Understanding Retail Investors: Evidence from China

Charles M. Jones, Donghui Shi, Xiaoyan Zhang and Xinran Zhang*

First version: September 2019
This version: May 2022

Abstract

Using comprehensive account-level data from 2016 to 2019, we examine retail investor trading behavior in the Chinese stock market. We separate millions of retail investors into five groups by their account sizes and document strong heterogeneity in their trading dynamics and performance. Retail investors with smaller account sizes cannot predict future price movements correctly, in the sense that they buy future losers and sell future winners. Their trading patterns are momentum over daily horizons, but become contrarian over weekly horizons. These investors fail to process public news and display behavioral properties such as overconfidence and gambling preferences. In sharp contrast, retail investors with larger account balances predict future returns correctly, display contrarian trading patterns, incorporate public news in their trading, and their return predictive power are stronger in stocks which are more attractive to investors with behavioral biases.

Keywords: Retail investors, Chinese stock market, Return predictability, Information content.
JEL code: G12, G14, G15

* Charles M. Jones is with Columbia Business School, Donghui Shi is with Fanhai International School of Finance, Fudan University, Xiaoyan Zhang is with PBC School of Finance, Tsinghua University, and Xinran Zhang is with School of Finance, Central University of Finance and Economics. Xiaoyan Zhang acknowledges financial support from the National Natural Science Foundation of China [Grant 71790605]. We thank Laruen Cohen, Ron Kaniel, Zhiguo He, Hao Zhou, Utpal Bhattacharya, and seminar participants at Tsinghua PBC School of Finance, Renmin University, Shanghai Jiaotong University, Fudan University, Shanghai University of Finance and Economics, and conference audiences at the 2019 CIFFP, 2021 CFRC, 2021 CICF for helpful comments and suggestions. All remaining errors are our own. Corresponding author: Xiaoyan Zhang, PBC School of Finance, 43 Chengfu Road, Beijing, China, 100083, zhangxiaoyan@pbcsf.tsinghua.edu.cn.

Understanding Retail Investors: Evidence from China

First version: September 2019
This version: May 2022

Abstract

Using comprehensive account-level data from 2016 to 2019, we examine retail investor trading behavior in the Chinese stock market. We separate millions of retail investors into five groups by their account sizes and document strong heterogeneity in their trading dynamics and performance. Retail investors with smaller account sizes cannot predict future price movements correctly, in the sense that they buy future losers and sell future winners. Their trading patterns are momentum over daily horizons, but become contrarian over weekly horizons. These investors fail to process public news and display behavioral properties such as overconfidence and gambling preferences. In sharp contrast, retail investors with larger account balances predict future returns correctly, display contrarian trading patterns, incorporate public news in their trading, and their return predictive power are stronger in stocks which are more attractive to investors with behavioral biases.

Keywords: Retail investors, Chinese stock market, Return predictability, Information content.
JEL code: G12, G14, G15

Retail investors are important participants in financial markets, and many studies are devoted to understanding their trading motives, their performances, and their roles in information transmission and price discovery. However, these studies provide seemly conflicting results. For instance, Barber and Odean (2000, 2001, 2008) document behavioral biases exhibited by retail investors, such as over-confidence and overtrading, and as a result, retail investors make sub-optimal investment choices. Later studies, such as Barber, Odean, and Zhu (2008), Kaniel et al. (2008), Kelley and Tetlock (2013), and Boehmer et al. (2021), suggest that retail investors correctly predict future stock returns and trade accordingly, which indicates that retail investors might know something about future stock price movements. Most recently, interests have shifted to a new generation of retail investors, who trade at zero-commission trading platforms such as Robinhood. Barber et al. (2021), Eaton et al. (2021) and Welch (2021) find that Robinhood investors perform well, demand liquidity and engage more in attention-induced trading. How can we reconcile the conflicting results from previous studies? One possibility is that retail investors are not born equal, so the above-mentioned empirical results could be dominated by subgroups of retail investors. However, due to data limitations, few previous studies directly examine the heterogeneity of retail investors.

China's equity market, the second largest in the world, provides an ideal setting for studying retail investors and their heterogeneity. According to the annual report of the Shanghai Stock Exchange, retail investors contribute 85% of daily trading volumes on the exchange, while

institutional investors only contribute 15%. The dominance of retail trading in this market clearly brings retail investors to the center stage. Behind the high trading volumes are tens of millions of retail investors in China, accounting for the largest population of retail investors in the global capital market. Given the dominant role and the large population of Chinese retail investors, it is crucial for researchers, regulators and practitioners to understand Chinese retail investors' investment choices and how these choices affect information transmission and price discovery.

With account-level data from one main stock exchange, we examine the rich cross-section of retail investors in China, which greatly helps us to investigate the heterogeneity of retail investors and how their trading interacts with stock returns and information flows. We obtain account-level trading and holdings data from 2016 to 2019 for over 53 million retail accounts. To comply with regulatory requirements, all Chinese retail accounts are categorized into five groups by account balances: less than 100,000 CNY (RT1), between 100,000 and 500,000 CNY (RT2), between 500,000 and 3,000,000 CNY (RT3), between 3,000,000 and 10,000,000 CNY (RT4), and greater than 10,000,000 CNY (RT5). The five groups account for 58.7%, 28.6%, 10.9%, 1.4% and 0.4% of total number of accounts, respectively. With additional gender and age information, we find the majority of Chinese retail investors are young or middle-aged males with account sizes below 500k CNY.

With this rich cross section of retail investor data, we first examine whether (some) retail investors are informed about future price movements, in the sense that whether their buy and sell

activities can predict future stock returns. If the market is perfectly efficient, stock prices would follow random walks, and trading would not predict future returns. If the market is not perfectly efficient, and if some investors have value-relevant information for future stock prices, their order flows would positively predict future returns. On the other hand, if some investors have information disadvantage or fail to incorporate timely information into their trades, their order flows might negatively predict future returns. Using daily retail order imbalances from each retail group, we predict future stock returns at horizons ranging from one to 60 days. The smaller retail investors, RT1-RT4, predict next-day returns with negative coefficients. That is, the prices of stocks they buy experience negative returns the next day, while the ones they sell experience positive returns. In contrast, the largest retail investors, RT5, positively predict next-day returns, indicating that they buy and sell stocks in directions consistent with future price movements. When we look at longer horizons, the above-mentioned predictive patterns persist for at least 12 weeks. These patterns are also quite robust when we form long-short strategies on order flow information, and for subsets of stocks with differences in size, value, liquidity and share price levels.

Previous literature provides multiple explanations for the trading motives for retail investors, such as order flow persistence, liquidity provision, behavioral biases and information (dis)advantages. These explanations also naturally connect with the predictive pattern of retail order flows for future returns. We adopt the two-stage decomposition procedure in Boehmer et al. (2021) to examine whether these hypotheses can explain the trading activities of different retail

investor groups, and how these trading activities contribute to the predictive patterns for future returns.

Our results show that order flows from all retail investors display persistence. Order flows from smaller retail investors show momentum patterns at a daily horizon and demand immediate liquidity. The smaller retail investors also display significant behavioral biases, such as over-confidence and gambling preferences, and they fail to predict and process earnings news. On the contrary, the largest retail investors display contrarian trading patterns; they trade against the behavioral biases of the other retail groups and are capable of predicting and processing earnings news. In explaining order flow's predictive power for future returns, order persistence, daily momentum trading, behavioral biases and information disadvantages all contribute to the negative predictive power of smaller retail investors, while contrarian trading, trading against behavioral biases and information skills contribute to the positive predictive power of the largest retail investors.

We also investigate other dimensions of the data and conduct several robustness checks. We find that male investors across all ages negatively predict returns, especially the older ones. These findings are generally in line with Barber and Odean (2001). Results from all other robustness checks are consistent with the main results.

Our study is closely related to the retail investor literature. Previous studies on retail investors mostly use data from the U.S. and other markets, and they mostly treat retail investors as one group.

For instance, using data from a discount broker in the U.S., Barber and Odean (2000, 2001, 2008) document many behavioral biases; Kaniel, Saar and Titman (2008), Barber, Odean, Zhu (2009), Kelley and Tetlock (2013), and Boehmer et al. (2021) use different datasets from the U.S. and find that retail trading can positively predict the cross-section of future returns; and Barber, Lin, and Odean (2021) explain why U.S. retail investors lose money despite their predictive power for stock returns. Recently, Barber et al. (2021), Eaton et al. (2021) and Welch (2021) study the trading behavior of Robinhood retail investors in the U.S. Outside of the U.S., Grinblatt and Keloharju (2000), Linnainmaa (2010), Grinblatt, Keloharju, and Linnainmaa (2012) focus on the Finland data; Bach, Calvet, and Sodinish (2020) focus on the Sweden data; Dorn, Huberman, Sengmueller (2008) study data from Germany; Barrot, Kaniel and Sraer (2016) study data from France; Fong, Gallagher, and Lee (2014) study data from Australia; Barber, Lee, Liu and Odean (2009) examine Taiwan data; and Balasubramaniam, Campbell, Ramadorai and Ranish (2021) make use of Indian data. All these papers provide important results regarding retail investors' trading activities.[1]

Our study is also related to studies on the rapidly growing Chinese stock market. Liu, Stambaugh and Yuan (2019), and Liu, Zhou, and Zhu (2021) establish asset pricing factors for stock returns. For Chinese retail investors, An, Lou and Shi (2022) study the wealth redistribution role of financial bubbles and crashes over July 2014 and December 2015, and they document a net transfer of 250 billion CNY from the poor to ultra-wealthy retail investors over this period. Liu,

[1] Please see the Appendix for the literature review of the studies on retail investors in different markets.

Peng, Xiong and Xiong (2021) and Liao, Peng and Zhu (2021) both focus on behavioral properties of Chinese retail investors and document overconfidence, gambling preferences and extrapolative expectations in these investors. Li et al. (2017), Titman et al. (2022), Hu et al. (2021) and Jiang et al. (2020) and other papers[2] focus on an earlier Chinese sample period and examine behavioral biases and reactions to corporate events.

The above studies mostly rely on low frequency data, or data from brokerage which covers a small part of the market, or investigate issues other than return predictability. As a result, there is still no direct study on the heterogeneity of retail investors' trading behavior, their return predictive power, and how they process information, using high frequency trading data in one major stock market where retail investors dominate. Therefore, our study makes two important contributions to the literature. First, we separate retail investors into groups based on account sizes and provide unique, and direct evidence on investor heterogeneity in terms of return predictability. Second, we examine different hypotheses for the return prediction patterns for different retail investor groups, and we provide clear evidence on the sources of the negative or positive predictive power of different retail investors. Our study, with its large coverage of the market for a recent sample period, is one of the most thorough and comprehensive studies for of Chinese retail investors, and it provides many important implications for regulators, practitioners, and academic researchers.

I. Data

[2] These papers include Li, Geng, Subrahmanyam and Yu (2017), Chen, Gao, He, Jiang and Xiong (2019), Jiang, Liu, Peng, and Wang (2020), Titman, Wei, and Zhao (2022), and Hu, Liu and Xu (2021).

A. Data on Stock Returns and Firm Characteristics

We obtain data on stock returns, volumes, and accounting information from Wind Information Inc. (WIND), the largest financial data provider in China. To be consistent with our retail data, our sample period runs from January 2016 to June 2019. We adopt the filters in Liu, Stambaugh and Yuan (2019) and exclude stocks with less than 15 days of trading records during the most recent month. Liu, Stambaugh and Yuan (2019) also eliminate stocks that have become public within the past six months, stocks with fewer than 120 days of trading records during the past 12 months, and the smallest 30% of total firms listed in the Chinese A-share market. We do not exclude these stocks for the main results, because retail investors trade actively in small stocks and during the IPO period. We present the results with all filters from Liu et al. (2019) in our robustness checks, and our findings are almost the same with the additional filters. Starting from March 31, 2010, margin buying and short selling are allowed on Chinese stock exchanges for subsets of stocks. We include these leveraged trades in our main results, and provide additional analysis excluding leveraged trading in our robustness checks. Our sample covers over 1.1 million stock-day observations, and on each day, we have an average of around 1,200 firms.

We present summary statistics on our sample firms in Panel A of Table 1. Daily stock returns are calculated using closing prices, which are dividend and split adjusted.[3] The average daily stock

[3] Previous literature using the U.S. data shows that microstructure frictions can generate noise in daily return measures. For instance, Blume and Stambaugh (1986) show that daily returns computed from the end-of-day closing prices can have an upward bias due to bid-ask bounce. To assess the potential magnitude of the bias, they measure the bias as $\left(\frac{P_A-P_B}{P_A+P_B}\right)^2$, where P_A and P_B are closing ask and bid prices. Blume and Stambaugh (1986) find that the average bias

return, *Ret*, is -0.01% for Chinese stocks, while the average daily stock return is 0.04% in the U.S stock market over the same sample period. Market capitalization, *Size*, is computed as the product of the previous month's closing price and total A shares outstanding. The average Chinese firm capitalization is 20.1 billion CNY or 3 billion USD, about half of the cross-sectional average in the U.S stock market during the same period, which is 6.9 billion USD. The earnings to price ratio, *EP*, is computed as the ratio of the most recently reported quarterly net profit excluding non-recurrent gains/losses over last month-end's market capitalization. According to Liu, Stambaugh and Yuan (2019), the EP ratio captures the value effect. The average EP ratio is 0.0075 in China, while the average EP ratio is 0.0272 in U.S stock market. This difference may be driven by high valuations in China. Finally, monthly turnover is calculated as monthly share trading volume divided by tradable shares outstanding at the end of the previous month. The average monthly turnover in China is 48.32%, which is much larger than the monthly turnover of 22% in the U.S. during the same period.

B. Data on Retail Investors

We obtain investors' daily trading and holding data of all A-share stocks listed on the Shanghai Stock Exchange between January 2016 and June 2019. Out of the two stock exchanges in China,

for small stocks is 0.051%, and for large stocks, the bias is 0.001%, which are sizable magnitudes for daily returns averaging at less than 1%. Therefore, they recommend using closing bid-ask average prices to compute daily returns. We compute this bias measure using the closing bid and ask prices for all A-share stocks listed on the SHSE. The average bias measure is generally below 0.0002% across all stocks, which is negligible compared to the bias computed in Blume and Stambaugh (1983). Therefore, we compute daily returns using daily close prices without the Blume and Stambaugh (1983) adjustments.

the Shanghai Stock Exchange (SHSE) and the Shenzhen Stock Exchange (SZSE), the former accounts for 60% of the total market capitalization in China and thus is a reasonable representation of the overall Chinese stock market.[4] Our data contains roughly 53 million accounts, and based on investor identities, they are first grouped into three major categories: retail (RT), institutional (INST), and corporations (CORP). Retail investors are further stratified into five groups based on their account sizes, which is the average portfolio value (including equity holdings in both SHSE and SZSE-listed firms, plus cash) over the previous twelve months. As mentioned in the introduction, there are five subgroups: below 100,000 CNY (RT1), 100,000-500,000 CNY (RT2), 500,000 - 3 million CNY (RT3), 3 million - 10 million CNY (RT4), and above 10 million CNY (RT5). Since our focus in this study is how retail trades are related to stock prices in the cross section, we sum up individual investors' trading information at seven investor group level (RT1-RT5, INST and CORP) for each stock each day.

We merge the exchange data and WIND data by stock ticker and present account summary statistics in Panel B of Table 1. During our sample period, the total number of active accounts for retail investors, institutions and corporations are 53.4 million, 40,000 and 47,000, respectively. Within the retail investor category, there are 31.4 million, 15.3 million, 5.8 million, 0.7 million and

[4] In June 2019, there are 1,471 A-share stocks listed on the SHSE, with a total market capitalization of $ 4.6 trillion. In comparison, 2,157 A-share stocks are listed on the SZSE, with a total market capitalization of $ 3 trillion. The Science and Technology Innovation Board (or STAR Market) was launched on the SHSE on July 22, 2019, and thus is not included in our study. Given the data accessibility, our results only cover Shanghai stock exchange, but not Shenzhen stock exchange (SZSE). Chen et al. (2019) use Shenzhen stock exchange data to examine retail investors trading around price limits events, and find consistent patterns to those in our study, which suggest that the retail investors at SZSE likely behave similarly to those at SHSE.

0.2 million accounts for RT1 to RT5. Clearly, most of the retail investors have accounts less than 500,000 CNY. The overall trading volume on the SHSE averages 201 billion CNY per day, and retail investors, institutions and corporations account for 81%, 17% and 2% of the total trading volume, respectively. Within the retail investor sector, trading volumes for RT1 to RT5 are 5%, 17%, 27%, 13% and 19% of the total trading volume, which is more evenly distributed than the numbers of accounts. For stock holdings, retail investors' holdings account for 22%, institutions 17% and corporations 62%. Within the retail investor sector, the account values for RT1 to RT5 are 1%, 4%, 6%, 3% and 7% of the total tradable market cap.

To understand the relative importance of different investment groups' trading over time, we plot the time series of cross sectional means of various investors' trading activity in Figure I. Panel A presents each group's trading volume as a percentage of total trading volume. The RT3 group has the highest trading volume, accounting for about 30% of total trading. Interestingly, institutional trading gradually increases over time, from 10% in 2016 to over 20% in 2019. The corporations barely trade and account for a negligible amount of trading volume. Panel B displays the shares held percentage by each group, and the time-series patterns in holdings are quite stable. Overall, in the Chinese stock market retail investors dominate in terms of trading, while corporations dominate in holdings. The retail investors' dominance in trading of Chinese stock market is likely the joint result of market development, regulations and investor preference. This pattern is not particularly rare for emerging markets, but it is quite different from that of developed

markets. This particular dominance in trading by retail investors also renders more relevance and significance of our study.

Finally, to have a rough idea about holding horizons, we make a simple assumption that shareholders within same investor group have identical holding horizons. Then we compute the holding period for stock *i*, type *G* investors as $1/TO_{i,G}$, where $TO_{i,G}$ is the turnover (shares traded/shares held by this type of investor) of stock *i* for type *G* investors. For example, if 1% of the shares trade each day, then it takes 100 days for the entire stock of tradable shares held by this group to turn over, and the average holding period would be 100 days. In the last row of Panel B, the average holding period for the five groups of retail investors ranges from 35 days to 50 days, reflecting their active trading and short holding horizons. Institutional holding periods in our sample are much longer at 109 trading days. Corporations barely trade in our sample period, and their estimated holding period is 6,319 trading days. In comparison, the market overall monthly turnover in the U.S. over the same period is 22%, indicating a holding period of 1/0.22 = 4.5 months, which is about 90 days.[5] These dramatic differences in holding horizons suggest that different types of Chinese investors might have quite different trading patterns and trading preferences.[6]

[5] We would also like to mention that in the U.S., there is a large amount of high frequency trading, including establishing and closing positions on the same day. China adopts the "*T+1* trading rule", which requires that if stocks are bought on day *T*, they cannot be sold on the same day. The reverse trade has to be executed on day *T+1* or later. That is, there is essentially a minimum holding period of one day.

[6] We present additional summary statistics on retail trading volumes and holdings in Appendix Table 1. The results show that small retail investors prefer to trade and hold small, low earning/price ratio, and high turnover firms, while the largest retail investors trading and holding are tilted towards larger, high EP firms. In terms of sectors, the small

We measure order flows from different groups of investors' using order imbalance measures, as in Chordia and Subrahmanyam (2004). For stock *i,* day *d*, and investor group *G*, we compute

$$Oib(i,d,G) = \frac{\sum_{j\in G} BuyVol(i,d,j) - \sum_{j\in G} SellVol(i,d,j)}{\sum_{j\in G} BuyVol(i,d,j) + \sum_{j\in G} SellVol(i,d,j)}, \quad (1)$$

where the numerator is the difference between buy and sell volumes summed up over all individual *j*'s within group *G*, and the denominator is the sum of buy and sell volumes of all individuals in group *G*. The order imbalance measure is an order flow measure, and we directly observe each trade's direction from the data. When a set of investors buys more than they sell, the order imbalance is positive, and vice versa. We compute the order imbalance measure for each investor group as *OibRT*1 to *OibRT*5, *OibINST* and *OibCORP*. The overall retail order imbalance measure, *OibRT*, is calculated by summing up all trades within the five retail groups.

Table I Panel C reports summary statistics for the order imbalance measures. The average order imbalance for RT1 to RT5, institutions and corporations are -0.021, -0.011, -0.006, 0.002, 0.019, -0.003, -0.011 and -0.004, respectively.[7] The small magnitude of these average order imbalance measures indicates that most buys and sells within each investor group cancel out each other. The standard deviations of order imbalances are larger for large retail investors and institutions compared to small and medium retail investors, indicating that there is more cross-stock variation

retail investors prefer to trade and hold the alternative energy sector, and prefer not to trade and hold banks and life insurance firms, while the institutions and corporates behave in opposite ways. Finally, small retail investors tend to use small order sizes, large retail investors tend to use large orders sizes, while institutions use all order sizes.

[7] We plot the time-series of the cross-sectional mean, median and 25th and 75th percentiles of different types of investors' order imbalance in Appendix Figure I. There are no obvious time trends or structural breaks in the time series observations.

in large retail investor and institutional trading activity. The one-day autocorrelation coefficient, AR1, for these *Oib* measures are 0.243, 0.259, 0.216, 0.059 and 0.102 for RT1 to RT5, suggesting that small and medium retail order imbalances are generally more persistent than large retail imbalances.

In terms of order flow correlations across the seven groups, order flows from smaller retail investors, *OibRT*1, *OibRT*2 and *OibRT*3, are highly correlated, with correlation coefficients mostly higher than 0.60. *OibRT*4 is still positively correlated with *OibRT*1-*OibRT*3, but with a much lower correlation of 0.20. The largest retail investors' order imbalance, *OibRT*5, is negatively correlated with all four other groups, with correlations around -0.15, indicating that this group of retail investors might have different trading patterns from the others.[8] Institutional order imbalances are negatively correlated with all five retail groups, with correlations ranging from -0.380 to -0.188, again implying different trading patterns from retail investors, even the largest retail investors. As we saw earlier, corporations barely trade and their correlations with the rest of the investor categories are all lower than 10%.[9]

In this data section, we include *OibINST* and *OibCORP* for the completeness of the summary statistics. Corporations are long-term investors and rarely trade, while our study focuses on trading behavior, so we also drop corporations from the remaining empirical results. In terms of

[8] In addition to the cross correlation analysis, we also estimate a VAR specification for the oib's from different group of investors. Results are similar and are available on request.

[9] Appendix Figure 1 present the time-series plot of the order imbalance measures for each investor group, and we find no evidence of time trends or breaks.

institutional investors, given that retail investors are commonly assumed to be less sophisticated than institutional investors, we keep institutions in our main empirical results for comparison purposes.

II. Can Retail Order Flows Predict Future Stock Returns?

Can retail investors' activity predict future stock returns in China? If they can, it is possible that these retail investors trading may contain information about future stock price movements. We start by investigating the whether retail investors could predict future short term and the long term returns with Fama-MacBeth regressions in Section II.A and II.B, respectively. In Section II.C, we examine the predictive patterns for different subsets of stocks based on firm and stock characteristics.

A. *Predicting Next Day Stock Returns Using Retail Order Flows*

To investigate the roles different retail investors play in the price discovery process, we first examine the predictive power of various order flow variables for next-day returns using the two-stage Fama-MacBeth regression. For the first stage, we estimate the following cross-sectional regression for each day *d*,

$$Ret(i,d) = a0(d) + a1(d)Oib(i,d-1) + a2(d)'Controls(i,d-1) + u1(i,d), \qquad (2)$$

where the dependent variable $Ret(i,d)$ is the stock return for firm *i* on day *d*, and the independent variables include order imbalance measures from the previous day, $Oib(i,d-1)$, and control variables, $Controls(i,d-1)$. We follow previous literature for the choices of control variables.

To control for potential momentum/reversal from past returns, we include returns from the previous day, *Ret*(-1), returns from the previous week, *Ret*(-6,-2), and returns from the previous month, *Ret*(-27, -7). For size, value and liquidity effects, we include log market size (*Size*), earnings-to-price ratio (*EP*), and turnover, all computed from the previous month-end.

From the first stage estimation, we obtain a daily time-series of coefficients, $\{a0(d), a1(d), a2(d)'\}$. For the second stage estimation, we conduct statistical inference based on the mean and standard errors of the first stage coefficients, and we compute Newey-West standard errors with 5 lags, which is the optimal lag number using a Bayesian Information Criterion (BIC). If the order flow variable from a specific investor group predicts future returns in the right direction, in the sense that more past purchases are associated with higher future returns, and more past sales are associated with lower future returns, we expect the coefficient $a1$ to be significantly positive, and vice versa. The set-up of other Fama-MacBeth regressions in this study are similar to this benchmark case. Therefore, we omit the similar details when introducing the other specifications.

The estimation results for equation (2) are reported in Panel A of Table 2, which displays distinctive predictive patterns across different groups of retail investors. For the smallest retail investor group, RT1, the coefficient on retail order flow variable is -0.0093, with a significant *t*-statistic of -24.98. The negative coefficient shows that if retail investors RT1 buy more than they sell on a given day, the next day return on that stock is significantly negative. To understand the economic magnitude of the coefficients, we report the inter-quartile range for *OibRT*1 at the bottom.

Multiplying the interquartile range, 0.2222, by the regression coefficient of -0.0093 generates an interquartile daily return difference of -21 basis points (more than 50% annualized!). For retail investors in groups RT2 to RT4, the predictive patterns are qualitatively similar. All coefficients are negative and statistically significant, and the daily interquartile return differences are -17, -11, and -2 basis points for RT2, RT3 and RT4, respectively. That is to say, the first four groups of investors all trade in the wrong direction vs. future price movements. Interestingly, when we move from the smaller account sizes to the larger ones, the negative coefficients become smaller, indicating that larger retail investors trade less incorrectly than smaller retail investors.

Indeed, for the largest retail investors, RT5, the coefficient on past day order imbalance is 0.0012, which is positive and significant with a *t*-statistic of 12.26. The interquartile daily return difference now is 5 basis points per day (over 12% per year). It seems that the largest retail investors' trading predicts the cross-section of future stock price movements in the correct direction.

As a comparison, the coefficient on the previous day order imbalance is 0.0016 for institutions, with a *t*-statistic of 20.34. That is to say, institutional order flows predict future stock price movements in the right direction, and the interquartile return difference is 10 basis points per day, about twice the magnitude of the RT5 estimate. This finding is consistent with many previous studies that institutional investors are more informed than retail investor in general.[10]

[10] Table 1 Panel C shows a negative correlation of -0.188 between OibRT5 and OibINST. Readers might find it confusing that both OibRT5 and OibINST positively predict returns while they have a negative correlation coefficient.

For the control variables, the coefficients on previous day return have mixed signs, while the coefficients on previous week and previous month returns are all negative and significant, indicating strong reversals over weekly and monthly horizons. Size is mostly insignificant, while the earnings-to-price variables are always positive and significant, indicating a strong value effect. The coefficients on turnover are always negative and significant, suggesting that higher turnover leads to lower returns in the future. The above findings are mostly consistent with previous studies of the Chinese stock market, such as Liu, Stambaugh and Yuan (2019). These results also confirm that the predictive power of various order flow variables for future stock returns is not a manifestation of size, value, liquidity or momentum/reversal effect.

B. Predicting Long Term Stock Returns Using Retail Order Flows

The exercise in Section II.A focuses on next-day return prediction. It is natural to ask whether the predictive patterns carry on for longer terms. If the predictive pattern quickly vanishes or reverses, what we observe might be driven by short-term noise. If the predictive pattern persists over longer horizons, it is more likely the return predictability is linked to firm fundamentals or persistent biases. Therefore, we extend the Fama-MacBeth specification in equation (2) to longer horizons up to 12 weeks:

$$Ret(i,w) = b0(w) + b1(w)Oib(i,d-1) + b2(w)'Controls(i,d-1) + u2(i,w). \qquad (3)$$

In our opinion, both OibRT5 and OibINST contain positive information for future returns, which leads to the positive predictive coefficients. However, the information contained in OibRT5 and OibINST are probably different, and the trading of RT5 and INST might be quite distinctive, which leads to the negative correlation between the two.

That is, we use previous day order imbalance, $Oib(i, d-1)$, to predict the cumulative returns over the next *w*-weeks. To be more specific, $Ret(i, w)$ is calculated as a cumulative return from day *d+1* to the end of week *w*, where w=1, …, 12. For instance, when *w*=1, $Ret(i, w)$ is the cumulative return over day *d*+1 to *d*+5; when *w*=12, $Ret(i, w)$ is the cumulative return over day *d*+1 to *d*+60. If order imbalances have only short-lived predictive power for future returns, we should observe the coefficient $b1$ decrease to zero quickly when *w* increases or even reverses. Alternatively, if the specified retail order imbalance has longer predictive power, the coefficient $b1$ should remain statistically significant for a longer period.

We present the estimates of coefficient $b1$ in equation (3) in Table II Panel B. To save space, we only report the coefficients and the statistical significance level by asterisks, with ***, ** and * indicating significance at 1%, 5% and 10% level, respectively. For the smallest retail investors RT1, the coefficient on *OibRT*1 monotonically increases from -0.0226 at week one to -0.0458 for a 12-week horizon, and all coefficients are statistically significant. Same patterns are also observed for *OibRT*2, *OibRT*3, and *OibRT*4. The positive predictive power of *OibRT*5 and *OibINST* also persist significantly for at least 12 weeks, and there are no obvious reversals. The persistence of cross-sectional predictability indicates that the predictive power is likely rooted in information related to fundamentals or from persistent noise trading or behavioral biases.

C. Predicting Patterns Across Firms with Different Characteristics

Previous studies show that stock returns can be significantly affected by firm and stock characteristics, such as size, EP ratio and liquidity. Do predictive patterns of retail order flows differ across firms with different characteristics? To answer this question, we modify the Fama-MacBeth specification in equation (2) and allow different coefficients for firms with different characteristics, by including interactions with characteristics dummies as follows,

$$Ret(i,d) = c0(d) + [c1(d)Dummy1(i,d-1) + c2(d)Dummy2(i,d-1) + c3(d)Dummy3(i,d-1)]Oib(i,d-1) + c4(d)'Controls(i,d-1) + u3(i,d). \quad (4)$$

Take size as an example. We first separate all firms on day *d* into three groups, based on previous month-end firm market capitalization. The dummy variable, $Dummy1(i,d-1)$, takes value 1 if firm *i* belongs to the smallest 1/3 of firms, zero otherwise; $Dummy2(i,d-1)$ takes value 1 if firm *i* belongs to the medium 1/3 of firms, zero otherwise; and $Dummy3(i,d-1)$ takes value 1 if firm *i* belongs to the largest 1/3 of firms, zero otherwise. The coefficients $c1, c2\ and\ c3$ provide information on whether the predictive pattern changes for firms with different sizes.

Estimation results for equation (4) are reported in Table III. In the first three rows, we separate firms by their market capitalization. The negative predictive pattern of order flow from RT1-RT4 for next day return, as observed in Table II, is quite robust for firms with different sizes. But it is interesting to notice that the magnitudes generally decrease from the smallest firms to the largest firms, indicating that the negative predictive pattern is the strongest for smaller firms. For the large retail investors, RT5, the positive predictive pattern remains for the small and medium-sized firms,

but not for large firms, indicating that their information advantage, if any, might be concentrated in smaller firms. As a comparison, order flows from institutions significantly predict next day returns in all three rows, and more so for the large firms, suggesting that their information advantage, if any, might be more prominent for larger firms.

When we separate firms by EP, turnover and stock price, we observe similar interesting patterns. That is, the predictive patterns in Table II are generally robust across firms with different characteristics, and the negative (positive) predictive power of smaller (larger) retail investors is stronger for small, low EP, and higher turnover firms, while the positive predictive power of institutional investors is stronger for large and high EP firms. [11] [12]

III. What Drives the Order Imbalance Predictive Power for Future Returns?

Given the large differences in predictive power for future returns of different investor groups' order flows, it is important to understand the driving forces for these differences. Previous literature provides several hypotheses for explaining investor order flows in general, and these might help to explain the heterogeneous predictive patterns from different investor groups for

[11] The Appendix Table 1 Panel A and Panel B show that small retail investors trade and hold more of smaller stocks, firms with lower EP and higher turnovers. Combining with the predictive patterns in Table 3, small retail investors likely have the largest information disadvantage or behavioral biases in these firms. In contrast, institutions have more trading and holding in larger stocks, firms with lower EP and higher turnovers. Those are also the firms that institutions have the highest predictive power, indicating they might have more information advantage over these firms.

[12] Given the positive and significant coefficients on OibRT5, one might wonder who are those larger retail investors. With a subsample between January 2019 to March 2019 with demographic information, we find RT5 are mostly male, with age above 45. Previous literature also provides additional information on the retail investors with larger accounts. For instance, An, Lou, Shi (2022) find that a transfer of 250 billion CNY from the poor to wealthy retail investors during bubbles and crashes period. Titman, Wei, and Zhao (2022) find larger retail investors tend to accumulate positions before suspicious split announcements and sell in the post-split period, indicating that they might have information advantage.

future returns. In Section III.A, we introduce a two-stage decomposition for the order flow's predictive power for future returns. We present the empirical results for the decomposition in Section III.B. We take a closer look at the information channel using event days in Section III.C.

A. A Two-Stage Decomposition to Explain Order Imbalance's Predictive Power

We consider four hypotheses for explaining the order flow dynamics and their predictive power for future stock returns. First, Chordia and Subrahmanyam (2004) state that order flows tend to be persistent, and persistent buying/selling pressure could lead directly to the predictability of future returns. Second, Kaniel, Saar, and Titman (2008) argue that retail traders in the U.S. are mostly contrarian, which provides liquidity to the market, and investors receive future positive returns. Following this logic, if the retail trades are momentum, which demand liquidity, then it is possible that the momentum trades might negatively predict future returns. Third, Liu et al. (2021) connect retail trading motives to behavioral biases, and they find that over-confidence about information advantage and gambling preferences are the two dominant behavioral biases that affect trades of Chinese retail investors.[13]. Finally, Kelley and Tetlock (2013) find that retail investors, especially the aggressive ones, may have valuable information about fundamental firm news, and thus their trading could correctly predict the direction of future returns. The above hypotheses are not mutually exclusive.

[13] There are many other interesting behavior biases of retail investors, such as disposition effects, extrapolation. Due to space limit of this study, here we choose the two most behavioral biases that affect trades of Chinese retail investors: over-confidence about information advantage and gambling preferences. We leave the others to future research.

To find out whether the above hypotheses help to explain the trading behavior of different retail investor groups, and their predictive power for future stock returns, we follow the two-stage decomposition method as in Boehmer et al. (2021). For the first stage, we use the above hypotheses to explain the retail flow measures to find out which ones are important drivers for the order flows. This step also helps to decompose the retail order flows into hypothesis-implied components for each hypothesis. For the second stage, we investigate which of the hypothesis-implied components contributes to the predictive pattern of different investor order flow measures.

To estimate the two-stage decomposition, we first identify proxies for each hypothesis. The proxies for the first two hypotheses are relatively easy to construct. For the order-persistence hypothesis, we adopt the previous day order imbalance measure, *Oib*(i,d-1), as the proxy. For the liquidity provision hypothesis, since it is directly linked to previous contrarian/momentum trading, we use returns from the previous day, week and month as proxies. For the overconfidence measure, we follow Barber et al. (2008) and Liu et al. (2021) and proxy it with corresponding investor group's turnover on that stock, which is the investor group's average of daily buy volume plus the daily sell volume divided by the investor group's holding shares at the end of previous day. Then we computed the final overconfidence variable, $Overconf(i,d)$, as an average of daily group turnover from the previous 20 days.[14] For gambling preferences at stock level, $Gamble(i,d)$, we

[14] For overconfidence proxy, Barber and Odean (2000) use each household's portfolio's turnover to proxy their overconfidence, which is aggregated at household level. Here we focus on investor groups rather than households, so we carry the spirit from the previous literature and use the stock-level turnover from each investor group. We also acknowledge that group turnover can potentially contain information other than over-confidence.

follow Bali et al. (2011) and compute the maximum daily returns from the previous 20 days as the proxy.[15]

For the information hypothesis, the most influential information at the firm level is earnings news, hence we follow Kelley and Tetlock (2013) and measure firm-level information by the cumulative abnormal returns (CAR) over the earnings announcement period. However, unlike the proxies for order persistence, liquidity provision and behavioral biases, which can be computed for each stock on each day, the news proxies are only available on earnings news days, which account for 1.58% of stock-days, and would render our two-stage estimation imprecise. To cope with this missing data issue for the news hypothesis, in this section we only consider the order persistence, liquidity provision and behavioral bias hypotheses, and we focus on the news hypothesis using an event-day approach in Section III.C.

After we collect all the proxies, we estimate the first stage for the two-stage decomposition. For each day *d*, we estimate a cross-sectional specification,

$$Oib(i,d) = d0(d) + d1(d)Oib(i,d-1) + d2(d)'Ret(i,d-1) + d3(d)Overconf(i,d-1) + d4(d)Gamble(i,d-1) + u4(i,d). \quad (5)$$

After we obtain the time-series of coefficients, $\{\widehat{d0}(d), \widehat{d1}(d), \widehat{d2}(d)', \widehat{d3}(d), \widehat{d4}(d)\}$, we

[15] An alternative measure for gambling preference proxy is introduced in Liu et al. (2021), which rely on events when the stock return hits 10% price limit. However, the 10% price limit hit only accounts for 0.07% of our total sample, and isn't suitable for our purpose on daily*stock frequency. Thus we choose the maximum daily return as our main gambling measure. We also consider other alternative proxies for gambling preferences, such as idiosyncratic volatility and skewness. These proxies deliver similar results to those using maximum daily returns, and are available on request.

conduct statistical inference using the time-series means and standard errors, which are adjusted using Newey-West with five lags, in order to understand how each of the four hypotheses contributes to retail order flows. Meanwhile, the first stage estimation allows us to decompose *Oib(i,d)* into five components:

$$Oib(i,d) = \widehat{Oib}_{i,d}^{persistence} + \widehat{Oib}_{i,d}^{liquidity} + \widehat{Oib}_{i,d}^{overconf} + \widehat{Oib}_{i,d}^{gamble} + \widehat{Oib}_{i,d}^{other}, \quad (6)$$

with $\widehat{Oib}_{i,d}^{persistence} = \widehat{d1}(d)Oib(i,d-1), \widehat{Oib}_{i,d}^{liquidity} = \widehat{d2}(d)'Ret(i,d-1), \widehat{Oib}_{i,d}^{overconf} = \widehat{d3}(d)Overconf(i,d-1), \widehat{Oib}_{i,d}^{gamble} = \widehat{d4}(d)Gamble(i,d-1)$ *and* $\widehat{Oib}_{i,d}^{other} = \widehat{u4}(i,d-1) + \widehat{d0}(d-1)$. That is, the "persistence" part is related to the order persistence hypothesis, the "liquidity" part is related to the liquidity provision hypothesis, the "overconf" and "gamble" are both related to behavioral biases, and the "other" component is the residual component, which potentially contains other relevant information about future returns.

For the second stage of the decomposition, we relate future returns to each individual component of order flow by estimating the following specification using the Fama-MacBeth methodology:

$$Ret(i,d+1) = e0(d+1) + e1(d+1)\widehat{Oib}_{i,d}^{persistence} + e2(d+1)\widehat{Oib}_{i,d}^{liquidity} + e3(d+1)\widehat{Oib}_{i,d}^{overconf} + e4(d+1)\widehat{Oib}_{i,d}^{gamble} + e5(d+1)\widehat{Oib}_{i,d}^{other} + e6(d+1)'Controls(i,d) + u5(i,d+1). \quad (7)$$

With the decomposition in equation (6), the coefficient estimates in equation (7) show how each component of various order flows helps to predict future stock returns.

According to Boehmer et al. (2021), the advantage of the two-stage decomposition approach is that it includes various components of $Oib(i,d)$ from alternative hypotheses in a unified and internally consistent empirical framework. The caveat of this approach is that we need to make empirical assumptions when choosing proxies for different hypotheses. Even though these assumptions seem to us to be reasonable, we still need to be cautions that the results depend on the validity of our empirical assumptions.

B. Estimation Results for the Two-Stage Decomposition

We report first-stage estimation results in Table IV Panel A. In the first row, the coefficients on lagged order flow variables are always positive and significant, indicating that order persistence is an important driver for order flows. For the next three rows, we connect order flows with returns from previous day, week and month, and the patterns are quite interesting. The order imbalances of RT1, RT2, and RT3 load positively and significantly on the previous day return, indicating that these investors buy more if the previous day return is positive, and sell more if the previous day return is negative. This corresponds to a daily momentum trading strategy, which demands immediate liquidity. For larger retail investors in RT4 and RT5, order imbalances load negatively and significantly on returns from the previous day, indicating that they are contrarian investors, buying low and selling high, and possibly providing immediate liquidity. If we extend the horizon to previous one week or one month, then the coefficients on all returns are negative and significant,

indicating that all retail investors follow contrarian strategies, buying losers and selling winners over the longer term.[16]

The next two rows present results on how behavioral biases are related to order flows. The coefficients on the overconfidence proxy are all positive and significant for RT1-RT4, indicating that overconfidence, proxied by group turnover, might be a strong driver for these retail investors' trading. Intriguingly, the magnitude of the coefficients gradually decreases from 0.0894 for RT1 to 0.0418 for RT4, implying a decreasing impact of overconfidence for retail investors as their account sizes increase. For the largest retail group, RT5, the coefficient becomes -0.0881 with a significant *t*-stat of -8.11. That is, the largest retail investors' trades are in the opposite direction of the overconfidence proxy. In terms of the gambling preference, for RT1-RT4, the coefficients are always positive and significant, indicating these retail investors like to buy stocks with lottery features. Interestingly, the coefficients gradually increase from 0.0330 for RT1 to 0.2423 for RT4, suggesting that larger retail investors trades have higher association with gambling preferences.[17]

[16] Our finding that large retail investors are contrarian and smaller ones are momentum traders over daily horizon is quite interesting and different from some previous studies. For instance, contrarian patterns have been documented in Kaniel, Saar and Titman (2008) using monthly horizons in the U.S., and Barrot, Kaniel and Sraer (2016) using daily and weekly horizons in France. Using U.S. data, Kelley and Tetlock (2013) and Boehmer et al. (2021) both find that retail trades follow momentum over daily horizons, but are contrarian at weekly horizons. In our setting, we find the trading patterns from investors with smaller account sizes are similar to those in Kelley and Tetlock (2013) and Boehmer et al. (2021), while the investors with the largest account sizes behave similarly to the patterns documented in Kaniel et al. (2008) and Barrot et al. (2016).

[17] The coefficients of gambling preferences monotonically increase from RT1 to RT4 are not necessary inconsistent with the finding that RT1 has the most negative return predictive power than other retail investor groups. There might be many rational and behavioral drivers for investors' trading behaviors, which potentially affect the overall predictive power of order flow for future returns. Here we only include the two most important behavior biases, as suggested by Liu et al. (2021), and the rest would be in the "other" component of order imbalance measures.

When we move on to RT5, the coefficient is -0.0671 with a significant *t*-stat of -3.09, which indicates that the largest retail investors trades are in the opposite direction of the gambling motive.

We report the second stage of the decomposition results in Panel B of Table IV. We take the first retail group, RT1, as an example. The coefficient estimate on *Oib(Persistence)* is -0.0333, with a *t*-statistic of -15.84, which implies that order persistence significantly and negatively contributes to the predictive power of RT1 trading flow. The coefficient estimate on *Oib(Liquidity)* is -0.0088, with a *t*-statistic of -2.61, which probably implies that daily momentum trading probably significantly and negatively contributes to the predictive power of RT1 trading flow. The coefficient of *Oib(Overconf)* is -0.1024, with a *t*-statistic of -2.84, and the coefficient for *Oib(Gamble)* is insignificant. For the *Oib(Other)* component, the coefficient is -0.0085, with a significant *t*-statistic of -27.11, indicating that there is other information, other than those incorporated in the three hypotheses, that significantly contributes to RT1's negative predictive pattern for future returns. In terms of economic magnitude, we compute the interquartile range of all five components of the order imbalance measure. For the smallest retail group RT1, if we move from the 25^{th} percentile to the 75^{th} percentile in the distribution, the interquartile differences in future one-day stock return, for the *Oib(Persistence)*, *Oib(Liquidity)*, *Oib(O*verconf*)*, *Oib(Gamble)* and *Oib(Other)*, are -0.1177%, -0.0290%, -0.0337%, -0.0314%, -0.1778%, respectively. That is to say, order persistence, liquidity demand, overconfidence, and gambling preferences all contribute to the negative predictive power of RT1 for next day returns, while the

first term has the largest magnitude. Similar patterns are observed for other smaller retail investor groups RT2-RT4.

If we turn our attention to the largest retail investors, RT5, the patterns are quite different. In terms of coefficient estimates, we find the order persistence and other are both positive and significant. In terms of economic magnitude, if we move from the 25th percentile to the 75th percentile in the distribution, the interquartile differences in future one-day stock return, for the *Oib(Persistence)*, *Oib(Liquidity)*, *Oib(O*verconf*)*, *Oib(Gamble)* and *Oib(Other)*, are 0.0281%, 0.0097%, 0.0257%, 0.0425%, 0.0490%, respectively. This indicates all three hypotheses contribute to RT5's positive predictive pattern for future returns, while only the first is significant.

Overall, our decomposition exercise shows that a substantial part of the negative predictive power of the retail investors with smaller account sizes comes from order persistence, liquidity demand, and behavioral biases, while the positive predictive power of the retail investors with larger account balances comes from order persistence and trading against overconfidence and gambling preferences. Across all investor groups, the significance and the large magnitude of the "other" component indicates that existing hypotheses cannot fully explain the trading behaviors and their predictive power for returns. So what does "other" stand for? One possibility is information, which we take a close look at in the next subsection.

C. A Close Look at the Information Channel

It is important to understand how various retail investors participate in the information discovery process. As mentioned earlier, the most influential information at the firm level is earnings news, hence we follow Kelley and Tetlock (2013) and measure firm-level information by the cumulative abnormal returns (CAR) over the earnings announcement period. Notice that earnings news only happens quarterly rather than daily, so the daily Fama-MacBeth estimation we adopt for the two stage estimation might not be proper for understanding how Chinese retail investors process information. As an alternative, in this section, we focus on event days to study this issue. To capture each retail investor groups' participation in the information discovery process, we proceed in three steps.

In the first step, we examine whether different retail investors can predict earnings news the next day. A positive answer indicates that these investors anticipate the information before the information becomes public, either because they have access to private information or they have better skills. In the second step, we check whether they can process contemporaneous earnings news to find out whether they have skills to gather information from available public news. In the third step, we investigate whether retail order flow's predictive power for future returns improves or deteriorates on earnings event days, to find out whether information is a significant contributor for the overall prediction pattern documented in previous sections.

For this first step, to find out whether retail order flows can predict earnings news, we estimate the following cross-sectional specification for each quarter q:

$$CAR(i, d-1, d) = f0(q) + f1(q)Oib(i, d-2) + f2(q)'Controls(i, q-1) + u6(i, q). \quad (8)$$

Assuming the earnings announcement day is day d, we compute the cumulative returns over day d-1 and day d, and subtract the market returns over the same period to obtain cumulative abnormal returns for each stock, $CAR(i, d-1, d)$.[18] The main predictive variable on the right hand of the equation is order imbalance measure from day d-2. Notice that each firm only has one earnings day each quarter, and equation (8) is estimated for each quarter in the cross section to make sure we cover all firms each quarter. As in a standard Fama-MacBeth setting, the statistical inferences are based on the quarterly time-series of the estimated coefficients, and standard errors are computed using Newey-West with 4 lags. If retail order flows can predict earnings surprises in the right direction, the coefficient $f1$ should be significantly positive, and vice versa.

We present the estimation results in Panel A of Table V. For retail investors RT1-RT3, the coefficients $f1$ are -0.0251, -0.0234, and -0.0166, respectively, all with highly significant *t*-statistics. These negative and significant coefficients indicate that these investors incorrectly predict earnings surprises. The coefficient $f1$ for RT4 is close to zero and insignificant. In contrast, the coefficients $f1$ for RT5 is 0.0023, positive and statistically significant, implying that these investors are able to correctly predict future earnings surprises.

For the second step, we examine whether different retail groups can process contemporaneous public news. Here the dependent variable is retail order flow, *Oib(i,d)*, and we connect it to

[18] We also examine wider window such as CAR(-1,1) and CAR(-3,3), the results are similar and available on request.

contemporaneous earnings news, *CAR(i,d-1,d)*. The specification is similar to equation (8), except the timeline is different:

$$Oib(i,d) = g0(q) + g1(q)CAR(i,d-1,d) + g2(q)'Controls(i,d-2) + u7(i,q). \quad (9)$$

If a particular type of retail order imbalance can process contemporaneous and public earnings news in the right direction, we expect the associated coefficient $g1$ to be significantly positive, and vice versa.

Panel B of Table V report the estimation results. For retail investors RT1-RT4, the coefficients $g1$ are -1.9225, -1.8291, -1.4349, -0.8781 respectively, all with highly significant *t*-statistics. These negative and significant coefficients indicate that these retail investor groups process the contemporaneous public earnings news in the wrong direction. In contrast, the coefficient $g1$ for RT5 is 0.1583, implying that RT5 might be able to correctly process contemporaneous public earnings news. However, the coefficient is statistically insignificant.

For the third step, we examine whether retail order flows' predictive power for future returns improves or deteriorates on event days to understand how much the information hypothesis helps to explain the return predictive patterns we observe in Section II. We estimate a modified version of equation (2), by adding the event day dummy and an interaction term:

$$Ret(i,d) = h0(d) + [h1(d) + h2(d)Event(i,d-1)]Oib(i,d-1) + \quad (10)$$
$$h3(d)Event(i,d-1) + h4'(d)Controls(i,d-1) + u8(i,d) \ .$$

Here the event dummy Event(i, d-1), is equal to one if the firm i has earnings news on day d-1, and zero otherwise. For non-news days, the predictive power of retail trades is measured by coefficient $h1$; for news days, the predictive power is measured by ($h1$+$h2$). If coefficient $h2$ is significantly different from zero, that group of retail investors anticipates future stock returns differently on these news days.

In the U.S., firm earnings announcements are chosen by firms and scattered throughout the year. In China, all firms are required to report their financial statements to regulators before four preset deadline dates each year. As a result, firms mostly announce their earnings within a short period before these deadline dates, and there would be zero announcements outside of these short periods. To make sure that we have enough observations to estimate the Fama-Macbeth coefficients in equation (10), we only include days with at least 5% of total number of firms with earnings announcements, which gives us 68 days, or 8% of the total days in our sample.

The results are presented in Table V Panel C. Here we take the smallest retail investors, RT1, as an example. The coefficients on order imbalance, $h1$, is -0.0079 and is statistically significant, indicating that on average the trades from RT1 negatively predict future returns. When there is earnings announcement news, the coefficient on the interaction of event dummy and the order imbalance is -0.0080, with a significant t-statistic of -3.20, implying that the negative prediction of RT1 for future stock returns doubles on earnings news days. This is consistent with our earlier finding that the smaller retail investors fail to predict and process the earnings news, which leads

to more negative prediction for returns on event days. We observe similar patterns for RT2, RT3 and RT4. For the largest retail investors, RT5, the coefficients $h1$ and $h2$ are 0.0005 and 0.0014, both statistically significant. That is to say, the large retail investors' predictive power for future returns quadruples on earnings news days, possibly because these retail investors can correctly predict and process the earnings news, which enhances their ability to predict future stock returns.

Overall, our results reveal interesting heterogeneous patterns of how retail investors predict and process public information. On one hand, smaller retail investors are unable to predict future news and lack skills to correctly process public news, while the largest retail investors and institutions are able to correctly anticipate future earnings news and incorporate the contemporaneous news into their trading. The differences in information-processing abilities of different retail investors clearly contribute to the differences in their predictive powers for future returns.

IV. Further Discussions and Robustness

A. Ages and Genders

In this section, we examine heterogeneity through demographic differences, such as gender and age, of retail investors. According to Barber and Odean (2001), male investors could be more susceptible to behavioral biases, such as overconfidence and lack of attention. Due to the limited access to data, we only have a three-month sample period from January 2019 to March 2019 on investor gender and age. We first present summary statistics on age and gender in Table VI Panel

A. Male investors contribute 67% of trading volume on average, and females account for 33%. Within the male group, the trading volume (%) across age groups below 45, and above 45 is 29% and 38% (summing to the 67% male total), while the trading volume (%) for the same age groups for females is 13% and 20% (summing to the 33% of volume traded by females). That is, across all gender-age groups, older male investors trade the most.

Next, we examine the determinants of return prediction for each gender-age group specified in equation (2). The results are reported in Table VI Panel B. For return predictions, we find male investors across all ages significantly and negatively predict returns, especially for older males. The predictive coefficients are insignificantly different from zero for female retail investors. These interesting patterns across age and gender are mostly consistent with previous findings in Barber and Odean (2001) provide further evidence regarding heterogeneity of retail investors.

B. Applying stricter filters from Liu, Stambaugh and Yuan (2019)

In this study, we apply a filter from Liu et al. (2019) and discard stocks with less than 15 days of trading during the most recent month. In addition, Liu et al. (2019) also eliminate stocks that have become public within the past six months, stocks with less than 120 days of trading during the past 12 months, and the smallest 30% of firms listed in SHSE and SZSE. We add all these additional filters and check the robustness of our results.

In Table VI Panel C, the order imbalance prediction directions are similar to the results in Table II. The first four groups of retail investors tend to trade in the wrong direction for future price

movements, while the largest retail investor group RT5 and institutions trade in the same direction as the cross-section of future stock returns. The economic magnitudes for the first four type of retail investors are quantitatively similar, while RT5's economic magnitude is only half as large when adding these additional filters, perhaps because RT5's positive return mainly comes from small stocks. The economic magnitude for institutions is still large. In conclusion, our main results are robust to the stricter filters from Liu, Stambaugh and Yuan (2019).

C. Leveraged positions

Our trade level data also identify investors' margin buys, short sales and collateral trades. Leveraged trading may be different from non-leverage trading. On each day, margin buys account for 10% of the trading volume, short sales account for 0.2% and collateral trading accounts for 15% during our sample period. We exclude the leverage trades and re-estimate equation (2).

Results are reported in Table VI Panel E. The order imbalance prediction directions are similar to the results in Table II. The first four groups of retail investors trade in the wrong direction of future price movements, while the largest retail investor group RT5 and institutions trade in a way that positively predicts the cross-section of future stock returns. The economic magnitudes are quantitatively similar. In conclusion, our results are robust to whether or not we include these leverage trades.

D. Price limits

One institutional feature of the Chinese market is the price limit restrictions. That is, investors can buy and sell stocks freely when the stock's price is within ±10% from previous day close price; if the price move out of the ±10% range, trading stops till the next day open. Chen et al. (2019) focus on the price limit days and find that large investors tend to trade differently on these days. Here we examine whether our results still hold when we exclude the price limit days.

We re-estimate equation (2) without the price limit days and present the results in Table VI Panel E. The order imbalance prediction directions are similar to the results in Table II. The first four groups of retail investors trade in the wrong direction of future price movements, while the largest retail investor group RT5 and institutions trade in a way that positively predicts the cross-section of future stock returns. The economic magnitudes are quantitatively similar. That is, our results are robust to whether or not we include these price limits days.

E. Forming Portfolios Using Retail Order Flows

Our main results in previous sections are based on Fama-MacBeth regressions, which assumes linear relations between the future returns and order flow variables. In this section, we adopt an alternative portfolio approach and examine whether our results still hold. To be more specific, we sort firms into five groups each day, based on previous day's order imbalance from a particular investor group, buying and selling the 20% of stocks with the highest and lowest order imbalance measures for that particular investor group. We report the risk adjusted returns (alphas) on this

long-short strategy for one to 60 days, where we conduct risk adjustment using the Liu, Stambaugh and Yuan (2019) three factor model.

From Table VI Panel F, the one-day long-short portfolio alpha, using the previous day order imbalance from RT1, is -0.0042, and highly significant. From one week to 12 weeks, the cumulative alphas for the long-short portfolio decrease from -0.0089 to -0.0183, and they are all highly significant. That is, the cumulative alphas using OibRT1 is consistently negative and significant, and there are no signs of reversal within 12 weeks, which echoes our earlier results in Table 2. Similar patterns exist for RT2 and RT3. For RT4, the one-day alpha is negative at -0.0007, but it quickly becomes insignificant when we extend the holding horizon to one week, indicating that RT4 for horizons longer than one day. For the largest retail investors in RT5, the one-day alpha is 0.0017, which is positive and significant. The 12-week cumulative alpha is 0.0057, still positive and significant, confirming the results in Table 2 that RT5 has both short and long term predictive power for future returns. Results on OibINST are similar to those for OibRT5.

F. News from CFNDS

Our earlier results show that the smaller retail investors lack skills to predict or process public earnings news, while the largest retail investors are able to correctly predict and process future earnings news and incorporate the contemporaneous news into their trading. In this section, we use an alternative public news dataset to investigate whether the results from earnings news can be extended to other news. We obtain news data from the Financial News Database of Chinese

Listed Companies (CFND), which includes news on all A-share stocks from over 400 internet media and over 600 newspapers. In comparison with earnings news, the data coverage is more substantial, but the news content is more diverse.

We estimate equation (10) and report the results in Table VI Panel G. For the smallest retail investors, RT1, the coefficient on order flow is -0.0075 and highly significant, confirming that their order flows predict returns negatively. The coefficient on the interaction between order flow and the event dummy is -0.0045, again highly significant, suggesting the negative predictive power is significantly stronger on news days, which is consistent with our results in Section III.C. Similar patterns are observed for RT2-RT4. For the largest retail investors, RT5, the coefficient on order flow is 0.0008, and on the interaction is 0.0011, both highly significant, indicating that the RT5 order flow on average predicts future returns in the correct direction, and their prediction becomes much stronger on news days. To summarize, we confirm with an alternative news dataset that smaller retail investors lack skills to process public earnings news, and their negative predictions for future returns are worse on news days, while the largest retail investors and institutions are able to correctly process future earnings news and enhance their predictive power for future returns.

V. Conclusion

Using comprehensive retail trading and holding data from 2016 to 2019, we separate tens of millions of retail investors into five groups by their account sizes, and examine heterogeneity in retail investors' return predictabilities, and sources of the return predictabilities.

We provide strong and direct evidence on retail investors' heterogeneity. Retail investors with account sizes less than 3mil CNY buy and sell stocks in the wrong directions. The prices of stocks they buy experience negative returns the next day, while the ones they sell experience positive returns. For retail investors with large account balances, their trading predicts returns in the correct direction. In tracing their differences in predicting future returns, we provide evidence that the negative predictive power of the retail investors with smaller account sizes are mostly related to their order persistence, daily momentum trading, behavioral biases and failures in processing earnings news. In contrast, the positive predictive power of the large retail investors is mostly associated with order persistence, contrarian trading, trading against behavioral biases and advantages in processing earnings news.

Our results on the heterogeneity of retail investors help to understand the conflicting empirical results in the previous literature regarding retail investors. In addition, the exchange itself acknowledges the heterogeneity in retail investors and is focused on adopting policies on investor education and suitability that restrict some kinds of trading for the smallest accounts. For example, the Shanghai Stock Exchange requires a retail investor to have at least 500k CNY holdings of stocks for at least 20 trading days to open a leverage trading account or to trade on the riskier Science and Technology Innovation Board (or STAR Market). These policies effectively exclude the smallest retail investors from leverage trading and trading on riskier start-ups, which could help protect these small retail investors from even worse losses.

Our study clearly leaves many interesting questions unsolved. For example, why retail investors dominate trading in the Chinese stock market? Is it the T+1 trading rule that discourage participation of institutional investors in trading? How much do retail investors gain or loss with their investments in stock market? We leave these interesting and important questions to future research.

References

An, Li, Dong Lou, and Donghui Shi, 2022, Wealth redistribution in bubbles and crashes. *Journal of Monetary Economics*, 126(3), 134-153.

Anagol, Santosh, Vimal Balasubramaniam, and Tarun Ramadorai, 2021, Learning from noise: Evidence from India's IPO lotteries. *Journal of Financial Economics*, 140(3), 965-986.

Bach, Laurent, Laurent E. Calvet, and Paolo Sodini, 2020, Rich pickings? Risk, return, and skill in household wealth. *American Economic Review*, 110(9), 2703-47.

Balasubramaniam, Vimal, John Y. Campbell, Tarun Ramadorai, and Benjamin Ranish, 2021, Who owns what? A factor model for direct stockholding. NBER Working paper.

Bali, Turan G., Nusret Cakici, and Robert F. Whitelaw, 2011, Maxing out: Stocks as lotteries and the cross-section of expected returns, *Journal of Financial Economics*, 99(2), 427-446.

Barber, Brad M., Xing Huang, Terrance Odean, and Christopher Schwarz, 2021, Attention induced trading and returns: Evidence from robinhood users. *Journal of Finance*, Forthcoming.

Barber, Brad M., Yi-Tsung Lee, Yu-Jane Liu, and Terrance Odean, 2009, Just how much do individual investors lose by trading?. *The Review of Financial Studies*, 22(2), 609-632.

Barber, Brad M., Yi-Tsung Lee, Yu-Jane Liu, and Terrance Odean, 2014, The cross-section of speculator skill: Evidence from day trading. *Journal of Financial Markets*, 18, 1-24.

Barber, Brad M., Shengle Lin, and Terrance Odean, 2021, Resolving a paradox: Retail trades positively predict returns but are not profitable. Working Paper.

Barber, Brad M., and Terrance Odean, 2000, Trading is hazardous to your wealth: The common stock investment performance of individual investors, *Journal of Finance* 55, 773−806.

Barber, Brad M., and Terrance Odean, 2001, Boys will be boys: Gender, overconfidence, and common stock investment, *The Quarterly Journal of Economics*, *116*(1), 261-292.

Barber, Brad M., and Terrance Odean, 2008, All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors, *Review of Financial Studies* 21, 785-818.

Barber, Brad M., Terrance Odean, and Ning Zhu, 2009, Do retail trades move markets? *Review of Financial Studies*, 22, 151−186.

Barrot, Jean-Noel, Ron Kaniel, and David Alexandre Sraer, 2016, Are retail traders compensated for providing liquidity? *Journal of Financial Economics*, 120, 146-168.

Black, Fischer, 1986, Noise, *Journal of Finance*, 41(3), 528-543.

Blume, Marshall E. and Robert F. Stambaugh, 1983, Biases in computed returns: An application to the size effect, *Journal of Financial Economics*, 12, 387-404.

Boehmer, Ekkehart, Charles M. Jones, Xiaoyan Zhang, and Xinran Zhang, 2021, Tracking retail investor activity, *Journal of Finance*, 76(5), 2249-2305.

Carpenter, Jennifer N., Fangzhou Lu, and Robert F. Whitelaw, 2021, The real value of China's stock market. Journal of Financial Economics, 139(3), 679-696.

Chiang, Yao-Min, David Hirshleifer, Yiming Qian, and Ann E. Sherman, 2011, Do investors learn from experience? Evidence from frequent IPO investors. *The Review of Financial Studies*, 24(5), 1560-1589.

Chen, Ting, Zhenyu Gao, Jibao He, Wenxi Jiang, and Wei Xiong, 2019, Daily price limits and destructive market behavior, *Journal of Econometrics*, 208(1), 249-264.

Chordia, Tarun, and Avanidhar Subrahmanyam, 2004, Order imbalance and stock returns: Theory and evidence, *Journal of Financial Economics*, 72, 485−518.

Dorn, Daniel, Gur Huberman, and Paul Sengmueller, 2008, Correlated trading and returns. *The Journal of Finance*, 63(2), 885-920.

Eaton, Gregory W., T. Clifton Green, Brian Roseman, and Yanbin Wu, 2021, Zero-commission individual investors, high frequency traders, and stock market quality. Working paper.

Fama, Eugene F., and James D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy*, 81, 607−636.

Fong, Kingsley YL, David R. Gallagher, and Adrian D. Lee, 2014, Individual investors and broker types. *Journal of Financial and Quantitative Analysis*, 49(2), 431-451.

Gao, Xiaohui, and Tse-Chun Lin, 2015, Do individual investors treat trading as a fun and exciting gambling activity? Evidence from repeated natural experiments. *The Review of Financial Studies*, 28(7), 2128-2166.

Grinblatt, Mark, and Matti Keloharju, 2000, The investment behavior and performance of various investor types: a study of Finland's unique data set. *Journal of Financial Economics*, 55(1), 43-67.

Grinblatt, Mark, Matti Keloharju, and Juhani T. Linnainmaa, 2012, IQ, trading behavior, and performance. *Journal of Financial Economics*, 104(2), 339-362.

Hu, Conghui, Yu-Jane Liu, and Xin Xu, 2021, The valuation effect of stock dividends or splits: Evidence from a catering perspective, *Journal of Empirical Finance*, 61, 163-179.

Jiang, Lei, Jinyu Liu, Lin Peng, and Baolian Wang, 2019, Investor attention and asset pricing anomalies, Working paper, Tsinghua University.

Kaniel, Ron, Saar Gideon, and Titman, Sheridan, 2008, Individual investor sentiment and stock returns, *Journal of Finance*, 63, 273−310.

Kaniel, Ron, Liu, Shuming, Saar, Gideon, and Titman, Sheridan, 2012, Individual investor trading and return patterns around earnings announcements, *Journal of Finance*, *67*, 639-680.

Kelley, Eric K. and Paul C. Tetlock, 2013, How wise are crowds? Insights from retail orders and stock returns, *Journal of Finance*, 68, 1229-1265.

Leippold, Markus, Qian Wang, and Wenyu Zhou, 2022, Machine learning in the Chinese stock market. *Journal of Financial Economics*, forthcoming.

Li, Xindan, Ziyang Geng, Avanidhar Subrahmanyam, Honghai Yu, 2017, Do wealthy investors have an informational advantage? Evidence based on account classifications of individual investors, *Journal of Empirical Finance* 44, 1-18.

Liao, Jingchi, Cameron Peng, and Ning Zhu, 2022, Extrapolative bubbles and trading volume, *Review of Financial Studies, ,* 35(4), 1682-1722.

Linnainmaa, Juhani T., 2010, Do limit orders alter inferences about investor performance and behavior?. *The Journal of Finance*, 65(4), 1473-1506.

Liu, Jianan, Robert F. Stambaugh, and Yu Yuan, 2019, Size and value in China, *Journal of Financial Economics*, 134(1), 48-69.

Liu, Hongqi, Cameron Peng, Wei A. Xiong, and Wei Xiong, 2021, Taming the bias zoo, *Journal of Financial Economics*, 143(2), 716-741.

Newey, Whitney K., and Kenneth D. West, 1987, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica*, 55, 703−708.

Ozik, Gideon, Ronnie Sadka, and Siyi Shen, 2021, Flattening the illiquidity curve: Retail trading during the COVID-19 lockdown, *Journal of Financial and Quantitative Analysis*, 56(7), 2356-2388.

Stoffman, Noah, 2014, Who trades with whom? Individuals, institutions, and returns. *Journal of Financial Markets*, 21, 50-75.

Titman, Sheridan, Chishen Wei, and Bin Zhao, 2022, Corporate actions and the manipulation of retail investors in China: An analysis of stock splits. *Journal of Financial Economics*, forthcoming.

Welch, Ivo, 2020, The wisdom of the Robinhood crowd, *Journal of Finance*, forthcoming.

Table I. Summary statistics

This table reports summary statistics for stock characteristics, trading and holdings by different investor groups. Our sample period covers January 2016 to June 2019, and our sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 days with trades during the previous month. Panel A reports the time series average of the cross-sectional distribution of stock level characteristics, daily stock return (*Ret*), market capitalization (*Size*), earnings to price ratio (*EP*), and monthly turnover (*Turnover*). Panel B shows the number of accounts, aggregate trading and holdings, and holding horizons by different investor groups. The holding horizon is the shares held by each type of investor divided by shares traded by this type of investor and captures how many days on average this type of investor takes to turn over a position. Panel C reports the time series average of the cross-sectional mean, standard deviation, autocorrelation (AR1), and cross correlations of order imbalance*s* by different investor groups. Order imbalances (*Oib*) are computed as the buy share volume minus sell share volume divided by buy share volume plus sell share volume for each investor group, as specified in equation (1).

Panel A. Stock characteristics

	Variable description	Mean	Std	P25	P50	P75
Ret	Daily Stock Return	-0.01%	2.17%	-1.09%	-0.22%	0.77%
Size	Market Capitalization (Billion CNY)	20.1	80.3	2.9	5.6	12.1
EP	Earnings to Price Ratio	0.0075	0.0155	0.0018	0.0060	0.0122
Turnover	Monthly Turnover (of tradable A shares)	48.32%	72.48%	14.09%	25.40%	49.97%

Panel B. Number of accounts, trading and holdings by different types of investors

Account value	RT1 <100K CNY	RT2 (100K,500K) CNY	RT3 (500K,3M) CNY	RT4 (3M,10M) CNY	RT5 >10M CNY	INST	CORP
Number of Accounts (thousands)	31,410	15,282	5,827	735	235	40	47
Aggregate Trading Volume (Bil. CNY)	9	35	54	27	37	35	3
Aggregate Trading Volume (% of total)	5%	17%	27%	13%	19%	17%	2%
Aggregate Holdings Value (Bil CNY)	336	951	1,566	840	1,794	4,201	15,547

Aggregate Holdings Value (% of total)	1%	4%	6%	3%	7%	17%	62%
Holding Horizon (Days)	50	36	35	35	49	109	6,319

Panel C. Order imbalance in the cross section by investor group

	Mean	Std	AR1	Correlations						
				OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST	OibCORP
OibRT1	-0.021	0.187	0.243	1						
OibRT2	-0.011	0.171	0.259	0.802	1					
OibRT3	-0.006	0.166	0.216	0.610	0.710	1				
OibRT4	0.002	0.250	0.059	0.194	0.244	0.256	1			
OibRT5	0.019	0.352	0.102	-0.151	-0.158	-0.163	-0.091	1		
OibINST	-0.011	0.455	0.205	-0.315	-0.365	-0.380	-0.263	-0.188	1	
OibCORP	-0.004	0.720	0.088	0.022	0.029	0.021	-0.007	-0.043	-0.044	1

Table II. Predicting Future Stock Returns Using Order Imbalances from Different Investor Groups

This table reports estimation results on whether trading activity by different investor groups can predict the cross section of one-day-ahead returns and returns over the returns over the next 12 weeks. Our sample period covers January 2016 to June 2019, and our sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 days with trading during the previous month. In Panel A, we present coefficient estimates from Fama-MacBeth (1973) regressions as specified in equation (2). Panel B reports coefficient estimates from Fama-MacBeth (1973) regressions specified in equation (3). The independent variables are the previous day order imbalance *Oib(-1)*, and the control variables are the previous day return *Ret(-1),* the previous week return *Ret(-6,-2)* and the previous month return *Ret(-27,-7),* previous month log market cap (*Size*), earnings to price ratio (*EP*) and monthly turnover (*Turnover*). For each regression in Panel A, we also provide the interquartile range for the relevant explanatory order imbalance to compute the difference in predicted future returns for the interquartile range (*Interquartile return*). To account for serial correlation in the coefficients, the standard errors of the time-series are adjusted using Newey-West (1987) with 5 lags. ***, ** and * indicate significance at the 1%, 5% and 10% level.

Panel A. Predict next day return

Oib.var		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
Oib(-1)	Estimate	-0.0093***	-0.0091***	-0.0065***	-0.0009***	0.0012***	0.0016***
	[*t*-stat]	[-24.98]	[-22.58]	[-18.50]	[-7.21]	[12.26]	[20.34]
Ret(-1)	Estimate	-0.0027	-0.0091**	0.0006	0.0189***	0.0190***	0.0132***
	[*t*-stat]	[-0.62]	[-2.07]	[0.13]	[4.06]	[4.13]	[2.79]
Ret(-6,-2)	Estimate	-0.0149***	-0.0132***	-0.0124***	-0.0120***	-0.0115***	-0.0113***
	[*t*-stat]	[-8.06]	[-7.07]	[-6.62]	[-6.37]	[-6.13]	[-6.04]
Ret(-27,-7)	Estimate	-0.0039***	-0.0036***	-0.0034***	-0.0033***	-0.0032***	-0.0034***
	[*t*-stat]	[-4.36]	[-4.04]	[-3.86]	[-3.72]	[-3.62]	[-3.85]
Size	Estimate	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	[*t*-stat]	[0.36]	[0.17]	[-0.16]	[-0.32]	[-0.18]	[-0.21]
EP	Estimate	0.0147***	0.0150***	0.0145***	0.0144***	0.0146***	0.0140***
	[*t*-stat]	[3.54]	[3.57]	[3.41]	[3.42]	[3.47]	[3.34]

Turnover	Estimate	-0.0007***	-0.0007***	-0.0007***	-0.0007***	-0.0007***	-0.0007***
	[*t*-stat]	[-3.47]	[-3.63]	[-3.69]	[-3.83]	[-3.83]	[-3.83]
Intercept	Estimate	-0.0012	-0.0006	0.0002	0.0005	0.0001	0.0002
	[*t*-stat]	[-0.48]	[-0.26]	[0.06]	[0.19]	[0.06]	[0.10]
Adj.R2		8.83%	8.68%	8.43%	8.10%	8.11%	8.25%
Interquartile		0.2222	0.1827	0.1678	0.2868	0.4536	0.6740
Interquartile return		-0.2062%	-0.1668%	-0.1089%	-0.0247%	0.0523%	0.1046%

Panel B. Predict returns over the next 12 weeks

Cumulative weeks	OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
1	-0.0226***	-0.0220***	-0.0144***	-0.0019***	0.0027***	0.0045***
2	-0.0291***	-0.0281***	-0.0183***	-0.0019***	0.0037***	0.0057***
3	-0.0329***	-0.0313***	-0.0201***	-0.0020***	0.0044***	0.0064***
4	-0.0353***	-0.0334***	-0.0216***	-0.0022***	0.0050***	0.0069***
5	-0.0368***	-0.0351***	-0.0229***	-0.0028***	0.0054***	0.0072***
6	-0.0398***	-0.0377***	-0.0250***	-0.0032***	0.0054***	0.0078***
7	-0.0424***	-0.0402***	-0.0271***	-0.0036***	0.0053***	0.0087***
8	-0.0437***	-0.0411***	-0.0278***	-0.0039***	0.0055***	0.0091***
9	-0.0442***	-0.0411***	-0.0278***	-0.0036***	0.0058***	0.0092***
10	-0.0448***	-0.0414***	-0.0281***	-0.0034***	0.0057***	0.0093***
11	-0.0455***	-0.0425***	-0.0287***	-0.0035***	0.0055***	0.0097***
12	-0.0458***	-0.0424***	-0.0282***	-0.0036***	0.0055***	0.0097***

Table III. Predicting Next Day Stock Returns for Different Subgroups of Firms

This table reports estimation results on whether trading activity by different investor groups can predict the cross section of one-day-ahead returns for firms with different characteristics. Our sample period covers January 2016 to June 2019, and our sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 days with trading during the previous month. We present coefficient estimates from Fama-MacBeth (1973) regressions in equation (4). The dependent variable is the return on day *d*. The independent variables are the previous day order imbalance *Oib(-1)*, and the control variables are the previous day return *Ret(-1),* the previous week return *Ret(-6,-2)* and the previous month return *Ret(-27,-7),* previous month log market cap (*Size*), earnings to price ratio (*EP*) and monthly turnover (*Turnover*). To account for serial correlation in the coefficients, the standard errors of the time-series are adjusted using Newey-West (1987) with 5 lags. ***, ** and * indicate significance at the 1%, 5% and 10% level.

Oib.var	Coefficients	OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
Small Size	$c1$	-0.0119***	-0.0131***	-0.0090***	-0.0006***	0.0020***	0.0014***
Medium Size	$c2$	-0.0096***	-0.0093***	-0.0066***	-0.0011***	0.0009***	0.0015***
Large Size	$c3$	-0.0067***	-0.0061***	-0.0044***	-0.0012***	0.0001	0.0021***
Low EP	$c1$	-0.0122***	-0.0131***	-0.0095***	-0.0010***	0.0020***	0.0015***
Medium EP	$c2$	-0.0097***	-0.0098***	-0.0069***	-0.0009***	0.0012***	0.0015***
High EP	$c3$	-0.0062***	-0.0056***	-0.0039***	-0.0007***	0.0001	0.0017***
Low Turnover	$c1$	-0.0061***	-0.0057***	-0.0041***	-0.0007***	0.0003***	0.0013***
Medium Turnover	$c2$	-0.0091***	-0.0092***	-0.0066***	-0.0010***	0.0011***	0.0014***
High Turnover	$c3$	-0.0159***	-0.0176***	-0.0128***	-0.0009***	0.0026***	0.0019***
Low Price	$c1$	-0.0088***	-0.0086***	-0.0067***	-0.0012***	0.0009***	0.0014***
Medium Price	$c2$	-0.0098***	-0.0095***	-0.0068***	-0.0007***	0.0013***	0.0016***
High Price	$c3$	-0.0094***	-0.0094***	-0.0060***	-0.0007***	0.0014***	0.0016***

Table IV. Two Stage Decomposition for Understanding the Predictive Patterns of Order Flow Variables

This table reports estimation results on a decomposition of the predictive power of different investor groups' order imbalance for the cross-section of future stock returns. Our sample period covers January 2016 to June 2019, and the sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 trading days in the previous month. We estimate two-stage Fama-MacBeth (1973) regressions. Panel A reports the first-stage estimation results, where the order imbalance measures are decomposed into five components as specified in equation (5). Variable overconf (-1) is measured as corresponding investor group's average turnover on the stock from the previous 20 days, as a proxy for each investor's overconfidence. Variable gamble (-1) is the maximum daily returns from previous 20 days, as a proxy for gambling preference. Panel B reports the second-stage decomposition of order imbalance's predictive power, as specified in equations (6) to (7). As independent variables, the variable *Oib(*-1,*Persistence)* is estimated in the first stage using past order imbalance and reflects price pressure. The variable *Oib(*-1,*Liquidity)* is estimated in the first stage using past returns over different horizons and is connected to the liquidity provision or liquidity demand hypothesis. The variable *Oib(*-1,*Overconf)* is estimated in the first stage, reflecting overconfidence. The variable *Oib*(-1,*Gamble*) is estimated in the first stage using maximum daily returns from previous 20 days and reflects a preference for gambling. The residual part of the previous day order imbalance from the first-stage estimation is denoted "other," which can be attributed to private information about future returns. As control variables, we include previous day return, *Ret*(-1), previous week return, *Ret*(-6,-2), and previous month return, *Ret*(-27, -7), previous month log market cap (*Size*), earnings to price ratio (*EP*) and monthly turnover (*Turnover*). To account for serial correlation in the coefficients, the standard errors of the time series are adjusted using Newey-West (1987) with five lags. For each regression, we also report the difference in predicted day-ahead returns for observations at the two ends of the interquartile range (*Interquartile return*) in Panel B. To account for serial correlation in the coefficients, the standard errors of the time-series are adjusted using Newey-West (1987) with 5 lags. ***, ** and * indicate significance at the 1%, 5% and 10% level.

Panel A. First stage of projecting order imbalance on persistence, past returns, overconfidence and gambling proxies

		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5
Oib(-1)	Estimate	0.1870***	0.1967***	0.1711***	0.0499***	0.1036***
	[*t*-stat]	[40.77]	[48.26]	[49.37]	[21.71]	[39.54]
Ret(-1)	Estimate	0.5159***	0.7269***	0.4482***	-0.2205***	-1.2968***
	[*t*-stat]	[15.68]	[28.95]	[21.65]	[-9.71]	[-39.38]
Ret(-6,-2)	Estimate	-0.4214***	-0.2196***	-0.1063***	-0.0804***	-0.0485***
	[*t*-stat]	[-28.69]	[-19.75]	[-11.01]	[-7.01]	[-3.73]
Ret(-27,-7)	Estimate	-0.0350***	-0.0196***	-0.0235***	-0.0399***	-0.0203***
	[*t*-stat]	[-7.56]	[-5.30]	[-7.81]	[-11.35]	[-4.38]
Overconf(-1)	Estimate	0.0894***	0.0611***	0.0657***	0.0418***	-0.0881***
	[*t*-stat]	[5.90]	[5.80]	[8.81]	[4.43]	[-8.11]
Gamble(-1)	Estimate	0.0330*	0.0784***	0.1731***	0.2423***	-0.0671***
	[*t*-stat]	[1.83]	[5.67]	[13.75]	[14.68]	[-3.09]
Intercept	Estimate	-0.0214***	-0.0120***	-0.0115***	-0.0078***	0.0231***
	[*t*-stat]	[-7.00]	[-5.89]	[-9.05]	[-5.19]	[12.04]
Adj.R2		7.08%	5.60%	3.89%	0.73%	2.00%

Panel B. Second stage decomposition of order imbalance's predictive power

Dep.var		Ret	Ret	Ret	Ret	Ret
Oib.var		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5
Oib(-1,Persistence)	Estimate	-0.0333***	-0.0279***	-0.0220***	0.0022	0.0063***
	[*t*-stat]	[-15.84]	[-15.66]	[-12.31]	[0.40]	[6.51]
Oib(-1,Liquidity)	Estimate	-0.0088***	-0.0183***	-0.0212***	-0.0075	-0.0020
	[*t*-stat]	[-2.61]	[-4.48]	[-3.11]	[-0.82]	[-0.63]
Oib(-1,Overconf)	Estimate	-0.1024***	-0.0407	-0.0658	-0.0151	0.0250
	[*t*-stat]	[-2.84]	[-0.73]	[-1.09]	[-0.30]	[1.17]
Oib(-1,Gamble)	Estimate	-0.0422	-0.0155	-0.0627*	0.0205	0.0047
	[*t*-stat]	[-1.49]	[-0.63]	[-1.94]	[0.93]	[0.23]
Oib(-1,Other)	Estimate	-0.0085***	-0.0082***	-0.0058***	-0.0008***	0.0010***
	[*t*-stat]	[-27.11]	[-24.87]	[-21.68]	[-7.66]	[13.15]
Adj.R2		10.46%	10.32%	10.04%	9.60%	9.51%
Interquartile return						
Oib(-1,Persistence)		-0.1177%	-0.0962%	-0.0591%	-0.0125%	0.0281%
Oib(-1,Liquidity)		-0.0290%	-0.0351%	-0.0261%	0.0091%	0.0097%
Oib(-1,Overconf)		-0.0337%	-0.0475%	-0.0484%	-0.0428%	0.0257%
Oib(-1,Gamble)		-0.0314%	-0.0254%	-0.0271%	-0.0312%	0.0425%
Oib(-1,Other)		-0.1778%	-0.1493%	-0.1008%	-0.0233%	0.0490%

Table V. A Closer Look at the Relation between Investor Order Flows and Public News

This table reports estimation results on the relation of different investor groups order flow and public news in the form of earnings announcements. Our sample period covers January 2016 to June 2019, and our sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 days with trades in the previous month. Panel A reports whether different investor groups' trading activity can predict earnings surprises. For each quarter, we estimate Fama-MacBeth regressions as specified in equation (8) to measure stock returns around the earnings announcement for firm *i* on day *d*. The dependent variable, earnings surprise, is proxied by the cumulative abnormal return from day *d*-1 to day *d*, *CAR[-1,0]*. As independent variables, we use order imbalance measures from day *d*-2, *Oib(-2)*, to avoid overlapping with the CAR calculation. Other control variables are same as those in Table III. Panel B reports whether trades from different retail groups can process contemporaneous news. For each quarter, we estimate Fama-MacBeth regressions as specified in equation (9) to measure investor trading on the earnings announcement day. The dependent variables are order imbalance measures *Oib(0)*. As independent variables, we use the cumulative abnormal return from day *d*-1 to day *d*, *CAR[-1,0]*. Other control variables are same as those in Table III. Panel C reports how earnings news days affect the return predictability of different investor group trades. We estimate Fama MacBeth regressions, as specified in equation (10). The dependent variable is the return on day *d*. The independent variables are the previous day's order imbalance *Oib(-1)*, the news dummies *Event(-1)* and the interaction terms *Oib(-1)*Event(-1).* The *Event(-1)* dummy is equal to 1 if there is earnings announcement for that firm-day and zero otherwise. Other control variables are the same as those in Table III; those coefficients are not reported. To account for serial correlation in the coefficients, the standard errors of the time-series are adjusted using Newey-West (1987) with 4 lags. ***, ** and * indicate significance at the 1%, 5% and 10% level.

Panel A. Investor order flow predicting future earnings announcement news events

Dep.var		CAR[-1,0]	CAR[-1,0]	CAR[-1,0]	CAR[-1,0]	CAR[-1,0]
Oib.var		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5
Oib(-2)	Estimate	-0.0251***	-0.0234***	-0.0166***	-0.0003	0.0023***
	[*t*-stat]	[-7.33]	[-5.38]	[-4.40]	[-0.29]	[3.50]
Adj.R2		6.33%	5.98%	5.57%	5.19%	5.15%

Panel B. Investor order flow regressed on contemporaneous earnings announcement news events

		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5
CAR[-1,0]	Estimate	-1.9225***	-1.8291***	-1.4349***	-0.8781***	0.1583
	[*t*-stat]	[-17.79]	[-16.00]	[-14.34]	[-8.34]	[1.56]
Adj.R2		13.66%	14.60%	10.14%	1.85%	0.60%

Panel C. Return predictive power of investor order flow interacted with earnings announcement news events

Dep.var		Ret OibRT1	Ret OibRT2	Ret OibRT3	Ret OibRT4	Ret OibRT5
Oib(-1)	Estimate	-0.0079***	-0.0070***	-0.0038***	-0.0008*	0.0005**
	[*t*-stat]	[-8.27]	[-7.04]	[-3.55]	[-1.89]	[2.04]
Oib(-1)*Event(-1)	Estimate	-0.0080***	-0.0093***	-0.0071***	-0.0007	0.0014**
	[*t*-stat]	[-3.20]	[-3.15]	[-3.64]	[-0.75]	[2.26]
Event(-1)	Estimate	0.0011**	0.0008*	0.0006	0.0005	0.0005
	[*t*-stat]	[2.57]	[1.89]	[1.59]	[1.41]	[1.50]
Adj.R2		7.36%	7.16%	6.82%	6.38%	6.35%

Table VI. Further Discussion and Robustness

This table reports robustness results. Panel A and Panel B report return prediction across different age and gender investor groups. The sample period covers January 2019 to March 2019. The sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 days with trades during the previous month. Since age and gender are applicable only for retail investors, we only include retail investors. Panel A reports the summary statistics of trading volume across different age and gender groups. Panel B reports return predictions across different gender and age groups. Panel C shows the results of predicting market return using aggregate order imbalances by different investor groups, as specified in equation (12). Panel D and Panel E report return predictions by adding more data filters. The sample period covers January 2016 to June 2019, and the sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 trading days in the previous month. Panel D reports return predictions by including all filters in Liu, Stambaugh and Yuan (2019). Panel E report return predictions by excluding leveraged trading, which consists of margin buys, short sales and collateral trading. Panel F reports return predictions by excluding days that hitting the price limits (+10% and -10%) Panel G reports the Liu, Stambaugh and Yuan (2019) three-factor adjusted alphas for long-short portfolios formed on order imbalances, with holding periods from 1 day to 60 days. Panel H reports how news from CFNDS affects the return predictability of different investor group trades, as specified in equation (10). ***, ** and * indicate significance at the 1%, 5% and 10% level.

Panel A. Summary Statistics of gender and age groups

Gender	Trading Volume (% of total)	
Age	<45	>=45
Male	29%	38%
Female	13%	20%

Panel B. Cross-sectional return predictions, by different gender and age groups

Dep.var		Ret	Ret	Ret	Ret
Gender		Male	Male	Female	Female
Age		<45	>45	<45	>45
Oib(-1)	Estimate	-0.0026***	-0.0060***	0.0007	-0.0002
	[*t*-stat]	[-3.76]	[-4.71]	[1.36]	[-0.30]

Interquartile return	-0.05%	-0.11%	0.02%	0.00%	

Panel C. Cross-sectional return predictions, including all filters from Liu Stambaugh and Yuan (2019)

	Dep.var	Ret	Ret	Ret	Ret	Ret	Ret
	Oib.var	OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibInst
Oib(-1)	Estimate	-0.0075***	-0.0071***	-0.0052***	-0.0011***	0.0004***	0.0017***
	[*t*-stat]	[-24.16]	[-20.81]	[-16.40]	[-7.91]	[4.49]	[17.88]
	Interquartile return	-0.17%	-0.14%	-0.09%	-0.03%	0.02%	0.11%

Panel D. Cross-sectional return predictions, excluding leveraged trading

Dep.var		Ret	Ret	Ret	Ret	Ret	Ret
Oib.var		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibInst
Oib(-1)	Estimate	-0.0092***	-0.0088***	-0.0056***	-0.0003***	0.0008***	0.0016***
	[*t*-stat]	[-24.62]	[-21.97]	[-17.20]	[-3.43]	[11.25]	[20.21]
	Interquartile return	-0.20%	-0.16%	-0.11%	-0.01%	0.05%	0.11%

Panel E. Cross-sectional return predictions, excluding days that hitting the price limits (+10%, and -10%)

Dep.var		Ret	Ret	Ret	Ret	Ret	Ret
Oib.var		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibInst
Oib(-1)	Estimate	-0.0087***	-0.0087***	-0.0064***	-0.0010***	0.0009***	0.0017***
	[*t*-stat]	[-27.92]	[-25.56]	[-21.37]	[-8.90]	[11.76]	[22.46]
	Interquartile return	-0.19%	-0.16%	-0.11%	-0.03%	0.04%	0.11%

Panel F. Liu, Stambaugh, and Yuan (2019) CH3 risk adjusted alphas for long-short portfolios formed on order imbalances over different holding periods

Holding Period	OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST

1 day	-0.0042***	-0.0036***	-0.0027***	-0.0007***	0.0017***	0.0025***
1 week	-0.0089***	-0.0068***	-0.0038***	-0.0004	0.0034***	0.0056***
2 weeks	-0.0115***	-0.0091***	-0.0048***	-0.0001	0.0046***	0.0075***
4 weeks	-0.0130***	-0.0104***	-0.0052***	0.0003	0.0056***	0.0098***
6 weeks	-0.0148***	-0.011***	-0.0056***	0.0005	0.0064***	0.0103***
8 weeks	-0.0183***	-0.0136***	-0.0069***	0.0001	0.0063***	0.0128***
10 weeks	-0.0187***	-0.0136***	-0.0063***	0.0005	0.0065***	0.0133***
12 weeks	-0.0183***	-0.0139***	-0.0072***	-0.0010	0.0057***	0.0137***

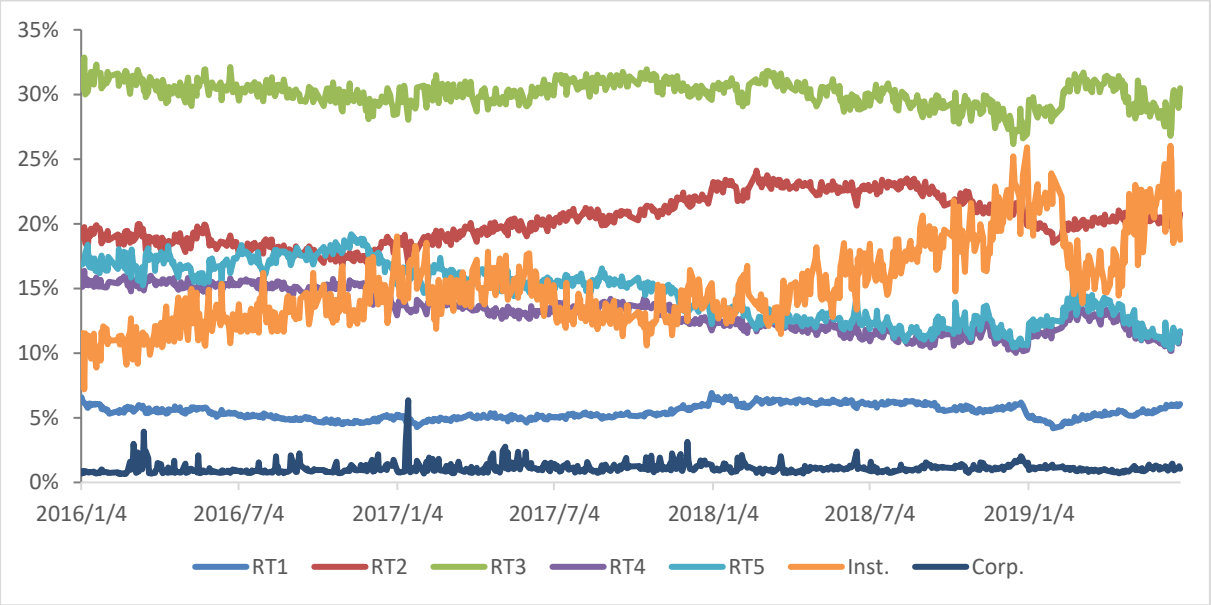
Panel G. Return predictive power of investor order flows interacted with CNFD news events

Dep.var		Ret	Ret	Ret	Ret	Ret	Ret
Oib.var		OibRT1	OibRT2	OibRT3	OibRT4	OibRT5	OibINST
Oib(-1)	Estimate	-0.0075***	-0.0070***	-0.0047***	-0.0005***	0.0008***	0.0014***
	[*t*-stat]	[-23.61]	[-21.16]	[-16.21]	[-3.81]	[9.82]	[17.27]
Oib(-1)*Event(-1)	Estimate	-0.0045***	-0.0053***	-0.0048***	-0.0013***	0.0011***	0.0007***
	[*t*-stat]	[-11.02]	[-11.92]	[-10.79]	[-6.28]	[6.45]	[6.32]
Event(-1)	Estimate	0.0002**	0.0001*	0.0001	0.0001	0.0000	0.0001
	[*t*-stat]	[2.24]	[1.83]	[1.17]	[0.72]	[0.37]	[1.20]
Adj.R2		9.01%	8.86%	8.58%	8.19%	8.21%	8.34%

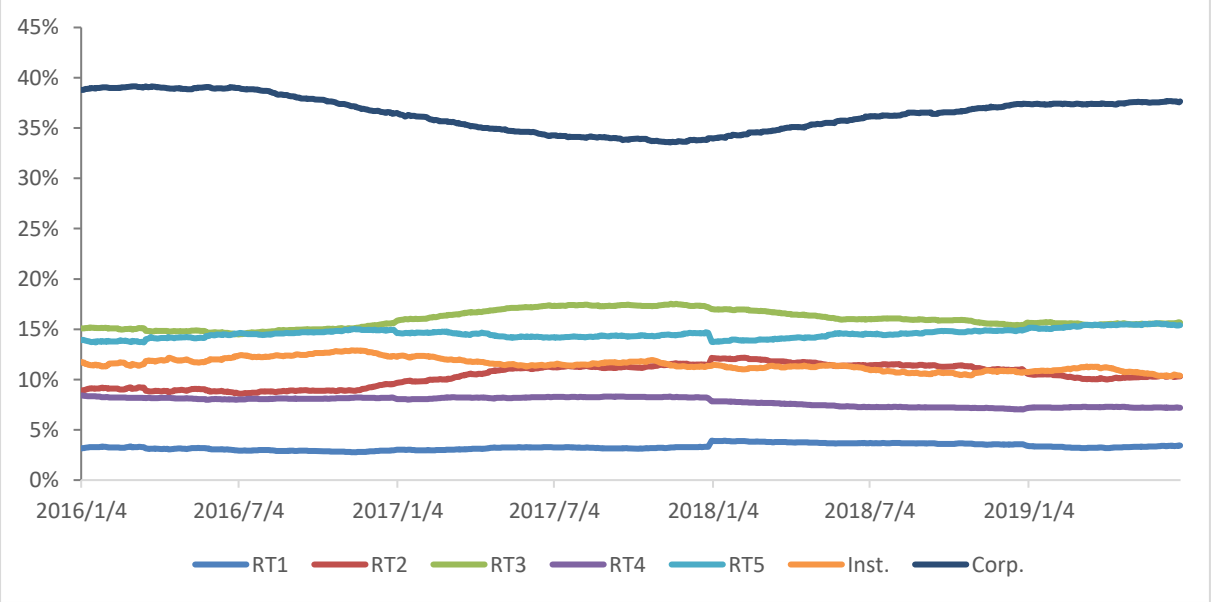
Figure I. Different Investor Type Order Flows between Jan 2016 and Jun 2019

There figures report the time series plot of the cross sectional mean for different types of investor trading activity from January 2016 to June 2019. Our sample firms are A-share stocks listed on the Shanghai Stock Exchange. In Panel A, we present the volume percentage by each type of investor. In Panel B, we show the shares held by each type of investor.

Panel A. Share volume (%), Cross Sectional Mean for Different Investor Types



Panel B. Shares Held (%), Cross Sectional Mean for Different Investor Types



Appendix Table I. Distributions of Investor Trading and Holdings

This table reports summary statistics for trading and holdings by different investor groups by different stock characteristics. Our sample period covers January 2016 to June 2019, and our sample firms are A-share stocks listed on the Shanghai Stock Exchange with at least 15 days with trades during the previous month. Panel A-C reports the time series average of the trading and holding volume statistics by different investor groups across stock characteristics. Panel D and E reports the bottom 2 and top 2 sectors, in terms of trading volume and holdings within each investor groups. We classify sectors by Datastream Level 4 sector classifications. Panel E reports the time series average of trading volume by different investor groups across different trade size.

Panel A. Volumes across stock characteristics

	RT1	RT2	RT3	RT4	RT5	INST	CORP
Small	6.3%	23.3%	32.5%	13.4%	14.0%	9.9%	0.7%
Medium	5.8%	21.2%	31.3%	13.3%	13.9%	13.7%	0.8%
Large	4.3%	16.3%	26.7%	12.9%	15.7%	22.4%	1.8%
Low EP	6.3%	22.0%	31.4%	13.7%	15.3%	10.5%	0.8%
Medium EP	5.6%	21.2%	30.7%	12.9%	13.8%	14.9%	0.9%
High EP	4.3%	17.1%	28.0%	13.1%	14.7%	21.2%	1.6%
Low Turnover	4.7%	17.5%	28.0%	12.9%	14.5%	20.7%	1.7%
Medium Turnover	5.2%	19.4%	29.9%	13.5%	15.0%	15.8%	1.1%
High Turnover	6.3%	23.4%	32.1%	13.2%	14.2%	10.1%	0.6%

Panel B. Holdings across stock characteristics

	RT1	RT2	RT3	RT4	RT5	INST	CORP
Small	4.1%	13.4%	20.2%	9.9%	17.4%	7.1%	27.9%
Medium	3.6%	10.8%	16.5%	8.0%	15.4%	11.1%	34.6%
Large	2.2%	6.6%	10.6%	5.3%	10.1%	16.7%	48.6%
Low EP	3.9%	11.2%	16.9%	8.5%	15.5%	8.2%	35.7%
Medium EP	3.4%	11.3%	17.3%	8.2%	15.6%	12.6%	31.6%
High EP	2.4%	7.3%	12.0%	6.2%	11.8%	14.5%	45.8%
Low Turnover	2.2%	5.8%	9.3%	4.9%	11.2%	11.9%	54.6%
Medium Turnover	3.1%	9.0%	14.7%	7.8%	15.5%	12.9%	37.0%
High Turnover	4.3%	15.0%	22.2%	10.1%	16.2%	10.5%	21.6%

Panel C. Sectors with lowest and highest trading volumes

	RT1	RT2	RT3	RT4	RT5	INST	CORP
Bottom1	**Banks & Life Insurance**	**Banks & Life Insurance**	**Banks & Life Insurance**	**Banks & Life Insurance**	**Alternative Energy**	**Alternative Energy**	**Alternative Energy**
	2.4%	10.5%	20.4%	11.9%	12.2%	6.5%	0.4%
Bottom2	Financial Services	Financial Services	Beverages	Health Care Equipment & Services	Gas, Water and Multiutilities	Industrial Metals & Mining	General Industrials
	3.6%	15.8%	27.4%	11.9%	12.3%	11.6%	0.7%
Top2	Industrial Metals & Mining	Industrial Engineering	Gas, Water and Multiutilities	Technology Hardware & Equipment	Software & Computer Services	Travel & Leisure	Financial Services
	6.5%	22.3%	32.0%	14.5%	17.7%	21.3%	2.9%
Top1	**Alternative Energy**	**Alternative Energy**	**Alternative Energy**	**Financial Services**	**Banks & Life Insurance**	**Banks & Life Insurance**	**Banks & Life Insurance**
	7.6%	25.5%	34.2%	14.5%	18.0%	32.8%	4.1%

Panel D. Sectors with highest and lowest holding volumes

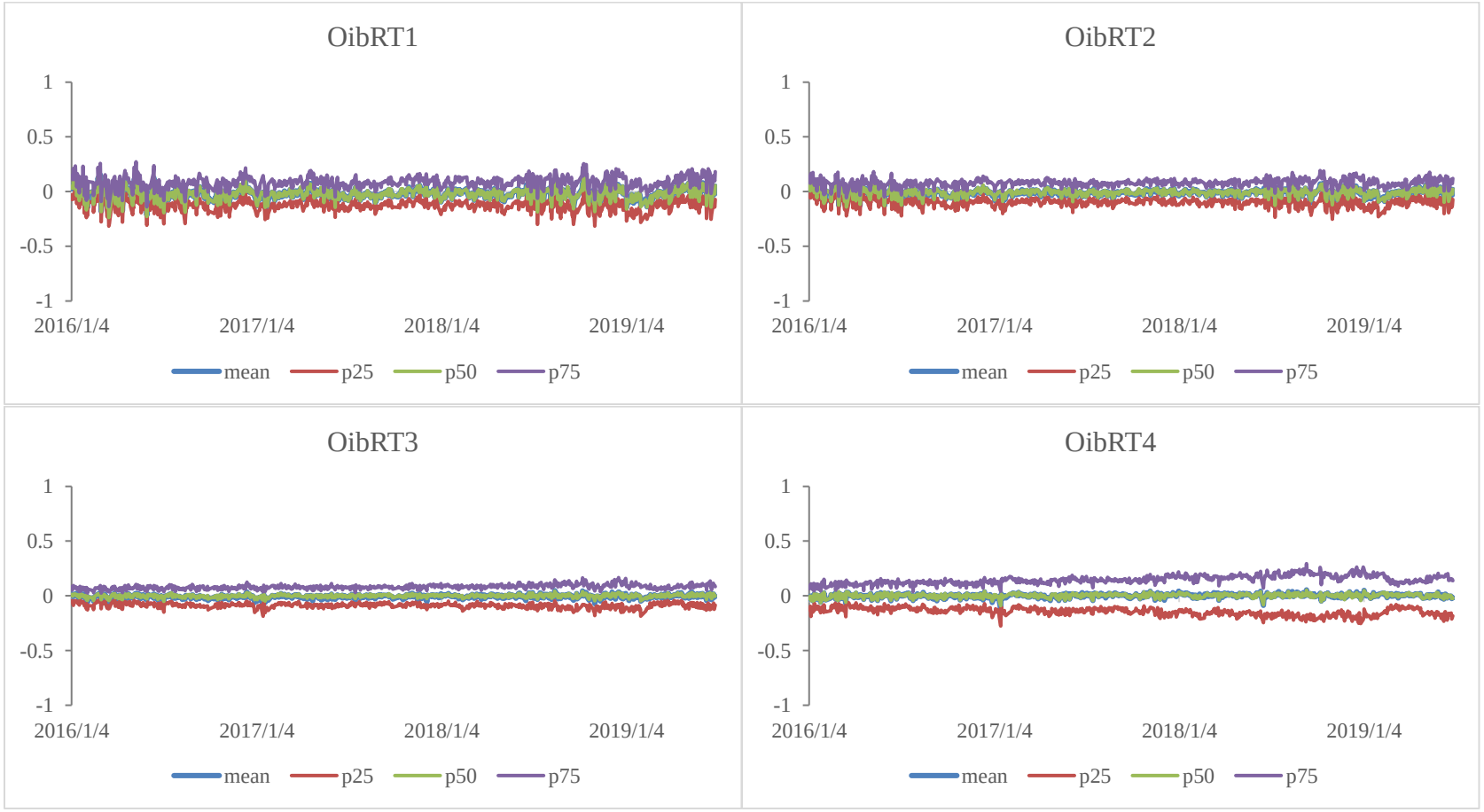
	RT1	RT2	RT3	RT4	RT5	INST	CORP
Bottom1	**Banks & Life Insurance**	**Banks & Life Insurance**	**Banks & Life Insurance**	**Banks & Life Insurance**	**Banks & Life Insurance**	**Alternative Energy**	**Household Goods & Home Construction**
	1.1%	4.0%	7.6%	4.2%	5.6%	2.9%	19.2%
Bottom2	Aerospace & Defense	Aerospace & Defense	Food & Drug Retailers	Industrial Transportation	Electricity	Mobile Telecommunications	Health Care Equipment & Services
	1.9%	7.2%	12.1%	5.6%	7.3%	4.8%	23.2%
Top2	Forestry & Paper	Household Goods & Home Construction	Support Services	Support Services	Support Services	Banks & Life Insurance	Banks & Life Insurance
	4.5%	12.8%	19.2%	9.7%	19.5%	23.3%	54.3%
Top1	**Alternative Energy**	**Alternative Energy**	**Alternative Energy**	**Forestry & Paper**	**Software & Computer Services**	**Health Care Equipment & Services**	**Mobile Telecommunications**
	6.1%	15.5%	20.9%	9.8%	23.8%	23.5%	55.1%

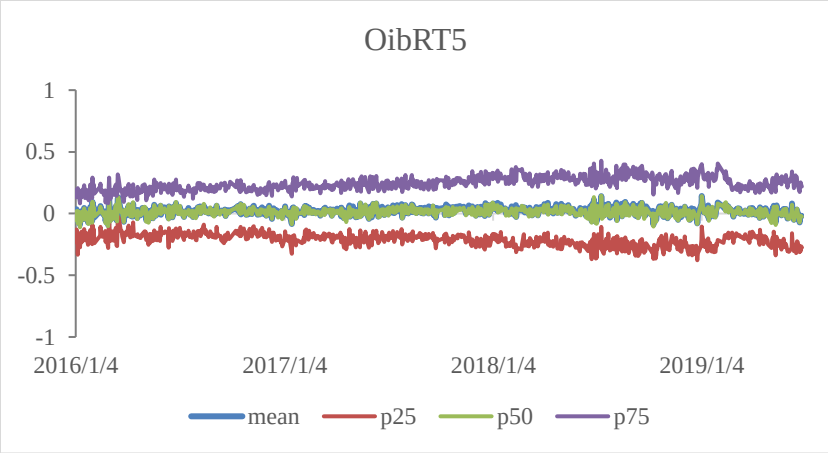
Panel E. Volumes across different trade sizes

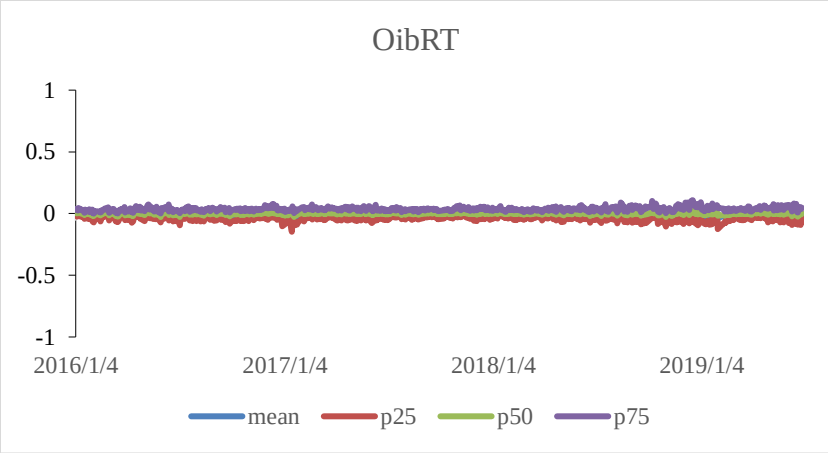
Trade Size	RT1	RT2	RT3	RT4	RT5	INST	CORP
<40,000 CNY	4.78%	10.86%	7.44%	1.22%	0.74%	5.35%	0.10%
40,000-200,000 CNY	0.69%	8.47%	14.54%	5.04%	3.65%	4.36%	0.27%
>200,000 CNY	0.00%	1.03%	8.19%	6.96%	10.30%	5.48%	0.73%

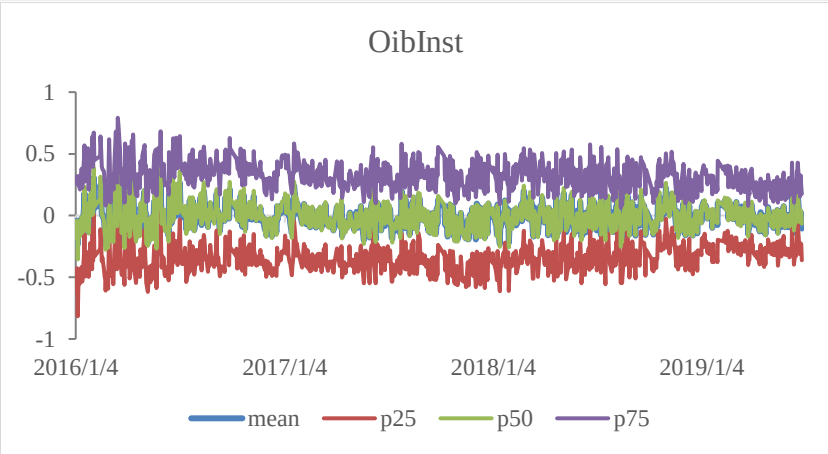
Appendix Figure I. Time Series of Different Types of Investor Order Imbalance

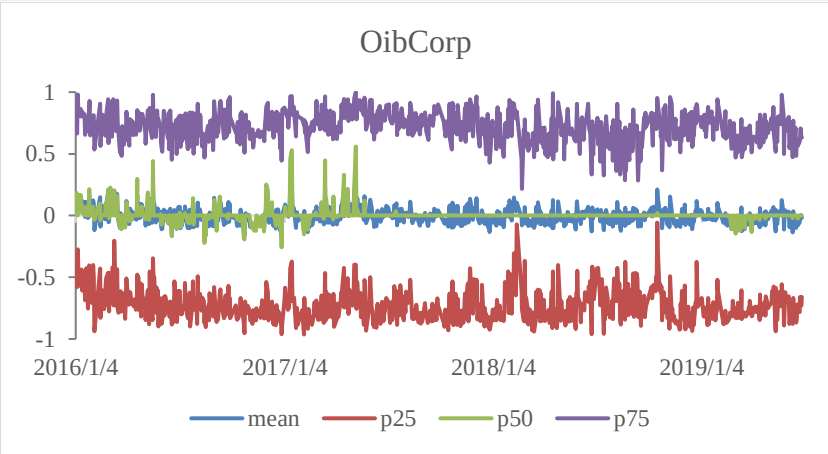
These figures reports time series of different types of investor trading activity. Our sample period covers January 2016 to June 2019, and our firms are A-share stocks listed on the Shanghai Stock Exchange. We present the cross-sectional mean, median, 25th percentiles and 75th percentiles of scaled daily order imbalances by each investor group each day. Order imbalance is computed as the buy share volume minus sell share volume divided by buy share volume plus sell share volume for each investor group, specified in Equation (1).




OibRT5
1
0.5
0
-0.5
-1
2016/1/4
2017/1/4
2018/1/4
2019/1/4
mean
p25
p50
p75


OibRT
1
0.5
0
-0.5
-1
2016/1/4
2017/1/4
2018/1/4
2019/1/4
mean
p25
p50
p75


OibInst
1
0.5
0
-0.5
-1
2016/1/4
2017/1/4
2018/1/4
2019/1/4
mean
p25
p50
p75


OibCorp
1
0.5
0
-0.5
-1
2016/1/4
2017/1/4
2018/1/4
2019/1/4
mean
p25
p50
p75

Appendix. Literature review of studies on retail investors in different markets

Markets	Title	Authors	Journal	Data	Research Questions and Main Findings
U.S.	Individual investor trading and stock returns	Kaniel, Saar, and Titman	Journal of Finance, 2008	NYSE CAUD file	1. Research Question: the relation between net individual investor trading and short-horizon cross-sectional stock returns 2. Main Findings: (1) Individuals buy stocks following declines in the previous month and sell following price increases. (2) Positive excess returns in the month following intense buying by individuals and negative excess returns after individuals sell.
U.S.	Do retail trades move markets?	Barber, Odean, and Zhu	Review of Financial Studies, 2008	TAQ trade size	1. Research Question: the trading of individual investors and stock returns 2. Main Findings: (1) Small trade order imbalance correlates well with order imbalance based on trades from retail brokers. (2) Retail investors herd. (3) Small trade order imbalance negatively forecasts future annual returns (4) Small trade order imbalance positively predicts future week returns
U.S.	How wise are crowds? Insights from retail orders and stock returns	Kelley and Tetlock	Journal of Finance, 2013	Retail brokerages	1. Research question: the role of retail investors in stock pricing by separately examine aggressive (market) and passive (limit) orders. 2. Main Findings: (1) Both market and limit order imbalance positively predict firms' monthly stock returns. (2) Market orders correctly predict firm news, including earnings surprises. (3) Limit orders following negative returns, consistent with traders providing liquidity.

U.S.	Tracking retail investor activity	Boehmer, Jones, Zhang and Zhang	Journal of Finance, 2021	TAQ price improvement algorithm	1. Research Question: Provide an algorithm to identify marketable retail purchases and sales using TAQ. 2. Main Findings: (1) Validate the algorithm. (2) Marketable retail order imbalance positive predict future weekly stock return. (3) Predictive power of marketable retail order imbalance could be attributable to order flow persistence, contrarian trading, public news sentiment and unexplained part.
U.S.	Resolving a paradox: Retail trades positively predict returns but are not profitable	Barber, Lin, and Odean	Working Paper	TAQ price improvement algorithm	1. Research Question: Retail order imbalance positively predicts returns, but in aggregate retail investor trades lose money. 2. Main Findings: (1) Order imbalance tests equally weight stocks, but retail purchases concentrate in stocks that subsequently underperform. (2) Trades by retail investors with less knowledge, experience, and wealth are more likely to underperform.
U.S.	The Wisdom of the Robinhood Crowd	Welch	Journal of Finance, forthcoming	Robinhood investors	1. Research Question: Robinhood investors trading behavior 2. Main Findings: (1) Robinhood investors increased their holdings in the March 2020 COVID bear market. (2) Robinhood investors tend to buy stocks with high past share volume and dollar-trading volume. (3) From mid-2018 to mid-2020, an aggregated Robinhood portfolio had both good timing and good alpha.
U.S.	Flattening the illiquidity curve: Retail trading during the COVID-19 lockdown.	Ozik, Sadka, and Shen	Journal of Financial and Quantitative Analysis, 2021	Robinhood investors	1. Research Question: The impact of retail investors on stock liquidity during the Coronavirus pandemic lockdown. 2. Main Findings: (1) Retail trading exhibits a sharp increase during Covid. (2) Retail trading attenuated the rise in illiquidity by roughly 40%, but less so for high-media-attention stocks.

Finland	The investment behavior and performance of various investor types: a study of Finland's unique data set	Grinblatt and Keloharju	Journal of Financial Economics, 2000	Finnish Central Securities Depository (FCSD)	1. Research Question: the extent to which past returns determine different types of investors' propensity to buy and sell. 2. Main Findings: (1) Foreign investors tend to be momentum, Domestic investors, particularly households, tend to be contrarians. (2) The portfolios of foreign investors seem to outperform the portfolios of households, even after controlling for behavior differences.
Finland	Do Limit Orders Alter Inferences about Investor Performance and Behavior?	Linnainmaa	Journal of Finance, 2011	Finnish Central Securities Depository (FCSD) registry	1. Research Question: Individual investors' trading behaviors' could be explained by investors' use of limit orders. 2. Main Findings: These patterns arise mechanically because limit orders are price-contingent and suffer from adverse selection.
Finland	IQ, trading behavior, and performance	Grinblatt, Keloharju, Linnainmaa	Journal of Financial Economics, 2012	Finnish Central Securities Depository (FCSD) registry and intelligence (IQ) test administered to Finnish male	1. Research Question: Whether IQ influences trading behavior. 2. Main Findings: (1) High-IQ investors are less subject to the behavior biases (2) High-IQ investors exhibit superior market timing, stock-picking skill, and trade execution.

Sweden	Rich pickings? risk, return, and skill in household wealth	Bach, Calvet and Sodini	American Economic Review, 2020	Administrative panel containing the full balance sheet of every Swedish resident between 2000 and 2007 (Annually)	1. Research Question: examine the wealth returns on balance sheets of Swedish residents. 2. Main Findings: (1) The expected return on household net wealth increases with net worth (2) The expected wealth return is driven by systematic risk-taking and exhibits strong persistence. Idiosyncratic risk is transitory but sufficiently large to generate substantial long-term dispersion in returns. (3) Heterogeneity in returns explains most of the historical increase in top wealth shares.
German	Correlated Trading and Returns	Dorn, Huberman, Sengmueller	Journal of Finance, 2008	One of the three largest German discount brokers, 37,000 clients	1. Research Question: Investors correlated trading and stock returns 2. Main Findings: (1) Investors tend to be on the same side of the market. (2) Correlated market orders lead returns due to persistent price pressure. (3) Correlated limit orders also predict subsequent returns, consistent with liquidity demands.
France	Are retail traders compensated for providing liquidity?	Barrot, Kaniel, and Sraer	Journal of Financial Economics, 2016	A leading European broker	1. Research Question: Whether individual investors provide liquidity to the stock market and whether they are compensated. 2. Main Findings: (1) The ability of aggregate retail order imbalances to predict short-term future returns is significantly enhanced during times of market stress. (2) Individual investors do not reap the rewards from liquidity provision because they experience a negative return on the trading day and reverse their trades for too long time after the liquidity provision dissipated.

Australia	Individual Investors and Broker Types	Fong, Gallagher, and Lee	Journal of Financial and Quantitative Analysis, 2014	Australian Securities Exchange (ASX) SIRCA	1. Research Question: examine the informativeness of trades via discount and full-service retail brokers. 2. Main Findings: (1) Trades via full-service retail brokers are more informative than are trades via discount retail brokers. (2) Past returns, volatility, and news announcements could explain the net volume of discount retail brokers but could not explain the net volume of full-service retail brokers.
Taiwan, China	Just how much do individual investors lose by trading?	Barber, Lee, Liu, and Odean	Review of Financial Studies, 2009	Taiwan stock exchange	1. Research Question: How much do individual investors lose by trading? 2. Main Findings: (1) Individual investor losses are equivalent to 2.2% of Taiwan's GDP. (2) Nearly all individual trading losses can be traced to their aggressive orders.
Taiwan, China	The cross-section of speculator skill: Evidence from day trading	Barber, Lee, Liu, and Odean	Journal of Financial Markets, 2014	Taiwan stock exchange	1. Research Question: examine the cross-sectional differences of returns earned by speculative day traders. 2. Main Findings: Less than 1% of the day trader population is able to predictably and reliably earn positive abnormal returns net of fees.
India	Who Owns What? A Factor Model for Direct Stock Holding	Balasubramaniam, Campbell, Ramadorai and Ranish	NBER 2021	10 million retail accounts in Indian stock market	1. Research Question: Build a cross-sectional factor model for retail investors' direct stock holdings 2. Main Findings: (1) Stock characteristics such as firm age and share price and account attributes such as account age, account size, and extreme under diversification have strong investor clienteles. (2) Coheld stocks have higher return covariance
India	Learning from noise: Evidence from India's IPO lotteries	Anagol, Balasubramaniam, and Ramadorai	Journal of Financial Economics, 2021	1.5 million investors participate in allocation lotteries for IPO stocks.	1. Research Question: Retail investors participate in allocation lotteries for Indian IPO stocks 2. Main Findings: Investors obtain IPO stocks that rise in value increase portfolio trading volume in non-IPO stocks relative to lottery losers. A learning model could explain the results.

China	Corporate actions and the manipulation of retail investors in China: An analysis of stock splits	Titman, Wei, and Zhao	Journal of Financial Economics, 2022	Shanghai Stock Exchange	1. Research Question: Corporate actions and the manipulation of retail investors in the stock splits events. 2. Main Findings: (1) Share prices temporarily increase after splits, and subsequently decline below their presplit levels. (2) Small retail investors buy shares in firms initiating suspicious splits, while more sophisticated investors buy before suspicious split announcements and sell in the postsplit period. (3) Insiders sell large blocks of shares and obtain loans using company stock as collateral before the suspicious splits.
China	Wealth redistribution in bubbles and crashes	An, Lou, and Shi	Journal of Monetary Economics, 2022	Shanghai Stock Exchange, 2014-15 bubbles and crash episode	1. Research Questions: What are the social-economic consequences of financial market bubbles and crashes? 2. Main Findings: The largest 0.5% households in the equity market gain, while the bottom 85% lose, 250B RMB through active trading in this period.
China	Extrapolative bubbles and trading volume	Liao, Peng, Zhu	Review of Financial Studies, 2021	one of the largest brokerage firms, account-level transaction data on the 2014–2015	1. Research Questions: Propose the extrapolative model to explain the sharp rise in prices and volume during financial bubbles. 2. Main Findings: (1) The model propose a novel mechanism for volume: because of extrapolative beliefs and disposition effects, investors are quick to buy assets with positive past returns and sell them if good returns continue. (2) Use Chinese account-level data to confirm the model's predictions.
China	Daily Price Limits and Destructive Market Behavior	Chen, Gao, He, Jiang, Xiong	Journal of Econometrics, 2019	Shenzhen Stock Exchange	1. Research Question: Daily price limits and investors trading behavior 2. Main Findings: Large investors tend to buy on the day when a stock hits the 10% upper price limit and then sell on the next day; and their net buying on the limit-hitting day predicts stronger long-run price reversal.

China	Do wealthy investors have an informational advantage? Evidence based on account classifications of individual investors	Li, Geng, Subrahmanyam and Yu	Journal of Empirical Finance, 2017	A brokerage firm in China provide one million investors' trading records from January 2007 to October 2009.	1. Research Question: Do wealthy investors have an informational advantage? 2. Main Findings: (1) Wealthy investors with portfolio values above the 99.5th percentile ("super" investors) outperform all other investors. (2) Part of their excess returns could be explained by informational advantages. These super investors profitably trade around companies' announcements of high stock dividends, particularly those registered in these super investors' localities.