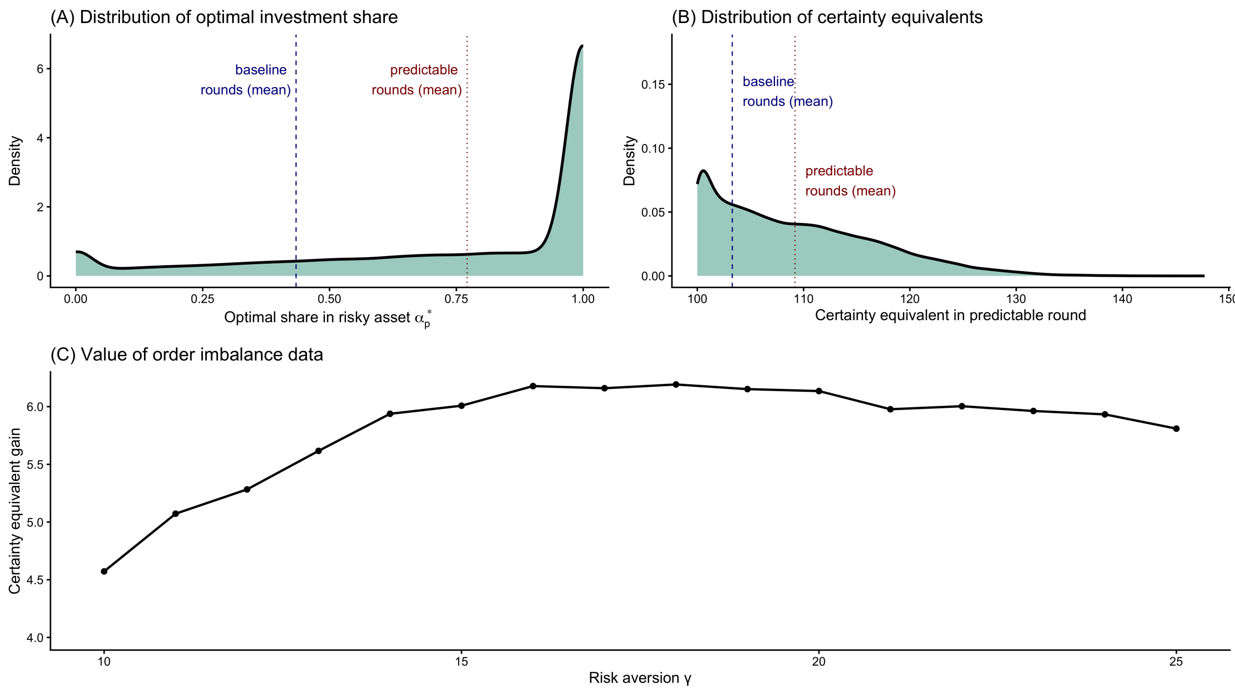
Figure 2 illustrates the simulation results for $\gamma = 24$. Panel (A) plots the distribution of optimal risky shares $\alpha_{p,i}^\star$, which vary substantially across simulations. The mean optimal share is 77%: that is, an increase of 33.6 percentage points relative to the baseline allocation. Predictive information therefore leads investors to take substantially larger positions on average, and it also generates considerable dispersion.

Panel (B) reports the corresponding distribution of certainty equivalents (CEs) across the 20,000 predictable-round simulations. The distribution is right-skewed: while most realizations yield moderate improvements over the baseline CE of E\$103.31, a meaningful subset delivers much larger utility gains when the order-flow signal strongly increases expected returns. The mean and median certainty equivalent in predictable rounds are E\$109.18, implying an ex-ante value of data of E\$5.87 (=E\$109.18-E\$103.31).

Panel (C) summarizes the ex ante value of order-flow imbalance data as a function of risk aversion γ. The certainty-equivalent gain is strictly positive for all γ considered, confirming that

Figure 2: **Optimal portfolios and the value of predictive order flow data**

This figure shows simulation results for the optimal risky share and certainty-equivalent gains under CRRA preferences. Panel (A) plots the distribution of optimal risky shares in predictable rounds. Panel (B) reports the corresponding distribution of certainty equivalents. Panel (C) shows the ex ante value of data as a function of risk aversion γ.



predictive order-flow information is always valuable, and it remains relatively stable between E$5 and E$6. On the one hand, data is valuable for risk-averse investors because it reduces the residual noise in future returns. On the other hand, as risk aversion increases, optimal positions shrink both with and without data, which dampens the incremental benefit of information.

3.3 Experiment design

Each participant plays 14 rounds, starting each round with an endowment of E$100. The first two rounds are training rounds in which all participants observe order-flow data.

We randomly assign participants to a **treatment** or **control** group. Control participants observe order-flow data at no cost in all twelve non-training rounds. In the treatment group, participants may purchase order-imbalance data for E$5 (the calibrated fair value from Section 3.2) at the start of each round — following, e.g., Huber et al. (2011); Fuster et al. (2022). After they make a choice, they are randomly assigned to one of two possible conditions:

- **Paid data round** (drawn with probability $\pi = 2/3$). In paid data rounds, the purchase decision is binding. Participants who chose to pay for data see the order-flow imbalance and have their balance reduced by E$5. Those who decline see the message: "*You do not have privileges to access order-flow data.*"

- **Free data round** (probability 1/3). In free data rounds, all participants see order-flow data at no cost regardless of their choice.

Order-flow data is informative in exactly half of the rounds; however, participants are not told the share of informative rounds.

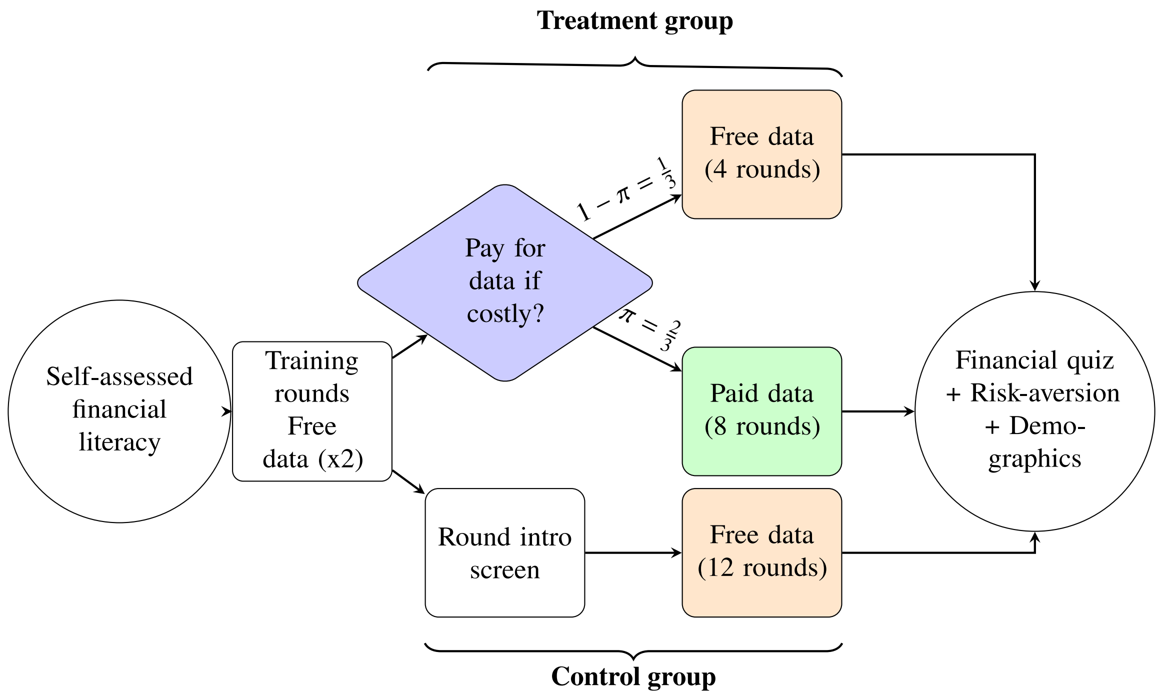
Finally, participants answer 13 financial literacy questions provided in Appendix A. The questions are adapted from Fernandes et al. (2014) and include the three-question measure developed by Lusardi and Mitchell (2011), as well as two trading specific questions following the MiFID questionnaire in Inghelbrecht and Tedde (2024). To measure overconfidence and distinguish between subjective and objective financial literacy, we also ask participants:

> *On a scale from zero to ten, where zero is not at all knowledgeable about personal finance and ten is very knowledgeable about personal finance, what number would you be on the scale?*

This question measures self-assessed (subjective) financial knowledge and is identical to the one used in Cupák et al. (2020). We pose this question before the experiment begins to avoid any influence from the trading game's monetary performance or the quiz's perceived difficulty.

Figure 3 illustrates the experiment's timeline. Appendix B reproduces the consent form. The experimental instructions given to participants are reproduced in Appendix C. Before the trading starts, participants need to correctly answer six comprehension questions, listed in Appendix D. This allows us to make sure that participants indeed understand the experiment before the trading rounds start. Finally, at the end of the experiment all participants are required to fill in a risk-aversion task (Holt and Laury, 2002) and a demographic questionnaire.

Figure 3: **Experiment timing**



Our framework isolates two forces that may drive demand for order-flow data. The first is a *sunk cost* or endowment effect: paying for data distorts beliefs and decisions *ex post*, holding the information set fixed. The second is a *selection* effect: participants differ in *ex ante* willingness to pay based on financial literacy, overconfidence, and risk aversion.[3]

Since participants who observe data in the treatment and control groups receive identical information in any given round, any differences in beliefs or investment behavior between the treatment and control groups must arise from the act of purchasing the data rather than from differential information sets.

[3] A third possibility is an *elicitation effect*: being asked whether to pay for data may itself shift beliefs, even among participants who decline.

4 Results

4.1 Cohort formation

We recruited 774 participants from Prolific, an online platform designed for academic research (Palan and Schitter, 2018). To obtain a diverse sample, we restricted recruitment to United States residents and used Prolific's stratified sampling, which matches U.S. census distributions for age, gender, and ethnicity. Data collection occurred on January 21–22, 2026; median completion time was 25 minutes. We excluded three participants who did not complete the demographics questionnaire, yielding a final sample of 771 subjects (578 in the treated condition and 193 in the control group). We targeted a 3:1 allocation ratio between treatment and control groups to balance the number of paid and free data rounds in our full sample.

Participants received a base payment of £8.81 per hour (£3.67 for the median duration of 25 minutes). They also earned bonuses tied to outcomes in a randomly selected round, forecast and belief accuracy across all rounds, and financial literacy quiz performance. Mean and median bonuses were £3.48 and £3.43, respectively, ranging from £3.04 to £4.40.

Table 1 reports the summary statistics of our sample. The median age is 45 years, with participants ranging from 18 to 82 years old. The sample is evenly split by gender. Participants are well-educated and skew towards financially literate: 78% hold a college degree, with 33% having taken a finance course. Sixty-four percent of subjects report trading experience, whereas the average financial literacy score is 76%.

On average, participants invest 45.8% of their endowment in the risky asset, close to the optimal baseline share of 43.4% derived in Section 3.2. This suggests that our risk-aversion estimates are well-calibrated, which in turn implies the data should be appropriately priced. The fact that 56% of participants choose to pay for the data when offered the option corroborates this. The intermediate take-up rate reflects meaningful heterogeneity in the sample, ensuring that we avoid corner solutions where all participants either purchase data or decline to do so.

Participants report a 72% average belief that data is informative. This is well above the true (undisclosed) probability of 50%, suggesting optimism about data quality. Participants do exhibit some ability to discriminate, with beliefs rising to 75% when data is informative and falling to 68% when it is not, though the gap remains modest. Unlike Andries et al. (2025), who restrict

Table 1: **Summary Statistics**

This table presents the summary statistics of our experimental dataset.

Statistic	Mean	St. Dev.	Pctl(25)	Median	Pctl(75)	Min	Max
Return forecast	12.66	11.22	5.00	12.00	18.00	−20.00	50.00
Investment share (%)	45.79	35.05	10.53	45.00	75.00	0.00	100.00
Belief informative	0.72	0.45	0	1	1	0	1
Belief informative | informative	0.75	0.43	1	1	1	0	1
Belief informative | uninformative	0.68	0.47	0	1	1	0	1
Choose to pay	0.56	0.50	0	1	1	0	1
Overconfidence	−0.21	0.27	−0.40	−0.22	−0.05	−0.90	0.67
Financial quiz	0.76	0.22	0.67	0.83	0.92	0.00	1.00
Female	0.50	0.50	0	0	1	0	1
Age	45.65	15.70	32	46	59	18	82
Finance course	0.33	0.47	0	0	1	0	1
College education	0.78	0.41	1	1	1	0	1
Trading experience	0.64	0.48	0	1	1	0	1

their Prolific sample to participants with strong records of predicting data informativeness, we intentionally impose no such filter because our goal is precisely to investigate mistakes.

4.2 Payment for data and beliefs

Hypothesis 1. ***Payment and beliefs.*** *Participants who pay for order-flow data are more likely to perceive order-flow imbalance as informative, particularly when imbalance is objectively uninformative.*

The intuition is straightforward. Once participants pay the data fee, they become motivated to believe the expenditure was worthwhile. The purchased signal is mentally coded as an "asset" they now own, which inflates its perceived value and predictive content. As a result, participants who pay are more likely to interpret noise as information.

To test Hypothesis 1, we estimate:

$$\text{Belief}_{it} = \alpha + \beta_1 \text{Treated}_i + \beta_2 \text{PaidRound}_t + \beta_3 (\text{PaidRound}_t \times (1 - \text{Informative}_t)) \\ + \beta_4 \text{Informative}_t + \beta_5 \text{PayChoice}_{it} + X_{it}' \gamma + \phi \text{Round}_t + \varepsilon_{it}, \tag{4}$$

where $Belief_{it}$ equals one if participant i perceives order flow as informative in round t; $Treated_i$ indicates treatment assignment; $PayChoice_{it}$ indicates choosing to pay for data; $PaidRound_{it}$ indicates a paid data round, and $Informative_t$ indicates whether order flow truly predicts returns. We further control for round number, last stock return in the round, and participant demographics.

The coefficient β_2 and β_3 identify the causal effect of paying for data. Identification relies on random assignment of paid versus free rounds. Hypothesis 1 predicts $\beta_3 > 0$. We also expect $\beta_2 > 0$ although the effect is likely to be weaker, as participants may more readily recognize informativeness when order flow is genuinely predictive.

Table 2 reports the results. Consistent with Hypothesis 1, participants who pay for data are 8 percentage points more likely to perceive uninformative order flow as informative. The effect is economically meaningful and represents 11% of the baseline belief of 0.72 in Table 1. Notably, this causal effect amounts to 61% of the selection effect (13 percentage points), indicating that both the payment itself and selection into payment substantially distort beliefs.

The sunk cost bias is consistent across financial literacy levels: the effect is nearly identical for high- and low-scoring participants (columns 3–6). However, the ability to discriminate between informative and uninformative data varies sharply. Overall, beliefs are 10 percentage points higher when data is truly informative.

Table 2: **Payment for Data and Perceived Informativeness**

This table presents estimation results of regression (4). Columns (1)–(2) include all participants; columns (3)–(4) restrict to participants with above-median financial literacy scores; columns (5)–(6) restrict to below-median. *Treated* indicates assignment to the treatment group (which includes paid data rounds). *Paid round* indicates a round in which the participant paid for data. *Informative round* indicates whether order flow truly predicts returns in round t. *Choose to pay* indicates the participant chose to pay when offered the option. *Round number* is the round index. *Last return* and *Last imbalance* are the last values stock return and order-flow imbalance in the graph for a particular round. *Overconfidence* is measured as the difference between self-assessed and actual financial literacy. *Financial quiz* is the participant's score on the financial literacy test. Additional controls include age, gender, trading experience, risk aversion, and finance course exposure. Standard errors, shown in parentheses, are double-clustered by participant and round number.

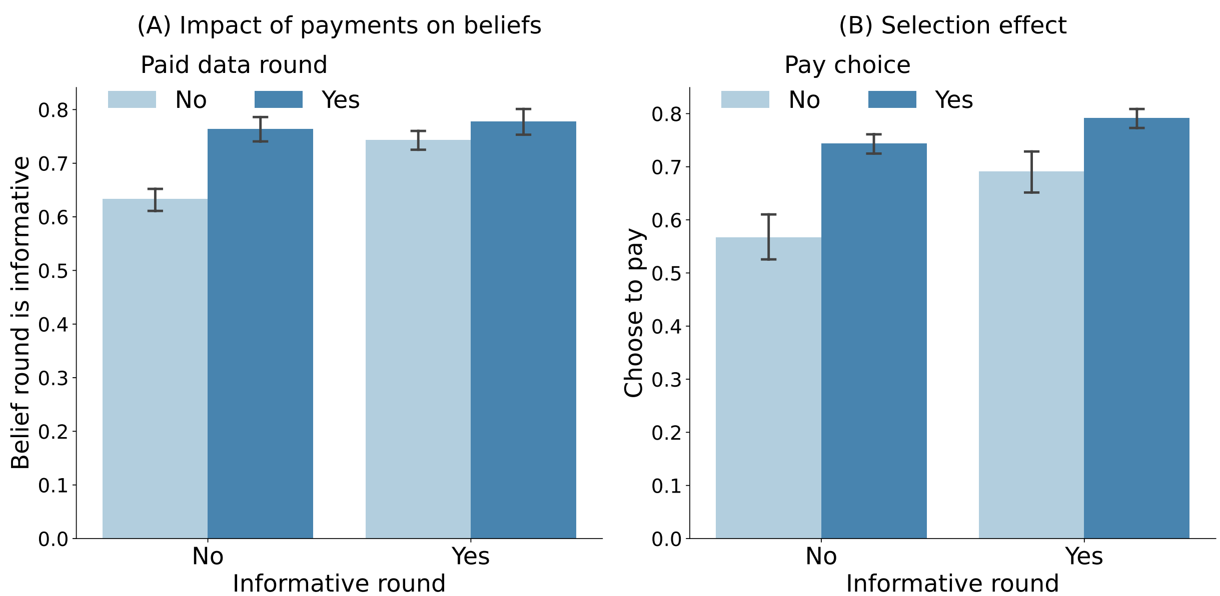
Dependent Variable:	Belief informative					
	All quiz scores		High quiz scores		Low quiz scores	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Paid × Not informative	0.08***	0.08***	0.08**	0.07**	0.09**	0.09**
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
Paid round	-0.02	-0.02*	-0.01	-0.01	-0.04	-0.05*
	(0.01)	(0.01)	(0.03)	(0.03)	(0.03)	(0.03)
Treated	-0.05*	-0.05**	-0.07*	-0.08**	-0.02	-0.02
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
Informative round	0.10***	0.10***	0.15***	0.15***	0.05**	0.05**
	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)
Choose to pay	0.13***	0.13***	0.14***	0.15***	0.10***	0.12***
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
Round number	-0.01***	-0.01***	-0.01	-0.01	-0.01	-0.01
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)
Last return	0.01	0.01*	0.00	0.00	0.01**	0.01**
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)
Last imbalance	0.01***	0.01***	0.01	0.01	0.01	0.01
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
Overconfidence	0.02**		-0.01		0.06***	
	(0.01)		(0.02)		(0.01)	
Financial quiz	0.01		0.04		0.07***	
	(0.02)		(0.05)		(0.02)	
Additional controls	Yes	No	Yes	No	Yes	No
Constant	0.62***	0.62***	0.53***	0.60***	0.69***	0.65***
	(0.02)	(0.02)	(0.05)	(0.03)	(0.04)	(0.02)
Observations	7,245	7,245	3,997	3,997	3,248	3,248
R^2	0.03	0.03	0.04	0.04	0.03	0.02

Among participants with high financial literacy, this gap widens to 15 percentage points; among the low literacy group, it shrinks to 5 percentage points. Thus, while financial literacy helps participants distinguish signal from noise, it offers no protection against sunk cost bias.

Figure 4 illustrates these findings. Panel (A) plots average beliefs by paid versus free rounds, separately for informative and uninformative data. The data purchase effect is salient: the gap between paid and free rounds is substantially wider when data is uninformative. Panel (B) illustrates the selection effect, comparing beliefs between participants who chose to pay and those who declined. Participants who select into purchasing data report higher beliefs, with the gap being persistent across both informative and uninformative rounds.

Figure 4: **Payment for Data and Beliefs**

This figure illustrates the relationship between payment and perceived informativeness of order-flow data. Panel (A) shows the causal effect of payment: average beliefs that data is informative, split by whether the participant paid for data in that round (randomly assigned) and whether data was truly informative. Panel (B) shows the selection effect: average beliefs split by whether the participant chose to pay when offered the option. Error bars represent 95% confidence intervals.



4.3 Payment for data and return forecasts

If payment inflates the perceived informativeness of data, the effect should propagate from beliefs to return forecasts. Among participants who view data as informative, those who actively pay for it may place greater weight on order-flow imbalance when forming return expectations.

Hypothesis 2. ***Payment and forecasts.*** *Conditional on perceiving data as informative, participants who pay for order-flow data adjust their return forecasts more strongly in response to order-flow imbalance.*

To test Hypothesis 2, we restrict the sample to rounds in which participants report believing data is informative and estimate:

$$\begin{aligned}\text{Forecast}_{it} = \alpha + \delta_1\left(\text{PaidRound}_t \times \text{Last imbalance}_t\right) + \delta_2\text{Last imbalance}_t + \delta_3\text{PaidRound}_t + \\ + \delta_4\,\text{PayChoice}_{it} + \delta_5\text{Treated}_i + \phi\,\text{Round}_t + X_{it}'\zeta + \nu_{it}.\end{aligned} \tag{5}$$

Hypothesis 2 predicts $\delta_1 > 0$: participants who pay for data adjust their forecasts more strongly to order-flow imbalance when they perceive the signal as informative.

We report the results in Table 3. Among participants who perceive data as informative, return forecasts respond significantly to order-flow imbalance: a one-standard-deviation increase in imbalance raises return forecasts by 1.85 percentage points, or 14.3% relative to the baseline forecast of 12.94 (column 1). Consistent with Hypothesis 2, this sensitivity is amplified by active payment for data. The interaction *Paid round* $\times$ *Last imbalance* is positive and significant ($\delta_1 = 0.57$), implying that payment for data increases the elasticity of forecasts to imbalance by 30.8%. The effect is even stronger among participants with high financial literacy (columns 3–4). For this subsample, payment increases the forecast response to a one-standard-deviation increase in imbalance from 2.09 to 2.95 (that is, 2.09+0.86) percentage points, equivalent to a 41.1% increase in responsiveness.

Table 3: **Payment for Data and Return Forecasts**

This table presents estimation results of regression (5). Columns (1)–(4) restrict to rounds in which participants report believing data is informative; columns (5)–(6) provide a placebo test on rounds where participants deem data uninformative. Columns (1)–(2) include all participants; columns (3)–(4) restrict to participants with above-median financial literacy scores. *Treated* indicates assignment to the treatment group (which includes paid data rounds). *Paid round* indicates a round in which the participant paid for data. *Choose to pay* indicates the participant chose to pay when offered the option. *Last return* and *Last imbalance* are the last values stock return and order-flow imbalance in the graph for a particular round. *Overconfidence* is measured as the difference between self-assessed and actual financial literacy. *Financial quiz* is the participant's score on the financial literacy test. Additional controls include age, gender, trading experience, risk aversion, and finance course exposure. Standard errors, shown in parentheses, are double-clustered by participant and round number.

Dependent Variable:	Return forecast					
	Belief informative				Belief uninformative	
	All quiz scores		High quiz scores		All quiz scores	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Paid round × Last imbalance	0.57**	0.60***	0.86***	0.90***	-0.38	-0.35
	(0.18)	(0.18)	(0.20)	(0.18)	(0.66)	(0.70)
Last imbalance	1.85***	1.83***	2.09***	2.07***	0.39	0.40
	(0.17)	(0.16)	(0.19)	(0.19)	(0.34)	(0.35)
Treated	-1.33*	-1.27*	-0.80	-0.74	-0.49	-0.43
	(0.71)	(0.70)	(0.87)	(0.84)	(0.75)	(0.73)
Choose to pay × Last imbalance	-0.12	-0.15	-0.37*	-0.41**	0.54	0.53
	(0.13)	(0.13)	(0.18)	(0.17)	(0.38)	(0.38)
Choose to pay	1.48***	1.57***	0.59	0.82	0.24	0.21
	(0.44)	(0.44)	(0.49)	(0.48)	(0.48)	(0.47)
Paid round	-0.85***	-0.85***	-0.33	-0.34	0.35	0.30
	(0.24)	(0.24)	(0.27)	(0.24)	(0.74)	(0.73)
Last return	3.96***	4.00***	3.64***	3.64***	3.40***	3.40***
	(0.23)	(0.22)	(0.27)	(0.27)	(0.29)	(0.27)
Overconfidence	1.35***		0.98**		1.33**	
	(0.42)		(0.42)		(0.47)	
Financial quiz	0.21		-0.37		0.26	
	(0.35)		(1.49)		(0.54)	
Additional controls	Yes	No	Yes	No	Yes	No
Constant	12.94***	13.37***	13.04***	12.83***	12.78***	12.34***
	(0.77)	(0.49)	(1.70)	(0.62)	(1.09)	(0.58)
Fit statistics						
Observations	5,198	5,233	2,843	2,843	2,047	2,072
R^2	0.15	0.13	0.15	0.13	0.10	0.09

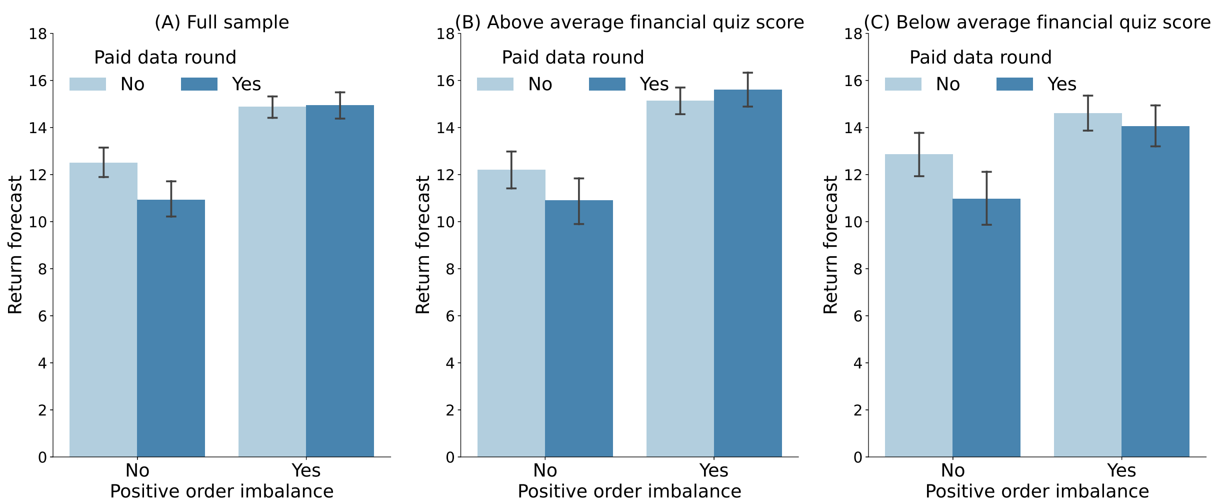
Columns (5)–(6) in Table 3 provide a placebo test, restricting the sample to rounds in which participants deem data uninformative. Reassuringly, both the baseline response to imbalance and the payment interaction are small and statistically insignificant.

Two additional patterns emerge. First, overconfident participants report higher forecasts overall. Second, forecasts load heavily on recent returns, suggesting participants extrapolate from past price movements (Andries et al., 2025).

Figure 5 illustrates these results. One salient result is that the effect of payment on forecast sensitivity is concentrated in rounds with negative order imbalance. When imbalance is negative, participants who paid for data lower their forecasts more aggressively than those in free data rounds. This asymmetry suggests that paying for data makes participants more willing to act on bad news.

Figure 5: **Forecast Sensitivity to Order-Flow Imbalance**

This figure illustrates the relationship between payment and return forecasts among participants who perceive order-flow data as informative. Each panel plots average return forecasts split by whether order imbalance is positive and whether the participant paid for data in that round (randomly assigned). Return forecasts are residualized with respect to lagged returns, overconfidence, treatment status, pay choice, financial quiz score, gender, age, and round number. Panel (A) shows the full sample; Panel (B) restricts to participants with above-median financial literacy; Panel (C) restricts to below-median. Error bars represent 95% confidence intervals.



When imbalance is positive, forecasts are similar across paid and free rounds, though the direction of adjustment differs by financial literacy. High-literacy participants who pay nudge their forecasts upward—the correct response to positive imbalance—while low-literacy participants who pay actually lower their forecasts (Panels B and C). However, both effects are small and statistically insignificant.

If investors are more likely to believe data is informative when it is not, and if they rely more heavily on data when forming forecasts, they should also make larger forecasting errors when paying for data: that is, a "noise overfitting" effect. We measure forecast error as the absolute difference between the return forecast of participant i in round t, and the true conditional expected return in that round:

$$\text{ForecastError}_{it} = \left|\text{Forecast}_{it} - \mathbb{E}\left[\text{Return}_t \mid \text{Informative}_t, \text{Imbalance}_t\right]\right|. \tag{6}$$

The conditional expectation equals 15% in uninformative rounds and 15% plus the realized imbalance in informative rounds.

In Table 4, we show supportive evidence for the noise overfitting channel. Participants who pay for data make 1.08 percentage points larger forecast errors when data is objectively uninformative, representing a 13.88% increase relative to the baseline error of 7.78 percent. The noise overfitting effect is remarkably consistent across financial literacy levels. High-literacy participants exhibit a 1.24 percentage point increase in forecast errors when paying for data (column 3); low-literacy participants exhibit a 0.90 percentage point increase (column 5). Both effects are statistically significant. This pattern mirrors our earlier finding that sunk-cost bias in beliefs is equally strong across literacy levels—even sophisticated participants overfit to noise when they have paid for data.

Financial literacy does reduce forecast errors overall: a one-standard-deviation increase in quiz score lowers errors by 0.71 percentage points. However, this benefit operates through unconditional forecast accuracy, not through immunity to payment-induced distortions.

Table 4: **Payment for Data and Return Forecast Errors**

This table reports estimates of regression model:

$$\text{ForecastError}_{it} = \alpha + \beta_1 \text{Treated}_i + \beta_2 \text{PaidRound}_t + \beta_3 \text{PayChoice}_{it} + X'_{it}\gamma + \varepsilon_{it}, \tag{7}$$

where $ForecastError_{it}$ is defined as in (6), *Paid round* indicates paid data rounds and *Choose to pay* indicates selection into purchase. *Last return* and *Last imbalance* are the last values stock return and order-flow imbalance in the graph for a particular round. The sample only includes rounds in which data is objectively uninformative. Columns (1)–(2) show the full sample; columns (3)–(6) split by financial literacy. Standard errors are double-clustered by participant and round number.

Dependent Variable:	Forecast error (%)					
	All quiz scores		High quiz scores		Low quiz scores	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Paid round	1.08**	1.09**	1.24**	1.25**	0.90**	0.89**
	(0.34)	(0.34)	(0.33)	(0.32)	(0.34)	(0.33)
Treated	-0.25	-0.15	-0.06	0.17	-0.61	-0.69
	(0.58)	(0.58)	(0.60)	(0.58)	(0.79)	(0.78)
Choose to pay	-0.30	-0.40	-0.40	-0.59	-0.19	0.00
	(0.56)	(0.54)	(0.59)	(0.54)	(0.72)	(0.70)
Last imbalance	-1.23***	-1.24***	-1.37***	-1.37***	-1.07***	-1.06***
	(0.21)	(0.21)	(0.24)	(0.23)	(0.17)	(0.14)
Last return	-0.64	-0.64	-0.75	-0.76	-0.50	-0.48
	(0.45)	(0.45)	(0.48)	(0.48)	(0.39)	(0.34)
Overconfidence	0.48		0.05		0.94*	
	(0.24)		(0.27)		(0.42)	
Financial quiz	-0.71**		-2.49**		-0.18	
	(0.28)		(0.69)		(0.62)	
Additional controls	Yes	No	Yes	No	Yes	No
Constant	7.78***	7.04***	8.73***	6.12***	8.13***	8.18***
	(0.63)	(0.49)	(1.22)	(0.51)	(1.01)	(0.64)
Fit statistics						
Observations	4,626	4,626	2,484	2,484	2,142	2,142
R^2	0.05	0.02	0.05	0.02	0.04	0.02

4.4 Payment for data and investments

In this section, we turn from beliefs and forecasts to trading choices. The sunk-cost literature suggests that paying for a good or service creates psychological pressure to justify the expenditure through action (Staw, 1976; Arkes and Blumer, 1985). What such action means in a financial context is not immediately obvious. Unlike a concert ticket, where attendance is the natural way to extract value, financial information can be used in multiple ways: one might trade more aggressively, or adjust beliefs more strongly in response to observed data.

To discipline the empirical predictions, we build a simple model based on Prelec and Loewenstein (1998), introducing the psychological burden of payment into the investor's utility function. Investors choose investment level θ to maximize

$$\max_{\theta} \mathbb{E}\left[U\left(W+\theta r\right) \mid m\right] - \lambda \ell\left(\theta\right), \tag{8}$$

where m denotes the signal used in forecasting returns, $\lambda \in \{0, 1\}$ indicates whether the data cost is incurred, and $\ell(\theta)$ captures the disutility from under-utilizing purchased data.

In our setting, the most tangible way to "use" purchased data is to trade on it, which provides an immediate sense of having extracted value from the purchase. We assume that investing more reduces the sense of waste ($\ell'(\theta) < 0$) and additional investments generate diminishing marginal relief ($\ell''(\theta) > 0$). The first-order condition is:

$$f\left(\lambda, \theta\right) = \mathbb{E}\left[U'\left(W+\theta r\right) r \mid m\right] - \lambda \ell'\left(\theta\right) = 0. \tag{9}$$

Applying the implicit function theorem:

$$\frac{\partial \theta^{\star}}{\partial \lambda} = \frac{\ell'\left(\theta\right)}{\mathbb{E}\left[U''\left(W+\theta r\right) r^2 \mid m\right] - \lambda \ell''\left(\theta\right)} > 0. \tag{10}$$

since both the numerator ($\ell' < 0$) and the denominator are negative (second-order condition). Thus, sunk-cost pressure increases equilibrium investment to justify the purchase of data.

Hypothesis 3. ***Sunk-cost investment.*** *Participants who pay for order-flow data invest larger amounts than participants who receive identical data for free.*

The sunk-cost mechanism not only raises average investment but also reduces the elasticity of investment to beliefs. The rationale is that the psychological pull toward "using" the data increases investment levels but at the same time dampens how much beliefs and forecasts translate into position changes.

To see this formally, note that the first-order condition implicitly defines optimal investment $\theta^\star$ as a function of the forecast m and sunk-cost indicator λ. Applying the implicit function theorem, the sensitivity of investment to beliefs is:

$$\left|\frac{\partial \theta^\star}{\partial m}\right| = \frac{\left|f_m\right|}{\lambda \ell'' (\theta) - \mathbb{E}\left[U'' (W + \theta r) r^2 \mid s\right]}, \tag{11}$$

where f_m captures how the marginal value of investment responds to beliefs. The denominator is positive by the second-order condition. For investors who incurred the data cost ($\lambda = 1$), the additional term $\ell''(\theta) > 0$ increases the denominator, strictly reducing belief sensitivity. The curvature of the sunk-cost term acts like additional risk aversion specific to deviating from a "justifying" investment level.

Hypothesis 4. ***Belief pass-through.*** *Participants who pay for order-flow data exhibit lower sensitivity of portfolio choices to their own return forecasts than participants who receive identical data for free.*

To test Hypotheses 3 and 4, we estimate the sensitivity of investment to return forecasts. Following Andries et al. (2025), we use an instrumental variable approach to extract "informed forecasts," instrumenting stated forecasts with the last observed order-flow imbalance. We then estimate:

$$\text{InvestmentShare}_{it} = \beta_1 \widehat{\text{Forecast}}_{it} + \beta_2 (\widehat{\text{Forecast} \times \text{PaidRound}})_{it} + \gamma \text{PaidRound}_{it} + \mathbf{X}_{it}' \delta + \varepsilon_{it}, \tag{12}$$

where $InvestmentShare_{it}$ is the share of endowment participant i invests in round t; $\widehat{Forecast}_{it}$ and $(\widehat{Forecast \times PaidRound})_{it}$ are the instrumented forecast and its interaction with payment status; $PaidRound_{it}$ indicates whether the participant paid for data; and $\mathbf{X}_{it}$ includes round fixed effects and controls. We restrict the sample to rounds in which participants report believing data is informative, since only then participants base their forecasts on imbalance data. Hypothesis 3 predicts $\gamma > 0$: that is, payment increases average investment. Hypothesis 4 predicts $\beta_2 < 0$: payment attenuates the responsiveness of investment to beliefs.

Table 5 reports the results. Investment share responds strongly to informed forecasts across all specifications: a one percentage point increase in the informed return forecast raises investment by 2.48 percentage points (column 1), an effect that is consistent across high- and low-literacy participants (columns 3 through and 6).

The results support the conjectured economic mechanism. Consistent with Hypothesis 3, paying for data raises investment share by 7.16 percentage points on average. Further, in line with Hypothesis 4, payment also reduces the response of investment to forecasts by 15%, from 2.48 to 2.11 percentage points (computed as 2.48-0.37 in the first column of Table 5).

Both effects are concentrated among participants with low financial literacy (columns 5–6). For this group, payment raises investment by 14.6 percentage points and reduces forecast elasticity by 0.80 percentage points. Among high-literacy participants (columns 3–4), we find no significant sunk-cost effect. This pattern suggests that the psychological burden of payment weighs more heavily on the less financially sophisticated.

Table 5: **Payment for Data and Forecast Pass-Through**

This table presents instrumental variable estimation results of regression (12). The sample is restricted to rounds in which participants report believing data is informative. *Return forecast* and *Return forecast* × *Paid round* are instrumented using last observed order-flow imbalance and its interaction with the *Paid round* dummy. Columns (1)–(2) include all participants; columns (3)–(4) restrict to participants with above-median financial literacy scores; columns (5)–(6) restrict to below-median. *Treated* indicates assignment to the treatment group (which includes paid data rounds). *Paid round* indicates a round in which the participant paid for data. *Choose to pay* indicates the participant chose to pay when offered the option. *Round number* is the round index. *Last return* is the most recent stock return observed in a given round. *Overconfidence* is measured as the difference between self-assessed and actual financial literacy. *Financial quiz* is the participant's score on the financial literacy test. *Female* is an indicator for female participants. Additional controls include age, trading experience, risk aversion, and finance course exposure. Standard errors, shown in parentheses, are double-clustered by participant and round number.

Dependent Variable:	Investment share (%)					
	All quiz scores		High quiz scores		Low quiz scores	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Constant	17.88***	17.70***	16.65**	26.76***	8.97**	8.14***
	(2.73)	(1.24)	(7.05)	(3.71)	(3.51)	(1.29)
Return forecast	2.48***	2.50***	2.30***	2.24***	2.68***	2.67***
	(0.09)	(0.13)	(0.11)	(0.09)	(0.23)	(0.19)
Return forecast × Paid round	-0.37***	-0.38**	-0.01	0.01	-0.80***	-0.84***
	(0.11)	(0.17)	(0.22)	(0.22)	(0.14)	(0.09)
Treated	-5.19	-6.62*	-7.30	-8.10	-2.45	-3.00
	(3.13)	(3.26)	(4.86)	(4.83)	(4.24)	(4.31)
Paid round	7.16***	7.33***	1.42	0.63	14.60***	15.02***
	(1.46)	(1.63)	(2.34)	(1.99)	(2.27)	(1.33)
Choose to pay	2.37	4.22*	7.80*	9.46**	-3.15	-3.98
	(2.15)	(2.30)	(4.00)	(4.20)	(3.00)	(3.07)
Round number	2.27***	2.28***	1.89***	1.87***	2.92***	2.94***
	(0.50)	(0.58)	(0.45)	(0.46)	(0.39)	(0.43)
Last return	-5.75***	-5.77***	-5.04***	-4.87***	-6.69***	-6.66***
	(0.43)	(0.43)	(0.41)	(0.30)	(0.87)	(0.75)
Last return × Paid round	2.92***	2.45***	2.18**	1.99***	4.11***	3.97***
	(0.61)	(0.61)	(0.80)	(0.55)	(0.41)	(0.31)
Overconfidence	-2.43		0.04		-5.58**	
	(1.58)		(2.14)		(2.26)	
Financial quiz	5.03***		14.19**		-1.03	
	(1.61)		(5.63)		(3.02)	
Female	-9.24***		-9.16**		-8.57**	
	(2.31)		(2.96)		(3.75)	
Additional controls	Yes	No	Yes	No	Yes	No
Observations	5,198	5,198	2,843	2,843	2,355	2,355
Adjusted R^2	0.09	0.03	0.06	0.03	0.07	0.03

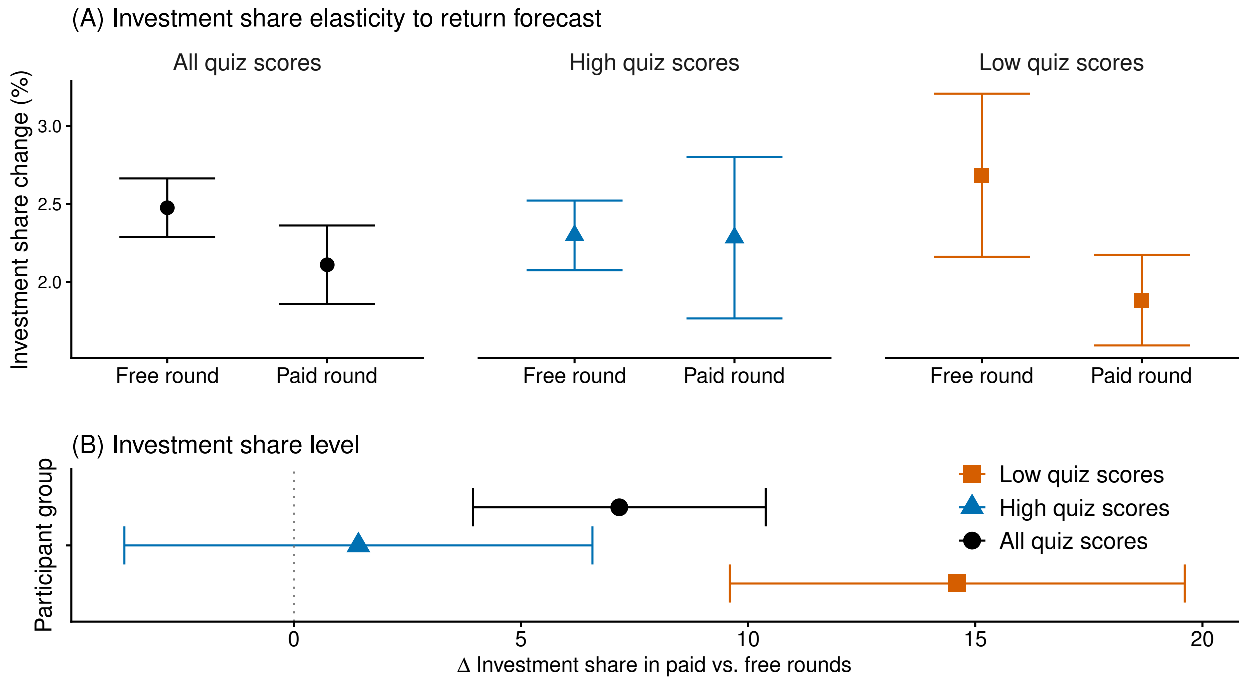
An alternative interpretation is that participants who pay for data increase risk-taking to recover the E$5 fee; that is, a "gambling for resurrection" channel rather than a sunk-cost bias. Under CRRA preferences, the optimal risky share is independent of wealth, so the E$5 fee does not alter the optimal portfolio allocation. Moreover, a pure wealth effect would predict uniform increases in risk-taking across literacy levels, whereas we observe the effect concentrated among low-literacy participants.

Two additional findings emerge. First, financial literacy predicts investment levels: a one-standard-deviation increase in quiz score is associated with 5 percentage points higher investment. Second, female participants invest 9.2 percentage points less than males, consistent with the well-documented gender gap in stock market participation.

Figure 6 summarizes the instrumental variable estimation results. Panel (A) shows that the elasticity of investment to forecasts drops in paid rounds, with the effect driven entirely by low-literacy participants. Panel (B) confirms the same pattern for investment levels: low-literacy participants invest nearly 15 percentage points more when they pay for data, while high-literacy participants show no significant difference.

Figure 6: **Payment for Data and Investment Decisions**

This figure summarizes the IV estimates from regression (12). Panel (A) plots the elasticity of investment to instrumented forecasts in free versus paid rounds; the decline in paid rounds reflects attenuated belief pass-through. Panel (B) plots the difference in investment share between paid and free rounds. Both effects are concentrated among participants with low financial literacy. The sample is restricted to rounds in which participants report believing data is informative. Error bars represent 95% confidence intervals.



One potential concern could be that free data rounds, in which treated participants expected to pay but received data at no cost, may generate a "house money" effect (Thaler and Johnson, 1990). However, the house money effect cannot explain the belief distortions in Table 2. More importantly, this alternative channel would predict higher investment in free data rounds, which is the opposite of our result.

4.5 Who pays for data?

Finally, we investigate the selection channel: which investor characteristics predict the decision to pay for data? We consider three candidate mechanisms. First, overconfident individuals may overestimate their ability to translate information into profitable trades, leading them to overvalue data. Second, financially literate participants may value data for rational reasons, recognizing that order-flow signals can improve forecasts. Third, risk-averse participants might purchase data to sharpen their signal and reduce uncertainty.

Hypothesis 5. ***Selection effects.*** *Overconfident, risk-averse, and financially literate participants are more likely to purchase order-flow data.*

To test Hypothesis 5, we estimate the following cross-sectional regression:

$$\text{PayChoice}_{it} = \alpha + \beta_1 \text{Overconfidence}_i + \beta_2 \text{FinQuiz}_i + \beta_3 \text{RiskAversion}_i + \phi\, \text{Round}_t + X_i' \zeta + \varepsilon_{it}, \quad (13)$$

where $PayChoice_{it}$ indicates whether participant i chose to purchase data in round t. Overconfidence is computed as the difference between the normalized self-assessed financial knowledge and the normalized financial quiz score, $FinQuiz_i$. Finally, $RiskAversion_i$ is proxied by the switching point in the post-experimental Holt and Laury (2002) task. Hypothesis 5 predicts that $\beta_1 > 0$ (overconfidence increases data demand), $\beta_2 > 0$ (financial literacy increases data demand), and $\beta_3 > 0$ (risk aversion increases data demand).

Table 6 reports the results. Consistent with Hypothesis 5, both overconfidence and financial literacy predict data purchase. A one-standard-deviation increase in overconfidence raises the probability of purchasing data by 5 percentage points; the same increase in financial quiz score raises it by 6 percentage points. These effects are robust across OLS and Probit specifications and to the inclusion of round fixed effects.

Contrary to our prediction, risk aversion does not increase data demand. The relationship may be more complex than hypothesized: while risk-averse participants might value data to reduce uncertainty, they also invest less in the risky asset, making the fixed data fee relatively expensive given their lower exposure. Other demographic characteristics such as gender, age, college education, and trading experience, appear to have no significant effect on data purchase.

Table 6: **Determinants of Data Purchase**

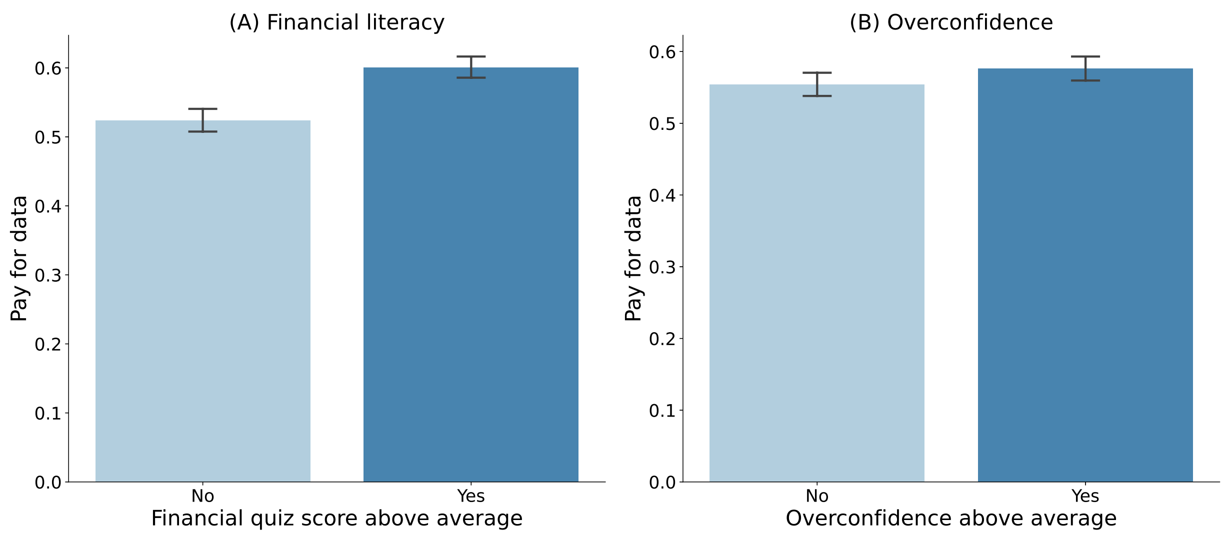
This table presents estimation results of cross-sectional regression (13). The dependent variable equals one if the participant chose to purchase data when offered the option. Columns (1)–(3) report OLS estimates; columns (4)–(6) report Probit estimates. Overconfidence is measured as the difference between self-assessed and actual financial literacy. Financial quiz is the participant's score on the financial literacy test. College education indicates whether the participant holds or is currently pursuing a college degree. Risk aversion is elicited from the Holt and Laury (2002) list choice task. Columns (2)–(3) and (5)–(6) include round fixed effects. Additional controls include gender, age, and trading experience. Standard errors, shown in parentheses, are clustered by participant.

Dependent Variable:	Choose to pay					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	OLS	OLS	Probit	Probit	Probit
Constant	0.55***			0.12		
	(0.05)			(0.12)		
Overconfidence	0.05**	0.07***	0.05**	0.14**	0.17***	0.14**
	(0.02)	(0.02)	(0.02)	(0.06)	(0.05)	(0.06)
Financial quiz	0.06**	0.07***	0.06**	0.15**	0.19***	0.15**
	(0.02)	(0.02)	(0.02)	(0.06)	(0.05)	(0.06)
College education	0.00	0.00	0.00	0.00	0.01	0.00
	(0.04)	(0.04)	(0.04)	(0.11)	(0.11)	(0.11)
Female	-0.01		-0.01	-0.04		-0.04
	(0.03)		(0.03)	(0.09)		(0.09)
Age	0.02		0.02	0.05		0.05
	(0.02)		(0.02)	(0.04)		(0.04)
Trading experience	0.04		0.04	0.10		0.10
	(0.04)		(0.04)	(0.10)		(0.10)
Risk aversion	-0.02		-0.02	-0.06		-0.06
	(0.02)		(0.02)	(0.04)		(0.04)
Fixed-effects						
Round number		Yes	Yes		Yes	Yes
Observations	6,936	6,936	6,936	6,936	6,936	6,936
Pseudo R^2	0.02	0.02	0.02	0.02	0.02	0.02

Figure 7 illustrates these patterns. Participants with above-median financial literacy are roughly 10 percentage points more likely to purchase data than their below-median counterparts (Panel A). The difference for overconfidence is smaller and not statistically significant in the raw means (Panel B), though it emerges clearly in the regression estimates that control for financial literacy.

Figure 7: **Who Pays for Data?**

This figure illustrates selection into data purchase. Panel (A) plots the share of participants who choose to pay for data, split by whether the participant scored above or below median on the financial literacy quiz. Panel (B) splits by above- or below-median overconfidence. Financially literate participants are approximately 10 percentage points more likely to purchase data; overconfident participants show a smaller, statistically insignificant difference. Error bars represent 95% confidence intervals.



5 Conclusion

The past decade has witnessed an explosion in retail participation in financial markets alongside a parallel rise in the availability of granular market data. The conventional wisdom holds that more information is always better: that is, democratizing access to institutional-grade data empowers retail investors and levels the playing field. Our evidence challenges this view. We show that when retail investors pay for market data, the act of payment itself distorts how they interpret and use that information, leading to systematic mistakes concentrated among the least sophisticated market participants.

Our experimental design isolates the causal effect of paying for data from selection effects by randomly assigning identical information environments to participants who chose to purchase data. We document three main findings. First, payment inflates perceived informativeness by 11%, with participants who pay for data substantially more likely to believe uninformative noise contains predictive content. Second, this belief distortion propagates to forecasts and trading behavior: payment increases forecast sensitivity to order flow by 31% and generates 14% larger forecast errors when data is objectively uninformative. Third, paying for data creates a sunk-cost mechanism that raises investment levels by 7 percentage points while simultaneously reducing the elasticity of investment to beliefs by 15%. Critically, the investment effects are concentrated among participants with low financial literacy, who exhibit nearly 15 percentage point higher investment when paying for data alongside sharply attenuated belief pass-through.

As retail participation in equity markets grows, platform design choices that shape how these investors access and interpret information become increasingly consequential for market quality and investor protection. Unbundling data from trading services may appear to offer consumer choice and flexibility. Our evidence suggests this architecture comes at a cost for less financially literate investors who are not only less able to extract value from complex market data, but are also more vulnerable to the behavioral distortions induced by paying for it.

More broadly, our findings speak to the economics of information in modern financial markets. The standard framework treats information as an input to decision-making whose value depends solely on its precision and the investor's processing capacity. We show that the mechanism of information acquisition: that is, whether data is actively purchased or not, can itself become a source of bias. Paying for information creates a sense of ownership that inflates perceived precision and distorts how the information influences beliefs and actions.

References

Andries, M., Bianchi, M., Huynh, K. K., Pouget, S., 2025. Return predictability, expectations, and investment: Experimental evidence. The Review of Financial Studies 38, 1687–1729.

Arkes, H. R., Blumer, C., 1985. The psychology of sunk cost. Organizational Behavior and Human Decision Processes 35, 124–140.

Arnold, M., Pelster, M., Subrahmanyam, M. G., 2022. Attention triggers and investors' risk-taking. Journal of Financial Economics 143, 846–875.

Barber, B. M., Huang, X., Odean, T., Schwarz, C., 2022. Attention-induced trading and returns: Evidence from robinhood users. The Journal of Finance 77, 3141–3190.

Bartlett, R. P., McCrary, J., O'Hara, M., 2023. The market inside the market: Odd-lot quotes. The Review of Financial Studies 38, 661–711.

Bryzgalova, S., Pavlova, A., Sikorskaya, T., 2023. Retail trading in options and the rise of the big three wholesalers. The Journal of Finance 78, 3465–3514.

Chapkovski, P., Khapko, M., Zoican, M., 2025a. Gamified risk-taking. Journal of Behavioral and Experimental Finance 46, 101049.

Chapkovski, P., Khapko, M., Zoican, M., 2025b. How do retail investors use order flow data? Working paper .

Chapkovski, P., Khapko, M., Zoican, M., 2026. Trading gamification and investor behavior. Management Science 72, 32–56.

Chen, D. L., Schonger, M., Wickens, C., 2016. oTree—an open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance 9, 88–97.

Cupák, A., Fessler, P., Hsu, J. W., Paradowski, P. R., 2020. Confidence, financial literacy and investment in risky assets: Evidence from the survey of consumer finances. Finance and Economics Discussion Series 2020.0.

Eaton, G. W., Green, T. C., Roseman, B. S., Wu, Y., 2022. Retail trader sophistication and stock market quality: Evidence from brokerage outages. Journal of Financial Economics 146, 502–528.

Farboodi, M., Singal, D., Veldkamp, L., Venkateswaran, V., 2024. Valuing financial data. The Review of Financial Studies 38, 938–980.

Fernandes, D., Lynch, J. G., Netemeyer, R. G., 2014. Financial literacy, financial education, and downstream financial behaviors. Management Science 60, 1861–1883.

Fuster, A., Perez-Truglia, R., Wiederholt, M., Zafar, B., 2022. Expectations with endogenous information acquisition: An experimental investigation. Review of Economics and Statistics 104, 1059–1078.

Holt, C. A., Laury, S. K., 2002. Risk aversion and incentive effects. American Economic Review 92, 1644–1655.

Huber, J., Angerer, M., Kirchler, M., 2011. Experimental asset markets with endogenous choice of costly asymmetric information. Experimental Economics 14, 223–240.

Inghelbrecht, K., Tedde, M., 2024. Overconfidence, financial literacy and excessive trading. Journal of Economic Behavior & Organization 219, 152–195.

Kahneman, D., Knetsch, J. L., Thaler, R. H., 1990. Experimental tests of the endowment effect and the coase theorem. Journal of Political Economy 98, 1325–1348.

Kalda, A., Loos, B., Previtero, A., Hackethal, A., 2021. Smart(phone) investing? a within investor-time analysis of new technologies and trading behavior. .

Kaniel, R., Saar, G., Titman, S., 2008. Individual investor trading and stock returns. The Journal of Finance 63, 273–310.

Kelley, E. K., Tetlock, P. C., 2013. How wise are crowds? insights from retail orders and stock returns. The Journal of Finance 68, 1229–1265.

Kwan, A., Philip, R., Shkilko, A., 2024. The conduits of price discovery: A machine learning approach. SSRN Electronic Journal .

List, J. A., 2011. Does market experience eliminate market anomalies? the case of exogenous market experience. American Economic Review 101, 313–317.

Lusardi, A., Mitchell, O., 2011. Financial literacy around the world: An overview .

Merton, R. C., 1969. Lifetime portfolio selection under uncertainty: The continuous-time case on JSTOR. Review of Economics and Statistics 51, 247–257.

Palan, S., Schitter, C., 2018. Prolific.ac—a subject pool for online experiments. Journal of Behavioral and Experimental Finance 17, 22–27.

Prelec, D., Loewenstein, G., 1998. The red and the black: Mental accounting of savings and debt. Marketing Science 17, 4–28.

Staw, B. M., 1976. Knee-deep in the big muddy: a study of escalating commitment to a chosen course of action. Organizational Behavior and Human Performance 16, 27–44.

Thaler, R. H., Johnson, E. J., 1990. Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. Management Science 36, 643–660.

Veldkamp, L., 2023. Valuing data as an asset*. Review of Finance 27, 1545–1562.

Yelagin, E., 2024. Gamification of stock trading: Losers and winners .

Appendix

April 6, 2026

A Financial literacy quiz

1. Suppose you had \$100 in a savings account and the interest rate was 3% per year. After 4 years, how much do you think you would have in the account if you left the money to grow?

   (a) More than \$112

   (b) Exactly \$112

   (c) Less than \$112

   (d) Don't know / Not sure

   (e) Prefer not to say

2. Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, with the money in this account, would you be able to buy

   (a) More than today

   (b) Exactly the same as today

   (c) Less than today

   (d) Don't know / Not sure

   (e) Prefer not to say

3. When placing an order, you may see several prices: the last traded price, the ask price, the bid price, and the closing price. What is the bid price?

   (a) The highest price a buyer is currently willing to pay.

   (b) The price you personally choose when submitting a buy order.

   (c) The lowest price a seller is currently willing to accept.

   (d) The most recent price at which the security traded.

   (e) Don't know / Not sure

   (f) Prefer not to say

4. A 15-year mortgage typically requires higher monthly payments than a 30-year mortgage, but the total interest paid over the life of the loan will be less.

   (a) True

   (b) False

   (c) Don't know / Not sure

   (d) Prefer not to say

5. Normally, which asset described below displays the highest fluctuations over time?

   (a) Savings accounts

   (b) Stocks

   (c) Bonds

   (d) Don't know / Not sure

   (e) Prefer not to say

6. When an investor spreads his money among different assets, does the risk of losing a lot of money:

   (a) Increase

   (b) Decrease

   (c) Stay the same

   (d) Don't know / Not sure

   (e) Prefer not to say

7. Considering a long time period (for example, 10 or 20 years), which asset described below normally gives the highest return?

   (a) Savings accounts

   (b) Stocks

(c) Bonds

(d) Don't know / Not sure

(e) Prefer not to say

8. Do you think that the following statement is true or false? "If you were to invest $1,000 in a stock mutual fund, it would be possible to have less than $1,000 when you withdraw your money."

    (a) True

    (b) False

    (c) Don't know / Not sure

    (d) Prefer not to say

9. When viewing a security's quote, you may see both a bid price and an ask price. What is the bid–ask spread?

    (a) The difference between the lowest ask price and the highest bid price.

    (b) The average of the bid price and the ask price.

    (c) The fee the broker charges each time you trade.

    (d) The change in the stock price from market open to market close.

    (e) Don't know / Not sure

    (f) Prefer not to say

10. Which of the following statements is correct?

    (a) Once one invests in a mutual fund, one cannot withdraw the money in the first year

    (b) Mutual funds can invest in several assets, for example invest in both stocks and bonds

    (c) Mutual funds pay a guaranteed rate of return which depends on their past performance

    (d) None of the above

    (e) Don't know / Not sure

(f) Prefer not to say

11. Which of the following statements is correct? If somebody buys a bond of firm B:

    (a) She owns a part of firm B

    (b) She has lent money to firm B

    (c) She is liable for the company's debts

    (d) None of the above

    (e) Don't know / Not sure

    (f) Prefer not to say

12. Suppose you owe \$3,000 on your credit card. You pay a minimum payment of \$30 each month. At an annual percentage rate of 12% (or 1% per month), how many years would it take to eliminate your credit card debt if you made no additional new charges?

    (a) Less than 5 years

    (b) Between 5 and 10 years

    (c) Between 10 and 15 years

    (d) Never

    (e) None of the above

    (f) Don't know / Not sure

    (g) Prefer not to say

B Consent form

Dear Participant,

You are invited to take part in a research study on financial decision-making funded by the Social Sciences and Humanities Research Council of Canada. The purpose of this study is to understand how people use data when making investment choices.

Your participation in this study is entirely voluntary. Participation presents no known risks or burdens. You will receive a compensation for participating in the study, announced on the Prolific platform, as well as a monetary compensation proportional to your performance in the experiment. You can withdraw from the study at any time without penalty. However, if you withdraw before the end of the experiment the compensation will be forfeit, as per Prolific's rules.

Please be assured that all data generated during this study will remain confidential. You will be asked to provide some basic demographic information about yourself, but none of this information will in any way be used to personally identify you and no personal information will appear in any published study.

Your decision to participate in this study will remain completely confidential and only the researchers will have access to any of the individual data. We do not collect any data that can personally identify you. Any demographics data will be stored electronically on a web-based server until it is downloaded for the purpose of analysis. This server will only be accessible by the research team using a secure password. Data will be retained in de-identifiable format for 7 years and discarded afterwards. At that time, electronic files will be deleted and written files will be shredded. This is a standard practice for the American Psychological Association. As required by the guidelines of American Economic Association, during the publication process the data might be shared with other investigators for re-analysis. However, to ensure security and confidentiality of the data files only details of the computations sufficient to permit replication might be shared.

This study has been reviewed by the Research Ethics Boards at the University of Calgary. If you are interested, you can receive a summary of the research study by emailing the researchers.

The research study you are participating in may be reviewed for quality assurance to make sure that the required laws and guidelines are followed. If chosen, (a) representative(s) of the Human Research Ethics Program (HREP) may access study-related data and/or consent materials as part of the review. All information accessed by the HREP will be upheld to the same level of confidentiality that has been stated by the research team.

If you wish to obtain further information regarding your rights as a participant you may contact the Conjoint Faculties Research Ethics Board (CFREB) at University of Calgary at cfreb@ucalgary.ca or (403) 220-8640. You may also contact the researchers for further information or to obtain answers to any questions about this research.

I agree to participate in research being conducted by the researchers.

I understand that I can withdraw my consent to participate at any time and by giving consent I am not giving up any of my legal rights. I have read and understood the information above.

C Instructions

Welcome to this trading experiment! Please read these instructions carefully.

Overview. In each round, you will see a graph showing past data for a traded stock, measured over half-hour intervals within a single trading day. The graph displays:

1. **The stock return**, expressed in percentage terms; and
2. **The buy/sell imbalance**, a measure of order flow that captures buying versus selling pressure, expressed in percentage terms. It is positive when buy order flow exceeds sell order flow, and negative when sell order flow exceeds buy order flow.

In some rounds, the buy/sell imbalance is informative and helps predict the next stock return. In other rounds, the two variables are independent, and the imbalance does *not* help predict returns.

Your task. In each round, you are endowed with 100 experimental dollars (E$). Your tasks are:

1. [Treatment only] At the start of each round, you choose whether to purchase access to the buy/sell imbalance data for E$5. After you make your choice, the computer randomly determines the round type:
   - **Paid data round** (probability 2/3): Your choice is binding. If you chose to purchase data, E$5 is deducted from your cash account and you see the imbalance data. If you declined, you pay nothing but do not see the imbalance data.
   - **Free data round** (probability 1/3): You see the imbalance data at no cost, regardless of your initial choice.

2. [Treatment Only] If you see order imbalance data, assess whether the imbalance can be used to predict the next stock return.

3. Provide a forecast for the next stock return.

4. Choose how much of your endowment you would like to invest in the stock.

There are twelve main rounds in the experiment. All rounds are independent, and the average stock return is 15%. Before the main rounds begin, you will complete two practice rounds to familiarize yourself with the interface.

After each round, you will see the realized stock return, whether the buy/sell order flow imbalance was actually informative, the accuracy of your forecast, and the total wealth you earned in that round.

Your payoff. Your final payoff comprises four parts:

1. The value of your portfolio (stocks plus cash) plus any bonus earned, from one randomly selected round.

2. [Treatment only] E$1 for every correct assessment of whether the order-flow imbalance can predict returns.

3. E$1 for every precise forecast (within ±1% of the actual stock return).

4. E$0.25 for each correct answer in a financial literacy quiz which you will complete after the main rounds.

At the end of the experiment, one of the twelve main rounds will be randomly selected to determine your investment payoff (component 1). Your total payoff in ECU is the sum of (1) from the selected round, plus [Treatment only] (2) and (3) accumulated across all rounds, plus (4) from the financial literacy quiz. Your final payment in pounds is your payoff in ECU divided by 20.

D Comprehension quiz

1. (Treatment only) If you choose to purchase data and the round is a **paid data round**, what happens?

(a) You see the imbalance data and E$5 is deducted from your balance.

(b) You see the imbalance data at no cost.

(c) You do not see the imbalance data and E$5 is deducted.

(d) You receive E$5 as a bonus.

2. (Treatment only) If you *decline* to purchase data and the round is a **free data round**, what happens?

   (a) You do not see the imbalance data.

   (b) You see the imbalance data and E$5 is deducted from your balance.

   (c) You see the imbalance data at no cost.

   (d) You must pay E$8 to see the data.

3. Is the buy/sell order flow imbalance always useful for predicting the next stock return?

   (a) Yes, it is always informative.

   (b) No, it is never informative.

   (c) It is informative in some rounds and uninformative in others.

   (d) It depends on your investment choice.

4. When do you earn the forecast bonus?

   (a) When your forecast is exactly equal to the realized stock return.

   (b) When your forecast is within $\pm 1\%$ of the realized stock return.

   (c) When your forecast is within $\pm 5\%$ of the realized stock return.

   (d) When your forecast is equal to the average stock return of 15%.

5. What happens to the portion of your endowment that you *do not* invest in the stock?

   (a) It earns interest.

   (b) It is lost.

(c) It remains in cash and stays unchanged.

(d) It is automatically invested for you.

6. Some participants skim instructions. To show you read this, pick the third option below.

   (a) I skimmed this question.

   (b) I did not read the instructions above.

   (c) I read this and am selecting the third option as instructed.

   (d) I prefer not to answer.

7. How is the portfolio return paid?

   (a) All 12 main days are paid in full.

   (b) One paid day is randomly selected; bonuses accumulate across paid days.

   (c) Only the last day is paid.

   (d) Only training days are paid.