Three Essays on Finance

by

Maria Andreea Vaduva

A dissertation submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy in

Business and Finance

Universidad Carlos III de Madrid

Advisor:

José Sebastian Penalva Zuasti

Tutor:

José Sebastian Penalva Zuasti

July, 2023

Acknowledgements

The PhD. road has been long and, in parts, an arduous one. Along the way, I discovered, learned and faced plenty of challenges. I am grateful for every moment lived during this journey, as it was an entirely enriching experience from a professional and personal point of view.

I would like to first thank to my advisor, tutor, Prof. Dr. José Sebastian Penalva Zusati. Words cannot express my gratitude for his patience, feedback, knowledge, and care throughout the Ph.D.

I would also like to thank to Prof. Dr. Sophie Moinas, my visiting sponsor and mentor at the Toulouse School of Management and Toulouse School of Economics for the effort, time dedicated to my work and great suggestions. A special thank to Prof. Dr. José Maria Marin Vigueras, who provided me with valuable knowledge, expertise, great suggestions and support.

I am also grateful to my committee: Prof. Dr. Pablo Ruiz Verdú, Prof. Dr. Jesper Rüdiger and Prof. Dr. Pedro Serrano for their constructive feedback and support.

But all this would not have been the same without the time spent with Jess over a glass of wine, the adventures experienced with Tan discovering new places, the nice chats with Stavri and with my flatmate Juliana. I will miss the time spent with all the colleges from the PhD Program.

I am deeply grateful to my family for their support, love, their efforts to make me happy and for giving me the freedom to choose my own path.

Published and submitted content

Chapter 1 was submitted to and accepted for presentation at the following conferences:

- The initial proposal was submitted to Universidad Carlos III de Madrid as the Master thesis, 2018.
- Finance Workshop TSE: 2020-2021 (21/05/2021) https://sites.google.com/site/financeworkshoptse/2020-2021
- Finance Workshop TSE: 2018-2019 (12.04.2019) https://sites.google.com/site/financeworkshoptse/2018-2019

The paper: Non-Standard Errors, which constitutes the second part of Chapter 2:

- Was submitted to and accepted for presentation at the following conferences:
  - Microstructure Exchange, 2021. (https://microstructure.exchange/)
  - Derivatives Forum Frankfurt, 2022. (https://www.eurex.com/ex-en/find/forum/frankfurt-2022)
  - Financial Intermediation Research Society (FIRS), 2022. (https://firsociety.org/wp-content/uploads/2022/06/firs2022.conferenceprogram.pdf)
  - Research in Behavioral Finance Conference (RBFC), 2022. (https://static1.squarespace.com/static/566856db69492e8025183f94/t/631737b4c5e6bd76b1b0d67c/1662465974448/RBFC+2022+program+booklet.pdf)

- Society for Experimental Finance (SEF), 2022.
- Society for Financial Econometrics (SoFiE), 2022. (https://www.janeway.econ.cam.ac.uk/events/2022/sofie2022/SoFiE-programme-summary-22jun22b.pdf) where the paper was runner-up for the best-paper prize (https://albertjmenkveld.com/text/sofie_2022_runner_up_best_paper_prize.pdf)
- Vienna-Copenhagen Conference on Financial Econometrics, 2022. (https://eventsignup.ku.dk/vieco2022/program)
- Western Finance Assocation (WFA), 2022. (https://westernfinance-portal.org/program/2022/WFA.2022.program.preconf.pdf)

• The original version of *Non-Standard Errors* contains the results of the analysis outlined in the pre-analysis plan. This original version is available as Tinbergen Institute Discussion Paper TI 2021-102/IV.

• Have been submitted to SSRN webpage. (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3961574)

• Have been submitted and accepted/in press - 15 Feb 2023 in the Journal of Finance. (forthcoming Journal of Finance https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3961574)

• Have been submitted in other webpages by the other co-authors of the project. (For instance: https://www.unige.ch/gfri/application/files/8016/7719/0235/SSRN-id3961574.pdf)

• The online appendix is available in: https://www.albertjmenkveld.com/text/fincap-paper-online-appendix.pdf

Contents

Abstract

This thesis is comprised of three chapters. The main objective of Chapter 1 is to assess empirically the impact an increase in the adverse selection risk has on competition between passive traders in different exchanges in the US. stock market (inter-market competition). By using macroeocnomic news as an exogenous shock, I show that increases in adverse selection risk is associated with a decrease in inter-market competition. Around macroeconomic news, despite the increase in adverse selection risk, passive traders take advantage and earn larger rents than during intra-days with no macroeconomic news. Furthermore, to the best of my knowledge this is the first paper to study the impact an increase in fragmentation has on inter-market competition as a function of adverse selection risk. In line with this, I show that the impact an increase in fragmentation has on inter-market competition is context-dependent.

Chapter 2 is written in the context of the Finance Crowed Analyses Project (FINCAP). With this aim, FINCAP launched a project to understand the mechanism behind differences in results across different Research Teams (RTs). In line with this, all RTs had to test the same hypothesis using the same data, and finally submit a short paper with the results (estimates) obtained. The dispersion in estimates across researchers (non-standard error) is the object of study of FINCAP. I participated in the project together with Sophie Moinas as a Research Team (RT). With this aim, we had to test 6 hypothesis related to market quality. The short paper written together with Sophie Moinas and submitted in the final stage of the experiment constitutes the first part of Chapter 2. The estimates provided by all RTs are used as an input in the final project Non-Standard Errors to study what drives differences in estimates across researchers. The final conclusions of the experiment are specified in the paper: Non-Standard Errors (forthcoming in the Journal of Finance), which constitutes the second part of Chapter 2. The paper Non-Standard Errors is

coauthored with researchers affiliated to a total of 207 institutions listed in Annex 1.

In the third chapter, a joint work with José Maria Marin Vigueras, we test performance evaluation in the context of the Rational Expectations Equilibrium (REE) Paradigm. Based on the extremely positive results obtained, we develop a strongly micro-founded new measure of performance evaluation which we call *Informed Alpha* that beats the standard Jensen alpha measure of performance. We show that *Informed Alpha* sorts truly talented managers who exhibit strong persistence in their performance.

Chapter 1

Inter-market competition in a highly fragmented market around macroeconomic news

1.1 Introduction

Passive traders maximize revenues by maximizing the probability of trading against liquidity traders. On the contrary, their revenues may be harmed if they trade against informed traders. If the probability of being adversely selected increases, then they may increase the spread to compensate from the losses of trading against an informed trader (Glosten and Milgrom (1985)). Competition between passive traders in different exchanges in the US. is governed since 2007 by Rule 611, which lies at the core of Reg NMS. It states that orders should be executed at best prices across all exchanges, at the National Best Bid or Offer (NBBO), with some exemptions, such as Intermarket Sweep Orders (ISOs), or Immediate or Cancel Orders (IOC).[1]

When adverse selection risk is low, Rule 611 creates incentives for passive traders to be at the NBBO, as it increases the probability of getting executed. This in turn may lead to an increase in competition between passive traders in different exchanges (from now on: inter-market competition). An increase in inter-market competition may reduce trading costs for aggressive traders. On the contrary, when adverse selection risk is high, despite the main objective of Rule 611, I expect passive traders in different exchanges to withdraw from the market or post quotes further away from the NBBO, leading to a decrease in inter-market competition. This is the hypothesis I want to test in this paper.

In line with the previous hypothesis, the main objectives of the current paper is to go a step further and assess empirically, the role an increase in adverse selection risk has on: i) inter-market competition; ii) on within exchange competition; and iii) on the rents passive traders extract. With the introduction of rule 611, many exchanges emerged leading to a highly complex fragmented market. Given the highly fragmented market, it is worth exploring if adverse selection risk is exacerbated by an increase in the number of exchanges.

Adverse selection risk in the standard framework refers mainly to asymmetric information

[1] If an exchange receives a market orders (aggressive order) to buy (sell), however all the limit orders (passive orders) are posted away from the NBB (NBO), then the exchange should reject the order, as trade-throughs are prohibited). In other words, Rule 611 is "forcing" passive traders in different exchanges to post limit orders at best prices if they want to get executed or increase their probabilities of getting executed.

across traders (Copeland and Galai (1983); Glosten and Milgrom (1985)). Recently, adverse selection risk has been acknowledged to arise also as a consequence of the inability of passive traders to update their stale quotes (the risk of being picked-off), with the arrival of public information in the market in the fashion of Budish et al. (2015) and Aıt-Sahalia and Sa (2017).

The release of public information is one situation in which adverse selection risk may increase. Adverse selection risk may increase at announcement due to: the endogenous increase in the participation of informed traders Crego (2020)[2]; and/or an increase in the risk faced by passive traders of being picked off. Therefore, the release of macroeconomic news offers the perfect setting to study the impact an increase adverse selection risk has on inter-market competition.

My empirical analyses is done using TAQ data over 2015 to 2017 for SPDR S&P500 (SPY): the exchange-traded fund. In order to capture changes in the level of inter-market competition, I calculate the proportion of exchanges at the NBBO (Quote Clustering). Lower Quote Clustering should suggest a higher adverse selection risk in the market, as passive traders prefer to stay away from the NBBO and free ride on those posting quotes at the NBBO just as in Foucault et al. (2003).

Quote Clustering may be interpreted as a measure of i) the ex-ante risk assessed by passive traders of not being adversely selected; or ii) indirectly, as the probability of finding a liquid exchange in the same way as in Biais et al. (2015). A liquid exchange refers to an exchange at the NBB (NBO). The probability of finding a liquid exchange is especially relevant for slow aggressive traders, as brokers have the obligation to route their orders to exchanges at the NBB (NBO). If there is a lower number of exchanges at the NBBO, there are more probabilities that order have to be re-routed, which in turn will increase explicit trading costs for slow aggressive traders. Some exchanges charge the maximum fee allowed of 0.30 cents per share to route orders to other exchanges (i.e. NYSE Li et al. (2023)).

For purposes of illustration, in Figure 1.1, in Panel A the average (time weighted) Quote Clustering within 1 minute interval has been plotted from 13:00 to 15:00, while in Panel B, I zoom in and plot the average (time weighted) Quote Clustering within 10 seconds interval from 13:50 to 14:10. The majority of the news released at 14:00 are news released by the Fed

[2]The release of public information may increase asymmetric information if informed traders are risk averse (Crego (2020))

(Federal Reserve Board) with a considerable impact on asset prices. The blue line represents Quote Clustering on days with news released at 14:00, while the gray line on days with no news. Around news released at 14:00, Quote Clustering decreases from 65% to less than 50% on average (See Panel A). By zooming in to a higher frequency, Quote Clustering reaches the value of 42% within the first 10 seconds after the release of macroeconomic news (See Panel B), which points out that adverse selection risk reaches the highest level at announcement which may be correlated with a high risk faced by passive traders of being picked-off.

I first document how Quote Clustering behaves around announcements. To do this, independent regressions are performed for intervals of time around the event time 0. First, I consider all the news released at different points in time within the trading day (Unrestricted Sample), and second, I restrict the sample to intra-days around 14:00 (Restricted Sample). I find that Quote Clustering experiences a significant decrease up to 5 minutes before and 15 minutes after the release of news. The drop reaches the highest level during the first minute after the release of news, with a drop of 6.2%. The magnitude of the decrease is larger when I restrict the sample to news released at 14:00, with a drop of 18.2%.

At the root of the previous results, one question that arises is if the decrease in liquidity is systematic across all exchanges, or liquidity concentrates in some exchanges at the expense of the other ones. By calculating the unconditional probability of being at the NBBO for each exchange within one minute interval, I show that during the first minute after the release of news the measure decreases on average by 8% as compared to intra-days with no news, and the decrease is systematic across all exchanges. By restricting the sample to intra-days at 14:00, the decrease in the probability of being at the NBBO is systematic and is on average of 18%.

The lower inter-market competition together with the higher adverse selection risk lead to ask the following questions: do passive traders manage to protect themselves against the high adverse selection risk? What role does inter-market competition play in this? An increase in adverse selection risk around news may lead to a decrease in competition between passive traders, which may lead passive traders to earn strictly positive rents Foucault et al. (2003). The decrease in competition may be attributed to some dealers monitoring only the market and free riding on those monitoring both the market and the news.

By calculating the realized spread of passive traders as a proxy for their rents, I find that passive traders manage to earn larger rents during intra-days with a high adverse selection risk in contrast to intra-days with a low adverse selection risk. The results reflect passive traders' ability to take advantage of pre-scheduled news. In particular, if passive traders unwind their positions within the first 0.1 seconds, the realized spread is larger by $2 ($3.5) per 100 shares traded for the Unrestricted (Restricted) sample. Furthermore, I show that inter-market competition may play a key role on the rents extracted by passive traders. The lower inter-market competition contributes to an increase in the rents extracted by passive traders around intra-days with news in contrast to intra-days with no news.

In addition, I show that passive traders usually monitor the market. Furthermore, with the release of public information, characterized by a high adverse selection risk and when speed is especially important, changes in the proportion of exchanges at the NBBO is informative and relevant for passive traders free riding on those monitoring both the market and the news.

Finally, I address the impact an increase in fragmentation may have on inter-market competition. The increase in adverse selection risk may be exacerbated by an increase in fragmentation, as the proportion of fast traders to liquidity traders changes to the detriment of passive trades, as fast traders will trade against all the active quotes in the market Baldauf and Mollner (2021). By using the entrance of Investors Exchange (IEX), I find that an increase in fragmentation leads to a slight increase in competition between passive traders across pre-existing exchanges during intra-days with no news. However, for the restricted sample to events around 14:00, I find that one minute after the release of macroeconomic news, the presence of the IEX exchange leads to a decrease in inter-market competition across the pre-existing exchanges around news in contrast to no news. The result points out that during intra-days with a high impact on asset prices, the entrance of IEX exacerbates the risk faced by passive traders of meeting a fast trader after the release of news.

As a robustness check, I distinguish between high impact and low impact news. The release of high impact news is an important source of adverse selection risk as they are characterized by fast price changes and an increase in volatility. High impact news have a higher impact on Quote Clustering then low impact news. For instance, by restricting the sample to intra-days at 14:00, I show that Quote Clustering experiences a decrease of 28.1% during the first minute after the release of high impact news on asset prices, while QuoteClustering experiences a decrease of 4.7% during the first minute after the release of low impact news on asset prices.

To show that news drive the results of this paper, I make use of a placebo test. The results are reassuring, as those intra-days with no macroeconomic news considered as treated are not different from other intra-days with no event (the control). The results suggests that the decrease in inter-market competition observed during intra-days with an event is the consequence of an increase in adverse selection risk.

The paper contributes to the literature in different ways: i) it proposes a new measure to gauge inter-market competition, Quote Clustering ; ii) it describes a new mechanism through which macroeconomic news may affect market quality; iii) it shows that passive traders earn larger rents during intra-days with news than during intra-days with no news, which suggests that they do a great job on protecting themselves of being adversely selected; iv) this is the first paper to the best of my knowledge to test if there is a contagion effect across passive traders in different exchanges; and v) to test the impact an increase in fragmentation has on inter-market competition as a function of adverse selection risk.

The rest of the document is structured as follows. The next section presents the literature review. Section 1.3 is devoted to describing the data and the variables used for the purpose of this study. Section 1.4 presents the results obtained by studying the impact news may have on inter-market competition and on the rents passive traders extract, and if an increase in fragmentation increases adverse selection risk. Section 1.5 is devoted to robustness check. Finally, in Section 1.6, I conclude.

1.2 Related literature

In the standard-setting framework in the field of economics an increase in the number of companies is associated with an increase in competition. When looking at stock trading, different dimensions may affect the level of competition between passive traders in different exchanges for order flow, such as i) the number of liquidity suppliers[3]; ii) the number of exchanges and iii) regulation. The current paper looks at these dimensions and its main objective is to assess empirically the impact increases in adverse selection risk may have on inter-market competition, and on the ability passive traders have to avoid being adversely selected.

Exchanges and liquidity suppliers have a common objective: to maximize profits. Liquidity suppliers maximizes revenues by maximizing the probability of trading against liquidity traders. On the contrary side, their revenues may be harmed if they trade against informed traders. If the probability of being adversely selected increases, then they may increase the spread to compensate from the losses of trading against an informed trader (Glosten and Milgrom (1985)). In line with this, there is a strand in the literature modeling the impact competition between market makers has on liquidity (i.e. Biais et al. (2000); Dennert (1993); Bernhardt et al. (1997); Mendelson (1987)). Depending on the assumptions made, competition between market makers may benefit or harm liquidity. In particular, Biais et al. (2000) show that the aggregated depth is higher, and the quoted spread decreases with an increase in the number of market makers. Dennert (1993) shows that the limited number of liquidity traders together with the incentives of informed traders to trade against every market maker leads to an increase in the probability of trading against an informed trader, which in turn leads to larger quoted spreads. Thus, the probability of trading against informed traders increases with the number of market makers. Despite the increased risk of trading against an informed trader, prices are better than in a monopoly.

Previous studies focus on competition between market makers without taking into account the level of fragmentation in the market in terms of exchanges. Given the current market landscape, Foucault and Menkveld (2008); Baldauf and Mollner (2021); Chowdhry and Nanda (1991) go a step further and assess the impact an increase in fragmentation has on liquidity.[4] An increase in fragmentation fosters competition leading to an increase in depth (Foucault and

[3] Liquidity suppliers are traders who usually post quotes/orders on both sides of the market within the trading day. For the purpose of this paper, I use the word passive traders instead of liquidity suppliers. I use "passive traders" instead of "liquidity suppliers" as I cannot distinguish between limit orders posted by liquidity suppliers from the ones posted by brokers just on one side of the market. Despite the fact that I cannot distinguish across them, I assume that the majority of limit orders are posted by liquidity suppliers

[4] In the context of this study fragmentation refers to the number of exchanges in the market.

Menkveld (2008)). On the contrary, an increase in fragmentation may increase "asymmetric information" (Chowdhry and Nanda (1991)), which is not always going to be compensated by the higher competition in the market (Baldauf and Mollner (2021)).[5] The increase in "asymmetric information" is based on the assumptions that i) the number of liquidity traders will be the same while faster traders in the market will take advantage of all the stale quotes in all the exchanges and ii) passive traders will post quotes in all the exchanges, as they cannot predict where liquidity takers will execute their orders. As the proportion of faster traders increases to the detriment of passive traders, adverse selection risk increases. The current paper adds up to this literature by studying the impact the entrance of IEX (which increases fragmentation) may have on inter-market competition, and to what extent the effect is a function of the adverse selection risk.

The current paper is closely related to the papers of Bessembinder (2003) and Biais et al. (2010). Bessembinder (2003) studied the relation between quote-based competition and trading costs. Bessembinder (2003) shows that increases in quote competition measured as the proportion of times one exchange is at (or improves) the NBB (NBO) is associated with a decrease in the spread. Recently, Madhavan (2012), used the HHI (Herfindahl-Hirschman index) to capture the level of Quote Competition in the market within one day interval, by using as an input the proportion of times one exchange is at the NBO (NBB). This measure reflects part of the behavior of limit orders submitters as it “captures the competition among high-frequency traders and aggressive quote behavior”.

Since I use macroeconomic news as a source of exogenous shock to adverse selection risk, the current paper is related to the literature on adverse selection risk and macroeconomics news. The adverse selection risk literature posits that there are two sources of risk: asymmetric information across traders and the risk passive traders face of being picked-off due to their inability to update

[5] Another group of papers find that market fragmentation in terms of volume traded across different exchanges has a positive impact on market quality (i.e. Battalio (1997); Branch and Freed (1977); Cohen and Conroy (1990); Foucault and Menkveld (2008); O’Hara and Ye (2011)), when it is measured by effective spreads, volatility, depth etc. These results suggest that at the root of an increase in market quality lies the increase competition for order flow across liquidity providers (in the different trading venues). Contrarily, Bennett and Wei (2006); Bernales et al. (2018); Gajewski and Gresse (2007) find that market quality is harmed by an increase in fragmentation. In addition, it is worth distinguishing between visible and dark fragmentation, as Degryse et al. (2015) point out that visible fragmentation has a positive impact on market quality, while dark fragmentation has a negative one.

their stale quotes. The period before prescheduled macroeconomic news announcements is the period when the market is waiting for the arrival of pending information, which may cause price changes. Thus, passive traders are reluctant to post competitive limit orders, which lead to an increase in spreads, as it has already been pointed out by USA treasury market studies: Fleming and Piazzesi (2005); Fleming and Remolona (1999) and by USA equity market studies: Erenburg and Lasser (2009); Scholtus et al. (2014).

When macroeconomic news is released to the market, information asymmetries decrease (Glosten and Milgrom (1985); Kyle (1985); Graham et al. (2006)), and price volatility increases and is persistent within the first 15 minutes (Ederington and Lee (1993)). On the contrary, Crego (2020) shows theoretically and empirically that asymmetric information increases. Besides the potential increase in asymmetric information, passive traders face two types of risk: i) the risk of being picked-off by trading against faster traders (Fleming and Piazzesi (2005)), and ii) the inventory control risk which increases with fast price changes (Fleming and Remolona (1999)).

In relation to the risk of being picked off, theoretically it has been acknowledged that the risk faced by passive traders stems from the risk of trading against faster traders even in the absence of informed traders (Biais et al. (2005); Copeland and Galai (1983); Foucault et al. (2003)). This risk is even higher in the current trading environment, which is characterized by the increased presence of a subgroup of Algorithmic Traders: High Frequency Traders (HFTs). In line with this, Biais et al. (2015); Foucault et al. (2013) and Martinez and Rosu (2011) show that passive traders face a high risk of being picked off, as some HFTs use market orders. The risk of being picked-off by faster traders is analogous to the risk of trading against informed traders in those papers that study inter-market competition, and hence I also consider it as a form of asymmetric information.

Alternatively, there is another view, where the new market makers HFT can update their quotes fast enough to reduce the probability of being picked off (Hoffmann (2014) . Although, according to Menkveld (2013) and Budish et al. (2015) the risk of being picked off cannot be completely eliminated, not even by informed HFT market makers. All these theoretical articles highlights the increased presence of HFTs around news. This is confirmed for the USA treasury market by Jiang et al. (2011) and Jiang et al. (2014), showing that liquidity is harmed by the activity of HFT. Furthermore, Cartea et al. (2019) show that ultra-fast machine-driven activity is associated with a lower market quality in the market, measured by: quoted spreads, depth; and effective spreads. Their measure of ultra-fast machine-drive in activity is designed to capture HFT liquidity provision strategies. Hence, the previous articles mentioned suggest that high levels of HFT is associated with lower liquidity.

If fast traders use aggressive orders to take advantage of their speed advantage, then passive traders at the NBBO are highly exposed. In the context of the theoretical model of Foucault et al. (2003), when adverse selection risk is high enough but below a threshold, some dealers may monitor both the quotes and the news, while another group of dealers may monitor just the market and free ride on the updates of quotes posted by those dealers monitoring both the news and the market.[6] Consequently, those passive traders monitoring only the market have incentives to stay away from the NBBO, and in this way reduce their adverse selection risk. Thus, the proportion of exchanges at the NBBO (QuoteClustering) should capture the adverse selection risk in the market. Lower Quote Clustering should suggest a higher adverse selection risk in the market, as passive traders prefer to stay away from the NBBO and free ride on those at the NBBO in the fashion of Foucault et al. (2003).

1.3 Institutional framework and data

Reg NMS, implemented in 2007, had the aim to foster competition between different exchanges. The Order Protection Rule lies at the core of Reg NMS and it states that orders should be executed at the NBBO, which means that trades executed at worse prices (trade-through) are prohibited. The prohibition of trade-throughs may encourage the entrance of new exchanges in the market, as there is no "adverse side of price fragmentation" (Pagnotta and Philippon (2018)). The Securities Information Processor (SIP) connects all the exchanges and calculates and disseminates the NBBO among all investors. The SIP is used widely by investors. In addition, exchanges have the obligation to avoid trade-throughs. The prohibition of trade-troughs is one specific characteristic of the USA stock market which may foster competition between passive traders in different

[6]Foucault et al. (2003) studies different equilibria depending on the adverse selection risk. They do not account for fragmentation.

exchanges.[7] In line with this, Foucault and Menkveld (2008) show theoretically and empirically that an increase in trade-through would discourage the provision of liquidity to the market.

1.3.1 Data

I use the TAQ database from 2015 to 2017 for the SPDR S&P 500 trust (ETF: exchange traded fund) which is listed on NYSE Arca. Since this ETF is the largest US ETF and it is highly liquid, it makes it suitable for this analysis, as it allows to isolate the effect macroeconomic news has on inter-market competition due to an increase in adverse selection risk and not due to a potential lack of liquidity in the market.

TAQ data contains information regarding the volume traded, the price and the exchange where the trade takes place. In addition, it contains for each exchange: the best bid (the highest bid) and the best ask (the lowest ask), the aggregate number of shares at the best prices, and the timestamp in milliseconds. I use the filters proposed by Holden and Jacobsen (2014) in order to clean-up the data. Days with short sessions are not considered. Thus, I am left with 746 trading days.

The ECONODAY database is used to identify the time when macroeconomic news are released to the market, while the FOREXFACTORY webpage is used to determine if macroeconomic news have a high, medium or low impact on asset prices. In the webpage, those macroeconomic news with a high impact on news are characterized by a red flag, while those macroeconomic news with a medium and low impact are characterized by an orange and yellow flag, respectively. From all the universe of events in the database, I choose the ones that have been acknowledged to have an impact on asset prices, as adverse selection experiences an increase as shown by Crego (2020) and Fleming and Piazzesi (2005)). Furthermore, Scholtus et al. (2014) show the importance speed has on high frequency trading strategies around macroeconomic news. Since the objective of this paper is to assess the impact an increase in adverse selection risk has on inter-market competition which may be caused on the one hand by speed advantage and/or on the other hand by an increase in asymmetric information, I use similar macroeconomic

[7] In The European stock market one of the main objectives was the implementation of the Order Protection Rule in June of 2022.

news to the ones used in Scholtus et al. (2014).

Table 1.1 reports all the events considered in this study with the corresponding time at which macroeconomic news were released to the market. In brackets, I report the impact it has on prices. The majority of macroeconomic news released at 14:00 are released by the FOMC (Federal Market Open Committee). News released by the FOMC have been knowledged to have a significant impact on asset prices. For instance, days when the FED announces changes in interest rates will have a big impact on the economy, and subsequently may affect the expectations investors may have on the future cash flows of the companies.

1.3.2 Variables

1.3.2.1. Measures of Inter-market Competition

With the aim to capture the level of inter-market competition for order flow, I propose to use Quote Clustering, which reflects the level of inter-market competition in terms of prices at the National Best Bid or Offer (NBBO).

The NBBO is key for the purpose of this study, as it lies at the core of Rule 611, which says that orders should be executed at the NBBO, with some exemptions such as Intermarket Sweep Orders (ISO) or Immediate or Cancel (IOC). The measure proposed is complementary to the measure of Quote Competition used by Bessembinder (2003) and Madhavan (2012). The input to calculate Quote Clustering is given by the following two equations:

$$BidQuoteClusteringNBB_s = \sum_{k=1}^{K} DNBB_{k,s} \tag{1.1a}$$

$$OfferQuoteClusteringNBO_s = \sum_{k=1}^{K} DNBO_{k,s} \tag{1.1b}$$

Where $DNBB_{k,s}(DNBO_{k,s})$ is a dummy variable that takes the value of 1 if exchange k is at the NBB (NBO), and 0 otherwise at a given point in time s. Quote Clustering NBB_t (Quote Clustering NBO_t) is the time weighted average of Bid Quote Clustering NBB_s (Offer Quote Clustering NBO_s) over the time interval t. Where the time interval can be either every 10 seconds or every minute. Thus, $QuoteClustering_t$ is defined as follows:

$$QuoteClustering_t = \frac{1/2(QuoteClusteringNBB_t + QuoteClusteringNBO_t)}{NoExchanges_{day}} \tag{1.1}$$

$NoExchanges_{day}$ is the total number of exchanges in the market.[8] The total number of public exchanges taken into account to calculate Quote Clustering is as follows:

- Period 1: January, 2015 to August 2016: 9 exchanges
- Period 2: September 2016 to July, 23, 2017: 10 exchanges (the IEX was launched as a public exchange in September).
- Period 3: From July, 24, 2017 to December, 2017 there are 11 exchanges (on July, 24, 2017 NYSE American, previously called AMEX starts reporting trades and orders for the ETF SPY).

In order to make sure that results are not driven by fast price changes, when the NBB or NBO have experienced one change, either driven by a cancellation of the quantity at the best prices or by an improvement in best prices, the number of exchanges at the NBBO is calculated always after at least 9, 10 or 11 updates, respectively, in quotes.[9]

QuoteClustering captures: i) the ex-ante risk passive traders in different exchanges assess of not being adversely selected. Ex-ante, if passive traders know that adverse selection risk is going to be low, then if they want to increase their probabilities of trading against an uninformed traders they should either 1) improve the NBB or NBO as long as the spread is not equal to the minimum tick size cent; or 2) join the NBBO. Thus, since the incentives to be at the NBBO are high when adverse selection risk is low, an increase in QuoteClustering would suggest that passive traders, ex-ante assess a low adverse selection risk; or ii) indirectly the probability of finding a liquid exchange in the fashion of Biais et al. (2015). Where a liquid exchange refers to

[8] The National Stock Exchange is taken into account for the calculation of the NBBO as it is a public exchange, however it is not considered for the calculation of the proxy of inter-market competition as the exchange does not provide quotes constantly over time. For instance on January, 2016 on: 4, 6, 7, 8 they did not post quotes. In addition, it is worth mentioning that it resumed on December 22 2015 and on February 1, 2017 it stopped its trading operation trading.https://en.wikipedia.org/wiki/National_Stock_Exchange_(Jersey_City,_New_Jersey)

[9] 9 for Period 1, 10 for Period 2, and 11 for Period 3. The minimum number of updates required after a change in NBB or NBO is justified by the number of exchanges in the market at one point in time.

an exchange posting attractive quotes (defined as limit orders posted at the NBBO), which is key due to the implications the Order Protection Rule entails. Higher levels of this measure reflect an increase in competition between passive traders in different exchanges for order flow, and an increase in the probability of finding a liquid exchange, which is especially relevant for slow aggressive traders which do not have the possibility to process the information fast enough.[10]

The measure I propose is slightly different from the one proposed by Madhavan (2012). QuoteClustering is calculated by using the number of exchanges at the NBB (NBO) at a given point in time, while the metric used by Madhavan (2012) uses the proportion of times one exchange is at the NBBO within one day interval. Observing the number of exchanges at the NBBO on a continuous basis provides a complete picture of the dynamics of changes in inter-market competition within the trading day.

1.3.2.2. Other variables and sample statistics

In this subsection I briefly discuss the construction of other variables. These variables can be grouped as i) variables calculated at the market level and ii) variables calculated at the exchange level.

Variables calculated at the market level across all exchanges:

- The time weighted total depth at the NBBO across all exchanges: calculated as the time weighted average of the total depth at the NBB and the total depth at the NBO.
- The time weighted Quoted Spread at the NBBO across all exchanges: calculated as the time weighted average of the difference between NBO and NBB per minute.

In addition, at the exchange level, I calculate the following variables:

- The Realized Spread: $2 * D_{ks} * (P_{ks} - M_{s+xseconds})$. Where D_{ks} is a dummy variable that takes the value of 1 if the transaction at s is a buy-initiated order and -1 if it is a

[10]The increase in the probability of finding a liquid increases: the larger the proportion of exchanges at the NBBO the higher the probability of finding a liquid exchange which is in line with the definition of liquid exchange in Biais et al. (2015)

sell-initiated order in exchange k.[11] P_{ks} is the price of transaction and $M_{s+xseconds}$ is the midpoint after x seconds of the transaction, which could be 0.1 seconds, 1 second, 60 seconds.[12] Finally, the Dollar Realized Spread is calculated by aggregating the Realized Spread volume weighted over all trades over one-minute interval.

- The Price Impact: $2 * D_{ks} * (M_{s+xseconds} - Ms)$: Where D_{ks} is a dummy variable that takes the value of 1 if the transaction at s is a buy-initiated order and -1 if it is a sell-initiated order. M_s is the midpoint of the NBBO assigned to the transaction s and $M_{s+xseconds}$ is the midpoint of the NBBO after x seconds of the transaction. Finally, the Dollar Price Impact is calculated by aggregating the Price Impact volume weighted over all trades over one-minute interval.

Furthermore, in order to control for the level of volatility the Realized Volatility is calculated (RV) as: the sum squared of continuously compounded rate of returns per minute.

Table 1.2 reports the summary statistics for the main variables used in this study. Panel A reports the summary statistics for the variables constructed at the market level. On average, the proportion of exchanges posting aggressive quotes is around 60.9% and the average spread is the minimum tick size one cent. In addition the Total Depth at the NBBO is on average of around 14 626.24 shares available. Panel B reports the summary statistics for the variables constructed at the exchange level. The average the rents extracted by passive traders per 1 share traded is positive as long as passive traders unwind their positions fast enough. The average price impact after 0.1 (1 or 60) seconds is of 0.010 (0.011 or 0.011) per 1 share traded. The average logarithm number of trades is of 3.701 and the average trade size is of 214.821 shares traded.

[11] In order to estimate the direction of the trade, I make use the algorithm suggested by Lee and Ready (1991). If the price of the transaction (P_{ks}) is larger than the midpoint (M_s: calculated as (NBB+NBO)/2) the transaction is classified as a buy initiated order and vice versa. If P_{ks}=M_s, then the tick convention is used, such that if the price of the transaction s in exchange k is larger than the price of the most recent transaction, then the trade is classified as a buy initiated order and vice versa.

[12] I assume that passive traders unwind their position within the following 0.1 seconds, 1 second, 60 seconds.

1.4 Empirical analysis

The current section has the objective to assess the impact macroeconomic news has on inter-market competition and if passive traders manage to protect themselves against an increase in adverse selection risk around macroeconomic news. In addition, it has the aim to assess if increases in fragmentation may exacerbate the adverse selection risk faced by passive traders.

In line with this, Subsection 1.4.1, is devoted to shed light on how inter-market competition reacts to macroeconomic news; Subsection 1.4.2 is devoted to assess if passive traders extract more or less rents at the root of an increase in adverse selection risk; Section 1.4.3 has the aim to proxy for quote monitoring. Finally, Subsection 1.4.4 has the aim to understand if an increase in fragmentation can lead to an increase in adverse selection risk.

1.4.1 Impact macroeconomic news has on inter-market competition

In order to assess if inter-market competition experiences changes around macroeconomic news, I use the differences in means approach. Intra-days characterized by macroeconomic news are considered as treated, while intra-days with no macroeconomic news are considered as the control group.[13] In line with this, the following benchmark model is estimated:

$$QC_t = \beta_0 + \beta_1 EVENT + \beta_t' X_t + \epsilon_t \quad (1.2)$$

Where QC_t is the dependent variable that stands for: Quote Clustering in the time interval t. The time subscript t refers to the time interval expressed in minutes. EVENT is a dummy variable that takes the value of 1 with the release of macroeconomic news and 0 otherwise. To characterize the complete dynamics, equation 1.2 is estimated independently for every minute around the event. The sample is restricted to 15 minutes before and up to 15 minutes after the release of news.[14]

[13]For instance, if during certain days macroeconomic news is released to the market at 14:00, I compare Quote Clustering at 14:00 (or around, for instance 14.01) during intra-days with news with Quote Clustering during intra-days with no macroeconomic news at 14:00 (or around, for instance 14.01). The same applies, when macroeconomic news are released at 10.00, 10.30 and 15.00.

[14]If within one trading day, there are news at 10:00 but not at other points in time, then the EVENT will take the value of one just around 10:00, while for the rest of the trading day will take the value of 0.

To assess the differences in means between intra-days with macroeconomic news and intra-days with no macroeconomic news during the first minute after the release of new information, the regression is performed for the time interval [0,60] (seconds).[15] The differences in means is captured by β_1 and it sheds light on the impact news has on inter-market competition. X_t includes control variables, in particular I control for intra-day patterns and year-month fixed effects.[16]

I consider two models: 1) the Unrestricted model, which considers the times during the day that have at least 10 events in the sample: 10:00, 10:30, 14:00 and 15:00; and 2) the Restricted model, which restricts the sample to those events at 14:00. In Figure 1.2 (a), I plot the β_1 coefficient which accompanies the EVENT together with the confidence interval over different time intervals for the unrestricted sample. The event starts having a statistically significant impact on Quote Clustering 5 minutes before the event, and it has the highest impact during the first minute after the release of news. In particular, the proportion of exchanges at the NBBO is lower by 6.2% during the first minute after the release of news relative to intra-days without news. Furthermore, the lower QuoteClustering is persistent for at least the next 15 minutes.[17]

For the restricted sample, in Figure 1.2 (b), I plot the β_1 coefficient which accompanies the EVENT together with the confidence interval over different time intervals. The decrease in QuoteClustering reaches the value of 18.2% during the first minute after the release of news, and it is considerably larger than in the unrestricted sample in Figure 1.2 (a). The large decrease may be attributed to the nature of the news released at 14:00, those released by the FED. The importance news released by the FED has on asset price may be responsible for the high adverse selection risk, leading to a higher decrease in inter-market competition in contrast to the case in which all four events are considered.[18]

To sum up, the results suggest that passive traders in different exchanges start withdrawing from the market or posting passive orders farther away from the NBBO before the release of new

[15] In other words, one regression is performed for the exact minute when macroeconomic news is released to the market, another one for 1 minute before the release of news, for 2 minutes before the release of news, etc.

[16] For intra-days patterns: ID_1 takes the value of 1 around 10:00: 15 minutes before and 15 minutes after 10:00, and 0 otherwise. ID_2 and ID_3 takes the value of 1 around 10:30 and 14:00, respectively: 15 minutes before and 15 minutes after 10:30 and 14:00, respectively, and 0 otherwise. For year-month fixed effects: MM_1 takes the value of 1 if the data is for January 2015, and the same logic applies for the following MM_i when i>1.

[17] For detailed results, see Table A1 in Appendix A.

[18] For detailed results, see Table A1 in Appendix A.

information, which lead to a decrease in inter-market competition. The results are consistent with the findings of Fleming and Piazzesi (2005), in the context of one single exchange. It is worth emphasizing that a high decrease in inter-market competition may have important implications on total trading costs, as i) searching cost for a liquid exchange are going to be higher; ii) and explicit costs may increase, as the lower number of exchanges at the NBBO may force some traders to send orders to be re-routed to other exchanges. Li et al. (2023) document that the re-routing fee charged by NYSE is the maximum fee allowed.

Patterns across different exchanges

The previous results suggest that, Quote Clustering experiences a decrease around news announcements. A question that arises is if the drop is systematic across all exchanges, or liquidity concentrates in some exchanges at the expense of the other exchanges. With the aim to shed light on this issue, following Bessembinder (2003) I calculate for every exchange the proportion of times one exchange k is at the NBB or NBO within one minute interval t ($PropNBBO_{k,t}$), which I call the probability of being at the NBBO. Finally across exchanges, I calculate the differences in means for the probability of being at the NBBO during intra-days with news vs intra-days with no news.

The results are reported in Table 1.3. During intra-days with no news, the average probability of being at the NBBO is of 62% for the Unrestricted sample, while the average for the Restricted sample is of 64%. There are important differences across exchanges, and even across exchanges belonging to the same operator. For instance, in the case of the exchanges ran by CBOE: BATS and EDGX are more than 78% of the times at the NBBO, while BATS-Y and EDGA are less than 53% at the NBBO. The main differences between these two groups of exchanges belonging to the same operator lies in the structure of the trading fees. BATS and EDGX are characterized by make-take fee models (rebates are paid to passive traders which is financed by the fees charged to aggressive traders) while BATS-Y and EDGA are characterized by inverted trading fees (provide rebate to aggressive traders which is financed by the fees charged to passive traders). The previous results document the fact that make-take fee exchanges have usually a higher probability of being at the NBBO than inverted trading fees.

In relation to changes in the probability of being at the NBBO during the first minute after the release of news in contrast to no news, when I consider the unrestricted sample, I find a systematic decrease in the probability of best prices of 8%, on average. Furthermore, by restricting the sample to intra-days at 14:00, the decrease in the probability of being at the NBBO is systematic and on average of 18%. The results suggest that across all the exchanges regardless of the structure, or idiosyncrasy in terms of trading fees, type of orders, etc, there is a systematic decrease in within exchange competition, as reflected by a lower probability of being at the NBBO. This in turn translates to a decrease in inter-market competition, as reflected by the lower Quote Clustering around announcements.

1.4.2 Adverse selection risk and rents extracted by passive traders around macroeconomic news

Around news, when adverse selection risk increases, passive traders compete less on prices in the different exchanges and this is reflected by a lower QuoteClustering. The questions that naturally arise are if passive traders manage to protect themselves against the increase in adverse selection risk, and what role inter-market competition plays on it. With the aim to shed light on these questions, I estimate the following regression:

$$y_{k,t} = \beta_0 + \beta_1 EVENT + \beta_2 QC_t + \beta_3 QC_t * EVENT + \beta_k' X_{k,t} + \epsilon_t \quad (1.3)$$

where, $y_{k,t}$ is the realized spread or the price impact for exchange k in t. The realized spread calculates the profits or loses for passive traders for closing out a position after a fixed period of time. I assume that passive traders unwind their position within 0.1, 1 and 60 seconds. The price impact calculates the impact a trade has on prices after a fixed period of time. I calculated it after 0.1, 1 and 60 seconds. While the realized spread gauges the rents extracted by passive traders, price impact gauges the adverse selection risk. EVENT is a dummy variable that takes the value of 1 during different time interval around the event: 15 minutes before and up to 15 minutes after the event, and 0 otherwise. QC_t stands for Quote Clustering in the time interval t, expressed at minutes. The interaction coefficient: β_3 between QC_t and $EVENT$ highlights how changes in inter-market competition around the EVENT affect the rents extracted by passive traders in different exchanges.[19]

The remaining variables in equation 1.3 control for other factors that can have an impact on the rents extracted by passive traders: the log of the number of trades in exchange k in t, the mean of the trade size on exchange k in t, the realized volatility in t, the one lag of the realized spread (price impact) to adjust for auto-correlation; and intra-day patterns and year-month fixed effects. The realized spread is the gross realized spread without considering the trading fees payed or received by passive traders. While in some exchanges passive traders receive a rebate (make-take fee models), in a lower number of exchanges (inverted fees) passive traders have to pay a fee. With the aim to control for heterogeneity across exchanges, exchange fixed effects are included in the model.[20]

Table 1.4 reports the results by estimating the regression in equation 1.3, where the dependent variable is the Realized Spread. I focus on the first minute after the release of macroeconomic news, as it is the moment when passive traders face a high adverse selection risk. The results are reported for both the Unrestricted and for the Restricted sample. The coefficient that accompanies the EVENT is positive and statistically significant if passive traders manage to unwind their positions within 0.1 or 1 seconds, for both the Unrestricted and the Restricted sample. If passive traders unwind their positions within the first 0.1 seconds, the realized spread is larger by \$2 per 100 shares traded for the Unrestricted sample. In the Restricted sample: if passive traders unwind their positions within the first 0.1 seconds, than the realized spread is larger by \$3.5 per 100 shares traded. The difference between the coefficients for the Restricted vs Unrestricted sample may be attributed to the nature of the news released at 14:00, as news released by the FED are characterized by a higher adverse selection risk.

In relation to the impact Quote Clustering may have on the rents extracted by passive traders around news, the coefficient which accompanies the interaction effect ($QuoteClutering_t * Event$) sheds lights on this issue. The interaction effect, $QuoteClutering_t * Event$, is negative and statistically significant for passive traders that unwind their position within 0.1 and 1 second. If there is a decrease of 1% in Quote Clustering, then the Realized Spread increases by \$0.033

[19] As the dependent variables are ex-post variables, all the independent variables and the controls are included contemporaneously, as it is unlikely that the dependent variable may affect the independent ones.

[20] The sample is restricted to 10 exchange. I exclude CSE exchange, as there is some missing data for this exchange.

(0.060) per 100 shares for the Unrestricted (Restricted sample). If we were to take into account the drop of 18.2% in Quote Clustering for the Restricted sample to intra-days at 14:00, this would imply an increase in the Realized Spread of $1.092 per 100 shares traded.[21]

Table 1.5 reports the results when the dependent variable is the price impact. During intra-days with no news, QuoteClustering has no effect on the price impact except for the Unrestricted sample after 60 seconds (and also after 1 second). The coefficient indicates that increases of 1% in QuoteClustering is characterized by an increase in the price impact after 60 seconds of $0.036 per 100 shares traded. The results may suggest that informed traders may wait to the right moment to trade, and they will trade when inter-market competition is high enough to mitigate the potential price impact. The results are consistent with the main conclusions in Kyle (1985), where informed traders take into account the impact their trading may have on prices, and they benefit from an increase in the depth in the market, as they can liquidate their positions at better prices on average. In relation to the interaction effect, around news in comparison to no news, when speed matters, lower inter-market competition leads to higher price impact after 60 seconds (unrestricted sample). The results point out to the ability of passive traders to assess correctly ex-ante the adverse selection risk, and compete less or more depending on the magnitude of adverse selection risk.[22]

To sum up, at times with high adverse selection risk, the rents extracted by passive traders are larger than at times with a low adverse selection risk as long as they manage to unwind their positions fast enough. A decrease in inter-market competition is associated with an increase in the rents extracted by passive traders. The decrease in inter-market competition is the result of the increase in adverse selection risk, which may lead some passive traders to compete less or to stop making a market. This is the first empirical paper, to the best of my knowledge, taking one step further to assess empirically how the rents of passive traders vary with changes in adverse

[21]In appendix Table A2 and Table A3, by looking at the rents extracted by passive traders one minute before and one minute after the release of macroeconomic news, respectively, I find similar results to the ones when I consider one minute during the first minute after the release of macroeconomic news. One minute after the release of macroeconomic news, passive traders earn rents regardless of the time it take them to unwind their positions.

[22]In appendix Table A4 and Table A5, I report the results for price impact one minute before and one minute after the release of macroeconomic news. One minute before the release of macroeconomic news, the Price Impact after 0.1 seconds is lower than during intra-days with no news, however, it is larger after 1 second. After one minute of the release PI is lower after 0.1 and 1 seconds, indicating that adverse selection risk decreases.

selection risk.

The larger rents extracted by passive traders may indicate that there are some passive traders monitoring the news and the market (those posting orders at the NBBO) while others decide to monitor just the market and to stay away from the NBBO. Those monitoring only the market may free ride on those monitoring both the news and the market. The results are consistent with the predictions of Foucault et al. (2003). While, I cannot identify which investors monitor the news, I can proxy for the extent to which passive traders monitor the market, and if the intensity of monitoring increases with adverse selection risk. By using a regression analyses approach, in the following sub-section 1.4.3. Quote Monitoring, I proceed to test the previous idea.

1.4.3 Quote monitoring

According to Foucault et al. (2003): "dealers never monitor news continuously because it is costly to do so and therefore their quotes may not reflect all publicly available information". This suggests that when adverse selection risk is low and no events are expected, passive traders will monitor just the market without monitoring the news. Aggressive traders with private information trade but in such a way as to mitigate price impact, as shown in the previous section and as pointed out by Kyle (1985). Therefore, the updates in quotes to any changes in the quotes by other traders will occur gradually. However, with the release of public information, characterized by a high adverse selection risk and when speed is especially relevant, changes in the proportion of exchanges at the NBBO may be more informative and relevant for passive traders free riding on those monitoring both the market and the news.

I expect that during intra-days with a low adverse selection risk passive traders will react to updates in quotes by other traders. Additionally, I expect this reaction to be exacerbated during intra-days with a higher adverse selection risk, when changes in the proportion of exchanges at the NBBO may be more informative and relevant.

If increases in within exchange competition is associated with an increase in inter-market competition across the other exchanges during intra-days with no news, then it may be argued that passive traders within one exchange choose to take more risk by posting more competitive quotes. This should be reflected by a lower spread and a higher probability of being at the

NBBO. In addition, if around news, changes in the proportion of exchanges at the NBBO is more information, then the previous effect should be exacerbated.

In order to test the previous hypothesis, the following regression is estimated:

$$Compet_{k,t} = \beta_0 + \beta_1 EVENT + \beta_2 QC_{-k,t-1} + \beta_3 QC_{-k,t-1} * EVENT + \beta_k' X_{k,t-1} + \epsilon_t \quad (1.4)$$

Within exchange competition ($Compet_{k,t}$) is calculated as follows: $QS_{k,t}$ stands for the time weighted average quoted spread over one minute interval t on exchange k, $PropNBBO_{k,t}$ is the unconditional probability of being at the NBBO which is calculated as the average between the proportion of times one exchange k is at the NBB and NBO within one minute interval t. The spread $QS_{K,t}$ reflects within exchange competition, while the $PropNBBO_{k,t}$ captures the level of aggressiveness in terms of quotes in exchange k at t. Both measures are complementary. EVENT is a dummy variable that takes the value of 1 during different time intervals around the event: 15 minutes before and up to 15 minutes after the event, and 0 otherwise. As I am interested on the impact inter-market competition across the other exchanges may have on within exchange competition in exchange k in time t, I recalculate Quote Clustering excluding one by one each exchange k, such that, the following two steps are followed: i) Calculate the National Best Bid (NBB') and the National Best Offer (NBO') excluding exchange k, and ii) then calculate Quote Clustering across the other exchanges ($QC_{-k,t-1}$), as the average of the sum of the proportion of exchanges at the NBB' and NBO', without considering exchange k in the calculation.

In order to make sure that the results are driven by the variables of interest, I control for the following additional variables: number of trades (log(trades)) for exchange k in t, the average size of trades for exchange k in t (SIZE), the realized volatility in t (RV). In addition, I include the lag of the dependent variable in order to control for serial auto-correlation. With the aim to make sure that the results are not driven by the specific characteristics of the exchange, exchange fixed effects are considered. Finally, I control for intra-day patterns and also for month by month variations, as if there a structural change in a given month, then the results should not be contaminated.

I am interested in: i) β_2 which captures the impact inter-market competition has on within exchange competition during intra-days without news; and ii) on the coefficient β_3 which captures the interaction effect: $QC_{-k,t-1} * EVENT$. The interaction effect sheds light on the impact changes in inter-market competition have on within exchange competition as a function of adverse selection risk.

Table 1.6, reports the results when the dependent variable is the SPREAD (PropDNBBO) during the first minute after the release of new information to the market. The results are reported by considering the Unrestricted and the Restricted sample. The event is statistically significant only for the restricted sample to intra-days at 14:00, when the dependent variable is the PropDNBBO.[23]. For the Unrestricted sample, during intra-days with no news, decreases in Quote Clustering are associated with decreases in within exchange competition, measured either by the SPREAD or PropDNBBO. A decrease in inter-market competition of 1% across the other exchanges is associated with an increase in the spread of \$0.00043 and a decrease in the probability of best prices of 0.130%. The interaction effect is not statistically significant.

However, the interaction effect is statistically significant at the 1% level of significance when within exchange competition is measured by the proportion of exchanges at the NBBO for the Restricted sample to intra-days at 14:00. I find a higher decrease in within exchange competition of 0.148%. Quote monitoring may be especially relevant when adverse selection risk is very high and when fast traders may take advantage of a higher number of arbitrage opportunities. When adverse selection risk is low, passive traders monitor the market and they learn from inter-market competition across other exchanges. However, when adverse selection risk is especially high, inter-market competition across the other exchanges may be more informative, so they react more to changes in inter-market competition across the other exchanges which is reflected by a significant and positive coefficient that accompanies the interaction effect.[24] The results are consistent with the idea that passive traders in different exchanges monitor the quotes. Inter-market competition across other exchanges play an important role in within exchange

[23]The fact that the event is not statistically significant in the other models may be due to the inclusion of the realized volatility and inter-market competition across the exchanges. Both the realized volatility and inter-market competition are both affect by the event. While the realized volatility increases with the release of news, inter-market competition decreases with the realize of news.

[24]In Table A6 and Table A7 reports the results, by considering one minute before the release of macroeconomic news and one minute after the release of macroeconomic, respectively. The results for one minute before the release are similar to the ones presented in Table 1.6. However, by considering one minute after the release of news, the results are less clear, and inter-market competition during intra-days with news is statistically significant and negative only for the restricted sample when the dependent variable is the probability of being at the NBBO.

competition.

1.4.4 Entrance of IEX

In a highly fragmented environment, my previous results suggest that intra-days with an increase in adverse selection risk are characterized by a decrease in inter-market competition and by an increase in the rents extracted by passive traders in contrast to intra-days with a low adverse selection risk.

Adverse selection risk may be exacerbated by an increase in the number of exchanges as shown theoretically and empirically by Baldauf and Mollner (2021).[25] Empirically, Baldauf and Mollner (2021) show that the shut down of Chi-X (during the entire trading day) led to lower spreads in the incumbent market ASX. On the contrary, Foucault and Menkveld (2008) show that an increase in fragmentation increases liquidity in the market.[26] Despite the high interest in the effects an increase in fragmentation may have on market quality, no one to the best of my knowledge has studied empirically if an increase in fragmentation may affect inter-market competition, and to what extent the effect is a function of the adverse selection risk. With the aim to test this, I use the entrance of IEX as a shock to fragmentation.

The current test is different in many aspects from the one in Baldauf and Mollner (2021) as: i) the level of fragmentation in the Canadian stock market (just 2 lit exchanges coexist) is lower than in the US. stock market (there are more than 9 exchanges); and ii) the current analysis aims to go a step further and assess empirically if for different levels of adverse selection risk, the entrance of a new exchange (IEX) may affect competition between passive traders in the pre-existing exchanges.

The Investors Exchange (IEX) was founded and operated as a dark pool from 2012 until September 2016, when it was launched as a public exchange. The introduction of IEX led to an increase in fragmentation within public exchanges in the US stock market. Increases in fragmentation may increase the percentage of faster traders in relation to liquidity traders to the

[25] At the root of the increase in the risk of being picked off lies the assumptions that i) the number of liquidity traders will be the same while faster traders in the market will take advantage of all the stale quotes in all the exchanges and ii) passive traders will post quotes in all the exchanges, as they cannot predict where liquidity takers will execute their orders.

[26] An increase in fragmentation allow traders to queue jump, by submitting limit orders in the competing exchange.

detriment of passive traders. With the aim of protecting passive traders, IEX imposes a speed bump of 350 microseconds on incoming orders. In other words, market orders would suffer a delay of 350 microseconds, time during which quotes should have the time to be updated automatically. The exchange favours passive traders while it may deter the incoming orders from fast traders. As the main objective of the exchange is to protect passive traders, I would expect slow traders to trade in all the exchanges, while fast aggressive traders to concentrate mainly in the pre-existing exchanges, increasing the adverse selection risk faced by passive traders in these exchanges.

The entrance of IEX offers the perfect setting to test the impact an increase in fragmentation has on different dimensions of inter-market competition when adverse selection risk varies.

1.4.4.1 Impact the entrance of IEX has on inter-market competition around news

With the aim to test if an increase in the number of exchanges may affect inter-market competition, I estimate the following model:

$$QC_{-IEX,t} = \beta_0 + \beta_1 EVENT + \beta_3 * EVENT + \beta_3 * IEX + \beta_4 * EVENT * IEX + \beta_t' X_t + \epsilon_t \tag{1.5}$$

$QC_{-IEX,t}$ stands for Quote Clustering excluding IEX. Since I am mainly interested on the impact the entrance of a new exchange has on competition between pre-existing exchanges, Quote Clustering has been calculated excluding IEX. IEX is a dummy variable that takes the value of 1 after the IEX was launched as an exchange, and 0 otherwise. β_2 captures the trend, the impact the entrance of IEX has on inter-market competition across the other exchanges during intra-days with no news. β_3 which accompanies the interaction effect: IEX*EVENT is key to assess if around intra-days with an EVENT and after the entrance of IEX inter-market competition across the other exchanges experienced any change in contrast to intra-days with no news and to the period before the entrance of IEX. In addition, given the importance the volatility has on passive trader's behaviour, it is imperative to control for the level of volatility in the market (RV). Controlling for the level of volatility is far from perfect, as the volatility is affected by the event which is one of our main variables of interest. Despite the previous econometric problem, since the event usually has a positive impact on volatility, the coefficients of interest will provide a lower bound for the variable of interest.[27] Thus, the lagged realized volatility RV_{t-1} has been considered in the model, which has been standardized with mean 0 and standard deviation equals 1. Since on 24 July 2017 AMEX starts making a market in the SPY, I restrict the sample to the period June 2015 to June 2017.

Figure 1.3 (in appendix B, see Table B1) reports the results by considering the Unrestricted sample, while Figure 1.4 (in appendix B, see Table B2) reports the results for the Restricted sample to intra-days around 14:00. The entrance of IEX is positive and statistically significant for both: the Unrestricted (See Figure 1.3 panel b) and for the Restricted sample (See Figure 1.4, panel b), indicating that the entrance of IEX increases inter-market competition by approximately 3% during inta-days with no news.

In relation to the impact the entrance of IEX has on inter-market competition during intra-days with news vs intra-days with no news is weak. The results point out to a slight increase in inter-market competition one minute before the release, and to a slight decrease in inter-market competition two minutes after the release of news (See Figure 1.3 panel c). The results indicate that one minute before the release of news, the entrance of a new exchange leads to an increase in competition which compensates for the increase in adverse selection risk. This is reflected by an increase in QuoteClustering of 1.9%. However, the contrary happens one minute after the release of news.

By restricting the sample to intra-days around 14:00, the interaction coefficient is statistically significant at the 1% level of confidence one minute after the release, leading to a decrease of 4.6% in inter-market competition. The results point out to a higher adverse selection risk which outweighs the benefits of an increase in competition.[28]

Empirically, Baldauf and Mollner (2021) use the spread as a proxy for the risk of being picked

[27]The period after the entrance of IEX is characterized by a lower volatility than the period before the entry of IEX. Thus, since I want to isolate the impact IEX has on inter-market competition, it is key to include the volatility level. The same logic does not hold for the main analyses in section 1.4.1. where the main objective is to identify the impact the event has on inter-market competition, on average, and not the trend in time.

[28]By distinguishing between high impact vs low impact around 14:00, it can be shown that the entrance of a new exchange in the market is especially relevant when adverse selection risk is very high, such as that during news released by the FOMC as inter-market competition is lower by 7.5% after one minute after the release of news (See appendix B, Table B8). In addition, this effect is persistent to up to 5 minutes after the release of news.

off in the context of the Canadian Stock Market. They claim that an increase in fragmentation increases the adverse selection risk, which is going to be reflected by larger spreads. Instead of using the spread, I propose a complementary measure to the spread: the proportion of exchanges at the NBBO. This goes a step further and provides information about: i) the ex-ante risk of not being adversely selected, such that a lower QuoteClustering would reflect an increase in the probability assessed by passive traders of being adversely selected; ii) the probability of finding a liquid exchange in the fashion of Biais et al. (2015), which in turn affects the trading costs aggressive traders may incur in. For instance, if there is a low number of exchanges at the NBBO, there is a higher probability that orders will be re-routed, which in turn would increase explicit trading costs, as re-routing of orders are usually charged with the maximum fee allowed.

1.5 Robustness check

1.5.1 High impact vs low impact news

Certain macroeconomic news has been shown to have a higher impact on asset prices and on volatility (i.e. Erenburg and Lasser (2009); Fleming and Piazzesi (2005); Fleming and Remolona (1999); Gilbert et al. (2010); Graham et al. (2006)). Since high impact news are characterized by larger price changes and higher volatility in contrast to low impact news, I expect high impact news to have higher adverse selection risk than low impact news, which should be reflected in a lower QuoteClustering and higher Quoted Spreads. In order to test the previous hypothesis, I distinguish between high impact (HI) vs low impact(LI) news as defined inTable 1.1 , and I estimate the regression in equation 1.2.

The estimated coefficients that accompanies HI and LI together with their corresponding confidence intervals are plotted in Figure 1.5, (a) Unrestricted sample (left panel: the impact it has on QuoteClustering). The magnitude of the decrease in Quote Clustering around the event is much higher for High Impact in contrast to Low Impact news. By restricting the sample to intra-days around 14:00, inter-market competition during the first minute after the release of new information is lower by 28.1 (3 exchanges less posting quotes on average) relative to intra-days with no news (see Figure 1.5 (b) 14:00 (left panel)). In addition, I estimate also the impact HI and LI has on Quoted Spread. High impact news lead to a higher increase in spreads than those news with a low impact on asset prices.[29]

The considerable decrease in QuoteClustering together with the increase in Quoted Spreads during high impact news may be attributed to the increased risk faced by passive traders around faster and larger price changes in comparison to low impact news which are not expected to have a big impact on asset prices. Around high impact news, the probability of finding a liquid exchange is much lower than during news with a low impact on asset prices which can increase trading costs for slow aggressive traders.

1.5.2 Placebo test

To ensure that just around macroeconomic news, inter-market competition experiences a significant decrease, a placebo test has been performed. The proxies for inter-market competition Quote Clustering with an event at: 10:00am, 10:30am, 14:00 and 15:00 pm are replaced with other intra-days within the same trading day without an event at: 11:00, 11:30, 12:00 and 13:00pm respectively. This group is called the “treated” group. The other group is the control group, which is constituted by those intra-days with no event at: 11:00, 11:30, 12:00 and 13:00pm, respectively. In addition, I distinguish between high impact (HI) vs low impact (LI) news. The estimated coefficients of the HI and LI along with the confidence interval around the “event” are plotted in Figure 1.6.

The results suggest that those intra-days with an “event” have no impact on inter-market competition, regardless of the impact of the "event". The evidence is supportive of the previous findings, suggesting that macroeconomic news have an impact on inter-market competition.

1.6 Conclusions

In the current paper, I analyse inter-market competition in a highly fragmented market. I show that inter-market competition experiences a sudden decrease around macroeconomic news, because the risk faced by passive traders of being adversely selected is high. The magnitude of

[29]For more detailed results, for the Restricted Sample see Table B5 and Table B6, and for the Unrestricted Sample see Table B3 and Table B4

the decrease depends on the ex-ante impact news may have on asset prices. For instance, news released at 14:00 with a high impact on asset prices are characterized by a decrease of 28% in Quote Clustering in contrast to intra-days with no news at 14:00.

I show that passive traders extract larger rents during intra-days with news in contrast to no news, which reflects the ability of passive traders to protect themselves against an increase in adverse selection risk. The higher rents extracted by passive traders, may be attributed to the reluctance of passive traders to post quotes at the NBBO and to the widening of spreads. In addition, I show that decreases in inter-market competition may contribute to larger rents extracted by passive traders during news in contrast to no news.

Furthermore, I contribute to the literature by showing how within platform competition responds to changes in inter-market competition. Changes in the proportion of exchanges at the NBBO may be more informative and relevant at times with a high adverse selection risk for passive traders free riding on those monitoring both the market and the news.

Finally, I use the entrance of IEX to assess how inter-market competition changes with an increase in fragmentation as a function of adverse selection risk. I show that inter-market competition experiences a slight increase with the entrance of IEX when adverse selection risk is low. On the contrary side, when adverse selection risk is relatively high, after the release of macroeconomic news, I show that inter-market competition is lower with the entrance of IEX, and it is persistent up to 5 minutes after the release of news. This suggests that the risk of being picked-off outweighs the potential benefit of an increase in competition, when adverse selection is high. Thus, the impact an increase in fragmentation has on inter-market competition is context-dependent.

References

Y. Aıt-Sahalia and M. Sa. High Frequency Market Making: Optimal Quoting. page 48, June 2017.

M. Baldauf and J. Mollner. Trading in Fragmented Markets. *Journal of Financial and Quantitative Analysis*, 56(1):93–121, Feb. 2021. ISSN 0022-1090, 1756-6916. doi: 10.1017/S0022109019000814.

R. H. Battalio. Third Market Broker-Dealers: Cost Competitors or Cream Skimmers? *The Journal of Finance*, 52(1):341–352, 1997. ISSN 1540-6261. doi: 10.1111/j.1540-6261.1997.tb03819.x.

P. Bennett and L. Wei. Market structure, fragmentation, and market quality. *Journal of Financial Markets*, 9(1):49–78, 2006.

A. Bernales, N. Garrido, S. Sagade, M. Valenzuela, and C. Westheide. A Tale of One Exchange and Two Order Books: Effects of Fragmentation in the Absence of Competition. *SSRN Electronic Journal*, 2018. ISSN 1556-5068. doi: 10.2139/ssrn.3276548.

D. Bernhardt, D. Bernhardt, E. Hughson, and E. Hughson. Splitting orders. *Review of Financial Studies*, 10(1), 1997. ISSN 08939454. doi: 10.1093/rfs/10.1.69.

H. Bessembinder. Quote-based competition and trade execution costs in NYSE-listed stocks. *Journal of Financial Economics*, 70(3):385–422, Dec. 2003. ISSN 0304-405X. doi: 10.1016/S0304-405X(03)00168-5.

B. Biais, D. Martimort, and J.-C. Rochet. Competing Mechanisms in a Common Value Environment. *Econometrica*, 68(4):799–837, 2000. ISSN 1468-0262. doi: 10.1111/1468-0262.00138.

B. Biais, L. Glosten, and C. Spatt. Market microstructure: A survey of microfoundations, empirical results, and policy implications. *Journal of Financial Markets*, 8(2):217–264, May 2005. ISSN 13864181. doi: 10.1016/j.finmar.2004.11.001.

B. Biais, C. Bisière, and C. Spatt. Imperfect Competition in Financial Markets: An Empirical Study of Island and Nasdaq. *Management Science*, 56(12):2237–2250, Dec. 2010. ISSN 0025-1909, 1526-5501. doi: 10.1287/mnsc.1100.1243.

B. Biais, T. Foucault, and S. Moinas. Equilibrium fast trading. *Journal of Financial Economics*, 116(2):292–313, May 2015. ISSN 0304-405X. doi: 10.1016/j.jfineco.2015.03.004.

B. Branch and W. Freed. Bid-Asked Spreads on the AMEX and the Big Board. *The Journal of Finance*, 32(1):159–163, 1977. ISSN 0022-1082. doi: 10.2307/2326910.

E. Budish, P. Cramton, and J. Shim. The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response*. *The Quarterly Journal of Economics*, 130(4): 1547–1621, Nov. 2015. ISSN 0033-5533, 1531-4650. doi: 10.1093/qje/qjv027.

Á. Cartea, R. Payne, J. Penalva, and M. Tapia. Ultra-fast activity and intraday market quality. *Journal of Banking & Finance*, 99:157–181, Feb. 2019. ISSN 03784266. doi: 10.1016/j.jbankfin.2018.12.003.

B. Chowdhry and V. Nanda. Multimarket Trading and Market Liquidity. *The Review of Financial Studies*, 4(3):483–511, July 1991. ISSN 0893-9454. doi: 10.1093/rfs/4.3.483.

K. J. Cohen and R. M. Conroy. An Empirical Study of the Effect of Rule 19c-3. *Journal of Law & Economics*, 33(1):277–306, 1990.

T. E. Copeland and D. Galai. Information Effects on the Bid-Ask Spread. *The Journal of Finance*, 38(5):1457–1469, 1983. ISSN 0022-1082. doi: 10.2307/2327580.

J. A. Crego. Why does public news augment information asymmetries? *Journal of Financial Economics*, 137(1):72–89, July 2020. ISSN 0304-405X. doi: 10.1016/j.jfineco.2019.05.020.

H. Degryse, F. de Jong, and V. van Kervel. The Impact of Dark Trading and Visible Fragmentation on Market Quality*. *Review of Finance*, 19(4):1587–1622, July 2015. ISSN 1573-692X, 1572-3097. doi: 10.1093/rof/rfu027.

J. Dennert. Price Competition between Market Makers. *The Review of Economic Studies*, 60(3): 735–751, 1993. ISSN 0034-6527. doi: 10.2307/2298133.

L. H. Ederington and J. H. Lee. How Markets Process Information: News Releases and Volatility. *The Journal of Finance*, 48(4):1161–1191, 1993. ISSN 0022-1082. doi: 10.2307/2329034.

G. Erenburg and D. Lasser. Electronic limit order book and order submission choice around macroeconomic news. *Review of Financial Economics*, 18(4):172–182, Oct. 2009. ISSN 1058-3300. doi: 10.1016/j.rfe.2009.06.001.

M. Fleming and M. Piazzesi. Monetary Policy Tick-by-Tick. page 59, Sept. 2005.

M. J. Fleming and E. M. Remolona. Price Formation and Liquidity in the U.S. Treasury Market: The Response to Public Information. *The Journal of Finance*, 54(5):1901–1915, 1999. ISSN 1540-6261. doi: 10.1111/0022-1082.00172.

T. Foucault and A. J. Menkveld. Competition for Order Flow and Smart Order Routing Systems. *The Journal of Finance*, 63(1):119–158, 2008. ISSN 1540-6261. doi: 10.1111/j.1540-6261.2008.01312.x.

T. Foucault, A. Röell, and P. Sandås. Market Making with Costly Monitoring: An Analysis of the SOES Controversy. *Review of Financial Studies*, 16(2):345–384, Apr. 2003. ISSN 0893-9454, 1465-7368. doi: 10.1093/rfs/hhg005.

T. Foucault, O. Kadan, and E. Kandel. Liquidity Cycles and Make/Take Fees in Electronic Markets. *The Journal of Finance*, 68(1):299–341, 2013. ISSN 1540-6261. doi: 10.1111/j.1540-6261.2012.01801.x.

J.-F. Gajewski and C. Gresse. Centralised order books versus hybrid order books: A paired comparison of trading costs on NSC (Euronext Paris) and SETS (London Stock Exchange). *Journal of Banking & Finance*, 31(9):2906–2924, Sept. 2007. ISSN 0378-4266. doi: 10.1016/j.jbankfin.2007.04.015.

T. Gilbert, C. Scotti, G. Strasser, C. Vega, T. Andersen, T. Bollerslev, D. Croushore, M. Mccracken, B. Rossi, and J. Wright. Why Do Certain Macroeconomic News Announcements Have a Big Impact on Asset Prices? July 2010.

L. R. Glosten and P. R. Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1):71–100, Mar. 1985. ISSN 0304405X. doi: 10.1016/0304-405X(85)90044-3.

J. R. Graham, J. L. Koski, and U. Loewenstein. Information Flow and Liquidity around Anticipated and Unanticipated Dividend Announcements*. *The Journal of Business*, 79(5): 2301–2336, Sept. 2006. ISSN 0021-9398, 1537-5374. doi: 10.1086/505236.

P. Hoffmann. A dynamic limit order market with fast and slow traders. *Journal of Financial Economics*, 113(1):156–169, July 2014. ISSN 0304-405X. doi: 10.1016/j.jfineco.2014.04.002.

C. W. Holden and S. Jacobsen. Liquidity Measurement Problems in Fast, Competitive Markets: Expensive and Cheap Solutions. *The Journal of Finance*, 69(4):1747–1785, 2014. ISSN 1540-6261. doi: 10.1111/jofi.12127.

G. Jiang, I. Lo, and G. Valente. High-Frequency Trading around Macroeconomic News Announcements: Evidence from the U.S. Treasury Market. 2014. doi: 10.34989/SWP-2014-56. URL https://www.bankofcanada.ca/2014/12/working-paper-2014-56/. Publisher: Bank of Canada.

G. J. Jiang, I. Lo, and A. Verdelhan. Information Shocks, Liquidity Shocks, Jumps, and Price Discovery: Evidence from the U.S. Treasury Market. *The Journal of Financial and Quantitative Analysis*, 46(2):527–551, 2011. ISSN 0022-1090.

A. S. Kyle. Continuous Auctions and Insider Trading. *Econometrica*, 53(6):1315–1335, 1985. ISSN 0012-9682. doi: 10.2307/1913210.

C. M. Lee and M. J. Ready. Inferring trade direction from intraday data. *Journal of Finance*, (46):733–746, 1991.

S. Li, M. Ye, and M. Zheng. Refusing the best price? *Journal of Financial Economics*, 147(2): 317–337, Feb. 2023. ISSN 0304-405X. doi: 10.1016/j.jfineco.2022.11.004.

A. Madhavan. Exchange-Traded Funds, Market Structure, and the Flash Crash. *Financial Analysts Journal*, 68(4):20–35, 2012. ISSN 0015198X. doi: 10.2469/faj.v68.n4.6.

V. Martinez and I. Rosu. Informed Traders, News and Volatility. page 27, May 2011.

H. Mendelson. Consolidation, Fragmentation, and Market Performance. *The Journal of Financial and Quantitative Analysis*, 22(2):189–207, 1987. ISSN 0022-1090. doi: 10.2307/2330712.

A. J. Menkveld. High frequency trading and the new market makers. *Journal of Financial Markets*, 16(4):712–740, Nov. 2013. ISSN 1386-4181. doi: 10.1016/j.finmar.2013.06.006.

M. O'Hara and M. Ye. Is market fragmentation harming market quality? *Journal of Financial Economics*, 100(3):459–474, June 2011. ISSN 0304-405X. doi: 10.1016/j.jfineco.2011.02.00 6.

E. S. Pagnotta and T. Philippon. Competing on Speed. *Econometrica*, 86(3):1067–1115, 2018. ISSN 1468-0262. doi: 10.3982/ECTA10762.

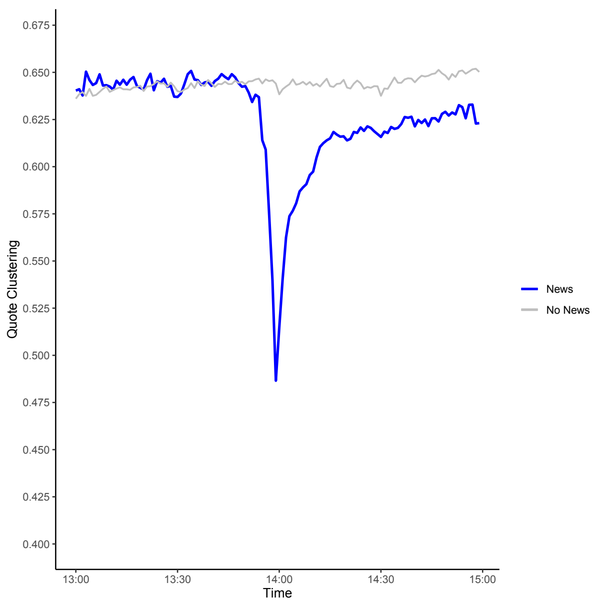
M. Scholtus, D. van Dijk, and B. Frijns. Speed, algorithmic trading, and market quality around macroeconomic news announcements. *Journal of Banking & Finance*, 38:89–105, Jan. 2014. ISSN 0378-4266. doi: 10.1016/j.jbankfin.2013.09.016.

Figures and Tables

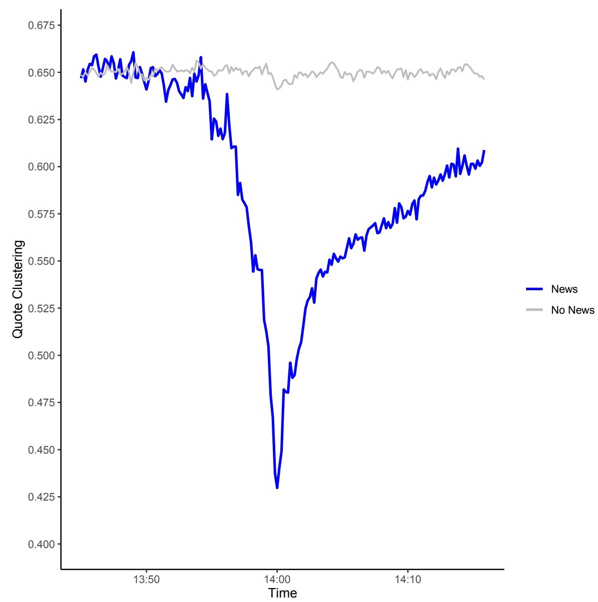
Figures

Figure 1.1: Quote Clustering intra-days with news vs intra-days with no news

Figure 1.1 plots the average (time weighted) of Quote Clustering for intra-days with news released at 14:00 in contrast to intra-days with no news released at 14:00. Panel A, plots the average (time weighted) of Quote Clustering within one minute interval. Panel B, zooms in to a higher frequency, and plots the average (time weighted) of Quote Clustering within 10 seconds interval.



(a) Quote Clustering: At 1 minute interval

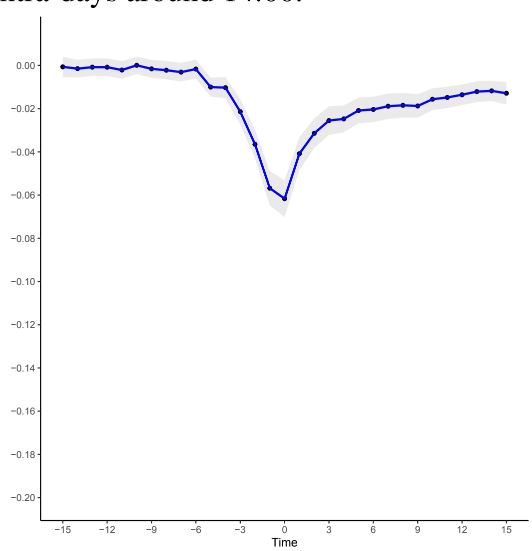


(b) Quote Clustering: At 10 seconds interval

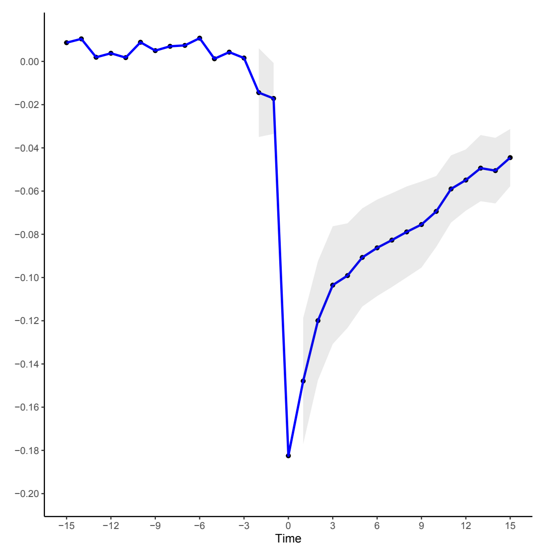
Figure 1.2: Estimated effect of news on Quote Clustering

$$y_t = \beta_0 + \beta_1 EVENT + \beta_t' X_t + \epsilon_t \quad (1.6)$$

The figure plots the estimated effect Macroeconomic news, captured by β_1, has on Quote Clustering. Panel (a) Unrestricted Sample: uses the four intra-days moments that have at least 10 events in the sample. While Panel b) Restricted Sample: restricts the sample just to those intra-days around 14:00.



(a) Unrestricted Sample

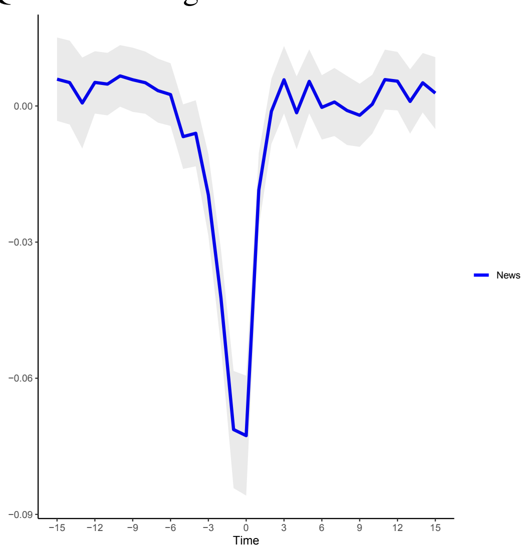


(b) Restricted Sample

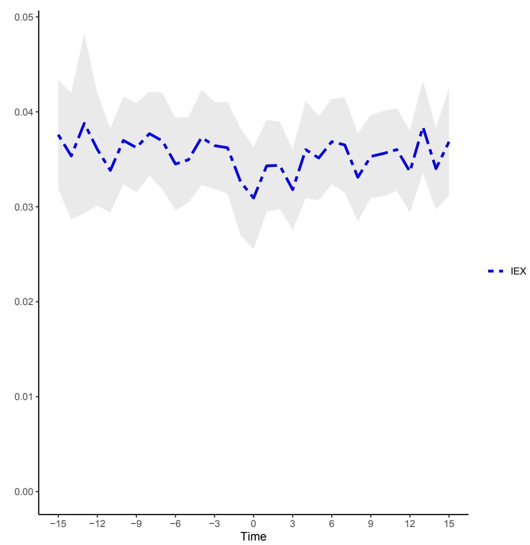
Figure 1.3: Estimated effect of the entrance of IEX on Quote Clustering: Unrestricted Sample

$$y_t = \beta_0 + \beta_1 EVENT + \beta_2 IEX + \beta_3 EVENT * IEX + \beta_t' X_t + \epsilon_t \quad (1.7)$$

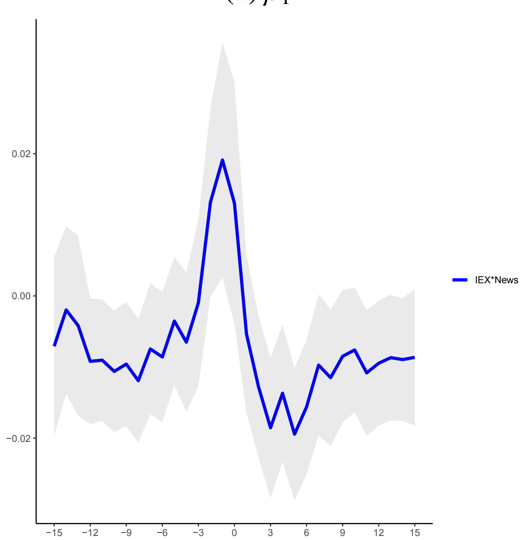
Estimated effect of the entrance of IEX on QuoteClustering around News. Panel (a) reports the β_1 which accompanies EVENT. Panel (b) reports the β_2 which accompanies IEX. Panel (c) reports the β_3 which accompanies the interaction term EVEN*IEX. Panel (d) reports the $\beta_1 + \beta_3$ which shed lights on if during intra-days with news, the entrance of IEX exerts any impact on QuoteClustering.



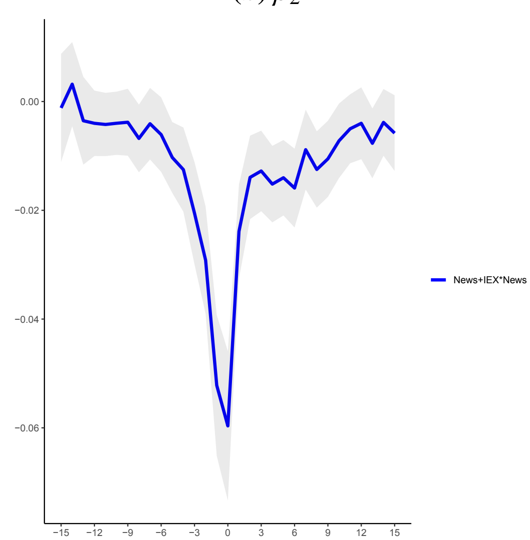
(a) β_1



(b) β_2



(c) β_3

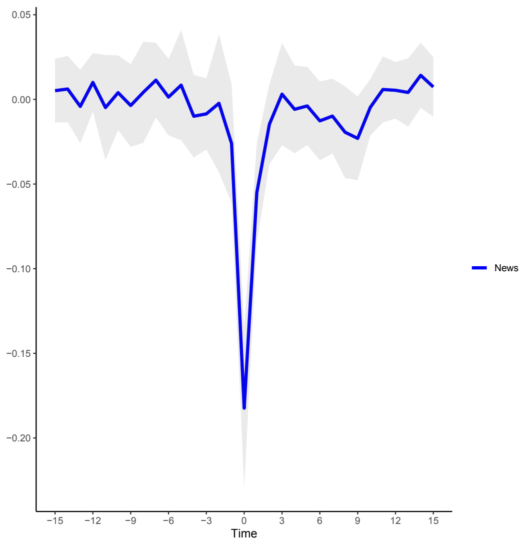


(d) $\beta_1 + \beta_3$

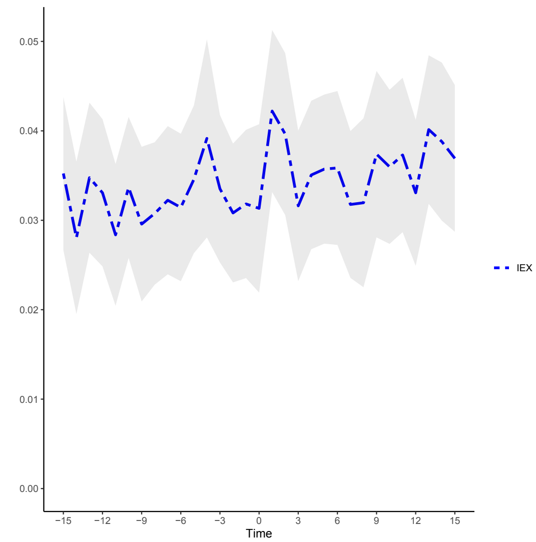
Figure 1.4: Estimated effect of IEX on Quote Clustering around news released at 14:00: Restricted Sample

$$y_t = \beta_0 + \beta_1 EVENT + \beta_2 IEX + \beta_3 EVENT * IEX + \beta_t' X_t + \epsilon_t \quad (1.8)$$

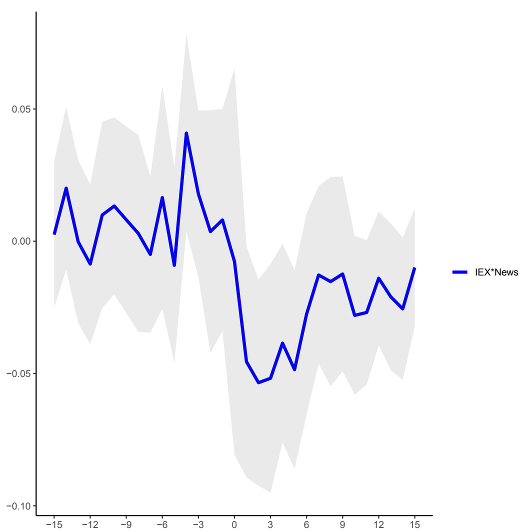
Estimated effect of the entrance of IEX on QuoteClustering around News. Panel (a) reports the β_1 which accompanies EVENT. Panel (b) reports the β_2 which accompanies IEX. Panel (c) reports the β_3 which accompanies the interaction term EVEN*IEXT. Panel (d) reports the $\beta_1 + \beta_3$ which shed lights on if during intra-days with news, the entrance of IEX exerts any impact on QuoteClustering. The sample has been restricted just to those intra-days with news released at 14:00.



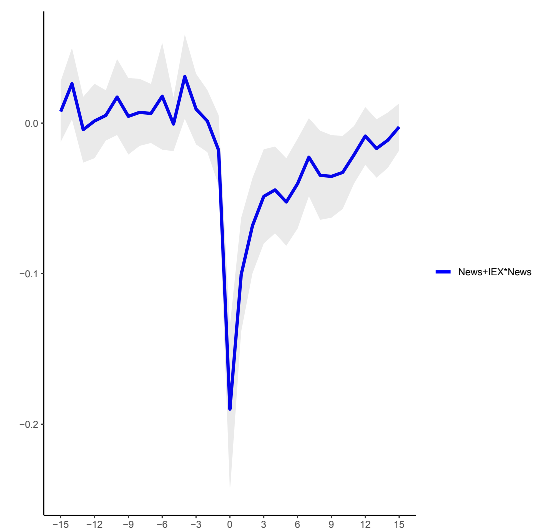
(a) β_1



(b) β_2



(c) β_3



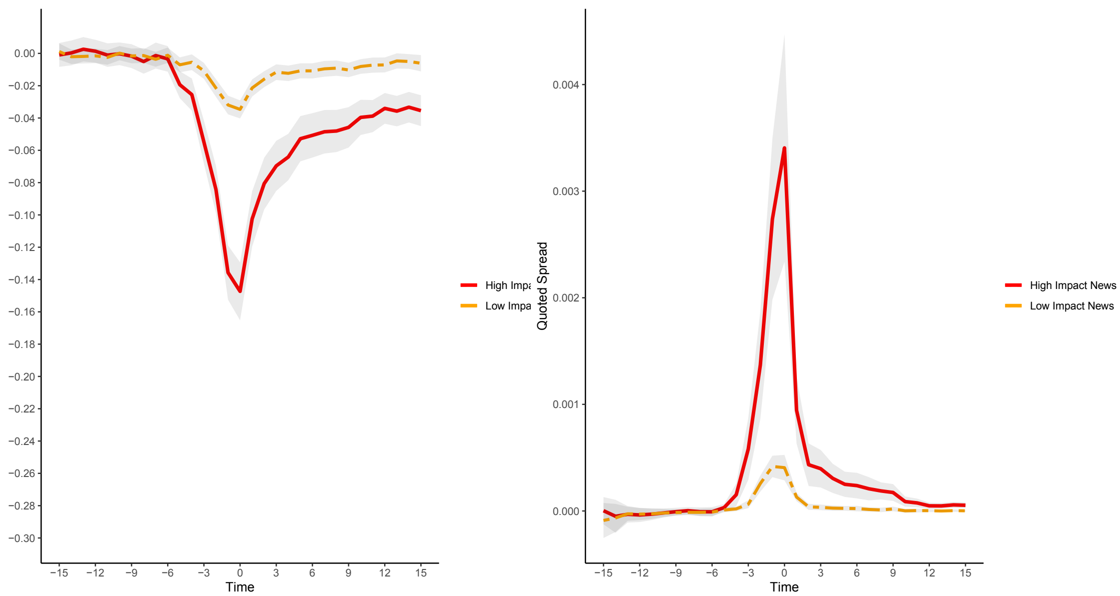
(d) $\beta_1 + \beta_3$

Figure 1.5: Estimated effect of HI vs. LI news on Quote Clustering and Quoted Spread

$$y_t = \beta_0 + \beta_1 HI + \beta_2 LI + \beta_t' X_t + \epsilon_t \quad (1.9)$$

Estimated effect of HI (High Impact: red line) vs LI (Low Impact: orange line) Macroeconomic news captured by β_1 and β_2 respectively has on Quote Clustering (left panel) and on Quoted Spread (right panel). Panel (a) reports the results for the all trading day, by considering the 4 intra-days with with news, while Panel (b) restricts the sample just to those intra-days around 14:00.

(a) Unrestricted Sample



(b) Restricted Sample: Around 14:00

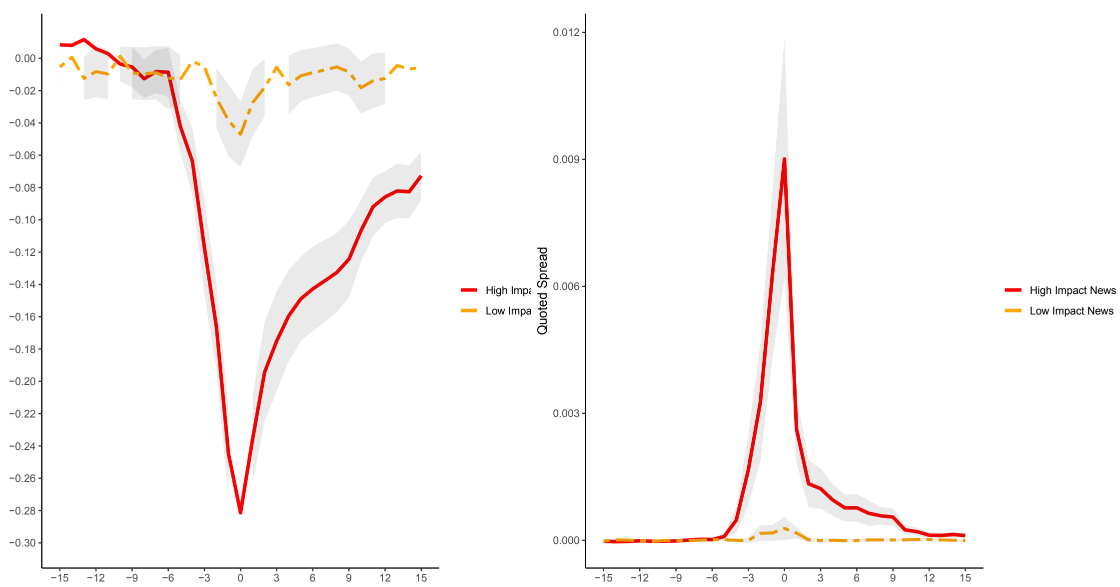
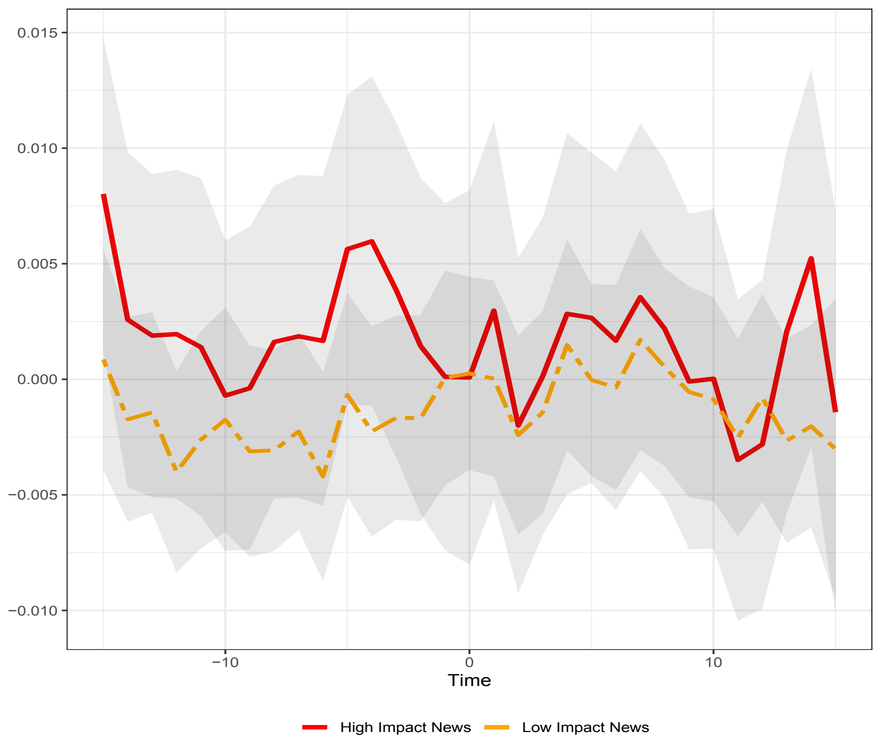


Figure 1.6: Placebo Test: Estimated effect of "High Impact" vs "Low Impact Macroeconomic News" on Quote Clustering



Tables

Table 1.1: Macroeconomic news

Time	Name	2015	2016	2017	Total
10:00	ISM Mfg Index (high)	11	11	11	33
10:00	Construction Spending (low)	12	22	11	45
10:00	Factory Orders (low)	12	12	12	36
10:00	ISM Non-Mfg Index (high)	12	12	12	36
10:00	Wholesale Trade (low)	11	12	12	35
10:00	Business Inventories (low)	12	11	11	34
10:00	Philadelphia Fed Business Outlook Survey (high)	10	0	0	10
10:00	Consumer Sentiment (medium)	24	24	24	72
10:00	Leading Indicators (low)	12	13	12	37
10:00	Existing Home Sales (medium)	12	36	12	60
10:00	Consumer Confidence (high)	11	12	12	35
10:00	New Home Sales (low)	12	12	12	36
10:00	Pending Home Sales Index (medium)	12	24	12	48
10:30	EIA Natural Gas Report (low)	49	51	51	151
14:00	FOMC Minutes (high)	7	8	8	23
14:00	Treasury Budget (low)	11	12	12	35
14:00	FOMC Meeting Announcement (high)	8	8	8	24
14:00	FOMC Forecasts (high)	4	4	4	12
15:00	Consumer Credit (low)	12	12	12	36

This table provides the frequency with which macroeconomic news are released to the market during the period: January 2015 to December 2017. In brackets, I report if macroeconomic news has a high, medium or low impact on asset prices. In order to assess the impact macroeconomic news may have on asset prices, I make use of the classification assesed by ForexFactory provided in https://www.forexfactory.com/.

Table 1.2: Summary Statistics

(a) Panel A: Variables at the market level

	mean	Std.Dev.	Q0.05	Q0.25	Q0.5	Q0.75	Q.095
Quote Clustering	0.609	0.081	0.469	0.563	0.616	0.665	0.722
Total Depth NBBO	14626.242	7275.260	5912.477	9959.071	13431.142	17592.411	27482.237
QS	0.010	0.002	0.010	0.010	0.010	0.010	0.011
RV_t	0.000	0.000	0.000	0.000	0.000	0.000	0.001

(b) Panel B: Variables at the exchange level

	mean	Std.Dev.	Q0.05	Q0.25	Q0.5	Q0.75	Q.095
$ $RS0.1sec_{k,t}$	0.000	0.008	-0.012	-0.004	0.001	0.005	0.010
$ $RS1sec_{k,t}$	-0.001	0.011	-0.018	-0.006	0.000	0.006	0.013
$ $RS60sec_{k,t}$	-0.001	0.057	-0.089	-0.022	0.002	0.023	0.079
$ $PI0.1sec_{k,t}$	0.010	0.009	0.000	0.004	0.009	0.014	0.023
$ $PI1sec_{k,t}$	0.011	0.012	-0.002	0.004	0.010	0.015	0.028
$ $PI60sec_{k,t}$	0.011	0.053	-0.062	-0.011	0.008	0.031	0.093
$log(trades)_{k,t}$	3.701	1.343	1.386	2.833	3.738	4.691	5.733
$SIZE_t$	214.821	117.763	103.846	148.148	190.840	246.220	397.462

This table This table reports the results obtained by estimating the regression specified in equation 5.reports the summary statistics for the variables constructed at the market level across all exchanges (Panel A: Variables at the market level) and for the variables constructed at the exchange level (Panel B: Variables at the exchange level). The summary statistics are reported for four intra-days moments: 10:00, 10:30, 14:00 and 15:00. Quote Clustering stands for the proxy capturing the level of inter-market competition. QS stands for the the time weighted average quoted spread. RV_t stands for the Realized Volatility. $RS 0.1 (1 or 60) seconds stands for the proxy for the gross rents earned by passive traders by assuming that passive traders unwind their positions after 0.1 (1 or 60) seconds interval. $PI 0.1 (1 or 60) seconds stands for the price impact after 0.1 (1 or 60) seconds interval. $log(trades)_{k,t}$ stands for log of the number of trades in t in exchange k ($log(trades)_{k,t}$); $SIZE_{k,t}$ stands for the mean of the trade size in t in exchange k. All the variables have been aggregated at 1 minute interval. For each variable, we report the mean (Mean), the standard deviation (Std. Dev.), the percentile 5% (Q.0.05), the percentile 25% (Q0.25), the median (Q0.5), the percentile 75% (Q0.75) and the percentile 95% (Q0.95). The time span is 2015 to 2017.

Table 1.3: Patterns across different trading platforms around the event

	Unrestricted sample			Restricted sample		
	EVENT	NO EVENT	DIFF	EVENT	NO EVENT	DIFF
ARCA-NYSE	0.87	0.92	-0.05 ***	0.77	0.93	-0.16 ***
NASDAQ	0.82	0.88	-0.05 ***	0.74	0.88	-0.14 ***
NASDAQ-BX	0.32	0.39	-0.07 ***	0.25	0.41	-0.15 ***
NASDAQ-PSX	0.46	0.63	-0.17 ***	0.34	0.65	-0.31 ***
BATS	0.83	0.90	-0.07 ***	0.69	0.91	-0.22 ***
BATS-Y	0.35	0.44	-0.09 ***	0.28	0.46	-0.18 ***
EDGA	0.40	0.50	-0.1 ***	0.31	0.53	-0.22 ***
EDGX	0.78	0.87	-0.09 ***	0.64	0.88	-0.24 ***
CSE	0.29	0.30	-0.01	0.26	0.30	-0.04 ***
IEX	0.24	0.32	-0.08 ***	0.21	0.34	-0.13 ***
AMEX	0.13	0.20	-0.07 ***	0.11	0.19	-0.09 *
ALL	0.55	0.62	-0.08 ***	0.46	0.64	-0.18 ***

This table reports the difference in t-test statistic across exchanges in the probability of being at the NBBO during intra-days with no news vs intra-days with news. Panel: Unrestricted sample uses the four intra-days moments that have at least 10 announcements (10:00, 10:30, 14:00 and 15:00), while panel: Restricted sample restricts the sample just to those intra-days at 14:00. The data has been aggregated by 1-minute interval. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 1.4: Estimated effect of increases in Quote Clustering around news on Realized Spread during the first minute after the release of news

	\$RS 0.1 seconds		\$RS 1 seconds		\$RS 60 seconds	
	Unrestricted	Restricted	Unrestricted	Restricted	Unrestricted	Restricted
EVENT	0.020***	0.035***	0.022***	0.029***	-0.000	0.006
	[0.001]	[0.002]	[0.001]	[0.002]	[0.007]	[0.011]
QC_t	0.000	0.006*	-0.010***	0.006	-0.046***	0.020
	[0.001]	[0.003]	[0.001]	[0.004]	[0.010]	[0.015]
QC_t*EVENT	-0.033***	-0.060***	-0.038***	-0.053***	-0.004	-0.021
	[0.002]	[0.004]	[0.002]	[0.004]	[0.012]	[0.021]
$log(trades)_{k,t}$	0.000	-0.000	0.000	-0.000**	-0.001**	-0.001
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.001]
$SIZE_t$	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
RV_t	-0.001***	-0.001**	-0.003***	-0.000	-0.004***	0.000
	[0.000]	[0.000]	[0.000]	[0.000]	[0.001]	[0.001]
$RS1sec_{k,t-1}$	0.114***	0.092***				
	[0.009]	[0.022]				
$RS1sec_{k,t-1}$			0.060***	0.066**		
			[0.012]	[0.022]		
$RS60sec_{k,t-1}$					0.011	0.016
					[0.011]	[0.025]
No Obs	25,275	6,273	25,275	6,273	25,275	6,273
R-squared	0.08	0.14	0.07	0.08	0.01	0.02

Number of clusters	10	10	10	10	10	10

This table reports the results obtained by estimating the following regression: $RS_{k,t} = \beta_0 + \beta_1 EVENT + \beta_2 QC_t + \beta_3 QC_t * EVENT + \beta_t' X_t + \epsilon_t$. $RS_{k,t}$ is the Realized Spread (RS) for exchange k at minute t. The columns: $RS 0.1 (1 or 60) seconds report the results by assuming that passive traders unwind their positions after 0.1 (1 or 60) seconds interval. EVENT is a dummy variable that takes the value of 1 during the first minute after the release of news and 0 otherwise. Quote Clustering is the average between QCNBB and QCNBO. QCNB (QCNBO) is calculated as the proportion of exchanges at the NBB (NBO). The interaction term $QC_t * EVENT$ stands for the interaction between: QC_t and $EVENT$. X_t contains a set of control variables: the log of the number of trades in t in exchange k ($log(trades)_{k,t}$); the mean of the trade size in t in exchange k ($SIZE_{k,t}$); the Realized Volatility (RV_t) in minute t; the one lag of the $RS_{k,t-1}$ to adjust for auto-correlation; and intra-day patterns and year-month fixed effects. In addition, exchange fixed effects has been considered. The data has been aggregated by 1-minute interval. The regressions are estimated during the first minute after the release of news. Robust standard errors are reported in brackets *** p<0.01, ** p<0.05, * p<0.1.

Table 1.5: Estimated effect of increases in Quote Clustering around news on Price Impact during the first minute after the release of news

	$PI 0.1 seconds		$PI 1 seconds		$PI 60 seconds	
	Unrestricted	Restricted	Unrestricted	Restricted	Unrestricted	Restricted
EVENT	0.003	-0.002	0.004	0.005	0.030***	0.030*
	[0.002]	[0.002]	[0.003]	[0.004]	[0.009]	[0.013]
QC_t	-0.001	-0.005	0.008**	-0.003	0.036***	-0.015
	[0.002]	[0.005]	[0.003]	[0.005]	[0.010]	[0.014]
QC_t*EVENT	-0.005	-0.001	-0.006	-0.009	-0.047**	-0.043
	[0.003]	[0.004]	[0.004]	[0.008]	[0.015]	[0.024]
$log(trades)_{k,t}$	-0.001***	-0.001***	-0.001***	-0.001**	0.000	-0.001
	[0.000]	[0.000]	[0.000]	[0.000]	[0.001]	[0.001]
$SIZE_{k,t}$	0.000**	0.000*	0.000**	0.000*	0.000	0.000
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
RV_t	0.003***	0.004***	0.004***	0.004***	0.005***	0.003*
	[0.000]	[0.000]	[0.000]	[0.001]	[0.001]	[0.001]
$PI1sec_{k,t-1}$	0.075***	0.053**				
	[0.008]	[0.018]				
$PI5sec_{k,t-1}$			0.028*	0.000		
			[0.014]	[0.031]		
$PI60sec_{k,t-1}$					0.007	-0.011
					[0.013]	[0.026]
No Obs	25,275	6,273	25,275	6,273	25,275	6,273
R-squared	0.13	0.21	0.13	0.19	0.02	0.04

Number of clusters	10	10	10	10	10	10

This table reports the results obtained by estimating the following regression: $PI_{k,t} = \beta_0 + \beta_1 EVENT + \beta_2 QC_t + \beta_3 QC_t * EVENT + \beta_t' X_t + \epsilon_t$. $PI_{k,t}$ is the Price Impact (PI) after 0.1, 1 or 60 seconds for exchange k. The columns: PI 0.1 (1 or 60) seconds report the results by considering the price impact after 0.1 (5 or 60) seconds interval. EVENT is a dummy variable that takes the value of 1 during the first minute after the release of news and 0 otherwise. Quote Clustering is the average between QCNBB and QCNBO. QCNB (QCNBO) is calculated as the proportion of exchanges at the NBB (NBO). The interaction term $QC_t * EVENT$ stands for the interaction between: QC_t and $EVENT$. X_t contains a set of control variables: the log of the number of trades in t in exchange k ($log(trades)_{k,t}$); the mean of the trade size in t in exchange k ($SIZE_{k,t}$); the Realized Volatility (RV_t) in minute t; the one lag of the $PI_{k,t-1}$ to adjust for auto-correlation; and intra-day patterns and year-month fixed effects. In addition, exchange fixed effects has been considered. The data has been aggregated by 1-minute interval. The regressions are estimated during the first minute after the release of news. Robust standard errors are reported in brackets *** p<0.01, ** p<0.05, * p<0.1.

Table 1.6: Estimated effect of increases in Quote Clustering on within platform competition during the first minute after the release of news

	$SPREAD_{k,t}$		$PropDNBBO_{k,t}$	
	Unrestricted	Restricted	Unrestricted	Restricted
EVENT	0.007	0.014	-0.021	-0.127***
	[0.005]	[0.009]	[0.013]	[0.026]
$QC_{-k,t-1}$	-0.043**	-0.032	0.130***	0.141***
	[0.018]	[0.018]	[0.026]	[0.034]
$QC_{k,t-1} * EVENT$	-0.012	-0.022	0.010	0.148**
	[0.008]	[0.014]	[0.024]	[0.051]
$log(trades)_{-k,t-1}$	-0.001*	-0.000	0.007***	0.006***
	[0.000]	[0.001]	[0.001]	[0.002]
$SIZE_{t-1}$	-0.000	0.000	0.000	-0.000
	[0.000]	[0.000]	[0.000]	[0.000]
RV_{t-1}	0.000	-0.001*	-0.004*	-0.003
	[0.001]	[0.001]	[0.002]	[0.002]
$SPREAD_{k,t-1}$	0.617***	0.927***		
	[0.156]	[0.159]		
$PropDNBBO_{k,t-1}$			0.756***	0.744***
			[0.031]	[0.025]
No Obs	25,174	6,241	25,174	6,241
R-squared	0.57	0.68	0.73	0.76
Number of clusters	10	10	10	10

Table 1.6: Estimated effect of increases in Quote Clustering on within platform competition during the first minute after the release of news *(continued)*

	Unrestricted	Restricted	Unrestricted	Restricted

This table reports the results obtained by estimating the regression: $Compet_{k,t} = \beta_0 + \beta_1 EVENT + \beta_2 QC_{-k,t-1} + \beta_3 QC_{-k,t-1} * EVENT + \beta_t' X_t + \epsilon_t$. Within platform competition ($Compet_{k,t}$) is calculated as: $SPREAD_{k,t}$: the time weighted average quoted spread on exchange k over one minute interval t; or $PropNBBO_{k,t}$ stands for the average between the proportion of times one platform k is at the NBB and NBO within one minute interval t. EVENT is a dummy variable that takes the value of 1 during the first minute after the release of news, and 0 otherwise. $QC_{-k,t-1}$ Quote Clustering has been calculated by excluding one by one each exchange in the market: i. I calculate the National Best Bid (NBB') and National Best Offer (NBO') excluding exchange k ii) by excluding exchange k, I calculate Quote Clustering ($_{-k,t-1}$) as the proportion of platforms at the NBB' (NBO') at minute t. $\beta_t' X_t$contains a set of control variables: the log of the number of trades on t on exchange k; the mean of the trade size on t on exchange k; the one lag of the $_{k,t}$ to adjust for auto-correlation; and intra-day patterns and year-month fixed effects. In addition exchange fixed effects has been considered. The data has been aggregated by 1-minute interval. The regression has been estimated during the first minute after the release of news. Robust standard errors are reported in brackets *** p<0.01, ** p<0.05, * p<0.1.

Appendices

Appendix A

Table A1: Estimated effect of News on Quote Clustering

	Unrestricted Sample			Restricted Sample		
Time	EVENT	R	No. Obs	EVENT	R	No. Obs
-15	-0.001	0.512	2489	0.009	0.434	746
	0.002					
-14	-0.001	0.504	2983	0.01	0.415	746
	0.002					
-13	-0.001	0.504	2983	0.002	0.439	746
	0.002					
-12	-0.001	0.507	2983	0.004	0.451	746
	0.002					
-11	-0.002	0.491	2984	0.002	0.417	746
	0.002					
-10	0	0.512	2984	0.009	0.430	746
	0.002					
-9	-0.002	0.488	2984	0.005	0.393	746
	0.002					
-8	-0.002	0.504	2984	0.007	0.453	746
	0.002					
-7	-0.003	0.495	2984	0.007	0.416	746
	0.002					
-6	-0.002	0.493	2984	0.011	0.398	746
	0.002					
-5	-0.01 ***	0.499	2984	0.001	0.392	746
	0.002					
-4	-0.01 ***	0.479	2984	0.004	0.402	746
	0.003					
-3	-0.021 ***	0.471	2984	0.002	0.399	746

Table A1: Estimated effect of News on Quote Clustering *(continued)*

Time	EVENT	R	No. Obs	EVENT	R	No. Obs
	0.003					
-2	-0.037 ***	0.463	2984	-0.014	0.415	746
	0.003					
-1	-0.057 ***	0.469	2984	-0.017 **	0.438	746
	0.004					
0	-0.062 ***	0.466	2984	-0.182 ***	0.530	746
	0.004					
1	-0.041 ***	0.449	2984	-0.148 ***	0.489	746
	0.004					
2	-0.031 ***	0.457	2984	-0.12 ***	0.463	746
	0.004					
3	-0.026 ***	0.457	2984	-0.104 ***	0.426	746
	0.003					
4	-0.025 ***	0.463	2984	-0.099 ***	0.445	746
	0.003					
5	-0.021 ***	0.473	2984	-0.091 ***	0.451	746
	0.003					
6	-0.02 ***	0.472	2984	-0.086 ***	0.470	746
	0.003					
7	-0.019 ***	0.471	2984	-0.083 ***	0.454	746
	0.003					
8	-0.018 ***	0.476	2984	-0.079 ***	0.460	746
	0.003					
9	-0.019 ***	0.472	2984	-0.075 ***	0.449	746
	0.003					
10	-0.016 ***	0.479	2984	-0.069 ***	0.474	746
	0.003					
11	-0.015 ***	0.484	2984	-0.059 ***	0.454	746
	0.003					
12	-0.014 ***	0.479	2984	-0.055 ***	0.453	746

Table A1: Estimated effect of News on Quote Clustering *(continued)*

Time	EVENT	R	No. Obs	EVENT	R	No. Obs
	0.002					
13	-0.012 ***	0.485	2984	-0.049 ***	0.452	746
	0.002					
14	-0.012 ***	0.484	2984	-0.051 ***	0.442	746
	0.002					
15	-0.013 ***	0.489	2731	-0.045 ***	0.493	746

This table reports the results obtained by estimating the regression specified in the benchmark model in equation 2. The results are reported for the Unrestricted Sample and for the Restricted Sample. The Unrestricted Sample uses the four intra-days moments that have at least 10 announcements. The Restricted Sample restricts the sample to those intra-days around 14:00. The dependent variable is Quote Clustering. Quote Clustering is calculated as: (Quote Clustering NBB +Quote Clustering NBO)/2, where Quote Clustering NBB (NBO) is defined as the proportion of exchanges at the NBB (NBO). The EVENT is a dummy variable that takes the value of 1 if there is an event and 0 otherwise. The data has been aggregated by 1-minute interval. Three dummies variables are included to control for intra-days patterns. In addition, 35 dummies are included to control for month by month variations. Robust standard errors are reported below the estimated coefficients.*** p<0.01, ** p<0.05, * p<0.1.

Table A2: Estimated effect of increases in Quote Clustering around news on Realized Spread one minute before the release of news

	\$RS 0.1 seconds		\$RS 1 seconds		\$RS 60 seconds	
	Unrestricted	Restricted	Unrestricted	Restricted	Unrestricted	Restricted
EVENT	0.011***	-0.007**	0.010***	-0.006	-0.010	0.001
	[0.001]	[0.003]	[0.002]	[0.004]	[0.008]	[0.010]
QC_t	-0.011***	-0.004	-0.021***	-0.018***	-0.080***	-0.023
	[0.001]	[0.003]	[0.001]	[0.002]	[0.015]	[0.014]
$QC_t * EVENT$	-0.019***	0.011**	-0.017***	0.009	0.009	0.004
	[0.002]	[0.005]	[0.004]	[0.006]	[0.013]	[0.015]
$log(trades)_{k,t}$	0.000***	0.000**	0.001***	0.001***	0.000	-0.002**
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.001]
$SIZE_t$	-0.000	-0.000	-0.000	-0.000	0.000	0.000
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
RV_t	-0.003***	-0.003***	-0.005***	-0.004***	-0.005***	0.003***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.001]	[0.001]
$RS1sec_{k,t-1}$	0.077***	0.088***				
	[0.008]	[0.015]				
$RS5sec_{k,t-1}$			0.044***	0.037**		
			[0.005]	[0.013]		
$RS60sec_{k,t-1}$					0.016	-0.031*
					[0.013]	[0.016]
No Obs	25,174	6,293	25,174	6,293	25,174	6,293
R-squared	0.12	0.14	0.11	0.11	0.01	0.02

Number of clusters	10	10	10	10	10	10

This table reports the results obtained by estimating the following regression: $RS_{k,t} = \beta_0 + \beta_1 EVENT + \beta_2 QC_t + \beta_3 QC_t * EVENT + \beta_t' X_t + \epsilon_t$. $RS_{k,t}$ is the Realized Spread (RS) for exchange k at minute t. The columns: $RS 0.1 (1 or 60) seconds report the results by assuming that passive traders unwind their positions after 0.1 (1 or 60) seconds interval. EVENT is a dummy variable that takes the value of 1 one minute before the release of news and 0 otherwise. Quote Clustering is the average between QCNBB and QCNBO. QCNB (QCNBO) is calculated as the proportion of exchanges at the NBB (NBO). The interaction term $QC_t * EVENT$ stands for the interaction between: QC_t and $EVENT$. X_t contains a set of control variables: the log of the number of trades in t in exchange k ($log(trades)_{k,t}$); the mean of the trade size in t in exchange k ($SIZE_{k,t}$); the Realized Volatility (RV_t); the one lag of the $RS_{k,t-1}$ to adjust for auto-correlation; and intra-day patterns and year-month fixed effects. In addition, exchange fixed effects has been considered. The data has been aggregated by 1-minute interval. The regressions are estimated one minute before the release of news. Robust standard errors are reported in brackets *** p<0.01, ** p<0.05, * p<0.1.

Table A3: Estimated effect of increases in Quote Clustering around news on on Realized Spread one minute after the release of news

	\$RS 0.1 seconds		\$RS 1 seconds		\$RS 60 seconds	
	Unrestricted	Restricted	Unrestricted	Restricted	Unrestricted	Restricted
EVENT	0.010***	0.022***	0.016***	0.030***	0.027***	0.051**
	[0.001]	[0.002]	[0.001]	[0.002]	[0.008]	[0.020]
QC_t	-0.005***	0.004*	-0.013***	-0.001	-0.004	-0.015
	[0.001]	[0.002]	[0.002]	[0.003]	[0.010]	[0.018]
$QC_t * EVENT$	-0.017***	-0.035***	-0.027***	-0.047***	-0.043***	-0.085**
	[0.001]	[0.003]	[0.001]	[0.004]	[0.013]	[0.035]
$log(trades)_{k,t}$	0.000***	0.000	0.001***	0.001***	0.000	0.001
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.001]
$SIZE_t$	-0.000*	-0.000	-0.000	-0.000	-0.000	0.000
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
RV_t	-0.002***	-0.002***	-0.004***	-0.003***	-0.004***	-0.004
	[0.000]	[0.000]	[0.000]	[0.000]	[0.001]	[0.002]
$RS1sec_{k,t-1}$	0.069***	0.051***				
	[0.006]	[0.011]				
$RS5sec_{k,t-1}$			0.039**	0.056**		
			[0.012]	[0.019]		
$RS60sec_{k,t-1}$					0.020**	-0.028*
					[0.008]	[0.014]
No Obs	25,229	6,259	25,229	6,259	25,229	6,259
R-squared	0.08	0.10	0.08	0.10	0.01	0.03

Number of clusters	10	10	10	10	10	10

This table reports the results obtained by estimating the following regression: $RS_{k,t} = \beta_0 + \beta_1 EVENT + \beta_2 QC_t + \beta_3 QC_t * EVENT + \beta_t' X_t + \epsilon_t$. $RS_{k,t}$ is the Realized Spread (RS) for exchange k at minute t. The columns: $RS 0.1 (1 or 60) seconds report the results by assuming that passive traders unwind their positions after 0.1 (1 or 60) seconds interval. EVENT is a dummy variable that takes the value of 1 one minute after the release of news and 0 otherwise. Quote Clustering is the average between QCNBB and QCNBO. QCNB (QCNBO) is calculated as the proportion of exchanges at the NBB (NBO). The interaction term $QC_t * EVENT$ stands for the interaction between: QC_t and $EVENT$. X_t contains a set of control variables: the log of the number of trades in t in exchange k ($log(trades)_{k,t}$); the mean of the trade size in t in exchange k ($SIZE_{k,t}$); the Realized Volatility (RV_t); the one lag of the $RS_{k,t-1}$ to adjust for auto-correlation; and intra-day patterns and year-month fixed effects. In addition, exchange fixed effects has been considered. The data has been aggregated by 1-minute interval. The regressions are estimated one minute after the release of news. Robust standard errors are reported in brackets *** p<0.01, ** p<0.05, * p<0.1.

Table A4: Estimated effect of increases in Quote Clustering around news on Price Impact one minute before the release of news

	\$PI 0.1 seconds		\$PI 1 seconds		\$PI 60 seconds	
	Unrestricted	Restricted	Unrestricted	Restricted	Unrestricted	Restricted
EVENT	-0.003***	0.009***	-0.001	0.013***	0.024**	0.009
	[0.001]	[0.002]	[0.002]	[0.003]	[0.009]	[0.009]
QC_t	0.001	-0.005	0.010***	0.008**	0.072***	0.024
	[0.001]	[0.003]	[0.001]	[0.003]	[0.014]	[0.015]
$QC_t" * EVENT$	0.005***	-0.014***	0.002	-0.019***	-0.033**	-0.018
	[0.001]	[0.003]	[0.003]	[0.004]	[0.013]	[0.013]
$log(trades)_{k,t}$	-0.000***	-0.000**	-0.001***	-0.000***	-0.001*	0.001**
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
$SIZE_{k,t}$	0.000	0.000	0.000	0.000	-0.000	-0.000
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
RV_t	0.003***	0.003***	0.005***	0.004***	0.005***	-0.001
	[0.000]	[0.000]	[0.000]	[0.000]	[0.001]	[0.001]
$PI1sec_{k,t-1}$	0.061***	0.084***				
	[0.006]	[0.015]				
$PI5sec_{k,t-1}$			0.038***	0.052***		
			[0.005]	[0.012]		
$PI60sec_{k,t-1}$					0.025	-0.011
					[0.014]	[0.014]
No Obs	25,174	6,293	25,174	6,293	25,174	6,293
R-squared	0.12	0.15	0.11	0.12	0.01	0.02

Number of clusters	10	10	10	10	10	10

This table reports the results obtained by estimating the following regression: $PI_{k,t} = \beta_0 + \beta_1 EVENT + \beta_2 QC_t + \beta_3 QC_t * EVENT + \beta_t' X_t + \epsilon_t$. $PI_{k,t}$ is the Price Impact (PI) after 0.1, 1 or 60 seconds for exchange k. The columns: PI 0.1 (1 or 60) seconds report the results by considering the price impact after 0.1 (5 or 60) seconds interval. EVENT is a dummy variable that takes the value of 1 one minute before the release of news and 0 otherwise. Quote Clustering is the average between QCNBB and QCNBO. QCNB (QCNBO) is calculated as the proportion of exchanges at the NBB (NBO). The interaction term $QC_t * EVENT$ stands for the interaction between: QC_t and $EVENT$. X_t contains a set of control variables: the log of the number of trades in t in exchange k ($log(trades)_{k,t}$); the mean of the trade size in t in exchange k ($SIZE_{k,t}$); the Realized Volatility (RV_t); the one lag of the $PI_{k,t-1}$ to adjust for auto-correlation; and intra-day patterns and year-month fixed effects. In addition, exchange fixed effects has been considered. The data has been aggregated by 1-minute interval. The regressions are estimated one minute before the release of news. Robust standard errors are reported in brackets *** p<0.01, ** p<0.05, * p<0.1.

Table A5: Estimated effect of increases in Quote Clustering around news on Price Impact one minute after the release of news

	\$PI 0.1 seconds		\$PI 1 seconds		\$PI 60 seconds	
	Unrestricted	Restricted	Unrestricted	Restricted	Unrestricted	Restricted
EVENT	-0.005***	-0.009***	-0.008***	-0.016***	-0.011	-0.025
	[0.001]	[0.001]	[0.001]	[0.001]	[0.011]	[0.020]
QC_t	-0.003**	-0.010***	0.005**	-0.004	-0.002	0.011
	[0.001]	[0.002]	[0.002]	[0.004]	[0.010]	[0.017]
QC_t*EVENT	0.008***	0.012***	0.014***	0.023***	0.016	0.045
	[0.002]	[0.002]	[0.002]	[0.003]	[0.017]	[0.034]
$log(trades)_{k,t}$	-0.000**	-0.000*	-0.001***	-0.001***	-0.001*	-0.002
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.001]
$SIZE_{k,t}$	0.000**	0.000**	0.000**	0.000*	0.000	0.000
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
RV_t	0.003***	0.003***	0.004***	0.005***	0.004***	0.005**
	[0.000]	[0.000]	[0.000]	[0.001]	[0.001]	[0.002]
$PI1sec_{k,t-1}$	0.053***	0.018				
	[0.009]	[0.011]				
$PI5sec_{k,t-1}$			0.018*	0.019		
			[0.008]	[0.014]		
$PI60sec_{k,t-1}$					0.018	-0.014
					[0.011]	[0.012]
No Obs	25,229	6,259	25,229	6,259	25,229	6,259
R-squared	0.13	0.21	0.11	0.18	0.01	0.03

Number of clusters	10	10	10	10	10	10

This table reports the results obtained by estimating the following regression: $PI_{k,t} = \beta_0 + \beta_1 EVENT + \beta_2 QC_t + \beta_3 QC_t * EVENT + \beta_t' X_t + \epsilon_t$. $PI_{k,t}$ is the Price Impact (PI) after 0.1, 1 or 60 seconds for exchange k. The columns: PI 0.1 (1 or 60) seconds report the results by considering the price impact after 0.1 (5 or 60) seconds interval. EVENT is a dummy variable that takes the value of 1 one minute after the release of news and 0 otherwise. Quote Clustering is the average between QCNBB and QCNBO. QCNB (QCNBO) is calculated as the proportion of exchanges at the NBB (NBO). The interaction term $QC_t * EVENT$ stands for the interaction between: QC_t and $EVENT$. X_t contains a set of control variables: the log of the number of trades in t in exchange k ($log(trades)_{k,t}$); the mean of the trade size in t in exchange k ($SIZE_{k,t}$); the Realized Volatility (RV_t); the one lag of the $PI_{k,t-1}$ to adjust for auto-correlation; and intra-day patterns and year-month fixed effects. In addition, exchange fixed effects has been considered. The data has been aggregated by 1-minute interval. The regressions are estimated one minute after the release of news. Robust standard errors are reported in brackets *** p<0.01, ** p<0.05, * p<0.1.

Table A6: Estimated effect of increases in Quote Clustering around news on within platform competition one minute before the release of news

	$SPREAD_{k,t}$		$PropDNBBO_{k,t}$	
	ALL	13:59	ALL	13:59
EVENT	0.013	0.042	-0.025*	-0.147***
	[0.017]	[0.025]	[0.011]	[0.025]
$QC_{-k,t-1}$	-0.040**	-0.102***	0.137***	0.156***
	[0.014]	[0.029]	[0.030]	[0.042]
$QC_{k,t-1} * EVENT$	-0.013	-0.070	0.001	0.155***
	[0.027]	[0.040]	[0.022]	[0.044]
$log(trades)_{-k,t-1}$	-0.003*	-0.000	0.003*	0.003*
	[0.001]	[0.001]	[0.001]	[0.001]
$SIZE_{t-1}$	0.000	-0.000*	0.000	0.000
	[0.000]	[0.000]	[0.000]	[0.000]
RV_{t-1}	0.003	-0.002*	-0.001	-0.002
	[0.003]	[0.001]	[0.001]	[0.002]
$SPREAD_{k,t-1}$	0.791***	0.677**		
	[0.166]	[0.258]		
$PropDNBBO_{k,t-1}$			0.770***	0.708***
			[0.035]	[0.034]
No Obs	25,135	6,208	25,137	6,209
R-squared	0.37	0.65	0.71	0.71
Number of clusters	10	10	10	10

Table A6: Estimated effect of increases in Quote Clustering around news on within platform competition one minute before the release of news *(continued)*

	ALL	13:59	ALL	13:59

This table reports the results obtained by estimating the regression: $Compet_{k,t} = \beta_0 + \beta_1 EVENT + \beta_2 QC_{-k,t-1} + \beta_3 QC_{-k,t-1} * EVENT + \beta_t' X_t + \epsilon_t$. Within platform competition ($Compet_{k,t}$) is calculated as: $SPREAD_{k,t}$: the time weighted average quoted spread on exchange k over one minute interval t; or $PropNBBO_{k,t}$ stands for the average between the proportion of times one platform k is at the NBB and NBO within one minute interval t. EVENT is a dummy variable that takes the value of 1 one minute before the release of news, and 0 otherwise. $QC_{-k,t-1}$ Quote Clustering has been calculated by excluding one by one each exchange in the market: i. I calculate the National Best Bid (NBB') and National Best Offer (NBO') excluding exchange k ii) by excluding exchange k, I calculate Quote Clustering ($_{-k,t-1}$) as the proportion of exchanges at the NBB' (NBO') at minute t. $\beta_t' X_t$contains a set of control variables: the log of the number of trades on t on exchange k; the mean of the trade size on t on exchange k; the one lag of the $_{k,t}$ to adjust for auto-correlation; and intra-day patterns and year-month fixed effects. In addition exchange fixed effects has been considered. The data has been aggregated by 1-minute interval. The regression has been estimated one minute before the release of news. Robust standard errors in brackets *** p<0.01, ** p<0.05, * p<0.1.

Table A7: Estimated effect of increases in Quote Clustering around news on within platform competition one minute after the release of news

	$SPREAD_{k,t}$		$PropDNBBO_{k,t}$	
	ALL	14:01	ALL	14:01
EVENT	-0.009	-0.009	0.027**	0.033
	[0.006]	[0.009]	[0.008]	[0.019]
$QC_{-k,t-1}$	0.010	0.007	0.000	0.153**
	[0.008]	[0.006]	[0.038]	[0.060]
$QC_{k,t-1} * EVENT$	0.012	0.015	-0.034*	-0.050
	[0.010]	[0.018]	[0.017]	[0.041]
$log(trades)_{-k,t-1}$	0.000	0.001**	0.005***	0.004**
	[0.000]	[0.000]	[0.001]	[0.001]
$SIZE_{t-1}$	-0.000	-0.000	-0.000	-0.000***
	[0.000]	[0.000]	[0.000]	[0.000]
RV_{t-1}	-0.000	-0.002*	-0.006***	-0.005***
	[0.000]	[0.001]	[0.002]	[0.001]
$SPREAD_{k,t-1}$	0.671***	0.708***		
	[0.042]	[0.081]		
$PropDNBBO_{k,t-1}$			0.759***	0.698***
			[0.019]	[0.023]
No Obs	25,272	6,273	25,272	6,273
R-squared	0.51	0.79	0.73	0.74
Number of clusters	10	10	10	10

Table A7: Estimated effect of increases in Quote Clustering around news on within platform competition one minute after the release of news *(continued)*

	ALL	14:01	ALL	14:01

This table reports the results obtained by estimating the regression: $Compet_{k,t} = \beta_0 + \beta_1 EVENT + \beta_2 QC_{-k,t-1} + \beta_3 QC_{-k,t-1} * EVENT + \beta_t' X_t + \epsilon_t$. Within platform competition ($Compet_{k,t}$) is calculated as: $SPREAD_{k,t}$: the time weighted average quoted spread on exchange k over one minute interval t; or $PropNBBO_{k,t}$ stands for the average between the proportion of times one platform k is at the NBB and NBO within one minute interval t. EVENT is a dummy variable that takes the value of 1 one minute after the release of news, and 0 otherwise. $QC_{-k,t-1}$ has been calculated by excluding one by one each exchange in the market: i. I calculate the National Best Bid (NBB') and National Best Offer (NBO') excluding exchange k ii) by excluding exchange k, I calculate Quote Clustering ($_{-k,t-1}$) as the proportion of exchanges at the NBB' (NBO') at minute t. $\beta_t' X_t$contains a set of control variables: the log of the number of trades on t on exchange k; the mean of the trade size on t on exchange k; the one lag of the $_{k,t}$ to adjust for auto-correlation; and intra-day patterns and year-month fixed effects. In addition exchange fixed effects has been considered. The data has been aggregated by 1-minute interval. The regression has been estimated one minute after the release of news. Robust standard errors are reported in brackets. *** p<0.01, ** p<0.05, * p<0.1.

Appendix B

Table B1: Estimated effect of the entrance of IEX around News on Quote Clustering excluding IEX: Unrestricted Sample

Time	EVENT	IEX	IEX*EVENT	RV_{t-1}	R2	No. Obs
-14	0.005	0.035 ***	-0.002	-0.037 ***	0.598	1491
	0.005	0.003	0.006	0.003		
-13	0.001	0.039 ***	-0.004	-0.032 ***	0.567	1491
	0.005	0.005	0.006	0.005		
-12	0.005	0.036 ***	-0.009 **	-0.034 ***	0.603	1987
	0.004	0.003	0.005	0.003		
-11	0.005	0.034 ***	-0.009 **	-0.038 ***	0.620	1987
	0.004	0.002	0.004	0.001		
-10	0.007 *	0.037 ***	-0.011 **	-0.036 ***	0.634	1988
	0.003	0.002	0.004	0.001		
-9	0.006	0.036 ***	-0.01 **	-0.037 ***	0.614	1988
	0.004	0.002	0.004	0.001		
-8	0.005	0.038 ***	-0.012 ***	-0.039 ***	0.626	1988
	0.003	0.002	0.004	0.001		
-7	0.003	0.037 ***	-0.007	-0.037 ***	0.610	1988
	0.004	0.003	0.005	0.002		
-6	0.003	0.035 ***	-0.009 *	-0.037 ***	0.624	1988
	0.004	0.002	0.005	0.002		
-5	-0.007 *	0.035 ***	-0.004	-0.038 ***	0.627	1988
	0.004	0.002	0.005	0.001		
-4	-0.006	0.037 ***	-0.007	-0.038 ***	0.615	1988
	0.004	0.003	0.005	0.002		
-3	-0.02 ***	0.036 ***	-0.001	-0.039 ***	0.589	1988
	0.005	0.002	0.006	0.001		
-2	-0.042 ***	0.036 ***	0.013 *	-0.04 ***	0.592	1988
	0.005	0.002	0.007	0.002		
-1	-0.071 ***	0.033 ***	0.019 **	-0.036 ***	0.532	1988
	0.007	0.003	0.008	0.002		

Table B1: Estimated effect of the entrance of IEX around News on Quote Clustering excluding IEX: Unrestricted Sample *(continued)*

Time	EVENT	IEX	IEX*EVENT	RV_{t-1}	R2	No. Obs
0	-0.073 ***	0.031 ***	0.013	-0.037 ***	0.525	1988
	0.007	0.003	0.009	0.002		
1	-0.019 ***	0.034 ***	-0.005	-0.032 ***	0.662	1988
	0.004	0.002	0.006	0.001		
2	-0.001	0.034 ***	-0.013 **	-0.037 ***	0.661	1988
	0.004	0.002	0.005	0.001		
3	0.006	0.032 ***	-0.019 ***	-0.04 ***	0.669	1988
	0.004	0.002	0.005	0.001		
4	-0.002	0.036 ***	-0.014 ***	-0.035 ***	0.635	1988
	0.004	0.003	0.005	0.002		
5	0.005	0.035 ***	-0.019 ***	-0.036 ***	0.648	1988
	0.004	0.002	0.005	0.001		
6	0	0.037 ***	-0.016 ***	-0.037 ***	0.640	1988
	0.004	0.002	0.005	0.001		
7	0.001	0.037 ***	-0.01 *	-0.036 ***	0.624	1988
	0.004	0.003	0.005	0.002		
8	-0.001	0.033 ***	-0.011 **	-0.038 ***	0.629	1988
	0.004	0.002	0.005	0.002		
9	-0.002	0.035 ***	-0.008 *	-0.039 ***	0.641	1988
	0.004	0.002	0.005	0.001		
10	0	0.036 ***	-0.008 *	-0.036 ***	0.633	1988
	0.003	0.002	0.004	0.001		
11	0.006 *	0.036 ***	-0.011 **	-0.038 ***	0.644	1988
	0.003	0.002	0.005	0.001		
12	0.005 *	0.034 ***	-0.009 **	-0.039 ***	0.637	1988
	0.003	0.002	0.004	0.001		
13	0.001	0.038 ***	-0.009 *	-0.036 ***	0.635	1988
	0.004	0.002	0.004	0.002		
14	0.005	0.034 ***	-0.009 **	-0.039 ***	0.636	1988

Table B1: Estimated effect of the entrance of IEX around News on Quote Clustering excluding IEX: Unrestricted Sample *(continued)*

Time	EVENT	IEX	IEX*EVENT	RV_{t-1}	R2	No. Obs
	0.003	0.002	0.004	0.001		

This table reports the results obtained by estimating the regression specified in equation: $QC_{-IEX,t} = \beta_0 + \beta_1 EVENT + \beta_3 * EVENT + \beta_3 * IEX + \beta_4 * EVENT * IEX + \beta_t' X_t + \epsilon_t$. The results are reported for the Unrestricted Sample. The Unrestricted Sample uses the four intra-days moments that have at least 10 announcements. The dependent variable is QuoteClustering. Quote Clustering is calculated as: (Quote Clustering NBB + Quote Clustering NBO)/2, where Quote Clustering NBB (NBO) is defined as the proportion of exchanges at the NBB (NBO). EVENT is a dummy variable that takes the value of 1 if there is an event and 0 otherwise. IEX is a dummy variable that takes the value of 1 after the entrance of IEX and 0 otherwise. The interaction term: EVENT*IEX stands for the interaction between: IEX and EVENT. RV_{t-1} stands for the standarized volatility in the previous minute (RV: was calculated as the sum of the 1-second squared continuously compounded returns over 1-minute interval). The data has been aggregated by 1-minute interval. and are robust. Three dummies variables are included to control for intra-days patterns. In addition, 35 dummies are included to control for month by month variations. Robust standard errors are reported below the estimated coefficients. *** p<0.01, ** p<0.05, * p<0.1.

Table B2: Estimated effect of the entrance of IEX around News on Quote Clustering excluding IEX: Restricted Sample

Time	EVENT	IEX	IEX*EVENT	RV_{t-1}	R2	No. Obs
-14	0.006	0.028 ***	0.02	-0.044 ***	0.622	497
	0.01	0.004	0.016	0.003		
-13	-0.004	0.035 ***	0	-0.043 ***	0.643	497
	0.011	0.004	0.016	0.002		
-12	0.01	0.033 ***	-0.009	-0.046 ***	0.635	497
	0.009	0.004	0.015	0.003		
-11	-0.005	0.028 ***	0.01	-0.043 ***	0.628	497
	0.016	0.004	0.018	0.002		
-10	0.004	0.034 ***	0.013	-0.044 ***	0.643	497
	0.011	0.004	0.017	0.002		
-9	-0.004	0.03 ***	0.008	-0.043 ***	0.598	497
	0.012	0.004	0.018	0.004		
-8	0.004	0.031 ***	0.003	-0.047 ***	0.632	497
	0.015	0.004	0.019	0.003		
-7	0.011	0.032 ***	-0.005	-0.046 ***	0.627	497
	0.011	0.004	0.015	0.002		
-6	0.001	0.031 ***	0.017	-0.043 ***	0.626	497
	0.012	0.004	0.021	0.002		
-5	0.008	0.035 ***	-0.009	-0.041 ***	0.621	497
	0.017	0.004	0.019	0.003		
-4	-0.01	0.039 ***	0.041 **	-0.038 ***	0.600	497
	0.012	0.006	0.019	0.007		
-3	-0.009	0.034 ***	0.018	-0.043 ***	0.624	497
	0.011	0.004	0.016	0.002		
-2	-0.002	0.031 ***	0.004	-0.048 ***	0.664	497
	0.021	0.004	0.023	0.003		
-1	-0.026	0.032 ***	0.008	-0.043 ***	0.623	497
	0.018	0.004	0.021	0.002		
0	-0.182 ***	0.031 ***	-0.008	-0.046 ***	0.636	497

Table B2: Estimated effect of the entrance of IEX around News on Quote Clustering excluding IEX: Restricted Sample ***(continued)***

Time	EVENT	IEX	IEX*EVENT	RV_{t-1}	R2	No. Obs
	0.024	0.005	0.037	0.004		
1	-0.055 ***	0.042 ***	-0.046 **	-0.03 ***	0.718	497
	0.015	0.005	0.022	0.003		
2	-0.015	0.04 ***	-0.053 ***	-0.038 ***	0.678	497
	0.012	0.005	0.02	0.003		
3	0.003	0.032 ***	-0.052 **	-0.045 ***	0.680	497
	0.015	0.004	0.022	0.002		
4	-0.006	0.035 ***	-0.039 **	-0.042 ***	0.668	497
	0.013	0.004	0.019	0.003		
5	-0.004	0.036 ***	-0.049 **	-0.037 ***	0.678	497
	0.012	0.004	0.019	0.003		
6	-0.013	0.036 ***	-0.028	-0.045 ***	0.665	497
	0.012	0.004	0.019	0.003		
7	-0.01	0.032 ***	-0.013	-0.047 ***	0.678	497
	0.011	0.004	0.017	0.003		
8	-0.019	0.032 ***	-0.015	-0.042 ***	0.607	497
	0.014	0.005	0.02	0.004		
9	-0.023 *	0.037 ***	-0.012	-0.044 ***	0.628	497
	0.013	0.005	0.019	0.003		
10	-0.005	0.036 ***	-0.028 *	-0.044 ***	0.652	497
	0.009	0.004	0.015	0.002		
11	0.006	0.037 ***	-0.027 *	-0.042 ***	0.629	497
	0.01	0.004	0.014	0.003		
12	0.005	0.033 ***	-0.014	-0.045 ***	0.642	497
	0.009	0.004	0.013	0.002		
13	0.004	0.04 ***	-0.021	-0.042 ***	0.626	497
	0.01	0.004	0.014	0.002		
14	0.014	0.039 ***	-0.026 *	-0.045 ***	0.621	497
	0.01	0.005	0.014	0.003		

Table B2: Estimated effect of the entrance of IEX around News on Quote Clustering excluding IEX: Restricted Sample *(continued)*

Time	EVENT	IEX	IEX*EVENT	RV_{t-1}	R2	No. Obs

This table reports the results obtained by estimating the regression specified in equation: $QC_{-IEX,t} = \beta_0 + \beta_1 EVENT + \beta_3 * EVENT + \beta_3 * IEX + \beta_4 * EVENT * IEX + \beta_t' X_t + \epsilon_t$. The results are reported for the Restricted Sample, to those intra-days around 14:00. The dependent variable is QuoteClustering. Quote Clustering is calculated as: (Quote Clustering NBB + Quote Clustering NBO)/2, where Quote Clustering NBB (NBO) is defined as the proportion of exchanges at the NBB (NBO). EVENT is a dummy variable that takes the value of 1 if there is an event and 0 otherwise. IEX is a dummy variable that takes the value of 1 after the entrance of IEX and 0 otherwise. The interaction term: EVENT*IEX stands for the interaction between: IEX and EVENT. RV_{t-1} stands for the standarized volatility in the previous minute (RV: is calculated as the sum of the 1-second squared continuously compounded returns over 1-minute interval). The data has been aggregated by 1-minute interval. In addition, 35 dummies are included to control for month by month variations. Robust standard errors are reported below the estimated coefficients. *** p<0.01, ** p<0.05, * p<0.1.

Table B3: Estimated effect of High Impact vs Low Impact News on Quote Clustering: Unrestricted Sample

Time	HI	LI	R	No. Obs
-14	0	-0.002	0.504	2983
	0.004	0.002		
-13	0.003	-0.002	0.504	2983
	0.004	0.002		
-12	0.001	-0.002	0.507	2983
	0.004	0.002		
-11	-0.001	-0.002	0.491	2984
	0.004	0.002		
-10	0	0	0.511	2984
	0.004	0.002		
-9	-0.002	-0.002	0.488	2984
	0.004	0.002		
-8	-0.005	-0.001	0.504	2984
	0.004	0.002		
-7	-0.001	-0.004	0.495	2984
	0.004	0.002		
-6	-0.003	-0.001	0.493	2984
	0.004	0.002		
-5	-0.019 ***	-0.007 ***	0.500	2984
	0.004	0.002		
-4	-0.026 ***	-0.005 **	0.482	2984
	0.005	0.002		
-3	-0.055 ***	-0.011 ***	0.488	2984
	0.007	0.003		
-2	-0.084 ***	-0.021 ***	0.492	2984
	0.008	0.003		
-1	-0.136 ***	-0.032 ***	0.535	2984
	0.009	0.003		

Table B3: Estimated effect of High Impact vs Low Impact News on Quote Clustering: Unrestricted Sample *(continued)*

Time	HI	LI	R	No. Obs
0	-0.147 ***	-0.035 ***	0.543	2984
	0.009	0.003		
1	-0.102 ***	-0.021 ***	0.497	2984
	0.009	0.003		
2	-0.081 ***	-0.016 ***	0.490	2984
	0.008	0.003		
3	-0.07 ***	-0.012 ***	0.484	2984
	0.008	0.002		
4	-0.064 ***	-0.012 ***	0.485	2984
	0.007	0.002		
5	-0.053 ***	-0.011 ***	0.488	2984
	0.007	0.002		
6	-0.051 ***	-0.011 ***	0.486	2984
	0.007	0.002		
7	-0.049 ***	-0.01 ***	0.484	2984
	0.007	0.002		
8	-0.048 ***	-0.009 ***	0.490	2984
	0.007	0.002		
9	-0.046 ***	-0.01 ***	0.483	2984
	0.006	0.002		
10	-0.04 ***	-0.008 ***	0.488	2984
	0.006	0.002		
11	-0.039 ***	-0.007 ***	0.493	2984
	0.005	0.002		
12	-0.034 ***	-0.007 ***	0.486	2984
	0.005	0.002		
13	-0.036 ***	-0.005 *	0.494	2984
	0.005	0.002		
14	-0.033 ***	-0.005 **	0.491	2984

Table B3: Estimated effect of High Impact vs Low Impact News on Quote Clustering: Unrestricted Sample *(continued)*

Time	HI	LI	R	No. Obs
	0.005	0.002		

This table reports the results obtained by estimating the regression specified in equation 2 by distinguishing between High vs Low Impact News: $y_t = \beta_0 + \beta_1 HI + \beta_2 LI + \beta_t' X_t + \epsilon_t$. The results are reported for the Unrestricted Sample. The Unrestricted Sample uses the four intra-days moments that have at least 10 announcements. The dependent variable is Quote Clustering. Quote Clustering is calculated as: (Quote Clustering NBB + Quote Clustering NBO)/2, where Quote Clustering NBB (NBO) is defined as the proportion of of exchanges at the NBB (NBO). HI is a dummy variable that takes the value of 1 if there is a high impact macroeconomic news and 0 otherwise. LI is a dummy variable that takes the value of 1 if there is a low impact macroeconomic news and 0 otherwise. The data has been aggregated by 1-minute interval. Three dummies variables are included to control for intra-days patterns. In addition, 35 dummies are included to control for month by month variations. Robust standard errors are reported below the estimated coefficients. *** p<0.01, ** p<0.05, * p<0.1.

Table B4: Estimated effect of High Impact vs Low Impact News on Quoted Spreads: Unrestricted Sample

Time	HI	LI	R	No. Obs
-14	0 ***	0	0.411	740
	0	0		
-13	0 **	0	0.357	743
	0	0		
-12	0	0	0.351	741
	0	0		
-11	0 **	0 **	0.327	743
	0	0		
-10	0 *	0	0.373	742
	0	0		
-9	0	0	0.395	742
	0	0		
-8	0	0	0.407	743
	0	0		
-7	0 *	0	0.400	739
	0	0		
-6	0	0	0.345	741
	0	0		
-5	0 ***	0	0.373	743
	0	0		
-4	0 ***	0	0.205	745
	0	0		
-3	0.002 ***	0	0.242	745
	0	0		
-2	0.003 ***	0 *	0.286	745
	0.001	0		
-1	0.006 ***	0 *	0.433	738
	0.001	0		
0	0.009 ***	0 **	0.438	743

Table B4: Estimated effect of High Impact vs Low Impact News on Quoted Spreads: Unrestricted Sample *(continued)*

Time	HI	LI	R	No. Obs
	0.001	0		
1	0.003 ***	0 ***	0.444	742
	0	0		
2	0.001 ***	0	0.315	741
	0	0		
3	0.001 ***	0	0.337	744
	0	0		
4	0.001 ***	0	0.359	743
	0	0		
5	0.001 ***	0	0.324	743
	0	0		
6	0.001 ***	0	0.304	743
	0	0		
7	0.001 ***	0	0.276	743
	0	0		
8	0.001 ***	0	0.391	745
	0	0		
9	0.001 ***	0	0.357	743
	0	0		
10	0 ***	0	0.322	741
	0	0		
11	0 ***	0	0.362	740
	0	0		
12	0 ***	0	0.343	742
	0	0		
13	0 ***	0	0.355	744
	0	0		
14	0 ***	0	0.384	740
	0	0		

Table B4: Estimated effect of High Impact vs Low Impact News on Quoted Spreads: Unrestricted Sample *(continued)*

Time	HI	LI	R	No. Obs

This table reports the results obtained by estimating the regression specified in equation 2 by distinguishing between High vs Low Impact News: $y_t = \beta_0 + \beta_1 HI + \beta_2 LI + \beta_t' X_t + \epsilon_t$. The results are reported for the Unrestricted Sample. The Unrestricted Sample uses the four intra-days moments that have at least 10 announcements. The dependent variable is Quoted Spread. Quoted Spread is defined as the time weighted average quoted spread. HI is a dummy variable that takes the value of 1 if there is a high impact macroeconomic news and 0 otherwise. LI is a dummy variable that takes the value of 1 if there is a low impact macroeconomic news and 0 otherwise. The data has been aggregated by 1-minute interval. Three dummies variables are included to control for intra-days patterns. In addition, 35 dummies are included to control for month by month variations. Robust standard errors are reported below the estimated coefficients. *** p<0.01, ** p<0.05, * p<0.1.

Table B5: Estimated effect of High Impact vs Low Impact News on Quote Clustering: Restricted Sample around 14:00

Time	HI	LI	R	No. Obs
-14	0.008	0.001	0.452	746
	0.006	0.008		
-13	0.012 **	-0.012 *	0.433	746
	0.006	0.007		
-12	0.006	-0.008	0.458	746
	0.005	0.008		
-11	0.003	-0.01	0.439	746
	0.005	0.008		
-10	-0.004	0.001	0.463	746
	0.005	0.008		
-9	-0.005	-0.009	0.441	746
	0.006	0.008		
-8	-0.013 **	-0.01	0.451	746
	0.006	0.008		
-7	-0.008	-0.009	0.447	746
	0.007	0.008		
-6	-0.009	-0.012	0.425	746
	0.008	0.01		
-5	-0.042 ***	-0.013 *	0.461	746
	0.008	0.008		
-4	-0.064 ***	-0.002	0.442	746
	0.011	0.01		
-3	-0.117 ***	-0.005	0.489	746
	0.015	0.01		
-2	-0.166 ***	-0.025 **	0.546	746
	0.013	0.009		
-1	-0.245 ***	-0.039 ***	0.647	746
	0.014	0.011		
0	-0.281 ***	-0.047 ***	0.693	746

Table B5: Estimated effect of High Impact vs Low Impact News on Quote Clustering: Restricted Sample around 14:00 *(continued)*

Time	HI	LI	R	No. Obs
	0.015	0.01		
1	-0.236 ***	-0.027 **	0.640	746
	0.014	0.011		
2	-0.194 ***	-0.018 **	0.586	746
	0.016	0.009		
3	-0.175 ***	-0.006	0.548	746
	0.016	0.01		
4	-0.159 ***	-0.016 *	0.535	746
	0.014	0.009		
5	-0.149 ***	-0.011	0.545	746
	0.013	0.008		
6	-0.143 ***	-0.009	0.555	746
	0.013	0.008		
7	-0.138 ***	-0.007	0.540	746
	0.013	0.008		
8	-0.133 ***	-0.005	0.543	746
	0.013	0.007		
9	-0.124 ***	-0.008	0.517	746
	0.012	0.007		
10	-0.107 ***	-0.018 **	0.518	746
	0.01	0.008		
11	-0.092 ***	-0.014	0.488	746
	0.009	0.009		
12	-0.086 ***	-0.012	0.484	746
	0.008	0.008		
13	-0.082 ***	-0.005	0.486	746
	0.009	0.009		
14	-0.083 ***	-0.007	0.472	746
	0.008	0.01		

Table B5: Estimated effect of High Impact vs Low Impact News on Quote Clustering: Restricted Sample around 14:00 *(continued)*

Time	HI	LI	R	No. Obs

makecell[l]This table reports the results obtained by estimating the regression specified in equation 2 by distinguishing between High vs Low Impact News: $y_t = \beta_0 + \beta_1 HI + \beta_2 LI + \beta_t' X_t + \epsilon_t$. The results are reported for the Restricted Sample, to those intra-days around 14:00.dependent variable is Quote Clustering. Quote Clustering is calculated as: (Quote Clustering NBB + Quote Clustering NBO)/2, where Quote Clustering NBB (NBO) is defined as the proportion of exchanges at the NBB (NBO. HI is a dummy variable that takes the value of 1 if there is a high impact macroeconomic news and 0 otherwise. LI is a dummy variable that takes the value of 1 if there is a low impact macroeconomic news and 0 otherwise. The data has been aggregated by 1-minute interval. Three dummies variables are included to control for intra-days patterns. In addition, 35 dummies are included to control for month by month variations. Robust standard errors are reported below the estimated coefficients. *** p<0.01, ** p<0.05, * p<0.1.

Table B6: Estimated effect of High Impact vs Low Impact News on Quoted Spreads: Restricted Sample around 14:00

Time	HI	LI	R	No. Obs
-14	0 ***	0	0.411	740
	0	0		
-13	0 **	0	0.357	743
	0	0		
-12	0	0	0.351	741
	0	0		
-11	0 **	0 **	0.327	743
	0	0		
-10	0 *	0	0.373	742
	0	0		
-9	0	0	0.395	742
	0	0		
-8	0	0	0.407	743
	0	0		
-7	0 *	0	0.400	739
	0	0		
-6	0	0	0.345	741
	0	0		
-5	0 ***	0	0.373	743
	0	0		
-4	0 ***	0	0.205	745
	0	0		
-3	0.002 ***	0	0.242	745
	0	0		
-2	0.003 ***	0 *	0.286	745
	0.001	0		
-1	0.006 ***	0 *	0.433	738
	0.001	0		
0	0.009 ***	0 **	0.438	743

Table B6: Estimated effect of High Impact vs Low Impact News on Quoted Spreads: Restricted Sample around 14:00 *(continued)*

Time	HI	LI	R	No. Obs
	0.001	0		
1	0.003 ***	0 ***	0.444	742
	0	0		
2	0.001 ***	0	0.315	741
	0	0		
3	0.001 ***	0	0.337	744
	0	0		
4	0.001 ***	0	0.359	743
	0	0		
5	0.001 ***	0	0.324	743
	0	0		
6	0.001 ***	0	0.304	743
	0	0		
7	0.001 ***	0	0.276	743
	0	0		
8	0.001 ***	0	0.391	745
	0	0		
9	0.001 ***	0	0.357	743
	0	0		
10	0 ***	0	0.322	741
	0	0		
11	0 ***	0	0.362	740
	0	0		
12	0 ***	0	0.343	742
	0	0		
13	0 ***	0	0.355	744
	0	0		
14	0 ***	0	0.384	740
	0	0		

Table B6: Estimated effect of High Impact vs Low Impact News on Quoted Spreads: Restricted Sample around 14:00 ***(continued)***

Time	HI	LI	R	No. Obs

This table reports the results obtained by estimating the regression specified in equation 2 by distinguishing between High vs Low Impact News: $y_t = \beta_0 + \beta_1 HI + \beta_2 LI + \beta_t' X_t + \epsilon_t$. The results are reported for the Restricted Sample, to those intra-days around 14:00. The dependent variable is Quoted Spread. Quoted Spread is defined as the time weighted average quoted spread. HI is a dummy variable that takes the value of 1 if there is a high impact macroeconomic news and 0 otherwise. LI is a dummy variable that takes the value of 1 if there is a low impact macroeconomic news and 0 otherwise. The data has been aggregated by 1-minute interval. Three dummies variables are included to control for intra-days patterns. In addition, 35 dummies are included to control for month by month variations. Robust standard errors are reported below the estimated coefficients. *** p<0.01, ** p<0.05, * p<0.1.

Table B7: Estimated effect of an increase in fragmentation around High Impact vs Low Impact on Quote Clustering excluding IEX: Unrestricted Sample

Time	HI	LI	IEX	IEX*HI	IEX*LI	RV_{t-1}	R2	No. Obs
-14	0.006	0.005	0.035 ***	0.005	-0.003	-0.037 ***	0.598	1491
	0.012	0.005	0.003	0.013	0.007	0.003		
-13	0.006	-0.001	0.039 ***	0.008	-0.007	-0.032 ***	0.567	1491
	0.014	0.005	0.005	0.016	0.007	0.005		
-12	0.014 **	0.001	0.036 ***	-0.011	-0.008	-0.034 ***	0.604	1987
	0.006	0.004	0.003	0.007	0.005	0.003		
-11	0.009	0.003	0.034 ***	-0.008	-0.009 *	-0.038 ***	0.620	1987
	0.006	0.004	0.002	0.007	0.005	0.001		
-10	0.006	0.007 *	0.037 ***	-0.005	-0.013 **	-0.036 ***	0.633	1988
	0.005	0.004	0.002	0.007	0.005	0.001		
-9	0.004	0.007	0.036 ***	-0.002	-0.013 **	-0.036 ***	0.614	1988
	0.006	0.004	0.002	0.007	0.005	0.001		
-8	0.006	0.005	0.038 ***	-0.01	-0.013 **	-0.039 ***	0.626	1988
	0.006	0.004	0.002	0.008	0.005	0.001		
-7	0.007	0.001	0.037 ***	-0.009	-0.006	-0.037 ***	0.610	1988
	0.007	0.004	0.003	0.008	0.005	0.002		
-6	0	0.004	0.034 ***	-0.01	-0.008	-0.037 ***	0.623	1988
	0.006	0.004	0.002	0.007	0.006	0.002		
-5	-0.016 **	-0.002	0.035 ***	0.001	-0.006	-0.038 ***	0.628	1988
	0.006	0.004	0.002	0.008	0.005	0.001		
-4	-0.014 **	-0.002	0.037 ***	-0.01	-0.005	-0.038 ***	0.616	1988
	0.006	0.004	0.003	0.009	0.005	0.002		
-3	-0.046 ***	-0.007 *	0.036 ***	0.004	-0.004	-0.039 ***	0.600	1988
	0.009	0.004	0.002	0.013	0.005	0.001		
-2	-0.087 ***	-0.021 ***	0.036 ***	0.022 *	0.008	-0.04 ***	0.619	1988
	0.009	0.005	0.002	0.013	0.006	0.002		
-1	-0.136 ***	-0.041 ***	0.032 ***	0.021	0.017 **	-0.036 ***	0.590	1988
	0.011	0.006	0.003	0.016	0.007	0.002		
0	-0.144 ***	-0.039 ***	0.03 ***	0.013	0.012 *	-0.038 ***	0.597	1988

Table B7: Estimated effect of an increase in fragmentation around High Impact vs Low Impact on Quote Clustering excluding IEX: Unrestricted Sample *(continued)*

Time	HI	LI	IEX	IEX*HI	IEX*LI	RV_{t-1}	R2	No. Obs
	0.011	0.006	0.003	0.016	0.007	0.001		
1	-0.016 *	-0.021 ***	0.034 ***	-0.025 **	0.005	-0.032 ***	0.664	1988
	0.008	0.004	0.002	0.012	0.005	0.001		
2	0.008	-0.006	0.034 ***	-0.033 ***	-0.002	-0.037 ***	0.663	1988
	0.007	0.004	0.002	0.01	0.005	0.001		
3	0.003	0.007 *	0.032 ***	-0.024 **	-0.016 ***	-0.04 ***	0.669	1988
	0.007	0.004	0.002	0.01	0.005	0.001		
4	-0.009	0.002	0.036 ***	-0.015	-0.013 **	-0.034 ***	0.635	1988
	0.007	0.004	0.003	0.009	0.005	0.002		
5	0.005	0.005	0.035 ***	-0.029 ***	-0.014 ***	-0.036 ***	0.649	1988
	0.006	0.004	0.002	0.009	0.005	0.001		
6	-0.001	-0.001	0.037 ***	-0.023 **	-0.011 **	-0.037 ***	0.640	1988
	0.007	0.004	0.002	0.01	0.005	0.001		
7	-0.008	0.005	0.037 ***	-0.005	-0.012 **	-0.036 ***	0.625	1988
	0.007	0.004	0.003	0.009	0.006	0.002		
8	-0.013 *	0.005	0.033 ***	-0.007	-0.014 ***	-0.038 ***	0.631	1988
	0.007	0.004	0.002	0.009	0.005	0.002		
9	-0.011 *	0.002	0.035 ***	-0.007	-0.009 *	-0.039 ***	0.643	1988
	0.007	0.004	0.002	0.009	0.005	0.001		
10	-0.003	0.002	0.036 ***	-0.01	-0.006	-0.036 ***	0.633	1988
	0.005	0.004	0.002	0.008	0.005	0.001		
11	0.004	0.006	0.036 ***	-0.015 **	-0.009	-0.038 ***	0.644	1988
	0.005	0.004	0.002	0.007	0.005	0.001		
12	0.004	0.006	0.034 ***	-0.012 *	-0.008	-0.039 ***	0.637	1988
	0.005	0.004	0.002	0.007	0.005	0.001		
13	-0.003	0.003	0.038 ***	-0.009	-0.008	-0.036 ***	0.635	1988
	0.006	0.004	0.002	0.007	0.005	0.002		
14	0.003	0.006	0.034 ***	-0.007	-0.01 *	-0.039 ***	0.636	1988
	0.005	0.004	0.002	0.007	0.005	0.001		

Table B7: Estimated effect of an increase in fragmentation around High Impact vs Low Impact on Quote Clustering excluding IEX: Unrestricted Sample *(continued)*

Time	HI	LI	IEX	IEX*HI	IEX*LI	RV_{t-1}	R2	No. Obs

This table reports the results obtained by estimating the regression specified in equation 5 by distinguishing between High vs Low Impact News. The results are reported for the Unrestricted Sample. The Unrestricted Sample uses the four intra-days moments that have at least 10 announcements. The dependent variable is Quote Clustering. Quote Clustering is calculated as: (Quote Clustering NBB + Quote Clustering NBO)/2, where Quote Clustering NBB (NBO) is defined as the proportion of exchanges at the NBB (NBO). High Impact is a dummy variable that takes the value of 1 if there is a high impact macroeconomic news and 0 otherwise. Low Impact is a dummy variable that takes the value of 1 if there is a high impact macroeconomic news and 0 otherwise. IEX is a dummy variable that takes the value of 1 after the entrance of IEX and 0 otherwise. The interaction term: EVENT*HI stands for the interaction between: IEX and HI.The interaction term: EVENT*LI stands for the interaction between: IEX and LI. RV_{t-1} stands for the standarized volatility in the previous minute (RV: was calculated as the sum of the 1-second squared continuously compounded returns over 1-minute interval). The data has been aggregated by 1-minute interval. Three dummies variables are included to control for intra-days patterns. In addition, 35 dummies are included to control for month by month variations. Robust standard errors are reported below the estimated coefficients. *** p<0.01, ** p<0.05, * p<0.1.

Table B8: Estimated effect of an increase in fragmentation around High Impact vs Low Impact on Quote Clustering excluding IEX: Restricted Sample

Time	HI	LI	IEX	IEX*HI	IEX*LI	RV_{t-1}	R2	No. Obs
-14	0.009	0.01	0.04 ***	-0.003	-0.011	-0.042 ***	0.572	497
	0.012	0.012	0.005	0.014	0.016	0.003		
-13	0.005	0.006	0.035 ***	0.005	-0.024 *	-0.042 ***	0.518	497
	0.015	0.011	0.005	0.017	0.013	0.003		
-12	-0.006	0.008	0.034 ***	0.014	-0.032	-0.046 ***	0.564	497
	0.009	0.012	0.005	0.012	0.02	0.003		
-11	-0.004	-0.001	0.038 ***	0.006	-0.02	-0.043 ***	0.564	497
	0.009	0.013	0.004	0.012	0.015	0.003		
-10	-0.008	0.013	0.038 ***	0.001	-0.016	-0.044 ***	0.560	497
	0.008	0.015	0.004	0.013	0.017	0.002		
-9	-0.014	-0.018	0.032 ***	0.018	0.022	-0.046 ***	0.572	497
	0.013	0.014	0.004	0.016	0.02	0.002		
-8	-0.019 *	-0.005	0.035 ***	0.003	-0.002	-0.049 ***	0.615	497
	0.011	0.01	0.004	0.018	0.015	0.003		
-7	-0.013	-0.005	0.036 ***	0.006	0	-0.045 ***	0.558	497
	0.012	0.013	0.005	0.02	0.022	0.003		
-6	-0.012	-0.003	0.025 ***	0.001	0.007	-0.052 ***	0.573	497
	0.015	0.009	0.004	0.022	0.02	0.003		
-5	-0.047 ***	0.01	0.033 ***	0.008	-0.022	-0.049 ***	0.539	497
	0.017	0.012	0.005	0.023	0.022	0.003		
-4	-0.056 ***	-0.001	0.035 ***	-0.01	-0.009	-0.048 ***	0.564	497
	0.017	0.012	0.005	0.03	0.017	0.003		
-3	-0.112 ***	0.009	0.035 ***	-0.021	-0.015	-0.048 ***	0.551	497
	0.026	0.014	0.005	0.04	0.019	0.003		
-2	-0.154 ***	-0.019	0.035 ***	-0.015	0.002	-0.044 ***	0.641	497
	0.02	0.015	0.005	0.03	0.023	0.003		
-1	-0.25 ***	-0.057 ***	0.028 ***	-0.012	0.015	-0.044 ***	0.715	497
	0.019	0.019	0.005	0.027	0.025	0.004		
0	-0.277 ***	-0.05 ***	0.032 ***	-0.018	0.006	-0.045 ***	0.779	497

Table B8: Estimated effect of an increase in fragmentation around High Impact vs Low Impact on Quote Clustering excluding IEX: Restricted Sample *(continued)*

Time	HI	LI	IEX	IEX*HI	IEX*LI	RV_{t-1}	R2	No. Obs
	0.017	0.016	0.005	0.026	0.021	0.003		
1	-0.086 ***	-0.038 **	0.044 ***	-0.075 ***	0.009	-0.027 ***	0.743	497
	0.022	0.016	0.005	0.025	0.023	0.003		
2	-0.022	-0.014	0.04 ***	-0.073 **	-0.021	-0.037 ***	0.682	497
	0.023	0.012	0.005	0.029	0.022	0.003		
3	-0.014	0.013	0.033 ***	-0.064 *	-0.027	-0.043 ***	0.686	497
	0.029	0.012	0.004	0.034	0.02	0.003		
4	-0.023	0.006	0.037 ***	-0.047 *	-0.02	-0.04 ***	0.673	497
	0.023	0.011	0.004	0.028	0.02	0.003		
5	-0.011	-0.001	0.037 ***	-0.068 **	-0.017	-0.036 ***	0.684	497
	0.02	0.012	0.004	0.028	0.018	0.003		
6	-0.032 *	0.003	0.038 ***	-0.039	-0.008	-0.042 ***	0.673	497
	0.019	0.01	0.004	0.029	0.014	0.003		
7	-0.029	0.006	0.034 ***	-0.026	0.009	-0.045 ***	0.688	497
	0.017	0.011	0.004	0.023	0.017	0.002		
8	-0.049 **	0.01	0.034 ***	-0.021	-0.004	-0.039 ***	0.625	497
	0.02	0.014	0.005	0.028	0.018	0.004		
9	-0.047 **	0.002	0.039 ***	-0.014	-0.007	-0.041 ***	0.638	497
	0.018	0.011	0.005	0.027	0.015	0.003		
10	0.001	-0.012	0.036 ***	-0.046 **	-0.003	-0.044 ***	0.653	497
	0.013	0.01	0.004	0.023	0.016	0.002		
11	0.003	0.008	0.038 ***	-0.03	-0.022	-0.041 ***	0.628	497
	0.014	0.013	0.004	0.019	0.018	0.003		
12	0.004	0.007	0.033 ***	-0.019	-0.006	-0.045 ***	0.641	497
	0.011	0.012	0.004	0.017	0.018	0.002		
13	0.002	0.006	0.04 ***	-0.022	-0.019	-0.042 ***	0.624	497
	0.014	0.014	0.004	0.018	0.021	0.003		
14	0.009	0.02	0.039 ***	-0.024	-0.027	-0.045 ***	0.620	497
	0.013	0.014	0.005	0.019	0.018	0.003		

Table B8: Estimated effect of an increase in fragmentation around High Impact vs Low Impact on Quote Clustering excluding IEX: Restricted Sample *(continued)*

Time	HI	LI	IEX	IEX*HI	IEX*LI	RV_{t-1}	R2	No. Obs

This table reports the results obtained by estimating the regression specified in equation 5 by distinguishing between High vs Low Impact News. The results are reported for the Restricted Sample, to those intra-days around 14:00. The dependent variable is Quote Clustering. Quote Clustering is calculated as: (Quote Clustering NBB + Quote Clustering NBO)/2, where Quote Clustering NBB (NBO) is defined as the proportion of exchanges at the NBB (NBO). High Impact is a dummy variable that takes the value of 1 if there is a high impact macroeconomic news and 0 otherwise. Low Impact is a dummy variable that takes the value of 1 if there is a high impact macroeconomic news and 0 otherwise. IEX is a dummy variable that takes the value of 1 after the entrance of IEX and 0 otherwise. The interaction term: EVENT*HI stands for the interaction between: IEX and HI.The interaction term: EVENT*LI stands for the interaction between: IEX and LI. RV_{t-1} stands for the standarized volatility in the previous minute (RV: was calculated as the sum of the 1-second squared continuously compounded returns over 1-minute interval). The data has been aggregated by 1-minute interval. In addition, 35 dummies are included to control for month by month variations. Robust standard errors are reported below the estimated coefficients. *** p<0.01, ** p<0.05, * p<0.1.

Chapter 2

Non-Standard Errors

2.1 Introduction

Empirical researchers usually come up with an hypothesis and they test it by using different samples, filters, methodologies, etc. Many times, the same question addressed by many authors yield different results. For instance, there is no consensus if an increase in fragmentation has on overall positive impact on market quality. In particular, Baldauf and Mollner (2021) show that an increase in fragmentation leads to an increase in adverse selection risk, which has a negative impact on market quality. However, Foucault and Menkveld (2008), show that an increase in fragmentation leads to an increase in the depth in the market, which is positive for market quality. One could claim that the field is too broad to reach similar conclusions, as different researchers end up testing an hypothesis by using different approaches, samples, filters, methodologies, etc.

Suppose that different researchers were to test the same hypothesis using the same data: how estimates would vary across them? To this end, Finance Crowed Analyses Project (FINCAP) launched a project to understand the mechanism behind the differences in results, or to be more precise to learn more about the scientific process. With this aim, different research teams (RTs) had to test the same hypothesis using the same data. The dispersion in estimates across researchers (non-standard error) is the object of study of FINCAP.

The project description would contain information about the type of data all RTs would receive. The sample would contain proprietary data on the EuroStox 50 futures traded in EUREX. Given my interest in topics related to market microstructure and the importance of understanding why there are discrepancies across results, I participated in the project together with Sophie Moinas as a RT. Given our participation as a RT, the rest of the introduction is devoted to describe briefly the experiment from the point of view of a Research Team.

Stage 1 (January 11 - March 23, 2021.): we received the detailed instructions along with access to the RT-sample. By the end of the deadline we should hand in the results (short paper plus code).

Market quality was at the heart of the hypothesis to be tested. The emergence of electronic order matching systems (automated exchanges) and electronic order generation systems (algorithms) marked a turning point in financial markets. Investors used to trade via broker-dealers while now investors trade in electronic limit-order markets. While with the entrance of the limit-order book, investors would send only aggressive orders (market or marketable orders), now investors can send passive orders too (limit orders). Over time exchanges not just upgraded their systems, but they offered a higher variety of services, such as co-location, which provides a speed advantage to certain investors. In addition, in the current trading environment there is an increased presence of a sub-group of Algorithmic Trading: High Frequency Traders, which takes advantage of their speed advantage. One question that arises is if market quality have experienced an improvement in time? In order to answer the previous question, we had to test 6 hypothesis.

In order to test the six hypothesis, we received proprietary data on the EuroStox 50 futures traded in EUREX. The data consists in buy and sell trade records and contains information about the account role: Agency trade (A), Market-maker principal trade (M), Proprietary (P). In addition, a flag identifies if a trade pertains to aggressive orders (market order or marketable limit order) or to a passive order (limit order). The flag is made available only from November 2009 onwords.

To test some of the hypothesis requires to compute the midpoint and to identify if an order was a buy or a sell initiated order.[1] Since the flag to distinguish between aggressive and passive orders is not available for the period before November 2009, we proceed to build an algorithm with this aim. In order to test the accuracy of our algorithm, we compare the outcome of the classification with the flag provided by Deutsche Börse after November 2009. We find that the algorithm has an accuracy level on average of 90% with a standard deviation of 2.5%. The algorithm enabled us to test the six hypothesis for the full sample.

With the aim to test the six hypothesis, we suggested a measure for each one, and in order to enable comparison across teams, we followed carefully the instructions we received in regards to the methodology. Thus, based on the full sample, we estimated the average per-year change in percentage terms and we submitted the short paper (benchmark model). By using the benchmark model, we did not reject any of the hypothesis related to market quality.

Since, the main objective of the experiment was to see how conclusions would change along the way, I will explain very briefly the next two submissions of the short paper and if there were

[1]The midpoint is calculated as the average between the best prices at the Bid and at the Ask.

any major changes with respect to the first submission of the short paper. And finally, I will devote the last part of the introduction to explain, how our main results using the benchmark model would change by using different methodologies.

- *Stage 2 (May 10 - May 28, 2021.)* RTs receive feedback from two anonymous peers and are allowed to update their analysis based on it: The results in the benchmark model did not change as a consequence of the peer evaluation. However, we not just improved the presentation of our results, but we presented also the results for some extra analyses. Hence, we distinguished between: i) benchmark model; ii) we run a time trend regression with monthly data instead of yearly data; and iii) we run different robustness checks.

- Stage 3 (May 31 - June 18, 2021.) RTs receive the receive best papers based on the average raw PE score. Friday, 2021-06-18 deadline for submitting the third version of the short paper by research teams. For this case, the same holds true. By using the benchmark model, the main conclusions did not change in comparison to the previous 2 versions submitted.

- Stage 4 (June 20 - June 28, 2021.) RTs are allowed to Bayesian update their results (i.e., estimates and standard errors) taking in all the information that has become available to them, in particular the five best papers. For this cases, we made a minor change in the algorithm to distinguish between aggressive and passive orders, leading us to update our results. Despite the change in the algorithm, the change in estimates was minor and it was such that the main conclusions for the hypothesis tested did not change for the benchmark model. In addition, for the time trend regressions, we computed Newey West standard errors with 12 lags.

Our main final conclusions vary with the frequency of the data used. When yearly data is used (benchmark model), the results are not statistically significant from zero. This points out to no change in trend in terms of market quality. However, when the monthly data is used, we found that while client share volume decreases over time by 1.3% on a monthly basis, the client realized spread increases by 1.2% on average on a monthly basis. In addition the fraction of client trades executed via aggressive orders have increased by 0.4% on a monthly basis. Our main contributions to the FINCAP project are as follows: 1. we construct an algorithm to distinguish between aggressive and passive orders, which could help researchers when the aggressor flag is not available (the 90% accuracy of the algorithm is outstanding); 2. we run the analyses over the full sample; 3. besides the benchmark model, we used a time trend regression, which would lead to different conclusions, suggesting that the methodology plays a key role on the conclusions reached; and finally 4. we did not limit the analyses to one metric, but to different variations of the metric suggested, which enable us to run some robustness check.

This chapter is structured as follows: Section 2.2 includes the short paper, Sophie Moinas and me submitted in the last round. The short paper contains a section: Methodology devoted to describe the filters, metrics and the methodology used to test the main hypothesis; Results: where the results obtained for the 6 hypothesis are reported and discussed. Finally, Section 2.3 contains the article: Non-Standard Errors (forthcoming in the Journal of Finance) which explains in more details the experiment and the main conclusions drawn.

2.2 Short Paper

2.2.1 Methodology

Filtering data. Deutsche Börse data that is made available by the #fincap team consist in buy and sell trade records on the EuroStoxx 50 futures from January 2002 to December 2018. At any given date, future contracts with maturity March, June, September, December are available. We focus on the lead month contract, that is the most liquid contract.[2] We focus on trades occurring during the opening hours of the equity markets on which the constituents of the EuroStoxx 50 index are traded, that is between 9:00am and 5:30pm.[3,4]

[2]We exclude the settlement date since the settlement price is established at 12pm.

[3]Futures contracts are traded from 8am to 10pm, but the daily settlement price is determined at 5:30pm.

[4]The first trade is often recorded after 9am. Since we are not allowed to match our data with external data sources to analyze the source of the delayed opening, we decided to keep all observations.

Identifying order aggressiveness. As of November 2009, the dataset includes a flag identifying whether the trade pertains to a market order or marketable limit order (hereafter aggressive orders), or to a limit order (hereafter passive orders). Addressing hypothesis 1, 2, 4 and 5 however requires to identify passive orders to compute the **midpoint** and **sign orders' aggressiveness** as of the beginning of the sample period. To this end, we build on the observation that aggressive orders have a larger trade size than executed passive orders, and we implement the following algorithm to classify order aggressiveness.[5] We first match the buy and sell sides of transactions by focusing on executed orders reported at the same price and at the same timestamp.[6] We then identify the order that has the larger size, and label it as aggressive. By extension, all the orders in the same direction will be labelled as aggressive, and all the orders in the opposite direction will be labelled as passive. When orders have the same size, we apply the tick rule (Lee and Ready (1991)) to classify aggressiveness.[7]

To analyze the accuracy of our algorithm, we compare the outcome of its classification with the flag provided by Deutsche Börse after November 2009. On a given day, we find that our algorithm classifies orders in the same category as the original data in 90% of the cases (with a standard deviation of 2.5%). Having checked the proper accuracy of our algorithm, we use it to classify all the trades over the sample period 2002-2018. We also compare our findings with those obtained using the flag when available, that is, over the period 2009-2018.

Using trade data to rebuild the midquote. Our data record multiple orders executed at precisely the same time-stamp, sometimes at different trade prices. We cannot be sure that observations are recorded in the dataset depending on the exact sequence of execution. Using the classification of order aggressiveness described above, we compute the trade imbalance at timestamp τ. If it is positive, it is likely that aggressive buy orders have walked up the book.

[5]Our data indeed do not report a single observation by trade, but the executed buy and sell orders. An incoming market buy order may be executed against a single sell limit order, in which case one would observe an executed buy order and an executed sell order of the same size; or it can be executed against multiple sell limit orders, in which case one would observe an executed buy order and a series of executed sell orders of smaller size.

[6]In the post 2009 data that include the aggressor's flag, sell executed orders are sometimes reported milliseconds after the corresponding buy trade, we therefore also screen sell orders reported at the consecutive time-stamp (within one second).

[7]Any buy order with a price larger (resp. lower) than the previous price is labelled as aggressive (resp. passive), and symmetrically for sell orders.

We consequently define the best ask prevailing after all trades occurring at time τ, A_τ, as the maximum trade price observed in the sequence of aggressive buy orders. Conversely, if trade imbalance is negative, we define the best bid, B_τ, as the minimum trade price of aggressive sell orders. For each τ, we therefore identify either the best ask prevailing just after the trades, or the best bid. To define the missing quote (i.e., the best bid or the best ask respectively), we propagate the best quote prevailing at the previous timestamp, provided that propagation would not result in a strictly negative bid-ask spread.[8,9] We are aware that our approach provides an upper bound on the quoted spread and may sometimes generate outliers, e.g., after a long sequence of buy order imbalances we may not capture an increase in the quoted bid price. We would thus measure an abnormally large spread; to avoid such outliers, we trim all the measures to 99%. We finally define the midquote as $m_\tau = \frac{A_\tau + B_\tau}{2}$.

Methodology. Finally, for all the measures X proposed below, we first compute the daily measure X_t on day t, then its average $\bar{X}_f$ across days at frequency f, where f is either year y or month m. To enable comparisons across teams and as requested, we compute the average per-year change in percentage terms as $\Delta X_y = 100\frac{\bar{X}_y - \bar{X}_{y-1}}{\bar{X}_{y-1}}$, and we use a Student test of the hypothesis $\Delta X_y = 0$ to test all six hypothesis. As a complement, we run the regression $\bar{X}_m = a_0 + a_1 \times Time_y + FE + \varepsilon_y$, where $Time_y$ captures a linear yearly time trend, and we test whether $a_1 = 0$. The regressions include twelve month fixed effects to capture seasonality and we compute Newey West standard errors with 12 lags.

2.2.2 Results

Hypothesis 1 *Market efficiency has not changed over time.*

We propose to test hypothesis 1 using the variance-ratio (VR) methodology, a standard approach to assess how closely prices follow a random walk (see Lo and MacKinglay (1988)).

[8] An alternative approach consists in rebuilding the quoted bid-ask spread prevailing at a given time τ using a forward-looking approach, i.e., considering that the missing quote is the best quote computed from the subsequent sequence. We have also implemented this approach and found similar results.

[9] We did not impose any condition on the time elapsed between time τ and that of the last observation of the missing quote. Since we observed that the time between two consecutive time-stamps decreased over the sample period, imposing a condition that may be more restrictive for the beginning of the sample may bias the results.

It exploits the fact that under the null hypothesis of a random walk, the variance of random walk increments is linear in all sampling intervals: the ratio $VR(1min, 5min)$ of the variance of five times 1-minute log midquote returns divided by the variance of 5-minute log midquote returns should be equal to one. Because we are interested in the gap between actual and efficient prices in either direction, we propose to measure price (in-)efficiency as the distance between the variance ratio and its theoretical value if prices were following a random walk:

$$Distance_t(1min, 5min) = |VR(1min, 5min)_t - 1|. \tag{2.1}$$

A reduction *Distance* captures an improvement in price efficiency. We do not reject the null hypothesis 1. We find that price efficiency declined as our measure *Distance* increased by 3.136% on average per year over the full sample period where the standard error of this change is 5.311% and the resulting t-statistic is 0.591.[10] Table 2.1.C however shows a significantly increasing yearly trend over the period 2009-2018.

Hypothesis 2 *The realized spread on market orders has not changed over time.*

For each market or marketable limit order i executed at time τ, we compute the daily average of the volume-weighted realized spread over one-minute intervals θ as Holden and Jacobsen (2014), then its daily average $RSpd_t$, as follows:

$$RSpd_t = \sum_{\theta \in t} \sum_{i_\tau, \tau \in \theta} 2 \times (p_{i_\tau} - m_{\tau+5mn})\, d_{i_\tau} \times \frac{q_{i_\tau}}{\sum_{i_{\tau'}, \tau \in \theta} q_{i_{\tau'}}}, \tag{2.2}$$

where p_{i_τ} is the trade price of the aggressive order i executed at time τ, $m_{\tau+5mn}$ is the midquote observed five minutes after the trade, d_{i_τ} takes value +1 if the aggressive order is a buy order and −1 if it is a sell order, q_{i_τ} is the transaction size. The realized spread could be thought of as the gross-profit component of the spread as earned by the limit order submitter if the trader undo its position later on.[11]

[10] Variance ratios can be calculated using other time intervals. We have repeated our analysis using longer intervals of VR(5min, 15min) and shorter intervals of VR(10sec, 60sec) and reported the corresponding results in Table 1, B.

[11] We have also measured the realized spread using the midquote after after 1 minute.The midquote after 5 minutes

We find an average realized spread of 0.217 euros. We do not reject the null hypothesis 2. We find that the realized spread on market orders increased as our measure $RS\,pd$ increased by 8.062% on average per year over the sample period where the standard error of this change is 6.734%, albeit not significantly as indicated by a t-statistic is 1.197. This result, consistent with the regression, shows that realized spread on market orders has not changed over time, which may be surprising as advancements in technologies such as algorithmic and high-frequency-trading have improved liquidity in equity markets in the 2000s (see e.g. Hendershott et al. (2011)).

Hypothesis 3 *Client share volume as a fraction of total volume has not changed over time.*

The dataset includes a variable that identifies Agency trades (i.e., a trade an exchange member does for a client). We compute the daily volume of orders submitted on behalf of clients V_t^A by summing the trade sizes of orders whenever the trade is flagged on the client account, and we divide it by the total daily volume to find the daily proportion of client volume CSV_t (in %).

We do not reject the null hypothesis 3. We find that client share volume declined as our measure CSV declined by 3.299% on average per year on the full sample where the standard error of this change is 1.934% and the resulting t-statistic is -1.705. Using the trend regression, we also reject the null hypothesis 3 against the alternative hypothesis that client share volume decreased at the 1%.

Hypothesis 4 *Client realized spreads have not changed over time.*

Building on the definition of the realized spread on an executed order i_τ in equation (2.2), the daily realized spread on clients' market or marketable limit orders $RS\,pd_t^A$ is computed on the subset on aggressive orders submitted on behalf of clients.

We do not reject the null hypothesis 4. We find that client realized spreads increased as our measure $RS\,pd^A$ increased by 7.094% on average per year on the full sample where the standard error of this change is 5.108% and the resulting t-statistic is 1.389. Figure 2.1 shows that realized

may be more relevant for the beginning of the sample period. In addition, the relative realized spread after 5 minutes has been calculated by dividing the realized spread at time τ by the midquote. Results reported in Table 2.1.B are consistent.

spread on clients' aggressive orders however increased between 2002 and 2018. Using time trend regressions, we reject the null hypothesis 4 against the alternative hypothesis that client realized spreads have increased at the 1% level of significance.

Hypothesis 5 *The fraction of client trades executed via market orders and marketable limit orders has not changed over time.*

Having determined orders' aggressiveness over the full sample following the procedure described above, we use the account flag to compute the fraction of client aggressive executed orders relative to the total number of clients' daily aggressive and passive executed orders $PMKT_t^A$.

We do not reject the null hypothesis 5. We find that the fraction of client trades executed via market orders and marketable limit orders increased as our measure $PMKT^A$ increased by 0.463% on average per year on the full sample where the standard error of this change is 1.930% and the resulting t-statistic is 0.240. Using the trend regression, we however reject the null hypothesis and find that the fraction of client trades executed via market orders and marketable limit orders has significantly increased over years.

Hypothesis 6 *Relative gross trading revenue for clients has not changed over time.*

We calculate clients' daily relative gross trading revenue as follows:

$$RGTR_t^A = \frac{\sum_{j=1}^{N_t} q_j \times d_j \times p_j + \left(\sum_{j=1}^{N_t} q_j \times d_j\right) RP_t}{\sum_{j=1}^{N_t} q_j \times p_j}, \tag{2.3}$$

where N_t is the number of clients' trades on day t, p_i, q_i and d_i are resp. the trade price, trade size and direction of the client's executed order i. We assume that investors have a zero position at the start of the day; the end-of-day position, captured by $\left(\sum_{j=1}^{N_t} q_j \times d_j\right)$, is evaluated at the daily settlement price. We follow Eurex rules and define the settlement price as the price-weighted average of the transactions occurring between 17:29 and 17:30 if there are more than five trades, and if not as the price-weighted average of the last five transactions occurring after 17:15.[12] We divide by the investor's total (euro) volume for that trading day to compute the relative measure.

We do not reject the null hypothesis 6. We find that the relative gross trading revenue for clients increased as our measure $RGTR$ increased by 33.133% on average per year on the full sample where the standard error of this change is 33.878% and the resulting t-statistic is 0.978.

Finally, as a robustness check, we test hypothesis 1, 2, 4 and 5 on the period 2009-2018 using the Deutsche Börse aggressor's flag. Table 2.1.C shows that our measure of realized spreads overestimates actual realized spreads, but yearly variations of the two measures are highly correlated in the range [65%; 97%] over the period and the conclusions of the tests are consistent. In addition, we added two measures: relative realized spread ($\%RSpd_{5min}$ and $\%RSpd^{A}_{min}$) and shorter horizon of price efficiency: Distance(10sec, 60sec) as suggested by the referees.

[12]See "Clearing Conditions for Eurex Clearing AG", 2006

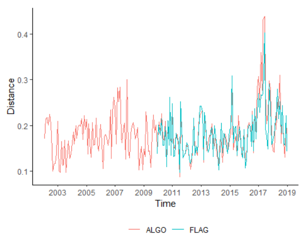
Table 2.1: Test of hypothesis 1 to 6 trimming 0.01 and 0.99

Table 1 reports the results of the test of hypothesis 1 to 6 and those of two series of robustness checks, based on alternative measures (Part B) ($\%RSpd_{5min}$: relative to the midquote), or comparing results using data flagged in the dataset those of our algorithm (Part C) over the period 2009-2018. Stars indicate significance at 10%, 5% and 1% respectively.

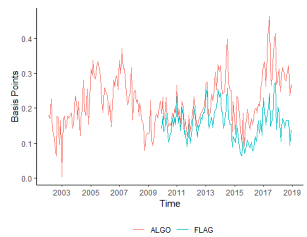
		Yearly data					Monthly data	
Hyp.	Measure	$\bar{X}_y$	$\Delta\bar{X}_y = 100.\frac{(\bar{X}_y - \bar{X}_{y-1})}{\bar{X}_{y-1}}$		Test $\Delta\bar{X} = 0$		$\bar{X}_y = a_0 + a_1 \times Time_y + FE_m + \varepsilon_y$	
		Mean	Mean	Std	t-stat	s.e.	a_0	a_1
A. Test of hypothesis 1 to 6								
H1	$Distance(1min, 5min)$	0.182	3.136	21.244	0.591	5.311	0.141***	0.003
H2	$RSpd_{5min}$	0.217	8.062	26.938	1.197	6.734	0.179***	0.004
H3	CSV	0.341	-3.299	7.737	-1.705	1.934	0.460***	-0.013***
H4	$RSpd^A_{5min}$	0.366	7.094	20.432	1.389	5.108	0.260***	0.012***
H5	$PMKT^A$	0.373	0.463	7.720	0.240	1.930	0.336***	0.004**
H6	$RGTR^A$	0.000	33.133	135.514	0.978	33.878	-0.001***	0.000
B. Robustness check: alternative measures								
H1	$Distance(10sec, 60sec)$	0.214	-1.002	19.182	-0.209	4.796	0.237***	-0.004
H1	$Distance(5min, 15min)$	0.300	-0.075	10.171	-0.029	2.544	0.325***	-0.002
H2	$RSpd_{1min}$	0.205	9.952	32.413	1.228	8.103	0.171***	0.004
H4	$RSpd^A_{1min}$	0.351	7.688	21.269	1.446	5.317	0.238***	0.013***
H2	$\%RSpd_{5min}$	0.000	7.829	29.166	1.074	7.292	0.000***	0.000
H4	$\%RSpd^A_{min}$	0.000	7.366	25.818	1.141	6.454	0.000***	0.000***
C. Robustness check: comparison between our algorithm and the flag, Subsample 2009-2018								
H1	$Distance$ (flag)	0.183	2.425	13.123	0.554	4.374	0.099***	0.006*
H1	$Distance$ (alg.)	0.186	4.246	20.872	0.610	6.957	0.070	0.008*
H2	$RSpd_{5min}$ (flag)	0.155	2.523	29.288	0.258	9.763	0.173***	-0.001
H2	$RSpd_{5min}$(alg.)	0.226	6.914	30.441	0.681	10.147	0.040	0.014**
H4	$RSpd^A_{5min}$(flag)	0.326	2.976	21.107	0.423	7.036	0.263***	0.005
H4	$RSpd^A_{5min}$ (alg.)	0.409	4.684	22.155	0.634	7.385	0.143***	0.020**
H5	$PMKT^A$ (flag)	0.361	-1.626	-0.812	2.002	1.978	0.413***	-0.004*
H5	$PMKT^A$ (alg.)	0.389	1.076	8.546	0.378	2.849	0.288***	0.007***

Figure 2.1: Monthly time series of our six measures

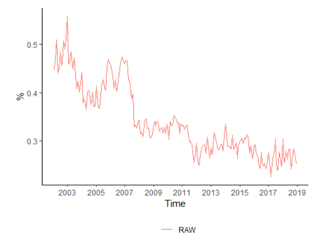
Figure 1 plots monthly measures of all 6 variables defined in the text and used to test Hypothesis 1 to 6 in the time trend regressions. In sub-figures (a), (b), (d) and (e) we compare the measure based on the algorithm used to compute the midquote and sign the trades and the measure using the flag after 2012.



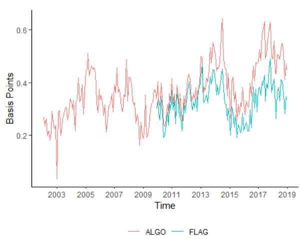
(a) H1: $Distance(1,5)$



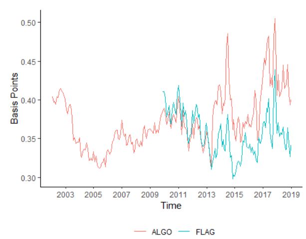
(b) H2: $RSpd$



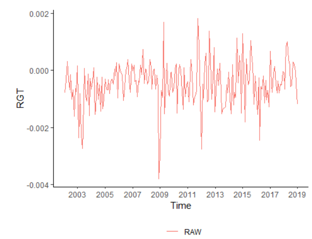
(c) H3: CSV



(d) H4: $RSpd^A$



(e) H5: $PMKT^A$



(f) H6: $RGTR^A$

References

Baldauf, M., Mollner, J., 2021. Trading in Fragmented Markets. Journal of Financial and Quantitative Analysis .

Foucault, T., Menkveld, A.J., 2008. Competition for Order Flow and Smart Order Routing Systems. The Journal of Finance .

Hendershott, T., Jones, C., Menkveld, A., 2011. Does algorithmic trading improve liquidity? Journal of Finance , 1–33.

Holden, C.W., Jacobsen, S., 2014. Liquidity measurement problems in fast, cometitive markets: Expensive and cheap solutions. Journal of Finance , 1747–1785.

Lee, C.M., Ready, M.J., 1991. Inferring trade direction from intraday data. Journal of Finance , 733–746.

Lo, A.W., MacKinglay, A.C., 1988. Stock market prices do not follow random walks: Evidence from a simple specification test. Review of Financial Studies , 41–66.

2.3 Non-Standard Errors

2.3.1 Introduction

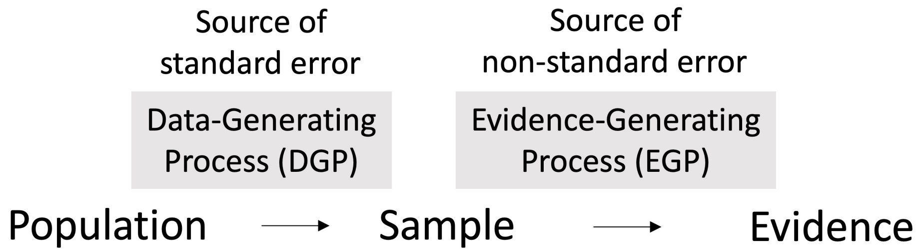
In their recent book, Kahneman, Sibony, and Sunstein (2021) (KSS) discuss variability in human judgment in terms of noise. They illustrate their analysis by judges passing sentence. They decompose total variation in sentencing into two canonical components: *level noise* and *pattern noise* (Ch. 6). Level noise captures the extent to which some judges are more lenient than others. Pattern noise, on the other hand, refers to variation in judgment when the same judge sentences similar cases. In statistical terms, this distinction can be defined as across-judge versus within-judge variation. Variation across judges is also referred as variation in judge *fixed effects*.

There are similarities to empirical science, where researchers analyze samples to test hypotheses. There is within-researcher variation due to sampling error. Re-sampling (or bootstrapping) yields different values of the estimator. The standard deviation (SD) of this distribution is referred to as *standard error* (SE) (Yule, 1897). It is a source of uncertainty that researchers are well aware of when conducting their tests.

Researchers are less aware that there is an additional level of uncertainty due to there not being a *standard* analysis path. Researchers vary in what they deem to be the most reasonable path in the "garden of forking paths" (Gelman and Loken, 2014). Conditional on the path, there is a well-defined estimator and standard error. Conditional on *the sample*, however, estimates may vary across researchers as they might pick different paths.[13] We refer to this additional variation as *non-standard error* (NSE). Note that the adjective, *non-standard*, emphasizes the lack of a standard approach. In other words, if all researchers agree on one path being the most reasonable one, then NSE is zero.

The schema below summarizes the overarching idea of non-standard errors. Statisticians use the term data-generating process (DGP) to convey the idea that samples are random draws from a population. Estimators, therefore, exhibit standard error.

[13] An important source of such variation is that researchers need to translate *conceptual research questions* to *empirical research questions* (Breznau et al., 2022).



Using the same language, one could say that scientists collectively engage in an evidence-generating process (EGP). Researchers potentially pick different analysis paths, which is a source of additional error: Non-standard error. Note that error in this case is to be understood as erratic as opposed to erroneous, in the sense that there simply is no right path in an absolute sense.[14]

Let us illustrate the idea with an example. In microstructure, market efficiency is conceptually defined as the extent to which a price process resembles a random walk. Suppose that one is interested in estimating the trend in market efficiency. To estimate, say, the mean annual change in market efficiency, a researcher faces many forks in the road: How to measure market efficiency, at what frequency to sample the data, how to define outliers, etc. Collectively, we refer to these decisions as the analysis path.

Our objective is to measure and analyze non-standard errors. The four questions that we focus on are:

1. How large are non-standard errors in finance?
2. Can they be "explained" in the cross-section of researchers? Are they smaller
   (a) for papers by higher quality teams?
   (b) for papers with better reproducible results?
   (c) for papers that score higher in peer evaluations?
3. Does peer feedback reduce non-standard errors?

[14]Variation in estimates reported in meta studies is of both types. The polar cases are the following. Estimates vary because researchers did the analysis in the exact same way, but on different samples (SE). Or, estimates vary because the sample is the same, but the analysis differs (NSE). Mavroeidis, Plagborg-Møller, and Stock (2014) is a special case, because they conduct their meta study by applying all observed analysis paths on all samples. They, unlike us, do not focus on distinguishing the two sources of variation explicitly. For a review of meta studies in finance, see Geyer-Klingeberg, Hang, and Rathgeber (2020).

4. Are researchers accurately aware of the size of non-standard errors?

The motivation for these questions is that non-standard errors are undesirable in the sense that they add uncertainty. Such uncertainty becomes particularly worrisome when some estimates are positive, while others are negative. It is reminiscent of the negative result known as the Sonnenschein-Mandel-Debreu "anything goes" theorem (Mas-Colell, Whinston, and Green, 1995, Ch. 17-E). We therefore want to learn if higher quality coincides with tighter NSEs, and if feedback reduces NSEs.

Finding answers to the four questions is extremely costly in terms of human resources. The core structure of an ideal experiment involves two sizable sets of representative researchers. A first set of researchers independently tests the same hypotheses on the same data, and writes a short paper presenting the results. A second, non-overlapping set of researchers obtains these papers, evaluates them, and provides feedback in a single-blind process.

We have run such an experiment under the #fincap tag (FINance Crowd Analysis Project). 164 research teams (RTs) and 34 peer evaluators (PEs) participated, with each PE evaluating about ten papers. The Deutsche Börse kindly made proprietary data available spanning 17 years of trading in Europe's most actively traded instrument: the EuroStoxx 50 index futures. This data enabled researchers to test pre-defined RT-hypotheses[15] on several important market trends. This unique opportunity might explain why participation was exceptionally high (at least double that of similar experiments elsewhere, discussed later in the introduction).[16] A back-of-the-envelope calculation shows that total human resources for #fincap span almost a single academic career: (164×2 months + 34×2 days ≈ 27 years).

Statistical framework. We define non-standard error as the interquartile-range (IQR) in estimates across researchers. The reason for picking a *robust* dispersion measure instead of SD, is that this distribution could exhibit fat tails, and thus be prone to outliers. #fincap itself is a case in point as will become clear. The distribution of estimates across researchers tends to the

[15]We refer to these hypotheses as RT-hypotheses to distinguish them from the hypotheses that we test when analyzing the #fincap results. Our hypotheses are based on the four overarching questions (Section 2).

[16]#fincap was presented to all involved by means of a dedicated website (https://fincap.academy) and a short video (https://youtu.be/HPtnus0Yu-o).

distribution of researcher fixed effects (RFEs), which could be any distribution. Using a robust dispersion measure, therefore, is a prudent choice.[17]

Statistical inference in #fincap needs to account for multiple hypothesis testing (MHT) (Bonferroni, 1936; Šidák, 1967). The critical values for individual tests need to account for multiple teams testing the same hypothesis. Put simply, if individual tests are performed at a five percent level, then the probability of at least one turning significant for multiple tests, (weakly) exceeds five percent. Harvey, Liu, and Zhu (2016) illustrate how to adjust levels in asset pricing tests. In his presidential address, Harvey (2017) emphasizes that MHT affects all of finance. We follow in his footsteps when applying MHT in #fincap.

Finally, to address the overarching questions, we need to analyze how NSEs co-vary with quality measures, and how they change across stages. Since NSE is defined in terms of quantiles, we will use quantile regression to conduct this analysis (Koenker and Bassett Jr., 1978). Note that ordinary least-squares only models conditional means, and it is therefore unfit for an analysis of dispersion. In addition to the first and the third quartile, we will also model the median, the first decile, and the ninth decile, in order to obtain a more complete view of the distribution, including results on the inter-decile range (IDR).

Summary of our findings. We first show that the group of #fincap participants is representative of the academic community in empirical finance/liquidity. About a third of the 164 research teams have at least one member with publications in the top-three finance, or the top-five economics journals.[18] For the group of peer evaluators, this share is 85%. 52% of RTs consist of at least one associate or full professor. For PEs, this is 88%. On a scale from 1 (low) to 10, the average self-ranked score on experience with empirical finance is 8.1 for RTs, and 8.4 for PEs. For experience with market liquidity, it is 6.9 for RTs, and 7.8 for PEs.

The evidence on the four overarching questions is as follows. First, the dispersion in

[17]The intuition is as follows. If the number of researchers tends to infinity, then the distribution of estimates tends to the distribution of RFEs, plus sampling errors. If, in addition, the sample size tends to infinity, then the distribution of estimates tends to the distribution of RFEs (because, for each analysis path, the group mean for this path tends to the RFE associated with this path). This distribution can be any distribution and might, therefore, exhibit fat tails. Section 2 provides a statistical framework.

[18]Finance: *Journal of Finance*, *Journal of Financial Economics*, and *Review of Financial Studies*. Economics: *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*.

estimates across RTs is sizable. All six RT-hypotheses had to be tested by proposing a measure and computing the average per-year percentage change. The first RT-hypothesis, for example, was "Market efficiency has not changed over time." The median estimate across RTs is -1.1% with a non-standard error (IQR) of 6.7%. The IDR is 27.3%.[19] The dispersion for the other RT-hypotheses is similar in magnitude, albeit smaller for RT-hypotheses that arguably involve fewer decisions on the analysis path (e.g., testing for a trend in market share).

Statistical tests show that, for all RT-hypotheses, at least a few estimates are significant (at a family level of 0.5%).[20] This number ranges from 6 (out of 164) for RT-H6 to 125 for RT-H3. We further test the null hypothesis of no dispersion in researcher fixed effects. We reject it for all RT-hypotheses. NSEs are therefore statistically significant for all RT-hypotheses.

Finally, it is worth noting that the uncertainty due to non-standard errors is similar in magnitude to that due to standard errors. For RT-H1, for example, the median standard error across RTs is 2.5%. For a Gaussian distribution, this implies an IQR of $1.35 \times 2.5\% = 3.4\%$, which compares to an NSE of 6.7%.

Second, the quantile regressions show that higher quality tends to coincide with smaller NSEs. A one SD increase in reproducibility significantly reduces NSEs by 25.0% and a one SD increase in peer-evaluator rating significantly reduces them by 33.3%. A one SD increase in team quality, however, significantly raises NSEs by 2.8%. This effect, however, is small in economic magnitude. If IDR were used instead of IQR, then a one SD increase significantly reduces IDRs for all quality measures: 13.3%, 17.9%, and 11.9%, respectively. Overall, higher quality seems to make extreme values less likely.

Third, peer feedback significantly reduces non-standard errors. The peer-feedback process involves multiple stages. We find that each stage reduces NSEs, albeit insignificantly. The reduction across *all* four stages is significant and amounts to 47.2%. This number for IDRs is also significant, and amounts to an even larger decline: 68.2%.

Fourth, RTs mostly underestimate the dispersion in estimates across RTs, which we tested in an incentivized belief survey. Such underestimation might well be the reason why non-standard

[19]This RT-hypothesis further illustrates the importance of robust statistics. One RT reports an estimate of +74,491%. This extreme outlier causes the mean and standard deviation to be 446.3% and 5,817.5%, respectively.

[20]We use the conservative significance levels advocated by Benjamin et al. (2018): 0.5% for significance and 5% for weak significance. They refer to the latter as "suggestive evidence."

errors never attracted much attention, until recently.

Finally, we dig deeper to discover what drives dispersion in estimates. A particularly useful tool for such analysis is a multiverse analysis (Liu et al., 2021). For key forks on the analysis path, the multiverse reveals how sensitive the distribution of estimates is to decisions at each particular fork.

It turns out that many of the key forks in #fincap add substantial noise. For RT-H1 on market efficiency, for example, it matters which frequencies teams choose for their variance ratio calculations. Some teams compare seconds to minutes, others days to months. A comparison of higher frequencies tends to find a decline in market efficiency, whereas for lower frequencies some find an increase in market efficiency.

The multiverse further reveals that Jensen's inequality can cause large dispersion. If a researcher is interested in assessing an N-period (long-term) trend in X_t, and estimates it based on one-period observations, then this could add substantial noise (Blume, 1974). Consider, for example, the expectation of a product of two independent and identically distributed relatives, where a relative is defined as X_t/X_{t-1}. Jensen's inequality implies that the expectation of this product is larger than the product of the expected relatives. The multiverse shows that the noise this adds can become particularly large for teams who sample at a daily frequency to estimate an annual trend, and use relatives instead of, for example, log-differences or a trend-stationary approach.

Contribution to the literature. The issue of variability in the research process is not new. Leamer (1983), for example, was troubled by the "fumes which leak from our computing centers." He called for studying "fragility in a much more systematic way."

Replication studies echo his concern as they typically find much weaker effects and less statistical strength (Ioannidis, 2005, Open Science Collaboration, 2015, Camerer et al., 2016, 2018). This is potentially the result of p-hacking: the process by which researchers try analysis paths until non-significant results turn significant.[21] We caution, however, that poor replication could also be demand-driven instead of supply-driven. This is the case when journals prefer to

[21] The p-value is the probability of observing an effect that is at least as large as the estimated effect, under the null hypothesis that there is no effect.

publish papers with low p-values. Munafò et al. (2017) survey the various threats to credible empirical science and propose several fixes.

The literature on replicability in finance is young, but growing rapidly. Examples are: McLean and Pontiff (2016), Hou, Xue, and Zhang (2018), Linnainmaa and Roberts (2018), Chordia, Goyal, and Saretto (2020), Harvey and Liu (2020), Ben-David, Franzoni, and Moussawi (2021), Black et al. (2021), Chen (2021), Mitton (2021), and Jensen, Kelly, and Pedersen (2022).

None of these replication studies focus on explaining the dispersion of estimates in a cross-section of researchers, or study the impact of peer feedback. We are the first to run an experiment, where this can be done in a clean way. Our objective is to study dispersion in estimates, short of a potential bias due to p-hacking. By design, there is no need to p-hack for #fincap researchers, because anyone who completes all stages of the project had been guaranteed co-authorship. Similarly, peer evaluators are guaranteed co-authorship to ensure clean feedback.

We are the first in finance to run an experiment to study dispersion in estimates, but we are not the first in science. Silberzahn et al. (2018) pioneered the multi-analyst study by letting multiple teams test whether soccer referees are more likely to draw red cards for players with a darker skin color. Other examples are Botvinik-Nezer et al. (2020) for neuroscience, Huntington-Klein et al. (2021) for economics, and Breznau et al. (2021) and Schweinsberg et al. (2021) for sociology. We innovate relative to these studies by explaining dispersion in estimates with quality attributes, by adding peer feedback stages, and by soliciting beliefs on dispersion *ex-ante*. A further strength of our study is the large cross-section of research teams: N=164. It is more than twice the size of any of the other multi-analyst samples.

The remainder of the paper is organized as follows. Section 2 provides an in-depth discussion of the project design.[22] It further presents the hypotheses associated with the four overarching questions, and develops an appropriate statistical framework to test them. Section 3 presents our results. Section 2 concludes.

[22]The design of #fincap follows the guidelines for multi-analyst studies proposed by Aczel et al. (2021).

2.3.2 Project design and hypotheses

This section first presents the details of the #fincap experiment, then presents hypotheses based on the four overarching questions, and finishes by discussing an appropriate statistical framework.

2.3.2.1 Project design

Before starting the #fincap experiment, we had filed a pre-analysis plan (PAP) with the Open Science Foundation (https://osf.io/h82aj/). The original version of *Non-Standard Errors* contains the results of the analysis outlined in the PAP. This original version remains available as Tinbergen Institute Discussion Paper TI 2021-102/IV. Subsequent feedback from various presentations and from reviewers at the *Journal of Finance* have led to the results presented here. Relative to the PAP, we now use robust methods to cope with unanticipated extreme outliers, we account for multiple testing, and we add a multiverse analysis to add deeper insight. Appendix 2 reconciles the current results with those in the original version.

In a nutshell, the #fincap experiment is about multiple research teams independently testing the same hypotheses on the same sample. We refer to these hypotheses as RT-hypotheses and to this sample as RT-sample. This is to distinguish them from the hypotheses that *we* will test based on the results generated by RTs and PEs (Section 2).[23]

The RT-sample is a plain-vanilla trade sample for the EuroStoxx 50 index futures with, added to it, a principal-agent flag.[24] For each side to a trade (i.e., buy and sell), we therefore know whether the exchange members traded for their own account, or for a client. The sample runs from 2002 through 2018 and contains 720 million trade records. These index futures are among the world's most actively traded index derivatives. They give investors exposure to Europe, or, more precisely, to a basket of euro-area blue-chip equities. With the exception of over-the-counter

[23]RTs and PEs have been recruited mostly by alerting appropriate candidates through suitable channels (e.g., the https://microstructure.exchange/). To inform them about #fincap, we created an online repository: https://fincap.academy. The repository remains largely unaltered (except for, e.g., adding FAQs).

[24]Trade records contain the following fields: Datetime, expiration, buy-sell indicator, size, price, aggressor flag, principal-agent flag, and a full- or partial-execution flag. Note that each side to a trade becomes a record, where the aggressor is the side whose incoming, say, buy order is matched with a resting sell order of the other side. The record is labeled *principal* if the exchange member trades for his own account, and *agent* when he trades for a client. More details on the sample are in Figure OA.6 of the Online Appendix.

activity, all trading is done through an electronic limit-order book (see, e.g., Parlour and Seppi, 2008, for details on limit-order book markets).

The RT-hypotheses are all statements about annual trends in the following market characteristics (with the null being no change):

RT-H1 market efficiency

RT-H2 realized bid-ask spread,

RT-H3 share of client volume in total volume,

RT-H4 realized bid-ask spread on client orders,

RT-H5 share of market orders in all client orders, and

RT-H6 gross trading revenue of clients.

The RT-hypotheses are presented only briefly here to conserve space. The full presentation of RT-H1, for example, characterizes informationally efficient prices as a random walk. Appendix 2 motivates and discusses all RT-hypotheses in detail. For the purpose of our analysis, we like to highlight two points. First, the RT-hypotheses are picked to address first-order questions in the field of empirical-finance/liquidity. These questions were used to market #fincap and convince appropriate candidates to join the project. Second, we ask for trends expressed as percentage changes to make them invariant to choice of unit (e.g., are measures expressed in thousands, or not).

Note that there is, purposefully, considerable variation across RT-hypotheses in the level of abstraction. RT-H1, for example, is on the relatively abstract notion of market efficiency. RT-H3, on the other hand, is on the share of client volume in total volume. Such share should be relatively straightforward to calculate because, in the RT-sample, each buy and sell trade is flagged agent (client) or principal (proprietary).

RTs are asked to test these RT-hypotheses by estimating an average yearly change for a self-proposed measure.[25] They are further asked to report standard errors for these estimates. We compute the ratio of the two, which we refer to as the implied t-value, or t-value for short.

[25] RTs are asked to express their results in annualized terms. To some, it was not clear. We therefore notified

RTs write a short academic paper in which they present and discuss their findings. These papers are evaluated by PEs who were recruited outside the set of researchers who registered as RTs. RT papers were randomly and evenly assigned to PEs in such a way that each paper is evaluated twice, and each PE evaluates nine or ten papers. PEs score the papers by providing an overall rating and a rating per RT-hypothesis. They do so in a single-blind process: PEs see the names of RTs, but not vice versa.[26] The reason for single-blind instead of double-blind is to incentivize RTs to exercise maximum effort.

PEs are asked to motivate their scores in a feedback form where they are encouraged to add constructive feedback. RTs receive this feedback unabridged, and are allowed to update their results based on it. Importantly, the design of #fincap was common knowledge to all because it had been available on a dedicated website before registration opened (see footnote 16).

More specifically, #fincap consists of the following four stages:

Stage 1 (January 11 - March 23, 2021.) RTs receive the detailed instructions along with access to the RT-sample. They conduct their analysis and hand in their results (short paper plus code). We emphasized in our emails and on the project website that RTs should work in *absolute secrecy* so as to ensure independence across RTs.

Stage 2 (May 10 - May 28, 2021.) RTs receive feedback from two anonymous PEs and are allowed to update their analysis based on it. They are asked to report their findings in the same way they did in stage 1.

Stage 3 (May 31 - June 18, 2021.) RTs receive the five best papers based on the average raw PE score. The names of the authors of these five papers were removed before distributing the papers.[27] Similar to stage 2, all RTs are allowed to update their analysis and resubmit their results.

everyone of the following clarification that we added to the FAQ section on https://fincap.academy: "Research teams are asked to report annualized estimates (and the corresponding standard errors); research teams are not required, however, to consider only annualized data."

[26] In our analysis, we remove PE fixed effects by demeaning (see Section 2).

[27] If two papers were tied in terms of their average score, then, following the pre-analysis plan, we picked the one that had highest reproducibility score provided by Cascad. For more information on Cascad, see the statement of H2 in Section 2.

Stage 4 (June 20 - June 28, 2021.) RTs report their final results, this time not constrained by delivering code that produces them. In other words, RTs are allowed to Bayesian update their results (i.e., estimates and standard errors) taking in all the information that has become available to them, in particular the five best papers. They could, for example, echo the results of one of these papers, simply because of an econometric approach that they believe is superior but that is beyond their capacity to code. This stage was added to remove all constraints and see how far the RT community can get in terms of reaching consensus.

The stages subsequent to the first one mimic the feedback researchers get from various interactions with peer researchers in the research process *before* a first journal submission. Stage 2 mimics, for example, immediate feedback from colleagues over lunch, during seminars, or in coffee breaks at conferences. Stage 3 mimics indirect feedback by means of seeing competitive papers that gain a lot of visibility through endorsements by colleagues, or by being presented in seminars or at conferences. Stage 4 solicits a final estimate whereby researchers are allowed to attach weight to estimates of others whom, for example, they believe implement a superior methodology that they are unable to code themselves. We like to emphasize that all these stages are designed in a way to keep the full dynamics of a refereeing process at a scientific journal out of scope.[28]

2.3.2.2 Hypotheses

Before running the experiment, we translated the project's four overarching questions into a set of pre-registered hypotheses. These hypotheses all center on the dispersion in estimates across RTs. Our main measure is the interquartile range, which we refer to as non-standard error. All hypotheses are stated as null hypotheses and tests will be two-sided.

The first set of three hypotheses focuses on how NSEs relate to various quality measures:

[28] Studying such dynamics requires a different experiment that involves "publishing" papers, *including* the names of the authors. Note that we do reveal the best five papers (according to PEs) to all RTs in stage 4, but the authors of these papers remain hidden. Our focus is narrowly on the pure findings and beliefs of the RTs, avoiding any possible corruption by "the publication game."

H1 NSE of stage-1 estimates does not co-vary with team quality. Team quality is proxied by the largest common factor in various candidate proxies for team quality. We prefer an appropriately weighted average over simply adding all proxies to maximize statistical power in the regressions. More specifically, we define team quality as the first principal component of the following standardized series:[29]

(a) *Top publications*: The team has at least one top-three publication in finance or one top-five publication in economics (0/1) (see footnote 18).

(b) *Expertise in the field*: Average of self-assessed experience in market liquidity and empirical finance (scale from 0 to 10).

(c) *Experience with big data*: The team has worked with samples at least as large as the sample they analyze in #fincap (0/1).

(d) *Academic seniority*: At least one team member holds an associate or a full professorship (0/1).

(e) *Team size*: The team size attains its maximum of two members (0/1).

H2 NSE of stage-1 estimates does not co-vary with reproducibility score. This score measures the extent to which RT estimates are reproducible from RT code. The scoring was done by the Certification Agency for Scientific Code and Data (Cascad). Cascad is a non-profit certification agency created by academics with the support of the French National Science Foundation (CNRS) and a consortium of French research institutes. The objective of Cascad is to provide researchers with a way to credibly signal the reproducibility of their research (used by, for example, the *American Economic Review*).[30]

H3 NSE of stage-1 estimates does co-vary with the average PE rating (RT-hypothesis level). To remove a possible PE fixed effect, we use demeaned PE ratings in all of our analysis.

[29] An important advantage of a principal-component analysis (PCA) is that the weighting is data-driven, thus avoiding subjective weights. Note that even the five proxies that enter were picked ex-ante in the pre-analysis plan filed at OSF. The PCA results will be discussed in Section 2.

[30] Cascad rates reproducibility on a five-category scale: RRR (perfectly reproducible), RR (practically perfect), R (minor discrepancies), D (potentially serious discrepancies), and DD (serious discrepancies). For #fincap, Cascad converted their standard categorical rating to an equal-distance numeric one: RRR, RR, R, D, and DD become 100, 75, 50, 25, 0, respectively.

The next hypothesis is about convergence in estimates across the four stages.

H4 NSE does not change across all feedback stages.

The final hypothesis focuses on RT *beliefs* about the dispersion in estimates across RTs.

H5 The average belief of RTs on the dispersion in estimates across RTs, is correct. The dispersion predictions were solicited in terms of the SD measure.

2.3.2.3 Statistical framework

To formalize the analysis of non-standard errors in a statistical sense, consider a set of researchers indexed by $j \in \{1, \ldots, J\}$. All researchers are given the same sample of size K. Researchers are asked to estimate the mean of a particular object (e.g., the per-year change in market efficiency). All researchers independently decide on the optimal analysis path and estimate the mean accordingly. Collectively, let these estimates, $X_1, \ldots, X_J$, be distributed as:

$$X_j = e_j + \varepsilon_j, \tag{2.4}$$

where e_j is a researcher-specific mean, henceforth referred to as a researcher fixed effect (RFE), and ε_j is a sampling error. The Central Limit Theorem (CLT) implies that, for large K, ε_j is approximately normal with mean zero and variance $\sigma^2_{j,K} = \sigma^2_j / K$, where σ^2_j is the path-specific variance of residuals.

Note that sampling errors are likely to correlate across researchers so that, collectively, the estimates are approximately distributed as:

$$\underset{(J\times 1)}{X} = \underset{(J\times 1)}{e} + \underset{(J\times 1)}{\varepsilon}, \text{ where } \underset{(J\times 1)}{\varepsilon} \sim N\left(\underset{(J\times 1)}{0}, \underset{(J\times J)}{\Sigma}\right), \tag{2.5}$$

where Σ is a positive semidefinite matrix. The off-diagonal elements of Σ are expected to be mostly positive since, if for a particular sample draw, X_i is above its (unconditional) mean e_i, then X_j is, most likely, also above its mean e_i.[31]

[31] For example, consider the case of estimating the mean of a distribution. If two researchers estimate this mean by taking the sample average, but one winsorizes the sample and the other does not, then a particular sample draw with unusually high values likely yields above-mean estimates for both researchers.

Non-standard error. Non-standard error is defined as the inter-quartile range in estimates:

$$\text{NSE} := Q_{0.75}(x) - Q_{0.25}(x), \tag{2.6}$$

where x denotes a realization of the random vector X, and $Q_\alpha(x)$ is the αth quantile of x. Note that NSE tends to the IQR of RFEs when J and K both tend to infinity:

$$\text{NSE} \overset{J,K\to\infty}{\to} Q_{0.75}(e) - Q_{0.25}(e). \tag{2.7}$$

We reiterate that for the distribution of RFEs (i.e., the distribution of e) could be any distribution. It is, therefore, prudent to pick a robust dispersion measure, which is why we use IQR instead of SD. The latter tends to get dominated by the size of extreme outliers.[32]

Testing for non-standard error. We test for "significance of non-standard errors" by testing whether or not there is any dispersion in RFEs. We do this by testing the following set of null hypotheses:

$$H_0 : e_j = \nu, \quad \forall j \in \{1, \dots, J\}, \tag{2.8}$$

where ν is the median RFE. Since X_j is an estimator of e_j, these hypotheses can be tested by verifying, for each $j \in \{1, \dots, J\}$, whether X_j is statistically different from ν. In the implementation, we set ν equal to the median estimate. If any of these tests rejects the null, then dispersion is non-zero, and we consider non-standard errors to be statistically significant.[33]

Conceptually, the distribution of X could be obtained by bootstrapping. Such procedure, however, is infeasible because it requires that researchers redo their analysis for every new draw

[32]#fincap is a case in point. For RT-H4, one team reports an estimate of -6,275,383%, whereas the estimates of other teams range from -2,897% to 870%. The SD based on all estimates is 490,024%, but it is only 245% if one leaves out the outlier.

[33]Two more technical points merit discussion. First, we prefer the median over the mean to have a robust location parameter. The asymptotic variance of the mean is smaller than that of the median for Gaussian distributions, but typically not for distributions with fat tails. The reason is that the former depends on variance and thus on extreme outliers, whereas the latter does not: σ^2/N and $1/(4Nf(m))$, respectively, where N is the sample size, f is the density function, and m is the median. Figure OA.1 in the Online Appendix shows that, in #fincap, the variance of the median is an order of magnitude smaller than the variance of the mean. Second, the proposed test assumes that sampling error is negligible for the median estimate as an estimator for the median RFE, because randomness in the median estimate is ignored. Figure OA.1 illustrates that, indeed, the variance of the median estimate is negligible for #fincap.

of the sample. Instead, we use multiple hypothesis testing (MHT) results to develop a feasible testing procedure.

Before turning to MHT, let us pause for a moment and take stock of what is available to us. The #fincap sample consists of estimates x_j, along with their standard error s_j. This is useful, but misses information on the covariance among all possible pairs of estimates across researchers.

To account for multiple testing, we rely on well developed statistical theory. If one aims to test at a level of 5% for a family of N tests, then individual tests should be performed with a $(5/N)\%$ critical value, if the test statistics are mutually independent (Bonferroni, 1936; Šidák, 1967; Harvey, Liu, and Zhu, 2016).[34]

In summary, we propose an NSE test where the null hypothesis is that there is no dispersion in RFEs. We use a Bonferroni adjustment of significance levels to account for multiple testing. The test is conservative, because Bonferroni assumes independence. As pointed out in footnote 31, estimates are likely to correlate across researchers, in which case the *effective* number of tests is likely to be smaller than the actual number of tests. In the implementation, we add a trivial extension where correlations between estimates are calibrated based on our multiverse analysis (Section 2). We close the section by discussing an alternative test and pointing out a caveat.

Alternative test. Note that a natural alternative to the proposed test is to simply test if IQR is statistically different from zero. We did not pick this shortcut, because our focus is on whether there is any dispersion at all in estimates across researchers. Although we pick IQR to express dispersion in a single number, the deeper interest is whether the distribution in estimates is non-degenerate.

Caveat. We like to point out one potential caveat. The procedure to obtain a conservative test on RFEs implicitly assumes that SEs reported by researchers are consistent estimators of the true SEs. This might not be true if (some) researchers report non-robust SEs. Non-robust SEs tend to be smaller, because they ignore commonalities. If true, then NSE tests tend to turn significant

[34] If the N tests statistics are independent, then the probability of *at least* one significant result is $(1-(1-\alpha)^N)$. For example, for $\alpha = 0.05$ and $N = 10$, this probability is 40 percent. Šidák (1967) proposes to adjust the significance level for the individual tests to $\alpha' = 1-(1-\alpha)^{1/N}$. A Taylor expansion of α' around zero yields $\alpha' \approx \alpha/N$, which is known as the Bonferroni correction (Bonferroni, 1936).

more often. NSEs themselves, however, remain consistent estimators.[35]

2.3.3 Results

This section presents all our findings. They are based on a balanced sample of 164 research teams who completed all stages of the project (out of 168 research teams). The first subsection presents various summary statistics and tests whether non-standard errors are statistically significant. The second subsection tests our hypotheses. The third subsection digs deeper by means of a multiverse analysis. The fourth and final subsection discusses alternative explanations.

2.3.3.1 Summary statistics

(Insert Table 2.2 about here.)

Table 2.2 summarizes our stage-1 sample by means of three sets of statistics, organized in three panels.[36] Panel (a) summarizes the qualities of the #fincap community. It consists of 164 research teams and 34 peer evaluators. Maximum RT size is two members, which is the size of 79% of RTs.

The statistics testify to the high quality of the #fincap community. 31% of RTs have at least one top publication in finance or economics (see footnote 18 for the list of journals). For PEs, this is 85%. The percentage of RTs who have at least one member who is tenured at the associate or full professor level is 52% for RTs. For PEs, this is 88%. Feedback seems to come from more established scholars, which likely mirrors reality.

RT members and PEs cover the global academic-finance community reasonably well (see Figure OA.2 in the Online Appendix). RT members reside in 34 countries with most residing in the US (51 out of 293). PEs reside in 13 countries with, again, most residing in the US (13 out of

[35] Unfortunately, we do not have precise information on the SEs reported in #fincap, because not all RTs provide detailed information on how they calculate SEs.

[36] Table OA.1 through OA.3 in the Online Appendix repeat panel (c) of Table 2.2 for the other stages. Panel (a) is the same for all stages, and panel (b) is only available for stage-1 results, since only these results are evaluated by peers and scored by Cascad on reproducibility.

34). The strong skew towards the US is not surprising given that the more senior, well-published finance scholars are predominantly affiliated with US universities.

Most RTs and PEs seem to have the appropriate background for testing the RT-hypotheses on the RT-sample. Their average self-reported scores on having experience in the field of empirical finance is 8.1 for RTs and 8.4 for PEs on a scale from 0 (low) to 10. For experience with market liquidity, these average scores are 6.9 for RTs and 7.8 for PEs. There is considerable variation around these averages as the SDs range from 1.7 to 2.4. When it comes to working with samples as large as the RT-sample, 720 million trade records, most RTs and PEs seem up to it. 65% of RTs have worked with samples at least as large. For PEs, this percentage is 88%.

Panel (b) of Table 2.2 shows that the average quality of the RT analysis is solid, but the dispersion is large. The average reproducibility score is 64.5 on a scale from 0 (low) to 100 (see footnote 30). This is high when benchmarked against other studies on reproducibility (Colliard, Hurlin, and Pérignon, 2021). The accompanying SD is 43.7, which shows that there is large variation across RTs: Most code either reproduces close to perfectly or barely at all. The paper-quality scores provided by PEs show a similar pattern, albeit with less dispersion. The average score across RTs is 6.2 on a scale from 0 (low) to 10, with an SD of 2.0.

Panel (c) provides descriptive statistics on the distribution of results across RTs. It does so by RT-hypothesis, and by type of result: Estimate, standard error, and *t*-value. Since our focus is on dispersion in estimates across RTs, we relegate a discussion of RT medians to Appendix 2. More specifically, this appendix discusses the RT-hypotheses in-depth and summarizes what RTs, as a group, seem to find with a focus on the across-RT median instead of the across-RT IQR (i.e., the non-standard error).

(Insert Figure 2.2 about here.)

Perhaps the most salient feature of the extensive panel (c) is that there is substantial variation across RTs for all RT-hypotheses, and for all types of results. Panel (a) in Figure 2.2 illustrates this result for estimates. For RT-H1 on market efficiency, for example, the median estimate across RTs is -1.1% with an IQR of 6.7%. Even for RT-H3, which is a seemingly straightforward calculation of a market share, the dispersion is sizable: an IQR of 1.2% around a median of

-3.3%. The figure further illustrates that there are extreme outliers for all RT-hypotheses, which motivates our analysis in terms of robust statistics.

(Insert Table 2.3 about here.)

NSE test results. Is the dispersion in estimates statistically significant? Table 2.3 presents the non-standard error test results. The null of no dispersion in researcher fixed effects is rejected for all RT-hypotheses at a 0.5% (family) significance level. The conservative Bonferroni adjustment in panel (a) yields at least 11 estimates that are individually significantly different from the median (RT-H6), and at most 38 significant differences (RT-H3). There are significant estimates both above and below the median for all RT-hypotheses.

If, instead of assuming zero correlation across test statistics as in Bonferroni, one calibrates them based on bootstrapping from the multiverse analysis (Section 2), results change to the ones presented in panel (b). The implied "effective" number of tests is much lower than the 164 tests used in Bonferroni. It ranges from 21 (RT-H3) to 86 (RT-H6). The factor by which significance levels are adjusted is, therefore, up to almost seven times smaller than what Bonferroni suggests (i.e., 164/24=6.8). The result is that, indeed, more differences become significant. The increases are moderate, though, with at most two more differences becoming significant.

In sum, the statistics presented thus far show that there is substantial dispersion across research teams, in terms of their estimates, but also in team quality, in reproducibility score, and in peer-evaluator rating. In the next subsection, we use this dispersion to test the first three hypotheses. Is there more dispersion in estimates for lower quality teams, for results that are harder to reproduce, or for lower quality papers?

2.3.3.2 Hypotheses tests

The results on the three sets of hypotheses are discussed in the next three subsections. Standard errors in the quantile regressions account for correlation in residuals by adding RT-hypothesis fixed effects, and by clustering per RT across all stages.

2.3.3.2.1 Co-variates for stage-1 dispersion (H1-3)

The first set of hypotheses relates NSEs to various quality variables. One of these is team quality, which we measure by picking the first principal component (PC1) of five standardized quality proxies (see H1 in Section 2). PC1 explains 38.3% of total variance, and loads positively on all quality proxies. It loads strongest on publications and weakest on big-data but, importantly, it loads positively on all of them. Table OA.4 in the Online Appendix provides detailed results on the PCA.

(Insert Table 2.4 and Figure 2.3 about here.)

Table 2.4 summarizes the results of the stage-1 quantile regressions, with as dependent variables, the 10th, 25th, 50th, 75th, and 90th percentile of the distribution in estimates across RTs. Figure 2.3 illustrates the results by showing how a one SD increase in each co-variate affects IQR (i.e., NSE) and IDR. Taken together, these results allow us to test the first three hypotheses that relate quality variables to dispersion in estimates.

First, we find that higher team quality coincides with somewhat larger IQR, but with smaller IDR. The effect of team quality on the 25th percentile is not significant, but for the 75th percentile, it is significantly positive. The economic magnitude is small, though, as can be seen in Figure 2.3. A one SD increase in quality raises IQR by only $(0.032 - 0.004) \times 7.2 = 0.2$ percentage points (pps), where 7.2 is the average IQR across hypotheses (see panel (c) of Table 2.2). This increase of 0.2 pps implies a relative increase of 2.8%.[37] In contrast, a one SD increase in team quality, *reduces* IDR by 6.7 pps (-11.9%, since average IDR is 56.3). This is the result of a significant increase in the first decile and a significant reduction in the ninth decile. These findings suggest that higher quality teams are less likely to report extreme estimates.

If one replaces team quality by the five quality variables on which it is based, then a more nuanced picture emerges (Table OA.5 in the Online Appendix). The statistically significant and sizable relationships are the following. A one SD increase in academic seniority (i.e., an

[37] A direct test on IQR, instead of separate tests on the 25th and 75th percentiles, requires jointly modeling these percentiles. Such multivariate modeling, combined with clustering on errors, is a non-trivial econometric challenge. Univariate modeling with clustering, on the other hand, is relatively standard. We use a python package to run these regressions: pyqreg.

associate/full professor in the team), reduces IQR by 1.4 pps (-19.4%). A one SD increase in team size reduces it by 0.9 pps (-12.5%).

A one SD increase in top publications, however, *increases* IQR by 1.9 pps (+26.4%). These three variables are positively correlated which explains why we find that the (aggregate) team variable has a relatively small effect on IQR. For IDR, the effects are of the same sign, but larger in magnitude: -19.4, -7.0, and +6.1 pps , respectively (-34.4%, -12.4%, and +10.8%). Note that now the negative effects really dominate, which explains that IDR co-varies negatively with team quality. In sum, these findings suggest that well published scholars seem to disagree more, but such effect is offset by the presence of a senior scholar or a second team member.

Second, all percentiles co-vary significantly with reproducibility, except for the median. The 10th and the 25th percentile co-vary positively and the 75th and the 90th percentile co-vary negatively. The figure shows that these changes are sizable. A one SD increase in reproducibility reduces IQR by 1.8 pps (-25.0%) and IDR by 7.5 pps (-13.3%). In sum, better reproducibility lowers overall dispersion.

Third, the results for paper quality mirror those of reproducibility, albeit a bit stronger in magnitude. The 10th and 25th percentile co-vary significantly positively, the 75th and 90th percentile co-vary significantly negatively. A one SD increase in paper quality reduces IQR by 2.4 pps (-33.3%) and IDR by 13.6 pps (-17.9%). Higher rated papers exhibit less dispersion in estimates.

In summary, the evidence on the first three hypotheses is such that the null of no co-variation is rejected for all three. Generally, higher quality is associated with less dispersion in estimates.

2.3.3.2.2 Convergence across stages? (H4)

The analysis of first-stage results has shown that dispersion in estimates is sizable and statistically significant. Does peer feedback create convergence? In other words, does dispersion in estimates decline in the three subsequent stages where teams get feedback from peers. This is the focus of the fourth hypothesis.

(Insert Table 2.5 and Figure 2.4 about here.)

Table 2.5 presents the results of quantile regressions to explain the dispersion in estimates in all four stages (thus far, only stage 1 has been analyzed). To account for heterogeneity in dispersion across RT-hypotheses, the explanatory variables are stage dummies that are multiplied by stage-1 (estimate) IQR *per* RT-hypothesis. The coefficients, therefore, measure a stage effect, expressed in IQR units. Figure 2.4 presents the results graphically.

The evidence makes us reject the null hypothesis of no convergence across all stages. All changes across consecutive stages are positive for the 10th and 25th percentile, and negative for the 75th and 90th percentile. The majority, however, is insignificant. However, the *total* change across stages is significant for all these percentiles at the 5% level, and, for all but one at the 0.5% level. Taken together, these results show that there is significant convergence from the first to the last stage, but a decomposition across the various stages lacks significance.

Figure 2.4 illustrates that the convergence is sizable. Panel (a) shows that the total decline in IQR is 3.4 pps (-47.2%). The decline seems evenly distributed across the stages, although this decomposition is mostly insignificant. Panel (b) shows that the total decline in IDR is even larger: 38.4 pps (-68.2%). More than half of it seems to happen from the first to the second stage, where RTs receive anonymized feedback from two PEs. However, this result is only weakly significant, since only the increase in the first decile is weakly significant (i.e., at a 5% level, not at a 0.5% level).

2.3.3.2.3 Are RT-beliefs on dispersion in estimates accurate? (H5)

The fifth and final hypothesis focuses on whether RTs are accurately aware of the dispersion in estimates across teams. Beliefs have been solicited in an incentivized way. All teams were asked to predict SDs in estimates across teams.[38] We randomly selected 20% of all RTs and paid each of them $300 if one of their predictions (randomly drawn) was within 50% of the realized SD. Details on the reward scheme are in the instruction sheet they received before reporting their beliefs (Figure OA.15 in the Online Appendix). The hypothesis pertains to stage-1 estimates, because beliefs are solicited for this stage only.

[38] In retrospect, we should have (also) asked for an IQR prediction, because SD is very sensitive to extreme outliers (see footnote 32). To assess whether RTs might have overlooked such outliers, we will compare their SD predictions with realized SDs, both on the full sample and on a trimmed sample.

As H5 is stated in terms of the average belief being correct, testing it requires a test on the equality of means: the mean belief about SDs in estimates across teams, and the SDs of these estimates in the population. Let us define a test statistic D that measures the relative distance between beliefs and realizations:

$$D = \frac{1}{6n} \sum_{i,j} \left(\frac{BeliefOnSD_{ij} - RealizationOfSD_j}{RealizationOfSD_j} \right), \tag{2.9}$$

where $BeliefOnSD_{ij}$ is the belief of team i on the SD in estimates across teams for RT-hypothesis j and $RealizationOfSD_j$ is the realized SD for this RT-hypothesis in the raw sample.[39] The distribution of D under the null of equal means is obtained by bootstrapping. For details on the bootstrap procedure, we refer to Appendix 2.

(Insert Figure 2.5 about here.)

Figure 2.5 plots the distribution of beliefs on SDs, along with realized SDs depicted by red dots. It illustrates that the vast majority of teams underestimate dispersion in estimates. The interquartile range denoted by the boxes is consistently below the red dot, which implies that at least 75% of the teams underestimate the dispersion.

One might think that teams simply overlook the extreme values that make realized SDs explode. This, however, does not seem to be the case, because even if one trims the estimates by removing the top and bottom 2.5%, the IQR box stays below these "trimmed" realized SDs, depicted by orange dots. The only exception is RT-H3, for which the orange dot is just within the top of the box.

The formal test results are in Table OA.6 of the Online Appendix. Pooling across all RT-hypothesis, the test statistic shows that the predicted SD is 71.7% below the realized SD. This underestimation is significant at a 0.5% level. Similar results holds for all RT-hypotheses individually, except for RT-H3, for which the underestimation is insignificant. Its value was also lowest of all, only 9.0% underestimation. RT-H3 is an hypothesis on market shares that, arguably,

[39]The benefit of a relative measure as opposed to an absolute one is that (i) it is easy to interpret as it allows for statements of RTs over- or underestimating by some percentage and (ii) it accounts for level differences across hypotheses (e.g., under the null of accurate beliefs, a uniform distribution of beliefs on the support 0.09 to 0.11 will exhibit the same dispersion as a uniform distribution of beliefs on 900 to 1100).

is relatively straightforward to test. In summary, the vast majority of tests show significant underestimation and we therefore firmly reject the null that beliefs on the dispersion in estimates are accurate.

2.3.3.3 Digging deeper: A multiverse analysis

Non-standard errors in #fincap are significant and sizable. Why? Can we somehow identify which forks on the analysis paths cause most of the dispersion? More specifically, can we rank key forks on the path according to the degree of refraction they cause in the light the sample sheds on the research question at hand? We turn to a multiverse analysis to address these questions.

Steegen et al. (2016) coined the term multiverse analysis to emphasize that data *construction* involves multiple decisions. The sample that enters the analysis, therefore, is a function of the set of reasonable choices. The sample becomes a (p. 702) "many worlds or *multiverse* of data sets." A particular result of an analysis then becomes a distribution of results (because samples vary). We generalize this approach by adding decision forks for the part of the analysis that follows the sample construction (e.g., the choice of econometric model).

The strength of a multiverse analysis is that it reveals how sensitive an estimate is to a particular fork on the analysis path. It does so by studying how much the estimates refract when varying across all reasonable alternatives at the fork. For example, let there be N reasonable analysis paths. Now suppose there are $k \leq N$ reasonable alternatives at the jth fork. Then the N estimates associated with the N paths are sorted into k sets, depending on the alternative picked at the fork. The degree to which the results differ across the k sets determines how sensitive results are to the jth fork. We measure the degree to which k distributions differ by a k-sample Anderson-Darlin (AD) test. Appendix 2 discusses the AD test in detail, including why it fits our application particularly well. AD is a standard option in the Boba software that we use (Liu et al., 2021).[40]

(Insert Table 2.6 about here.)

To make the multiverse feasible, we identify key forks on the analysis path and, for each

[40] The Boba software is available at https://github.com/uwdata/boba.

fork, we ask RTs to select the alternative they picked among a set of pre-defined alternatives. This was done by means of a questionnaire that all filled out after the experiment. The choice of forks and the alternatives at each fork is informed by the short papers RTs wrote for #fincap. The discretization of the decision space enables us to project the large space of realized analysis paths, onto a manageable space of "representative" paths. Table 2.6 provides an overview of all forks for the six RT-hypotheses. It lists the alternatives at each fork, along with the fraction of RTs that picked them (depicted in Figure OA.5 of the Online Appendix).

For each fork, we also asked RTs to rate the fit between the alternative they picked from the set, and what they actually did in #fincap. Their average rating ranges between 4.0 for RT-H6 and 4.4 for RT-H3 on a scale from 1 "Far from what we did" to 5 "Very close to what we did" (see Figure OA.4 in the Online Appendix). We, therefore, believe that the multiverse analysis is representative of the #fincap analysis itself.

A multiverse analysis is powerful, but resource intensive. The table illustrates that the analysis becomes very large very quickly. For RT-H6, for example, the nine forks generate $2 \times 2 \times 3 \times 4 \times 3 \times 4 \times 2 \times 3 \times 2 = 6,912$ possible paths. Not all *possible* paths are equally reasonable, and the #fincap data help us select the most reasonable ones. The result is a weighted multiverse, where untraveled paths get zero weight. The other ones get weights proportional to the number of teams who picked the path. The vast majority of paths, however, was picked by only one team so the size of the multiverse is slightly less than 164 (the actual number varies across RT-hypotheses).

The analysis is done for the original sample as well as for 1000 bootstrapped samples. These additional samples are needed to estimate the correlations in test statistics across paths. These correlations are used to adjust significance levels when accounting for MHT. This is used in assessing whether NSEs are statistically significant, and whether individual estimates are statistically significant (see panels (b) in Table 2.3 and Table OA.7, where the latter is in the Online Appendix, respectively). Each RT-hypothesis, therefore, requires processing the 720 million trade records almost 164,000 times.[41]

[41]To keep the multiverse analysis feasible, we optimized the code by identifying commonalities across paths and use these to economize on loops. For example, for a particular day, realized spread calculations can iterate once over all trades to obtain realized spreads both for the path that retains all trading and the path that excludes the first and last 30 minutes of trading. Efficient coding further involves identifying opportunities for parallel processing.

(Insert Figure 2.6 about here.)

Results. Figure 2.6 illustrates that the multiverse is able to generate dispersion in estimates that is on par with the dispersion in reported estimates. The box plots for reported estimates are drawn in gray, overlaid by the multiverse box plots in color. The large dispersion in multiverse is remarkable, since they are based on a few decisions only.[42]

(Insert Figure 2.7 and Figure 2.8 about here.)

Figure 2.7 illustrates how sensitive the distribution of estimates is to variation across alternatives at the various forks. The plots reveal that two common strong refractors are the (econometric) model choice and the sampling frequency. A well-known force that drives a wedge between high- and low-frequency relatives is Jensen's inequality (Blume, 1974):

$$\Pi_{t=1}^{T} \underbrace{E\left(M_t\right)}_{\substack{\text{Expected}\\ \text{high}\\ \text{frequency}\\ \text{relative}}} < \underbrace{E\left(\Pi_{t=1}^{T} M_t\right)}_{\substack{\text{Expected low}\\ \text{frequency relative}}}, \tag{2.10}$$

if $M_t \in \mathbb{R}^+$ are identical independently distributed random variables, since $f(x) = x^T$ is a convex function. First-order Taylor expanding the left-hand side around one, and then subtracting one from both sides, yields:

$$T\left(E\left(M_t\right) - 1\right) \lesssim E\left(\Pi_{t=1}^{T} M_t\right) - 1. \tag{2.11}$$

If there are T high-frequency periods in a low-frequency period, then T times the average high-frequency return is expected to be lower than the average low-frequency return. Figure 2.8 illustrates the effect of this inequality. The three right-most bars illustrate how, for the relative-change model, the median annualized return is -23,000% for data sampled at the daily frequency,

The multiverse analysis has been implemented on Snellius, a national supercomputer available to Dutch scientists (128 cores and 200 GB internal memory). With all this help, the code took a few days instead of a few months to run for each RT-hypothesis.

[42]The multiverse models only a few forks and its estimates, therefore, are unlikely to accurately predict reported estimates. The explanatory power of regressions with reported #fincap estimates as dependent variables and multiverse estimates as explanatory variables is low. The larger point of the multiverse is to illustrate that, for a subset of forks, variation across paths can generate large non-standard errors. It further allows researchers to drill down and identify the forks that generate most of the dispersion in estimates.

-200% for the monthly frequency, and only -4.56% for the yearly frequency. The left-most six bars that correspond to the trend-stationary or log-difference model do not show such discrepancy across frequencies. The reason is that both these models are linear and, therefore, do not suffer from Jensen's inequality. The trend-stationary model features a linear trend and in a log-difference model, the log of a product of relatives becomes a sum of log relatives.

Figure 2.7 further highlights some idiosyncratic sensitivities. For RT-H1, for example, the second-most sensitive fork is the frequencies that are picked to assess the deviation from a random walk. Further analysis reveals that when comparing high frequencies, such as one-second returns to one-minute returns, then almost all analyses exhibit a decline in market efficiency. But, when comparing low frequencies, such as daily returns to monthly returns, then about half of the analyses show an increase in market efficiency whereas the other half show a decline.

Another example is the retain-negative-sign fork, which is the most sensitive one for RT-H6. The decision each team had to make is whether a negative number that becomes more negative yields a positive percentage change, or a negative percentage change. The first one emphasizes that a (negative) number becomes magnified, whereas the second one emphasizes a negative trend (i.e., "retain a negative sign"). 21% of the teams picked the first option, 79% picked the second one. It is not surprising that mapping an estimate from the positive to the negative domain causes strong refraction in estimates. This is an example of how a decision that each team might have thought was a trivial one (in sense that there is only one option) can generate non-standard error.

2.3.3.4 Alternative explanations

After having presented all our results, it is useful to discuss alternative explanations. Might the sizable non-standard errors be due to the presence of inexperienced researchers testing unsuitable hypotheses with little effort? We believe this is unlikely to be the case for the following reasons.

Experience. Aware of this potential pitfall, we selectively approached researchers (for RTs and PEs), whom we knew were sufficiently experienced in the field. When signing up, they ticked a box that they understood that participating in #fincap requires research expertise and experience in empirical finance/liquidity and the analysis of large datasets. Ticking the box further meant that they acknowledge that one of the team members held a PhD in finance or economics. After ticking the box, researchers had to motivate in an open text box why they believe they meet these requirements. We parsed the content of this box to make sure that the team qualifies before accepting them into #fincap (see Figure OA.7 in the Online Appendix for the sign-up sheet).

Hypotheses. We proceeded with care when designing RT-hypotheses. Early versions were shared with senior scholars, and their feedback helped us fine-tune RT-hypotheses. We, therefore, feel comfortable that the RT-hypotheses are suitable and well motivated hypotheses to test with the RT-sample (see Figure OA.11 in the Online Appendix for the RT instruction sheet, which shows how RT-hypotheses were presented to RTs).

Related to the suitability question, one might wonder whether vagueness of an RT-hypothesis might be a viable alternative explanation for sizable NSEs. To address this concern, we included a very precise RT-hypothesis: RT-H3 on client volume share. The results for RT-H3 show that NSEs can be sizable, even for relatively precise hypotheses. It is true, however, that NSEs tend to be lower for the more precise RT-hypotheses.

Effort. We incentivized research teams to exert effort by providing them with the following information (before they sign up): the deadlines of the various stages so that they could plan for it; their *non-anonymized* paper would be evaluated by senior peer reviewers; the top-five (anonymized) papers would be announced to all others;[43] and, only those who complete all stages become co-authors. In addition to these incentives, we believe that most scientists are propelled by an intrinsic motivation to do good research.

Looking back, we have various reasons to believe that researchers did indeed exert serious effort. First, only four out of 168 research teams failed to complete all stages. 123 out of 168 teams (73.2%) handed in their stage-1 report at least a day early, and none of the teams seriously breached any deadline. The average reproducibility score was 64.5 on a scale from 0 (low) to 100, which is high in comparison to what has been reported in other reproducibility studies

[43] Individuals obtain “ego utility” from positive views about their ability to do well and they exert more effort (or take more risks) when they are informed about their rank in non-incentivized competitions (Köszegi, 2006; Tran and Zeckhauser, 2012; Kirchler, Lindner, and Weitzel, 2018).

(Colliard, Hurlin, and Pérignon, 2021). Finally, the average paper quality was 6.2 on a scale from zero (low) to 10. As for peer evaluators, we also believe they exerted serious effort, because all who signed up as a PE completed their reviews on time.

2.3.4 Conclusion

Researchers need to take many decisions when testing hypotheses on a particular sample: pick an appropriate measure, treat outliers, select a statistical model, etc. If researchers are not perfectly aligned on these decisions, their estimates likely differ. This potential dispersion in estimates therefore adds uncertainty to an estimate reported by a *single* team. Other teams might have reported other estimates based on the same data.

We measure dispersion in estimates across researchers robustly with an inter-quartile range, and refer to it as non-standard error. We study NSEs in an experiment where 164 teams test the same six RT-hypotheses on the same sample. We find NSEs to be substantial, even for a relatively straightforward market-share hypothesis. For this RT-hypothesis, we find it to be 1.2% around a median of -3.3%. A more opaque RT-hypothesis on market-efficiency yields larger variation with an NSE of 6.7% around a median of 1.1%. We further find that NSEs are smaller for better reproducibility and higher quality papers as rated by peers.

A multiverse analysis based on key forks sheds light on how important each fork is in generating dispersion in estimates. It turns out that many forks add substantial dispersion in estimates. Two particularly powerful ones are sampling frequency and the statistical model. Using a non-linear model at high frequency to estimate a low frequency trend can add substantial noise (Jensen's inequality).

NSEs being substantial is worrisome. An encouraging result, however, is that peer feedback reduces NSEs by half. In the real-world, published papers likely have gone through more stages of feedback, which makes #fincap NSEs an upper bound for real-world dispersion in estimates. Published results might further be affected by *p*-hacking (scoped out in #fincap), which is a selective process and thus likely further reduces dispersion, and potentially introduces bias. Overall, we believe the full process towards published empirical research deserves further scrutiny.

Finally, our multiverse analysis provides guidance on what threshold to use in individual tests when accounting for multiple testing. Bonferroni assumes independence among test statistics and adjusts significance levels by the number of tests: 164 in the case of #fincap. Bootstrapped multiverse results show that there is substantial correlation among test statistics and finds adjustment factors that range between 13 and 91 (depending on RT-hypothesis). The threshold for two-sided testing at 5% therefore should be at least $\Phi(1 - 0.025/13) = 2.9$. This is in line with the 3.0 lower bound recommended by Harvey, Liu, and Zhu (2016) for factor tests in asset pricing.

Appendices

2.A Reconciliation with pre-analysis-plan results

The original version of *Non-Standard Errors* contains the results of the analysis outlined in the pre-analysis plan. This original version is available as Tinbergen Institute Discussion Paper TI 2021-102/IV. Most tables and figures have not changed.[44]

The only two tables that have changed are Table 3 and 4. The reason is that these are the only two regression tables. In the original version, we estimate a heteroskedasticity model with ordinary least-squares (OLS). The dependent variable is log squared error. However, OLS estimates are notoriously sensitive to extreme outliers, which turn out to be a feature of the #finap sample (see footnote 32 or Figure 2.2). Quantile regressions are robust to the presence of extreme outliers and are, therefore, more appropriate for the analysis of our sample. Moreover, they model the entire distribution instead of just a conditional mean (as emphasized in the introduction). In the remainder, we compare results across the two tables in the original version and the current version to reconcile previous findings with current ones.

Table 3 in the original version has become Table 2.4 in the current version. These tables

[44]More specifically, Table 1, 2, and 5, and Figure 1, 2, 3, 4, and 5 have not changed. In the current version, they appear as Table 2.2, OA.4, and OA.6, and Figure OA.2, 2.2a, 2.2b, OA.3, and 2.5, respectively, where the OA prefix indicates that they are in the Online Appendix.

both relate dispersion in estimates to quality variables in order to test the first hypothesis. In the original version, most results are insignificant. The only significance is for reproducibility when using a 2.5%-97.5% winsorized sample. The coefficient of -0.24 implies that a 10% increase in reproducibility coincides with a reduction in the standard deviation of estimates by $1/2 \times 0.24 \times 10\% = 1.2\%$ (the coefficient 1/2 converts variance to SD, see footnote 21 in original paper). In the current version, the first quartile (Q1) co-varies significantly *positively* with reproducibility and paper quality, whereas the third quartile co-varies significantly *negatively* with them. They, therefore, co-vary significantly negatively with IQR. A 10% increase in reproducibility coincides with a reduction in IQR by $10\% \times (0.109 + 0.142) \times 0.44 = 1.1\%$.[45] Note that this effect is in the same ballpark as the 1.2% in the original paper.

Table 4 in the original version has become Table 2.5 in the current version. In the original version, the unwinsorized sample shows a weakly significant decline in dispersion of estimates across all stages. The effect is also relatively small in magnitude since the SD decline is only 9%. With extreme outliers removed in the 2.5%-97.5% winsorized sample, the decline becomes both significant and larger in magnitude. The SD now declines by 53.5% across all stages. The results in the current version show that Q1 of the estimate distribution increases significantly across all stages and Q3 declines significantly. The result is a decline of 47.2% (depicted in Figure 2.4). Again, the numbers in both versions are in the same ballpark.

2.B RT-sample, RT-hypotheses, and results

This appendix presents the RT-hypotheses in detail and the test results of #fincap RTs as a group. The instruction sheet itself is available as Figure OA.11 in the Online Appendix. We start by providing the context that motivates the RT-hypotheses.

2.B.1 Context

Electronic order matching systems (automated exchanges) and electronic order generation systems (algorithms) have changed financial markets over time. Investors used to trade through

[45] The square root of the average variance of reproducibility (de-meaned by RT-hypothesis) is 0.44.

broker-dealers by paying dealer ask prices when buying, and accepting dealer bid prices when selling. The wedge between these bid and ask prices, the bid-ask spread, was a useful measure of trading cost, and often still is.

Now, investors more commonly trade in electronic limit-order markets (as is the case for EuroStoxx 50 futures). They still trade at bid and ask prices. They do so by submitting so-called market orders and marketable limit orders. However, investors can now also quote bid and ask prices themselves by submitting (non-marketable) standing limit orders. And, investors increasingly use agency algorithms to automate their trades. Concurrently, exchanges have been continuously upgrading their systems to better serve their clients. Has market quality improved, in particular when taking the viewpoint of non-exchange members: (end-user) clients?

2.B.2 RT-hypotheses and test results

The RT-hypotheses and results are discussed based on estimates in the final stage of the project (available as Table OA.3 in the Online Appendix). We therefore base our discussion on the results that RTs settle on after receiving all feedback. What do RTs find after having shown some convergence across the stages? And, consistent with the main text, we base our discussion on robust location and dispersion statistics: the median and IQR, respectively. Finally, we note that such discussion is meaningful, because Table OA.7 in the Online Appendix shows that, for all RT-hypotheses, the null of a zero trend is rejected at a 0.5% significance level. This significance level is used for all tests in the remainder of the subsection.

(The first two hypotheses focus on all trades.)

RT-H1. Assuming that informationally-efficient prices follow a random walk, did market efficiency change over time?
Null hypothesis: Market efficiency has not changed over time.

Findings. The median estimate is -1.1% with an IQR of 2.6%. The third quartile is -0.2% and the vast majority therefore finds a negative trend in efficiency. The Bonferroni tests show that 31 RTs find a significant negative trend against only four who find a significant positive trend. The decline seems modest as the across-RT median[46] is -1.1% per year. The small changes add up, though, to a total change in the 2002-2018 sample of approximately $(0.989^{17} - 1) = -17.1\%$. This might reflect a trend of declining depth in the market, possibly due to new regulation in the aftermath of the global financial crisis of 2007-2008. Post-crisis regulation constrains the supply of liquidity by sell-side banks (e.g., Bao, O'Hara, and Zhou, 2018; Jovanovic and Menkveld, 2021). If these banks incur higher inventory costs as a result, then, in equilibrium, one observes larger transitory price pressures thus reducing market efficiency (e.g., Pastor and Stambaugh, 2003; Hendershott and Menkveld, 2014). In the interest of brevity, we discuss all remaining hypotheses in the same way.

RT-H2. Did the (realized) bid-ask spread paid on market orders change over time? The realized spread could be thought of as the gross-profit component of the spread as earned by the limit-order submitter.

Null hypothesis: The realized spread on market orders has not changed over time.

Findings. The median estimate is -2.3% with an IQR of 4.3%. The third quartile is -0.1% and the vast majority therefore finds a negative trend in realized spread. The tests show that 38 RTs find a significant negative trend, whereas only three RTs find a significant positive trend. The median decline of 2.3% per year implies a 32.7% decline over the full sample. This trend might be due to the arrival of high-frequency market makers who operate at low costs. They do not have the deep pockets that sell-side banks have, but they will offer liquidity for regular small trades by posting near the inside of the market. Their arrival is typically associated with a tighter bid-ask spread, but not necessarily with better liquidity supply for large orders (e.g., Jones, 2013; Angel, Harris, and Spatt, 2015; Menkveld, 2016).

(The remaining hypotheses focus on agency trades only.)

RT-H3. Did the share of client volume in total volume change over time?

Null hypothesis: Client share volume as a fraction of total volume has not changed over time.

[46] The across-RT median includes all RTs, thus also those who report insignificant results.

Findings. The median estimate is -2.9% with an IQR of 1.7%. The ninth decile is -1.1%, which shows that almost all RTs report a negative trend. The tests show that 123 RTs find a significant negative trend against only two RTs documenting a significant positive trend. An median decline of 2.9% per year implies a total decline of 39.4% for the full sample. Intermediation, therefore, seems to have increased which should surprise those who believe that the arrival of agency algorithms enables investors to execute optimally themselves, thus reducing the need for intermediation.[47]

RT-H4. On their market orders and marketable limit orders, did the realized bid-ask spread that clients paid, change over time?
Null hypothesis: Client realized spreads have not changed over time.

Findings. The median estimate is -0.2% with an IQR of 2.4%. The third quartile, however, is positive suggesting that a modest majority finds a negative trend. The tests show a bit stronger evidence for a negative trend, because 15 RTs find it to be significantly negative against only eight who find a significant positive trend. The median decline of 0.2% per year translates to a 3.3% decline for the full sample. The decline in client realized spread is therefore only about a tenth of the total realized spread decline, which suggests that market orders of intermediaries benefited most from the general realized-spread decline.

RT-H5. Realized spread is a standard cost measure for market orders, but to what extent do investors continue to use market and marketable limit orders (as opposed to non-marketable limit orders)?
Null hypothesis: The fraction of client trades executed via market orders and marketable limit orders has not changed over time.

Findings. The median estimate is 0.0% with an IQR of 0.6%. 13 RTs find a significantly negative trend, whereas nine find a significantly positive trend. The results seem rather balanced

[47] We verified with Deutsche Börse that this change is not purely mechanical in the sense that, in the sample period, many institutions became an exchange member and, with it, the status of their volume changes from agency to principal.

between a negative and a positive trend. The results therefore seem to suggest that clients neither increased their share of market orders, nor did they decrease it. One might have expected the latter because an increased use of agency algorithms should allow them to execute more through non-marketable limit orders as opposed to market orders or marketable limit orders. The benefit of execution via a non-marketable limit order is that one earns half the bid-ask spread as opposed to paying it.

RT-H6. A measure that does not rely on the classic limit- or market-order distinction is gross trading revenue (GTR). Investor GTR for a particular trading day can be computed by assuming a zero position at the start of the day and evaluating an end-of-day position at an appropriate reference price. Relative investor GTR can then be defined as this GTR divided by the investor's total (euro) volume for that trading day. This relative GTR is, in a sense, a realized spread. It reveals what various groups of market participants pay in aggregate for (or earn on) their trading. It transcends market structure as it can be meaningfully computed for any type of trading in any type of market (be it trading through limit-orders only, through market-orders only, through a mix of both, or in a completely different market structure).
Null hypothesis: Relative gross trading revenue (GTR) for clients has not changed over time.

Findings. The median estimate is 0.0% with an IQR of 1.1%. Three RTs find a significantly positive trend and another three find a significantly negative one. The significance, therefore, is rather weak and balanced. We cautiously conclude that GTR has stayed mostly at the same level throughout the sample.

2.C Explanatory variables for error variance

2.C.1 Team quality

The quality measures for research teams are based on the survey that participants filled out upon registration (see Figure OA.7 in the Online Appendix). To keep the regression model both concise and meaningful, we reduce the ordinal variable "current position" and the logarithmic interval-based variable “size of largest dataset worked with” to binary variables. The academic position variable is one if a researcher is either associate or full professor. The dataset variable is one if the researcher has worked with datasets that are contained at least 100 million observations, because the #fincap sample contains 720 million observations. We aggregate these binary variables to research team level by taking the maximum across the team members.

As for self-assessed experience, we asked for both empirical finance and market liquidity, which we deem equally relevant for testing the RT-hypotheses. Thus, and because of the anticipated high correlation, we use the average of these two measures to obtain the individual score. And, in the interest of consistency, we again aggregate to the team level by taking the maximum across the team members.

2.C.2 Workflow quality

We proxy for workflow quality with an objectively obtained score of code quality provided by Cascad (see footnote 30). The scale ranges from 0 (serious discrepancies) to 100 (perfect reproducibility).

2.C.3 Paper quality

Papers are rated by an external group of peer evaluators. They rate the analyses associated with each RT-hypothesis individually, but also the paper in its entirety (see Figure OA.16 in the Online Appendix). The ratings range from from 0 (very weak) to 10 (excellent). Each paper is rated by two PEs and the paper rating is the average of the two (after removing a PE fixed effect as discussed in Section 2).

2.D Bootstrap procedure for belief statistic D

The distribution of D under the null of equal means is obtained by bootstrapping as follows. For each RT-hypothesis, we subtract the difference between the average belief on standard deviation and the observed standard deviation, from the beliefs:

$$
\begin{aligned}
&AdjBeliefOnSD_{ij} = \\
&BeliefOnSD_{ij} - \left[\left(\frac{1}{n}\sum_{i} BeliefOnSD_{ij}\right) - RealizationOfSD_{j}\right]
\end{aligned}
\tag{2.12}
$$

In this new sample with adjusted beliefs, the average belief about dispersion equals the observed dispersion, by construction. This sample is input to the bootstrapping procedure which iterates through the following steps 10,000 times:

1. As we have n RTs, in each iteration we draw n times from the new sample, with replacement. Each draw picks a particular RT and stores its beliefs and its results for all of the six RT-hypotheses. The result of these n draws therefore is a simulated sample that has the same size as the original sample.

2. The simulated sample is used to compute the test statistic D in (2.9). This statistic for iteration k, a scalar, is stored as D_k.

The bootstrap procedure yields 10,000 observations of the test statistic under the null. For a significance level of 0.005, the statistic observed in the #fincap sample is statistically significant if it lands below the 25th lowest simulated statistic or above the 25th highest simulated statistic. Its p-value is:[48]

$$
\begin{aligned}
&2\min(EmpiricalQuantileFincapStatistic, \\
&1 - EmpiricalQuantileFincapStatistic).
\end{aligned}
\tag{2.13}
$$

[48]Note that the procedure accounts for within-RT correlations (i.e., including possible non-zero correlations among a particular RT's results and the beliefs that it reports). The reason the procedure accounts for these correlations is that the bootstrap uses block-sampling where, when an RT is drawn, all of its beliefs and all of its estimates are drawn. One therefore only assumes independence across RTs which holds by construction given the design of #fincap.

2.E Anderson-Darlin test

The sensitivity of dispersion to a particular fork is measured by a k-sample Anderson-Darling test (Scholz and Stephens, 1987). This test was designed to verify whether k separate samples are drawn from the same distribution. The AD test statistic T_{k-1} measures the distance between the empirical distribution functions of k separate samples. It does not rely on parametric assumptions. It is, therefore, particularly attractive for our application as distributions are unknown ex-ante. In case of independence, the percentiles of the asymptotic distributions are known (Scholz and Stephens, 1987, Table 1 with $m = k - 1$). T_{k-1} converges to a standard normal for k tending to infinity.

The AD approach builds on tests previously proposed by Kolmogorov, Smirnov, Cramér, and von Mises. It adds a weight function to allow the researcher to attach differential importance to various portions of the distribution function (Anderson and Darling, 1964a). It nests the Cramér-von Mises ω^2 statistic which is based on equal weighting. The AD default weighting is one that equalizes the sampling error across the (empirical) support of the distribution function (Anderson and Darling, 1964b, p. 767). It effectively attaches more weight to the tails of the distribution. Scholz and Stephens (1987, p. 919) argue that among alternatives, the AD test statistic has attractive small sample (i.e., small k) properties.

References

Aczel, Balazs, Barnabas Szaszi, Gustav Nilsonne, Olmo R. van den Akker, Casper J. Albers, Marcel A.L.M. van Assen, Jojanneke A. Bastiaansen, Dan Benjamin, Udo Boehm, Rotem Botvinik-Nezer, Laura F. Bringmann, Niko A. Busch, Emmanuel Caruyer, Andrea M. Cataldo, Nelson Cowan, Andrew Delios, Noah N.N. van Dongen, Chris Donkin, Johnny B. van Doorn, Anna Dreber, Gilles Dutilh, Gary F. Egan, Morton Ann Gernsbacher, Rink Hoekstra, Sabine Hoffmann, Felix Holzmeister, Juergen Huber, Magnus Johannesson, Kai J. Jonas, Alexander T. Kindel, Michael Kirchler, Yoram K. Kunkels, D. Stephen Lindsay, Jean-Francois Mangin, Dora Matzke, Marcus R Munafò, Ben R. Newell, Brian A. Nosek, Russell A Poldrack, Don van Ravenzwaaij5, Jörg Rieskamp, Matthew J. Salganik, Alexandra Sarafoglou, Tom

Schonberg, Martin Schweinsberg, David Shanks, Raphael Silberzahn, Daniel J. Simons, Barbara A. Spellman, Samuel St-Jean, Jeffrey J. Starns, Eric L. Uhlmann, Jelte Wicherts, and Eric-Jan Wagenmakers. 2021. "Consensus-Based Guidance for Conducting and Reporting Multi-Analyst Studies." *eLife* 10:1–13.

Anderson, Theodore W. and Donald A. Darling. 1964a. "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes." *The Annals of Mathematical Statistics* 23:193–212.

———. 1964b. "A Test of Goodness of Fit." *Journal of the American Statistical Association* 49:765–769.

Angel, James J., Lawrence E. Harris, and Chester S. Spatt. 2015. "Equity Trading in the 21st Century: An Update." *Quarterly Journal of Finance* 5:1–39.

Bao, Jack, Maureen O'Hara, and Xing (Alex) Zhou. 2018. "The Volcker Rule and Corporate Bond Market Making in Times of Stress." *Journal of Financial Economics* 130:95–113.

Ben-David, Itzhak, Francesco Franzoni, and Byungwook Kim Rabih Moussawi. 2021. "Competition for Attention in the ETF Space." Manuscript, Ohio State University.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony G. Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P.A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don A. Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E. Johnson. 2018. "Redefine Statistical Significance." *Nature Human Behavior* 2:6–10.

Black, Bernard S., Hemang Desai, Kate Litvak, Woongsun Yoo, and Jeff Jiewei Yu. 2021. "Specification Choice in Randomized and Natural Experiments: Lessons from the Regulation SHO Experiment." Manuscript, Northwestern University.

Blume, Marshall E. 1974. "Unbiased Estimators of Long-Run Expected Rates of Return." *Journal of the American Statistical Association* 69:634–638.

Bonferroni, Carlo E. 1936. "Teoria Statistica delle Classi e Calcolo delle Probabilità." *Pubblicazioni del R Istuto Superiore di Scienze Economiche e Commerciali di Firenze* 8:1–62.

Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A. Mumford, R. Alison Adcock, Paolo Avesani, Blazej M. Baczkowski, Aahana Bajracharya, Leah Bakst, Sheryl Ball, Marco Barilari, Nadège Bault, Derek Beaton, Julia Beitner, Roland G. Benoit, Ruud M.W.J. Berkers, Jamil P. Bhanji, Bharat B. Biswal, Sebastian Bobadilla-Suarez, Tiago Bortolin, Katherine L. Bottenhorn, Alexander Bowring, Senne Braem, Hayley R. Brooks, Emily G. Brudner, Cristian B. Calderon, Julia A. Camilleri, Jaime J. Castrellon, Luca Cecchetti, Edna C. Cieslik, Zachary J. Cole, Olivier Collignon, Robert W. Cox, William A. Cunningham, Stefan Czoschke, Kamalaker Dadi, Charles P. Davis, Alberto De Lucas, Mauricio R. Delgado, Lysia Demetriou, Jeffrey B. Dennison, Xin Di, Erin W. Dickie, Ekaterina Dobryakova, Claire L. Donnat, Juergen Dukart, Niall W. Duncan, Joke Durnez, Amr Eed, Simon B. Eickhoff, Andrew Erhart, Laura Fontanesi, G. Matthew Fricke, Shiguang Fu, Adriana Galván, Remi Gau, Sarah Genon, Tristan Glatard, Enrico Glerean, Jelle J. Goeman, Sergej A.E. Golowin, Carlos Gonzáez-García, Krzysztof J. Gorgolewski, Cheryl L. Grady, Mikella A. Green, Joao F. Guassi Moreira, Olivia Guest, Shabnam Hakimi, J. Paul Hamilton, Roeland Hancock, Giacomo Handjaras, Bronson B. Harry, Colin Hawco, Peer Herholz, Gabrielle Herman, Stephan Heunis, Felix Hoffstaedter, Jeremy Hogeveen, Susan Holmes, Chuan-Peng Hu, Scott A. Huettel, Matthew E. Hughes, Vittorio Iacovella, Alexandru D. Iordan, Peder M. Isager, Ayse I. Isik, Andrew Jahn, Matthew R.

Johnson, Tom Johnstone, Michael J.E. Joseph, Anthony C. Juliano, Joseph W. Kable, Michalis Kassinopoulos, Cemal Koba, Xiang-Zhen Kong, Timothy R. Koscik, Nuri Erkut Kucukboyaci, Brice A. Kuhl, Sebastian Kupek, Angela R. Laird, Claus Lamm, Robert Langner, Nina Lauharatanahirun, Hongmi Lee, Sangil Lee, Alexander Leemans, Andrea Leo, Elise Lesage, Flora Li, Monica Y.C. Li, Phui Cheng Lim, Evan N. Lintz, Schuyler W. Liphardt, Annabel B. Losecaat Vermeer, Bradley C. Love, Michael L. Mack, Norberto Malpica, Theo Marins, Camille Maumet, Kelsey McDonald, Joseph T. McGuire, Helena Melero, Adriana S. Méndez Leal, Benjamin Meyer, Kristin N. Meyer, Glad Mihai, Georgios D. Mitsis, Jorge Moll, Dylan M. Nielson, Gustav Nilsonne, Michael P. Notter, Emanuele Olivetti, Adrian I. Onicas, Paolo Papale, Kaustubh R. Patil, Jonathan E. Peelle, Alexandre Pérez, Doris Pischedda, Jean-Baptiste Poline, Yanina Prystauka, Shruti Ray, Patricia A. Reuter-Lorenz, Richard C. Reynolds, Emiliano Ricciardi, Jenny R. Rieck, Anais M. Rodriguez-Thompson, Anthony Romyn, Taylor Salo, Gregory R. Samanez-Larkin, Emilio Sanz-Morales, Margaret L. Schlichting, Douglas H. Schultz, Qiang Shen, Margaret A. Sheridan, Jennifer A. Silvers, Kenny Skagerlund, Alec Smith, David V. Smith, Peter Sokol-Hessner, Simon R. Steinkamp, Sarah M. Tashjian, Bertrand Thirion, John N. Thorp, Gustav Tinghög, Loreen Tisdall, Steven H. Tompson, Claudio Toro-Serey, Juan Jesus Torre Tresols, Leonardo Tozzi, Vuong Truong, Luca Turella, Anna van â€~t Veer, Tom Verguts, Jean M. Vettel, Sagana Vijayarajah, Khoi Vo, Matthew B. Wall, Wouter D. Weeda, Susanne Weis, David J. White, David Wisniewski, Alba Xifra-Porxas, Emily A. Yearling, Sangsuk Yoon, Rui Yuan, Kenneth S. L. Yuen, Lei Zhang, Xu Zhang, Joshua E. Zosky, Thomas E. Nichols, Russell A. Poldrack, and Tom Schonberg. 2020. "Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams." *Nature* 582:84–88.

Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, Henrik K. Andersen, Daniel Auer, Flavio Azevedo, Oke Bahnsen, Dave Balzer, Gerrit Bauer, Paul C. Bauer, Markus Baumann, Sharon Baute, Verena Benoit, Julian Bernauer, Carl Berning, Anna Berthold, Felix S. Bethke, Thomas Biegert, Katharina Blinzler, Johannes N. Blumenberg, Licia Bobzien, Andrea Bohman, Thijs Bol, Amie Bostic, Zuzanna Brzozowska, Katharina Burgdorf, Kaspar Burger, Kathrin Busch, Juan Carlos-Castillo, Nathan

Chan, Pablo Christmann, Roxanne Connelly, Christian S. Czymara, Elena Damian, Alejandro Ecker, Achim Edelmann, Maureen A. Eger, Simon Ellerbrock, Anna Forke, Andrea Forster, Chris Gaasendam, Konstantin Gavras, Vernon Gayle, Theresa Gessler, Timo Gnambs, Amélie Godefroidt, Max Grömping, Martin Groß, Stefan Gruber, Tobias Gummer, Andreas Hadjar, Jan Paul Heisig, Sebastian Hellmeier, Stefanie Heyne, Magdalena Hirsch, Mikael Hjerm, Oshrat Hochman, Andreas Hövermann, Sophia Hunger, Christian Hunkler, Nora Huth, Zsófia S. Ignácz, Laura Jacobs, Jannes Jacobsen, Bastian Jaeger, Sebastian Jungkunz, Nils Jungmann, Mathias Kauff, Manuel Kleinert, Julia Klinger, Jan-Philipp Kolb, Marta Kołczyńska, John Kuk, Katharina Kunißen, Dafina Kurti Sinatra, Alexander Langenkamp, Philipp M. Lersch, Lea-Maria Löbel, Philipp Lutscher, Matthias Mader, Joan E. Madia, Natalia Malancu, Luis Maldonado, Helge-Johannes Marahrens, Nicole Martin, Paul Martinez, Jochen Mayerl, Oscar J. Mayorga, Patricia McManus, Kyle McWagner, Cecil Meeusen, Daniel Meierrieks, Jonathan Mellon, Friedolin Merhout, Samuel Merk, Daniel Meyer, Leticia Micheli, Jonathan Mijs, Cristóbal Moya, Marcel Neunhoeffer, Daniel Nüst, Olav Nygård, Fabian Ochsenfeld, Gunnar Otte, Anna Pechenkina, Christopher Prosser, Louis Raes, Kevin Ralston, Miguel Ramos, Arne Roets, Jonathan Rogers, Guido Ropers, Robin Samuel, Gregor Sand, Ariela Schachter, Merlin Schaeffer, David Schieferdecker, Elmar Schlueter, Regine Schmidt, Katja M. Schmidt, Alexander Schmidt-Catran, Claudia Schmiedeberg, Jürgen Schneider, Martijn Schoonvelde, Julia Schulte-Cloos, Sandy Schumann, Reinhard Schunck, Jürgen Schupp, Julian Seuring, Henning Silber, Willem Sleegers, Nico Sonntag, Alexander Staudt, Nadia Steiber, Nils Steiner, Sebastian Sternberg, Dieter Stiers, Dragana Stojmenovska, Nora Storz, Erich Striessnig, Anne-Kathrin Stroppe, Janna Teltemann, Andrey Tibajev, Brian Tung, Giacomo Vagni, Jasper Van Assche, Meta van der Linden, Jolanda van der Noll, Arno Van Hootegem, Stefan Vogtenhuber, Bogdan Voicu, Fieke Wagemans, Nadja Wehl, Hannah Werner, Brenton M. Wiernik, Fabian Winter, Christof Wolf, Yuki Yamada, Nan Zhang, Conrad Ziller, Stefan Zins, Tomasz Żółtak, and Hung H.V. Nguyen. 2021. "Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty." Manuscript, University of Bremen.

Breznau, Nate, Deadric T. Williams, Katrin Auspurg, Josef Bruederl, Felix Holzmeister, Gustav

Nilsonne, Balázs Aczel, Cory J. Clark, and Eric Uhlmann. 2022. "Is the Variability in Social Science Research Results Produced by Different Researchers Explicable? An Adversarial Collaboration and Joint Effort to Parse the Dispersion in Estimates." Manuscript, University of Bremen.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351:1433–1436.

Camerer, Colin F., Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, and Hang Wu. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science." *Nature Human Behaviour* 2:637–644.

Chen, Andrew Y. 2021. "The Limits of *p*-Hacking: Some Thought Experiments." *Journal of Finance* 76:2447–2480.

Chordia, Tarun, Amit Goyal, and Alessio Saretto. 2020. "Anomalies and False Rejections." *Review of Financial Studies* 33:2134–2179.

Colliard, Jean-Edouard, Christophe Hurlin, and Christophe Pérignon. 2021. "The Economics of Research Reproducibility." Manuscript, HEC Paris.

Gelman, Andrew and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102:460–465.

Geyer-Klingeberg, Jerome, Markus Hang, and Andreas Rathgeber. 2020. "Meta-analysis in finance research: Opportunities, challenges, and contemporary applications." *International Review of Financial Analysis* 7:1–15.

Harvey, Campbell R. 2017. "Presidential Address: The Scientific Outlook in Financial Economics." *Journal of Finance* 72:1399–1440.

Harvey, Campbell R. and Yan Liu. 2020. "False (and Missed) Discoveries in Financial Economics." *Journal of Finance* 75:2503–2553.

Harvey, Campbell R., Yan Liu, and Heqing Zhu. 2016. "…and the Cross-Section of Expected Returns." *Review of Financial Studies* 29:5–68.

Hendershott, Terrence and Albert J. Menkveld. 2014. "Price Pressures." *Journal of Financial Economics* 114:405–423.

Hou, Kewei, Chen Xue, and Lu Zhang. 2018. "Replicating Anomalies." *Review of Financial Studies* 33:2019–2133.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, , Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and Yani Stopnitzky. 2021. "The Influence of Hidden Researcher Decisions in Applied Microeconomics." *Economic Inquiry* 59:944–960.

Ioannidis, John P.A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2:696–701.

Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Pedersen. 2022. "Is There a Replication Crisis in Finance." *Journal of Finance (forthcoming)* .

Jones, Charles M. 2013. "What Do We Know About High-Frequency Trading?" Manuscript, Columbia University.

Jovanovic, Boyan and Albert J. Menkveld. 2021. "Equilibrium Bid-Price Dispersion." *Journal of Political Economy (forthcoming)* .

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment*. 1 London Bridge Street, London: William Collins.

Kirchler, Michael, Florian Lindner, and Utz Weitzel. 2018. "Rankings and Risk-Taking in the Finance Industry." *Journal of Finance* 73:2271–2302.

Koenker, Roger and Gilbert Bassett Jr. 1978. "Regression Quantiles." *Econometrica* 46:33–50.

Köszegi, Botond. 2006. "Ego Utility, Overconfidence, and Task Choice." *Journal of the European Economic Association* 4:673–707.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73:31–43.

Linnainmaa, Juhani T. and Michael R. Roberts. 2018. "The History of the Cross-Section of Stock Returns." *Review of Financial Studies* 31:2606–2649.

Liu, Yang, Alex Kale, Tim Althoff, and Jeffrey Heer. 2021. "Boba: Authoring and Visualizing Multiverse Analyses." *IEEE Transactions on Visualization and Computer Graphics* 27:1753–1763.

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*. New York: Oxford University Press.

Mavroeidis, Sophocles, Mikkel Plagborg-Møller, and James H. Stock. 2014. "Empirical Evidence on Inflation Expectations in the New Keynesian Phillips Curve." *Journal of Economic Literature* 52:124–188.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251).

McLean, R. David and Jeffrey Pontiff. 2016. "Does Academic Research Destroy Stock Return Predictability." *Journal of Finance* 71:5–31.

Menkveld, Albert J. 2016. "The Economics of High-Frequency Trading: Taking Stock." *Annual Review of Financial Economics* 8:1–24.

Mitton, Todd. 2021. "Methodological Variation in Empirical Corporate Finance." *Review of Financial Studies (forthcoming)* .

Munafò, Marcus R., Brian A. Nosek, Dorothy V.M. Bishop, Katerine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P.A. Ioannidis. 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1:1–9.

Parlour, Christine A. and Duane J. Seppi. 2008. "Limit Order Markets: A Survey." In *Handbook of Financial Intermediation and Banking*, edited by Arnoud W.A. Boot and Anjan V. Thakor. Amsterdam, Netherlands: Elsevier Publishing.

Pastor, L. and R.F. Stambaugh. 2003. "Liquidity Risk and Expected Returns." *Journal of Political Economy* 111:642–685.

Scholz, Fritz W. and Michael A. Stephens. 1987. "K-Sample Anderson-Darling Tests." *Journal of the American Statistical Association* 82:918–924.

Schweinsberg, Martin, Michael Feldman, Nicola Staub, Olmo R. van den Akker, Robbie C.M. van Aert, Marcel A.L.M. van Assen, Yang Liu, Tim Althoff, Jeffrey Heer, Alex Kale, Zainab Mohamed, Hashem Amireh, Vaishali Venkatesh Prasad, Abraham Bernstein, Emily Robinson, Kaisa Snellman, S. Amy Sommer, Sarah M.G. Otner, David Robinson, Nikhil Madan, Raphael Silberzahn, Pavel Goldstein, Warren Tierney, Toshio Murase, Benjamin Mandl, Domenico Viganola, Carolin Strobl, Catherine B.C. Schaumans, Stijn Kelchtermans, Chan Naseeb, S. Mason Garrison, Tal Yarkoni, C.S. Richard Chan, Prestone Adie, Paulius Alaburda, Casper Albers, Sara Alspaugh, Jeff Alstott, Andrew A. Nelson, Eduardo Ariño de la Rubia, Adbi Arzi, Štěpán Bahník, Jason Baik, Laura Winther Balling, Sachin Banker, David AA Baranger, Dale J. Barr, Brenda Barros-Rivera, Matt Bauer, Enuh Blaise, Lisa Boelen, Katerina Bohle Carbonell, Robert A. Briers, Oliver Burkhard, Miguel-Angel Canela, Laura Castrillo, Timothy Catlett, Olivia Chen, Michael Clark, Brent Cohn, Alex Coppock, Natàlia Cugueró-Escofet, Paul G. Curran, Wilson Cyrus-Lai, David Dai, Giulio Valentino Dalla Riva, Henrik Danielsson, Rosaria de F.S.M. Russo, Niko de Silva, Curdin Derungs, Frank Dondelinger, Carolina Duarte de Souza, B. Tyson Dube, Marina Dubova, Ben Mark Dunn, Peter Adriaan Edelsbrunner, Sara Finley, Nick Fox, Timo Gnambs, Yuanyuan Gong, Erin Grand, Brandon Greenawalt, Dan Han, Paul H.P. Hanel, Antony B. Hong, David Hood, Justin Hsueh, Lilian Huang, Kent N.

Hui, Keith A. Hultman, Azka Javaid, Lily Ji Jiang, Jonathan Jong, Jash Kamdar, David Kane, Gregor Kappler, Erikson Kaszubowski, Christopher M. Kavanagh, Madian Khabsa, Bennett Kleinberg, Jens Kouros, Heather Krause, Angelos-Miltiadis Krypotos, Dejan Lavbič, Rui Ling Lee, Timothy Leffel, Wei Yang Lim, Silvia Liverani, Bianca Loh, Dorte Lønsmann, Jia Wei Low, Alton Lu, Kyle MacDonald, Christopher R. Madan, Lasse Hjorth Madsen, Christina Maimone, Alexandra Mangold, Adrienne Marshall, Helena Ester Matskewich, Kimia Mavon, Katherine L. McLain, Amelia A. McNamara, Mhairi McNeill, Ulf Mertens, David Miller, Ben Moore, Andrew Moore, Eric Nantz, Ziauddin Nasrullah, Valentina Nejkovic, Colleen S Nell, Andrew Arthur Nelson, Gustav Nilsonne, Rory Nolan, Christopher E. O'Brien, Patrick O'Neill, Kieran O'Shea, Toto Olita, Jahna Otterbacher, Diana Palsetia, Bianca Pereira, Ivan Pozdniakov, John Protzko, Jean-Nicolas Reyt, Travis Riddle, Amal (Akmal) Ridhwan Omar Ali, Ivan Ropovik, Joshua M. Rosenberg, Stephane Rothen, Michael Schulte-Mecklenbeck, Nirek Sharma, Gordon Shotwell, Martin Skarzynski, William Stedden, Victoria Stodden, Martin A. Stoffel, Scott Stoltzman, Subashini Subbaiah, Rachael Tatman, Paul H. Thibodeau, Sabina Tomkins, Ana Valdivia, Gerrieke B. Druijff-van de Woestijne, Laura Viana, Florence Villesèche, W. Duncan Wadsworth, Florian Wanders, Krista Watts, Jason D Wells, Christopher E. Whelpley, Andy Won, Lawrence Wu, Arthur Yip, Casey Youngflesh, Ju-Chi Yu, Arash Zandian, Leilei Zhang, Chava Zibman, and Eric Luis Uhlmann. 2021. "Same Data, Different Conclusions: Radical Dispersion in Empirical Results when Independent Analysts Operationalize and Test the Same Hypothesis." *Organizational Behavior and Human Decision Processes* 165:228–249.

Šidák, Zbyně. 1967. "Rectangular confidence regions for the means of multivariate normal distributions." *Journal of the American Statistical Association* 62 (318):626–633.

Silberzahn, Raphael, Eric L. Uhlmann, Dan P. Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Štepán Bahník, Feng Bai, Colin Bannard, Evelina Bonnier, Rickard Carlsson, Felix Cheung, Garret Christensen, Russ Clay, Maureen A. Craig, Anna Dalla Rosa, Lammertjan Dam, Mathew H. Evans, Ismael Flores Cervantes, Nathan Fong, Monica Gamez-Djokic, Andreas Glenz, Shauna Gordon-McKeon, Tim J. Heaton, Karin Hederos, Moritz Heene, Alicia J. Hofelich Mohr, Fabia Högden, Kent Hui, Magnus Johannesson, Jonathan Kalodimos, Erikson Kaszubowski, Deanna M. Kennedy, Ryan Lei, Thomas A. Lindsay, Silvia Liverani, Christopher R. Madan, Daniel Molden, Eric Molleman, Richard D. Morey, Laetitia B. Mulder, Bernard R. Nijstad, Nolan G. Pope, Bryson Pope, Jason M. Prenoveau, Floor Rink, Egidio Robusto, Hadiya Roderique, Anna Sandberg, Elmar Schlüter, Felix D. Schönbrodt, Martin F. Sherman, S. Amy Sommer, Kristin Sotak, Seth Spain, Christoph Spörlein, Tom Stafford, Luca Stefanutti, Susanne Tauber, Johannes Ullrich, Michelangelo Vianello, Eric-Jan Wagenmakers, Maciej Witkowiak, Sangsuk Yoon, and Brian A. Nosek. 2018. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results." *Advances in Methods and Practices in Psychological Science* 1:337–356.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing Transparency Through a Multiverse Analysis." *Psychological Science* 11:702–712.

Tran, Anh and Richard Zeckhauser. 2012. "Rank as an Inherent Incentive: Evidence from a Field Experiment." *Journal of Public Economics* 96:645–650.

Yule, G. Udny. 1897. "On the Significance of Bravais' Formulae for Regression." *Proceedings of the Royal Society of London* 60:477–489.

Table 2.2: Summary statistics

This table presents summary statistics. Standard deviations are in parentheses.

Panel (a): Quality of the #fincap community

	Research teams	Peer evaluators
Fraction with top finance/econ publications (see footnote 18)	0.31	0.85
Fraction including at least associate/full professor	0.52	0.88
Experience empirical-finance research (low-high, 1-10)	8.1 (1.7)	8.4 (1.8)
Experience market-liquidity research (low-high, 1-10)	6.9 (2.4)	7.8 (2.3)
Relevant experience (average of the above two items)	7.5 (1.3)	8.1 (1.7)
Fraction with "big data" experience (>#fincap sample)	0.65	0.88
Fraction teams consisting of two members (maximum team size)	0.79	
Number of observations	164	34

Panel (b): Quality of the analysis of research teams

	Research teams
Reproducibility score according to Cascad (low-high, 0-100)	64.5 (43.7)
Paper quality as judged by peer evaluators (low-high, 0-10)	6.2 (2.0)

(continued on next page)

(continued from previous page)

Panel (c): Dispersion across teams of stage-1 results: Estimates, SEs, and *t*-values

	RT-H1 Efficiency	RT-H2 RSpread	RT-H3 Client Volume	RT-H4 Client RSpread	RT-H5 Client MOrders	RT-H6 Client GTR
Estimate (yearly change, %)						
Mean	446.3	-1,093.4	-3.5	-38,276.1	-3.5	-87.1
SD	5,817.5	14,537.2	9.4	490,024.2	37.6	728.5
Min	-171.1	-186,074.5	-117.5	-6,275,383.0	-452.9	-8,254.5
Q(0.10)	-23.7	-6.9	-3.8	-6.7	-1.6	-192.1
Q(0.25)	-6.2	-3.6	-3.5	-2.1	-0.6	-18.2
Median	-1.1	-0.0	-3.3	0.1	-0.0	0.0
Q(0.75)	0.5	3.9	-2.4	3.8	0.2	3.2
Q(0.90)	3.7	21.5	-0.1	20.4	1.0	56.5
IQR (i.e., NSE)	6.7	7.5	1.2	5.9	0.8	21.4
IDR	27.3	28.4	3.7	27.1	2.5	248.5
Max	74,491.1	4,124.0	8.7	870.2	69.5	1,119.0
Standard error						
Mean	468.7	1,195.3	3.7	38,302.0	6.2	148.2
SD	5,810.6	14,711.9	29.5	489,929.5	40.1	526.0
Min	0.0	0.0	0.0	0.0	0.0	0.0
Q(0.10)	0.1	0.2	0.1	0.2	0.1	0.0
Q(0.25)	0.5	1.1	0.3	1.2	0.2	0.7
Median	2.5	5.0	1.4	4.4	1.0	9.7
Q(0.75)	9.3	13.9	2.0	14.3	2.4	77.1
Q(0.90)	44.7	39.6	2.2	31.2	3.1	235.4
IQR	8.8	12.8	1.7	13.1	2.2	76.4
IDR	44.6	39.4	2.1	31.0	3.1	235.4
Max	74,425.5	188,404.1	378.8	6,274,203.0	463.7	4,836.2
t-value						
Mean	-3.6	35.3	-47.1	24.3	-5.7	-2.0
SD	28.4	541.2	269.9	406.0	60.1	21.2
Min	-322.3	-764.6	-2,770.6	-852.6	-631.6	-191.7
Q(0.10)	-4.7	-5.7	-37.4	-3.5	-2.3	-1.7
Q(0.25)	-1.9	-1.5	-11.5	-1.0	-0.6	-1.0
Median	-0.7	-0.1	-1.8	0.1	0.0	0.0
Q(0.75)	0.3	0.8	-1.6	1.0	0.8	0.7
Q(0.90)	1.7	1.5	-0.3	1.6	1.7	1.2
IQR	2.2	2.3	9.9	1.9	1.3	1.7
IDR	6.4	7.2	37.1	5.2	3.9	2.9
Max	51.6	6,880.5	29.5	5,119.5	89.6	100.6

Table 2.3: Non-standard error test

This table tests for the presence of non-standard errors in stage 1. It does so by testing whether estimates provided by researchers deviate from the median across researchers. Critical values of the individual tests are raised to achieve the desired significance level at the family of tests. The number of significantly negative tests and significantly positive tests is reported in brackets. The reported family p-value is the probability that out of all test statistics, at least one is larger than the reported value, under the null of a multivariate normal with means equal to the realized #fincap medians, and a covariance matrix with squared SEs (reported by the RTs) on the diagonal and off-diagonals that are either zero (Bonferroni) or based on the multiverse analysis (Section 2).

Panel (a): Multiple tests (Bonferroni)

	Reject no-NSE at 0.5%?	p-value of family test	Mean (SD) correlation test statistics	Effective number of tests
RT-H1	Yes (8, 25)	< 0.0001	0.00 (0.00)	164
RT-H2	Yes (24, 10)	< 0.0001	0.00 (0.00)	164
RT-H3	Yes (13, 25)	< 0.0001	0.00 (0.00)	164
RT-H4	Yes (22, 4)	< 0.0001	0.00 (0.00)	164
RT-H5	Yes (13, 10)	< 0.0001	0.00 (0.00)	164
RT-H6	Yes (8, 3)	< 0.0001	0.00 (0.00)	164

Panel (b): Multiple tests (based on multiverse analysis)

	Reject no-NSE at 0.5%?	p-value of family test	Mean (SD) correlation test statistics	Effective number of tests
RT-H1	Yes (8, 26)	< 0.0001	0.03 (0.21)	77
RT-H2	Yes (24, 10)	< 0.0001	0.05 (0.22)	81
RT-H3	Yes (13, 26)	< 0.0001	0.22 (0.34)	21
RT-H4	Yes (22, 4)	< 0.0001	0.08 (0.24)	67
RT-H5	Yes (13, 10)	< 0.0001	0.20 (0.34)	31
RT-H6	Yes (8, 3)	< 0.0001	0.02 (0.21)	86

Table 2.4: Stage-1 quantile regressions

This table presents the results of quantile regressions that characterize how the distribution of stage-1 estimates co-varies with various quality metrics. These metrics are team quality, reproducibility score, and (de-meaned) peer rating. The three quality variables have been standardized and, subsequently, multiplied by the IQR per RT-hypothesis. Their coefficient therefore measures the result of a one-standard deviation change, expressed in terms of interquartile-range units. $^{*}/^{**}$ correspond to significance at the 5/0.5% level, respectively.

	Q(0.10)	Q(0.25)	Q(0.50)	Q(0.75)	Q(0.90)
Team quality (standardized/scaled)	0.597**	0.004	0.002	0.032**	-0.325**
	(0.030)	(0.014)	(0.007)	(0.012)	(0.030)
Reproducibility score (standardized/scaled)	0.473**	0.109**	-0.001	-0.142**	-0.555**
	(0.033)	(0.014)	(0.007)	(0.011)	(0.028)
Average rating (standardized/scaled)	0.766**	0.230**	-0.001	-0.097**	-0.626**
	(0.034)	(0.014)	(0.007)	(0.011)	(0.028)
Dummy RT-H1 Efficiency	-29.592**	-6.099**	-1.132**	0.939**	9.057**
	(0.813)	(0.340)	(0.166)	(0.269)	(0.708)
Dummy RT-H2 RSpread	-15.933**	-3.930**	-0.017	3.674**	22.451**
	(0.849)	(0.342)	(0.166)	(0.268)	(0.705)
Dummy RT-H3 Client Volume	-5.629**	-3.789**	-3.319**	-2.386**	0.221
	(0.836)	(0.339)	(0.166)	(0.268)	(0.721)
Dummy RT-H4 Client RSpread	-12.089**	-2.437**	0.162	4.161**	19.619**
	(0.837)	(0.340)	(0.166)	(0.266)	(0.704)
Dummy RT-H5 Client MOrders	-2.479**	-0.744*	-0.001	0.297	1.625*
	(0.837)	(0.339)	(0.166)	(0.268)	(0.721)
Dummy RT-H6 GTR	-194.457**	-21.385**	0.022	5.137**	65.203**
	(0.806)	(0.337)	(0.167)	(0.268)	(0.679)
#Observations	984	984	984	984	984

Table 2.5: All-stages quantile regressions

This table presents the results of quantile regressions that characterize how the distribution of estimates varies across all stages of the #fincap project. The stage dummies have been multiplied by the (stage-1) IQR per RT-hypothesis. Their coefficient therefore measures the effect in terms of interquartile-range units. Standard errors account for correlation in residuals by adding RT-hypothesis fixed effects and by clustering per RT across all stages. $^{*}/^{**}$ correspond to significance at the 5/0.5% level, respectively.

	Q(0.10)	Q(0.25)	Q(0.50)	Q(0.75)	Q(0.90)
Dummy Stage 2 - Dummy Stage 1	2.44*	0.07	-0.00	-0.06	-0.73
	(1.18)	(0.14)	(0.01)	(0.06)	(0.64)
Dummy Stage 3 - Dummy Stage 2	0.94*	0.15	0.00	-0.09	-0.73
	(0.41)	(0.09)	(0.01)	(0.05)	(0.40)
Dummy Stage 4 - Dummy Stage 3	0.21*	0.06*	0.00	-0.04	-0.25*
	(0.09)	(0.03)	(0.01)	(0.03)	(0.11)
Dummy Stage 4 - Dummy Stage 1	3.59**	0.28*	-0.00	-0.19**	-1.71**
	(1.23)	(0.14)	(0.01)	(0.05)	(0.50)
RT-hypotheses dummies	Yes	Yes	Yes	Yes	Yes
#Observations	3,936	3,936	3,936	3,936	3,936

Table 2.6: Analysis paths

This table summarizes all analysis paths by spelling out all forks and all alternatives at these forks. It further presents the empirical distribution of decisions at all forks.

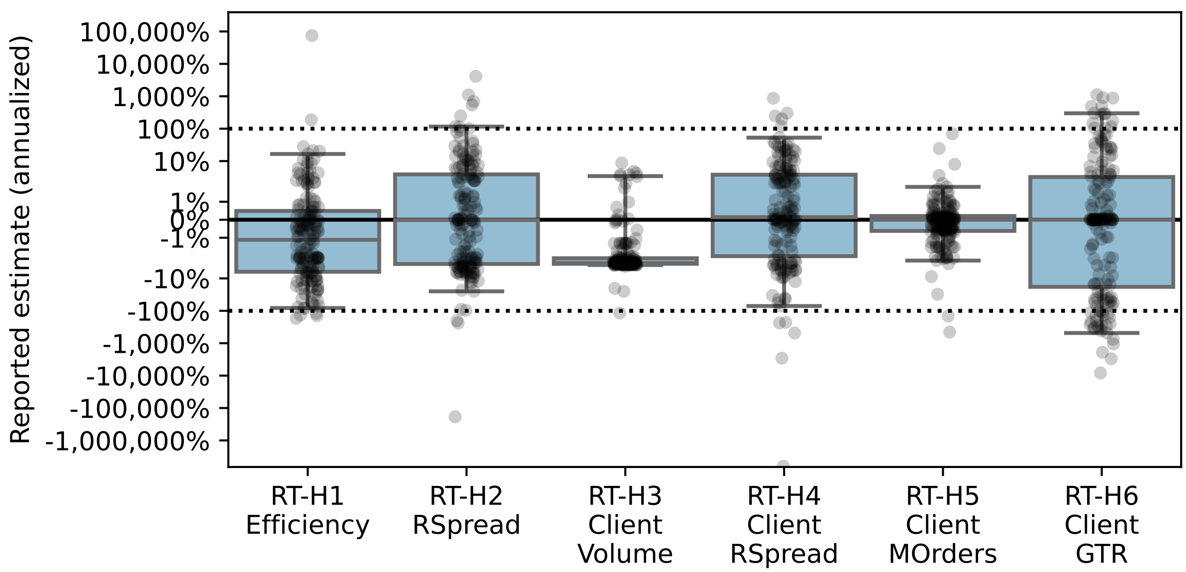
RT-hypothesis	Fork	Fork description	Alternatives	Frequency
All	1	Remove open/close	No	79%
			Yes, 30 minutes	21%
All	2	Days excluded	None	81%
			Settlement weeks	19%
All	3	Outlier treatment	None	65%
			Winsorize measure at 2.5 and 97.5 percentile[a]	20%
			Trim measure at 2.5 and 97.5 percentile[a]	14%
All	4	Frequency analysis	Daily	37%
			Weekly	1%
			Monthly	21%
			Annual	41%
All	5	Model	Trend stationary (regresion with linear trend)	35%
			Log difference (trivial regression, i.e., intercept only)	5%
			Relative difference (trivial regression)	60%
1	6	Measure	Variance ratio (low-frequency in numerator)	63%
			Autocorrelation (R^2 of AR model for returns)	37%
1	7	Measure frequencies	Second to minute	18%
			One to five minutes	26%
			Five to thirty minutes	34%
			Day to week	13%
			Day to month	10%
2,4,5	6	Tick test or aggressor flag	Aggressor flag (available only for part of the sample)	84%
			Tick test	16%
2,4	7	Post-trade value	Price 5 minutes after trade	81%
			Price 10 minutes after trade	6%
			Price 30 minutes after trade	13%
2,4	8	Aggregation	Equal-weighted average	47%
			Trade-size-weighted average	53%
3	6	Units...	Volume expressed in #contracts	70%
			Volume expressed in euro	30%
6	6	Reference price	Last trade price in the day	62%
			Last trade price one day later	1%
			Volume-weighted-average-price (VWAP) full-day	24%
			VWAP based on last five trades in the day	0%
6	7	Mean or median	Mean	96%
			Median	4%
6	8	Handle non-negatives	Translate and transform ($\varepsilon = 0.001$)	14%
			Translate and transform ($\varepsilon = 1$)	7%
			Set to missing	79%
6	9	Retain negative-trend sign	Yes	79%
			No	21%

[a]: Winsorization is applied at the frequency of analysis (fork 4).

Figure 2.2: Dispersion of stage-1 estimates across research teams

This plot illustrates the dispersion of stage-1 estimates across research teams. These estimates all focus on a trend in the sample, expressed in terms of a yearly percentage change. The six box plots correspond to the six trends RTs were asked to estimate. The boxes depict the first and third quartile. The horizontal line in the box corresponds to the median. The whiskers depict the 2.5% and 97.5% quantile. All estimates are also plotted individually as gray dots.

Panel (a): Dispersion of all estimates (N=164)



Panel (b): Dispersion of highest quality estimates (N=9)

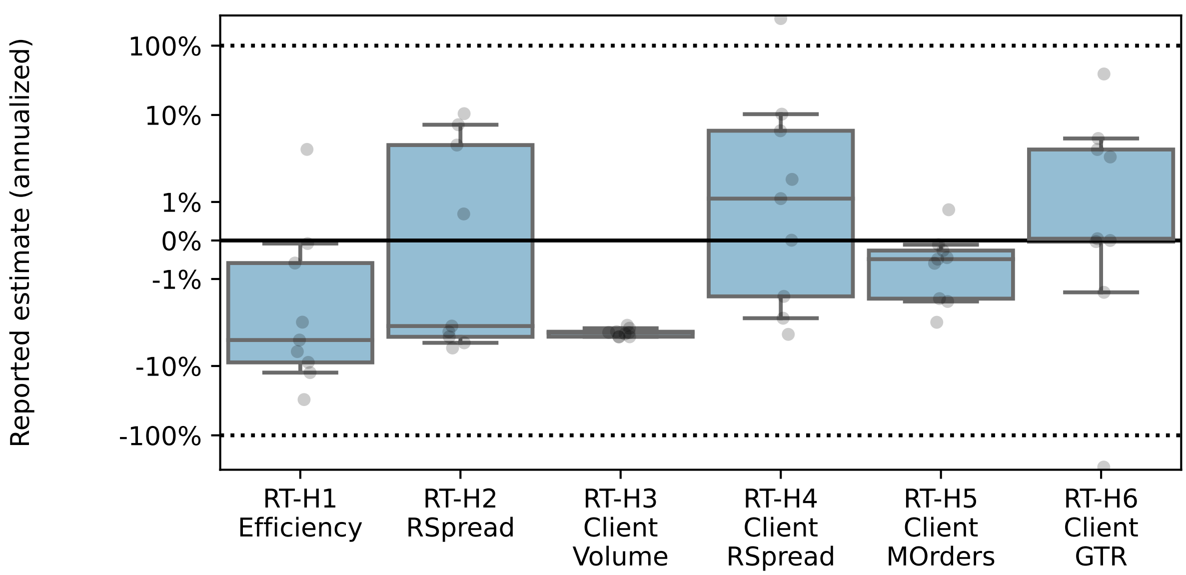


Figure 2.3: Dispersion in estimates related to quality measures

This figure plots how the dispersion in stage-1 estimates co-varies with various quality measures. The top graph uses the interquartile range (IQR) as a dispersion measure and the bottom graph uses the interdecile range (IDR). The quality variables are team quality, reproducibility score, and the rating by peer evaluators. The IQR and IDR estimates are taken from Table 2.4, where relative changes are averaged across RT-hypotheses. The baseline level is the average dispersion across RT-hypotheses.

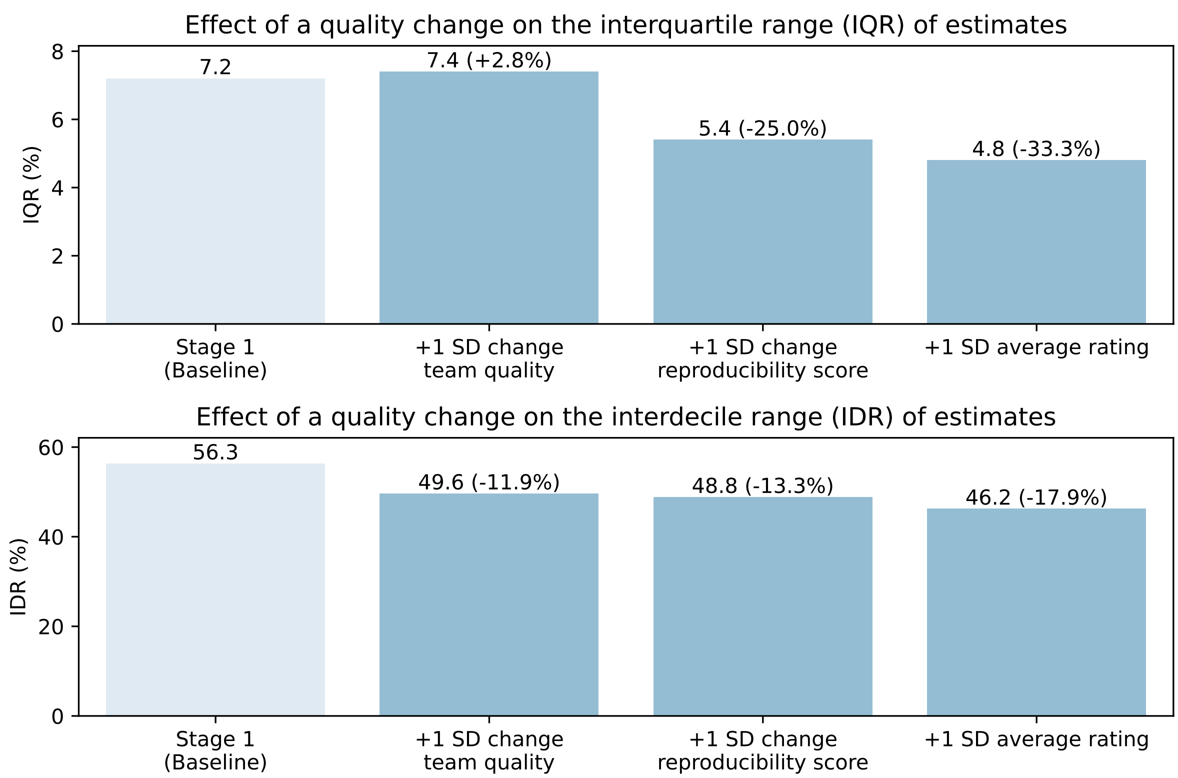


Figure 2.4: Dispersion in estimates related to feedback stages

This figure plots how the dispersion in estimates changes across feedback stages. Stage 1 is the baseline stage, which is the stage before any feedback. The top graph uses the interquartile range (IQR) as a dispersion measure, whereas bottom graph uses the interdecile range (IDR). The IQR and IDR values are based on the estimates in Table 2.5, where relative changes are averaged across all RT-hypotheses. The baseline level is the average dispersion in stage-1 estimates across RT-hypotheses.

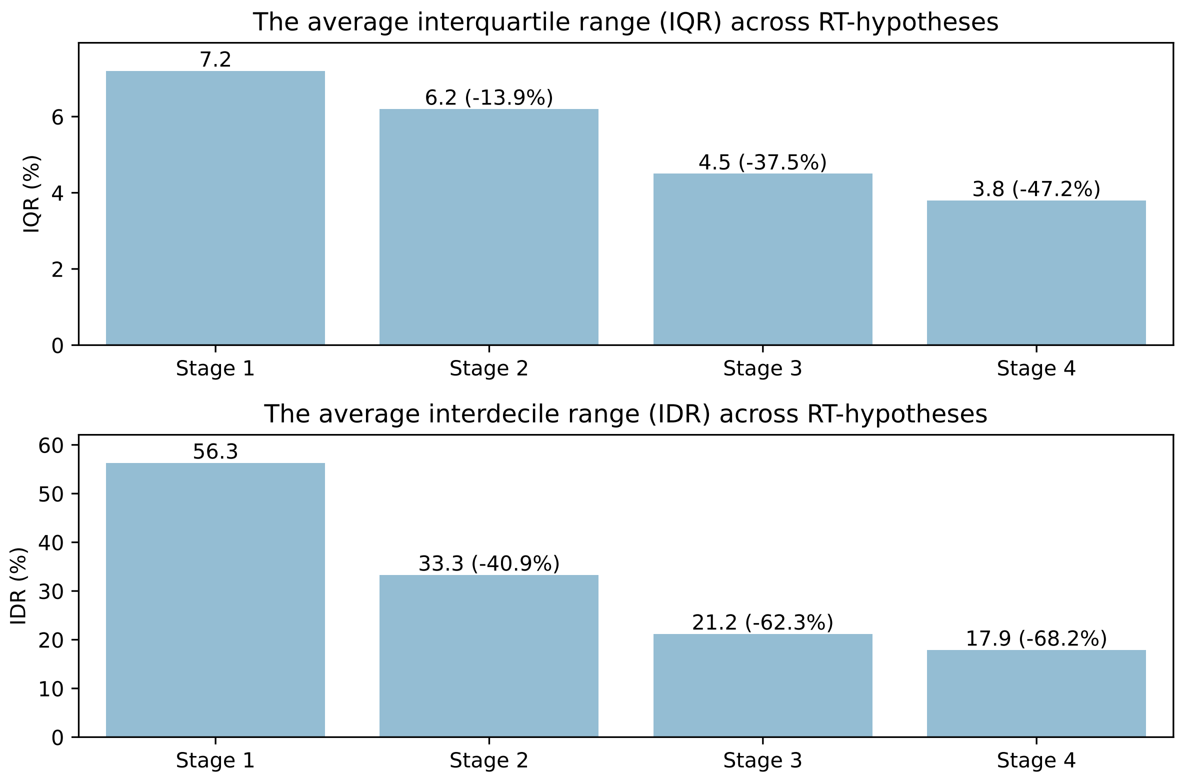


Figure 2.5: Research team beliefs on dispersion stage-1 estimates

This plot illustrates the dispersion in beliefs across research teams, for all six RT-hypotheses. All teams were asked to predict the SD in estimates across all RTs. The boxes depict the first and third quartile. The horizontal line in the box corresponds to the median. The whiskers depict the 2.5% and 97.5% quantile. All estimates are also plotted individually as gray dots. The red dots show the realized SD in estimates across RTs. The orange dots do the same, but are based on a 2.5%-97.5% trimmed sample.

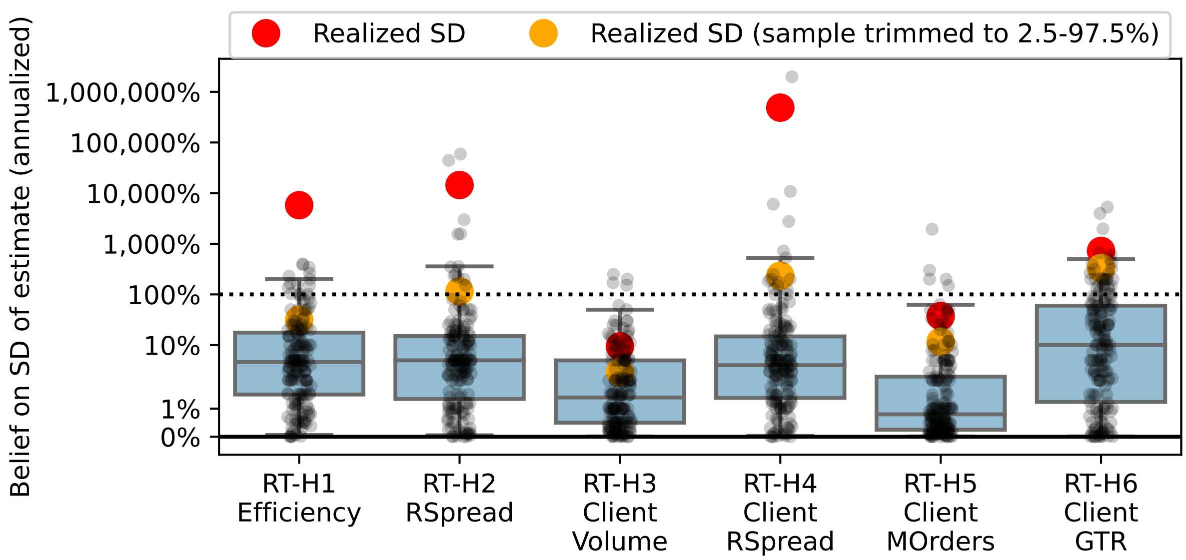


Figure 2.6: Dispersion in stage-1 estimates of multiverse analysis

This plot illustrates the dispersion in stage-1 estimates obtained from the multiverse analysis. The dispersion in *reported* estimates appears in gray and corresponds to panel (a) in Figure 2.2. The boxes depict the first and third quartile. The horizontal line in the box corresponds to the median. The whiskers depict the 2.5% and 97.5% quantile.

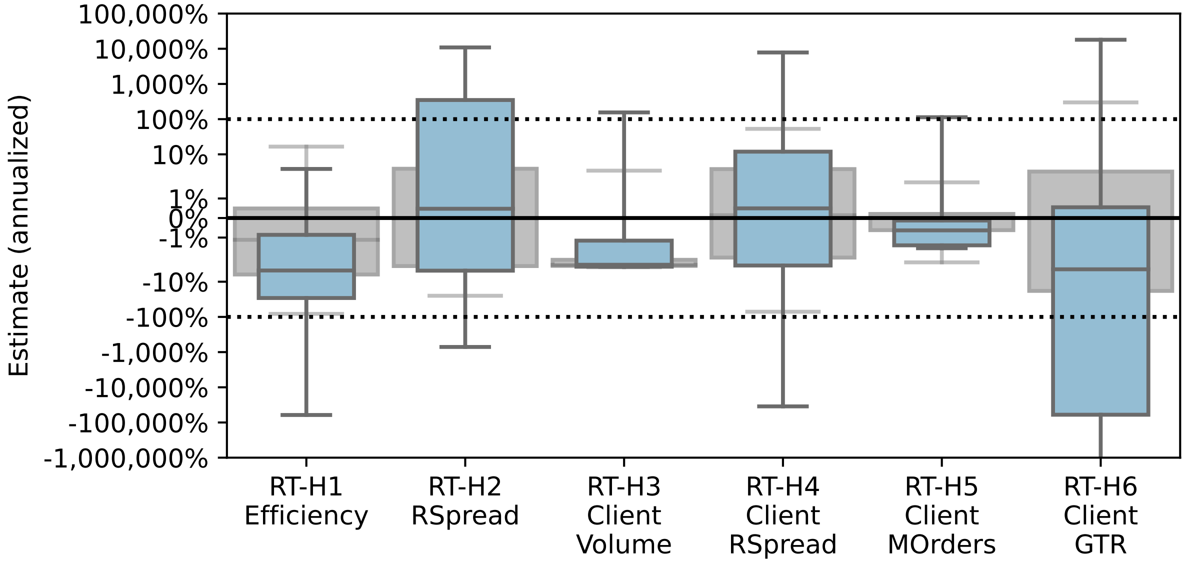


Figure 2.7: Fork sensitivity of estimates in multiverse analysis

This figure plots how sensitive the distribution in estimates is to the alternatives available at a fork in the multiverse analysis. The sensitivity is measured by the standardized Anderson-Darling test statistic. Higher values of the statistic imply that distributions become more dissimilar across alternatives at the fork.

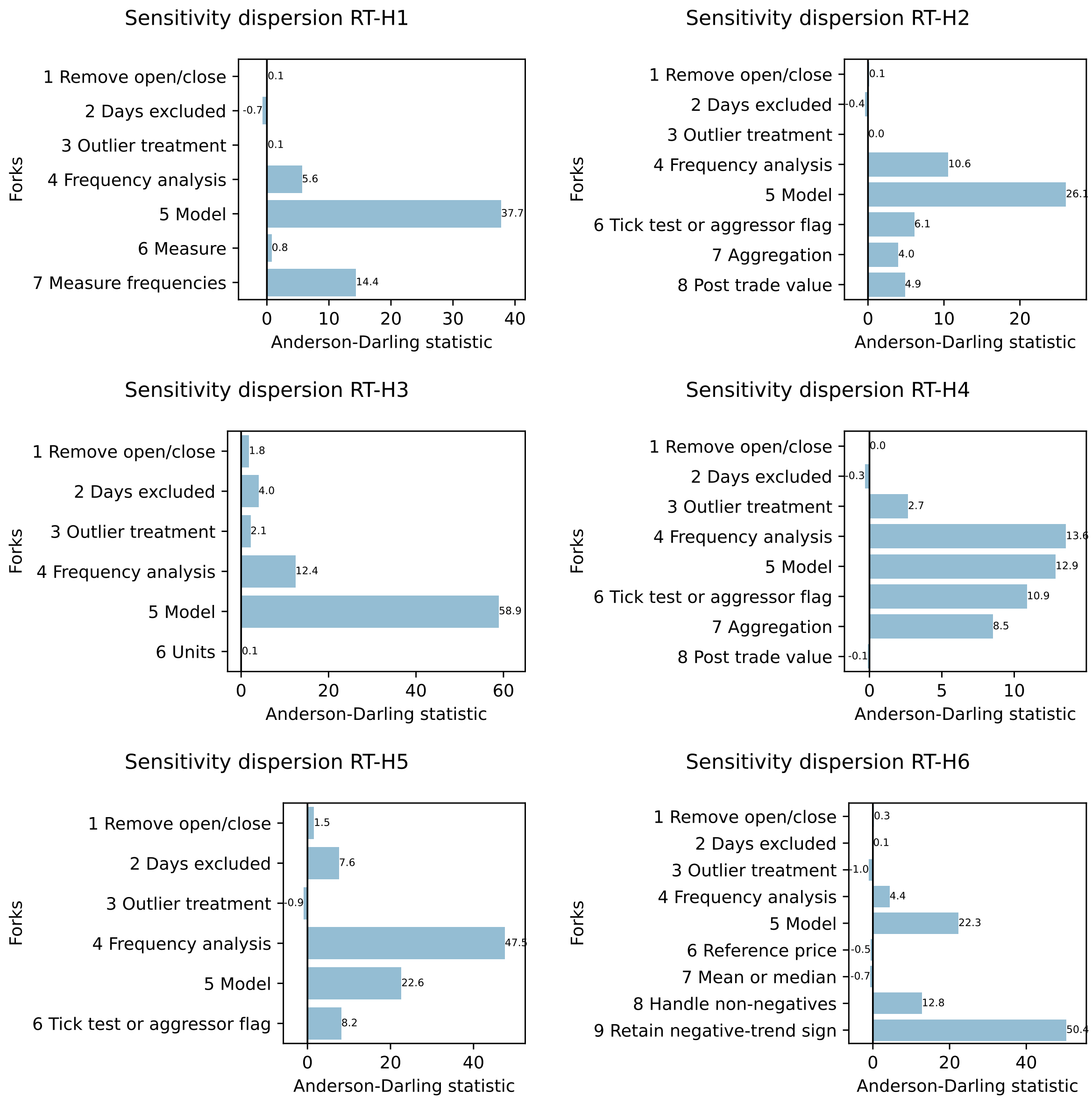
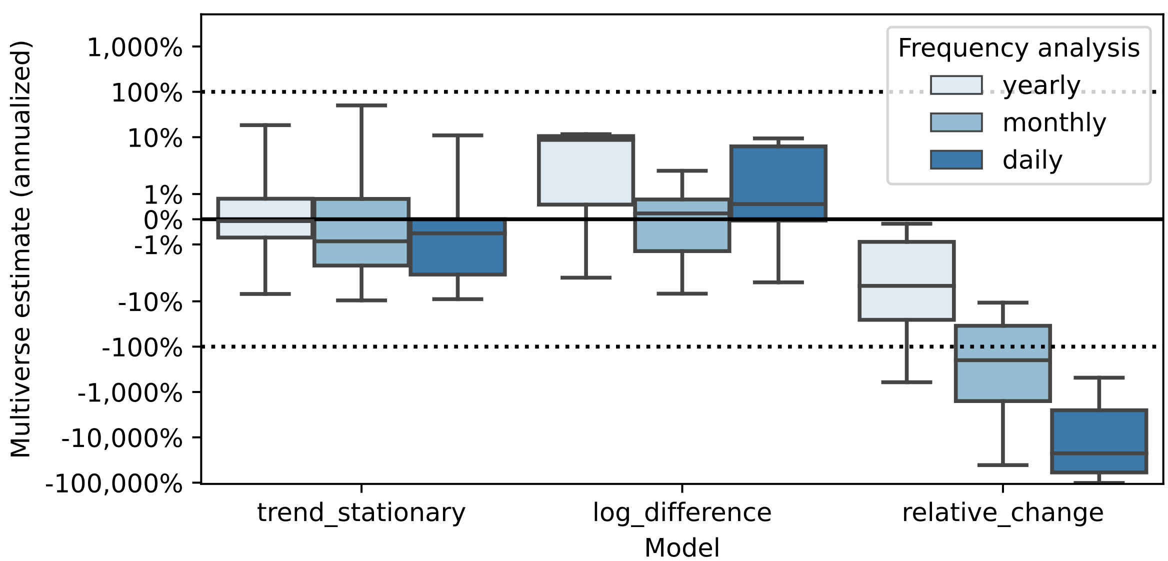


Figure 2.8: Sensitivity of estimates in multiverse analysis of RT-H1

This plot illustrates how the distribution of RT-H1 estimates depends on two influential forks in the multiverse analysis: (i) the model and (ii) the frequency of the analysis. Distributions are obtained by bootstrapping 1000 times from the original sample for each analysis path. To avoid clutter, the weekly frequency is dropped since it is used by only one team (out of 164).



Online appendix to

The online appendix contains the following additional material:

- Table OA.1: Summary statistics stage-2 dispersion in RT-results.
- Table OA.2: Summary statistics stage-3 dispersion in RT-results.
- Table OA.3: Summary statistics stage-4 dispersion in RT-results.
- Table OA.4: Principal component analysis team quality.
- Table OA.5: Stage-1 quantile regressions with all team quality variables.
- Table OA.6 Dispersion in research team beliefs.
- Table OA.7: Statistical significance under multiple testing.
- Figure OA.1: Distribution of the median/mean estimate.
- Figure OA.2: Countries of #fincap community.
- Figure OA.3: Dispersion estimates across all feedback stages.
- Figure OA.4: Multiverse fit.
- Figure OA.5: Distribution decisions at the forks.
- Figure OA.6: Details on the data sample.
- Figure OA.7: Sign-up sheet research team.
- Figure OA.8: Non-disclosure agreement research teams.
- Figure OA.9: Non-disclosure agreement peer evaluators.
- Figure OA.10: Registration as a team member.
- Figure OA.11: Instruction sheet for research teams.
- Figure OA.12: Form to submit stage 1 results.

- Figure OA.13: Form to submit stage 2 results.
- Figure OA.14: Form to submit stage 3 results.
- Figure OA.15: Belief elicitation sheet.
- Figure OA.16: Instruction sheet peer evaluators.
- Figure OA.17: The form that RTs completed for multiverse analysis.

Table OA.1: Summary statistics stage-2 dispersion in RT-results

This table mirrors panel (c) in Table 2.2, except this dispersion pertains to stage-2 results instead of stage 1 results.

	RT-H1 Efficiency	RT-H2 RSpread	RT-H3 Client Volume	RT-H4 Client RSpread	RT-H5 Client MOrders	RT-H6 Client GTR
Estimate (yearly change, %)						
Mean	451.2	-1,122.2	-3.6	-38,254.7	0.4	-37.1
SD	5,817.1	14,531.4	9.2	490,025.9	7.5	264.7
Min	-291.3	-186,074.5	-117.5	-6,275,383.0	-30.8	-3,024.9
Q(0.10)	-13.1	-7.7	-3.8	-5.8	-1.8	-83.4
Q(0.25)	-4.4	-4.7	-3.7	-2.7	-0.6	-9.4
Median	-1.2	-0.9	-3.3	0.0	-0.0	-0.0
Q(0.75)	0.3	2.5	-2.1	3.5	0.2	2.1
Q(0.90)	3.4	12.6	-0.3	14.2	1.1	33.7
IQR (i.e., NSE)	4.7	7.1	1.7	6.2	0.8	11.6
IDR	16.5	20.2	3.5	20.0	2.8	117.1
Max	74,491.1	1,098.0	4.8	870.2	86.1	486.5
Standard error						
Mean	462.2	1,166.4	3.5	38,279.4	2.0	86.3
SD	5,811.0	14,710.6	29.6	489,931.2	8.0	308.4
Min	0.0	0.0	0.0	0.0	0.0	0.0
Q(0.10)	0.1	0.0	0.0	0.0	0.0	0.0
Q(0.25)	0.4	0.5	0.1	0.7	0.1	0.3
Median	1.4	3.3	0.4	3.0	0.5	4.9
Q(0.75)	5.8	8.4	1.9	9.0	1.4	45.5
Q(0.90)	29.0	28.5	2.0	24.7	2.8	160.7
IQR	5.4	7.8	1.8	8.3	1.2	45.2
IDR	28.9	28.4	2.0	24.7	2.8	160.7
Max	74,425.5	188,404.1	378.8	6,274,203.0	86.1	2,740.2
t-value						
Mean	-4.8	24.9	-99.4	28.8	-3.5	-0.1
SD	28.4	434.5	627.1	400.5	73.8	4.4
Min	-322.3	-907.7	-7,208.7	-160.0	-876.2	-38.3
Q(0.10)	-7.3	-8.4	-45.4	-5.0	-3.7	-2.0
Q(0.25)	-2.3	-2.5	-17.4	-1.2	-0.8	-0.8
Median	-0.8	-0.4	-3.5	0.0	-0.1	-0.0
Q(0.75)	0.5	0.7	-1.7	1.0	1.0	1.0
Q(0.90)	2.3	1.6	-0.4	2.5	3.1	1.4
IQR	2.8	3.3	15.7	2.3	1.7	1.8
IDR	9.6	10.0	45.0	7.5	6.8	3.4
Max	30.8	5,479.5	56.1	5,120.8	318.8	25.2

Table OA.2: Summary statistics stage-3 dispersion in RT-results

This table mirrors panel (c) in Table 2.2, except this dispersion pertains to stage-3 results instead of stage 1 results.

	RT-H1 Efficiency	RT-H2 RSpread	RT-H3 Client Volume	RT-H4 Client RSpread	RT-H5 Client MOrders	RT-H6 Client GTR
Estimate (yearly change, %)						
Mean	453.4	-1,130.6	-3.1	-38,263.4	-2.4	-2.4
SD	5,816.9	14,530.5	10.7	490,025.2	36.2	105.9
Min	-70.5	-186,074.0	-117.5	-6,275,383.0	-452.9	-898.7
Q(0.10)	-6.7	-8.0	-3.9	-7.1	-1.7	-15.1
Q(0.25)	-3.2	-5.7	-3.8	-3.4	-0.6	-0.5
Median	-1.0	-1.8	-3.3	-0.3	-0.0	0.0
Q(0.75)	-0.0	0.0	-1.3	0.8	0.2	1.4
Q(0.90)	2.2	5.5	-0.4	5.3	0.9	12.8
IQR (i.e., NSE)	3.2	5.8	2.4	4.1	0.7	1.8
IDR	8.9	13.6	3.5	12.4	2.6	28.0
Max	74,491.1	1,098.0	66.7	302.4	86.1	486.5
Standard error						
Mean	458.6	1,156.2	3.0	38,264.0	4.2	40.5
SD	5,811.3	14,711.3	29.6	489,932.4	36.8	149.5
Min	0.0	0.0	0.0	0.0	0.0	0.0
Q(0.10)	0.1	0.0	0.0	0.0	0.1	0.0
Q(0.25)	0.3	0.4	0.1	0.4	0.1	0.0
Median	0.6	1.1	0.2	1.2	0.3	1.6
Q(0.75)	2.1	3.8	0.7	3.4	0.7	8.0
Q(0.90)	7.9	10.0	2.0	10.1	1.9	58.5
IQR	1.8	3.5	0.6	3.0	0.6	8.0
IDR	7.8	9.9	1.9	10.1	1.9	58.5
Max	74,425.5	188,404.0	378.8	6,274,203.0	463.7	1,149.3
t-value						
Mean	-3.7	25.2	-56.5	29.8	-3.3	0.3
SD	12.5	434.4	363.3	400.3	73.2	3.3
Min	-131.7	-908.3	-3,800.0	-160.0	-876.2	-10.2
Q(0.10)	-7.9	-8.8	-36.5	-5.5	-4.2	-1.6
Q(0.25)	-3.8	-5.1	-28.5	-2.9	-1.5	-0.6
Median	-1.8	-1.2	-11.1	-0.3	-0.0	0.1
Q(0.75)	0.1	0.3	-3.1	0.9	1.0	1.0
Q(0.90)	2.2	1.6	-1.4	2.5	3.5	1.6
IQR	3.8	5.4	25.4	3.9	2.5	1.6
IDR	10.0	10.5	35.0	8.0	7.7	3.2
Max	11.9	5,479.5	56.1	5,120.8	318.8	25.2

Table OA.3: Summary statistics stage-4 dispersion in RT-results

This table mirrors panel (c) in Table 2.2, except this dispersion pertains to stage-4 results instead of stage 1 results.

	RT-H1 Efficiency	RT-H2 RSpread	RT-H3 Client Volume	RT-H4 Client RSpread	RT-H5 Client MOrders	RT-H6 Client GTR
Estimate (yearly change, %)						
Mean	453.5	1,138.6	-1.8	-38,263.2	-2.9	4.9
SD	5,816.9	14,529.9	9.6	490,025.2	35.4	71.2
Min	-70.5	-90.1	-6.9	-6,275,383.0	-452.9	-360.7
Q(0.10)	-6.2	-6.9	-3.8	-5.8	-1.3	-5.0
Q(0.25)	-2.8	-4.4	-3.8	-2.0	-0.5	-0.2
Median	-1.1	-2.3	-2.9	-0.2	0.0	0.0
Q(0.75)	-0.2	-0.1	-2.0	0.4	0.1	0.8
Q(0.90)	1.2	2.2	-1.1	3.6	0.8	5.7
IQR (i.e., NSE)	2.6	4.3	1.7	2.4	0.6	1.1
IDR	7.4	9.1	2.7	9.5	2.1	10.8
Max	74,491.1	186,074.5	117.5	302.4	7.1	486.5
Standard error						
Mean	457.9	1,155.3	3.0	38,261.5	3.5	24.7
SD	5,811.4	14,711.4	29.6	489,932.6	36.2	88.8
Min	0.0	0.0	0.0	0.0	0.0	0.0
Q(0.10)	0.1	0.1	0.1	0.0	0.1	0.0
Q(0.25)	0.3	0.6	0.1	0.6	0.1	0.1
Median	0.5	1.2	0.3	1.3	0.3	2.0
Q(0.75)	1.5	3.0	0.5	2.7	0.6	5.2
Q(0.90)	5.2	7.0	1.8	5.5	1.2	46.6
IQR	1.2	2.4	0.4	2.1	0.4	5.1
IDR	5.1	7.0	1.8	5.5	1.1	46.6
Max	74,425.5	188,404.1	378.8	6,274,203.0	463.7	786.1
t-value						
Mean	-3.7	25.1	-54.7	29.7	-3.6	0.6
SD	12.1	434.4	363.4	400.2	73.0	6.4
Min	-131.7	-911.2	-3,801.4	-159.8	-876.2	-9.0
Q(0.10)	-7.5	-8.0	-33.5	-3.6	-3.6	-1.3
Q(0.25)	-3.1	-4.0	-18.5	-1.8	-1.1	-0.3
Median	-2.0	-1.9	-11.5	-0.2	0.0	0.1
Q(0.75)	-0.4	-0.4	-4.1	0.3	1.0	0.8
Q(0.90)	1.8	0.8	-1.7	1.8	2.9	1.4
IQR	2.7	3.6	14.3	2.1	2.1	1.1
IDR	9.4	8.8	31.8	5.4	6.5	2.7
Max	8.0	5,479.5	19.5	5,120.8	318.8	80.2

Table OA.4: Principal component analysis team quality

This table presents the results of a principal component analysis of the standardized team quality variables.

Panel (a): Correlation team quality measures

	Publications	Experience	Big Data	Position	#Members
Publications		0.34	0.10	0.54	0.30
Experience			-0.18	0.25	0.12
Big Data				0.14	0.14
Position					0.16

Panel (b): Fraction of variance explained

	PC1	PC2	PC3	PC4	PC5
Variance explained	38.3%	23.6%	17.1%	12.4%	8.6%

Panel (c): Loading of principal components on variables

	Publications	Experience	Big Data	Position	#Members
PC1	0.61	0.40	0.13	0.55	0.37
PC2	-0.01	-0.55	0.79	0.05	0.26
PC3	-0.10	0.06	-0.21	-0.46	0.86
PC4	-0.20	0.71	0.56	-0.35	-0.12
PC5	-0.76	0.14	-0.02	0.60	0.22

Table OA.5: Stage-1 quantile regressions with all team quality variables

This table replicates Table 2.4 in main paper. The only difference is that the first principal component of the team quality variables is replaced by the quality variables themselves. */** correspond to significance at the 5/0.5% level, respectively.

	Q(0.10)	Q(0.25)	Q(0.50)	Q(0.75)	Q(0.90)
Top publications (standardized/scaled)	-0.200**	-0.213**	0.000	0.056**	0.648**
	(0.063)	(0.020)	(0.009)	(0.013)	(0.037)
Experience in field (standardized/scaled)	-0.370**	0.031*	0.000	0.002	-0.106**
	(0.040)	(0.015)	(0.007)	(0.012)	(0.028)
Experience with big data (standardized/scaled)	0.074	-0.025	0.000	0.030*	0.138**
	(0.048)	(0.015)	(0.007)	(0.011)	(0.027)
Academic seniority (standardized/scaled)	1.666**	0.171**	0.001	-0.026*	-1.022**
	(0.058)	(0.019)	(0.008)	(0.013)	(0.035)
Team size (1 or 2 members) (standardized/scaled)	0.819**	0.068**	0.004	-0.061**	-0.150**
	(0.044)	(0.015)	(0.007)	(0.011)	(0.028)
Reproducibility score (standardized/scaled)	0.494**	0.186**	-0.001	-0.116**	-0.521**
	(0.043)	(0.015)	(0.007)	(0.011)	(0.028)
Average rating (standardized/scaled)	0.521**	0.158**	-0.001	-0.067**	-0.486**
	(0.041)	(0.014)	(0.007)	(0.011)	(0.029)
Dummy RT-H1 Efficiency	-31.553**	-6.779**	-1.107**	0.844**	9.262**
	(1.032)	(0.354)	(0.164)	(0.271)	(0.664)
Dummy RT-H2 RSpread	-22.463**	-4.543**	-0.032	3.846**	20.961**
	(0.990)	(0.356)	(0.165)	(0.270)	(0.647)
Dummy RT-H3 Client Volume	-6.548**	-3.776**	-3.315**	-2.361**	0.023
	(1.007)	(0.354)	(0.164)	(0.270)	(0.661)
Dummy RT-H4 Client RSpread	-18.079**	-2.875**	0.158	4.365**	18.189**
	(0.952)	(0.351)	(0.164)	(0.270)	(0.645)
Dummy RT-H5 Client MOrders	-3.169**	-0.853*	-0.003	0.286	1.817**
	(1.007)	(0.354)	(0.164)	(0.270)	(0.625)
Dummy RT-H6 GTR	-178.868**	-20.853**	-0.020	5.672**	60.704**
	(0.993)	(0.351)	(0.166)	(0.265)	(0.676)
#Observations	984	984	984	984	984

Table OA.6: Dispersion in research team beliefs

This presents test statistics on whether the beliefs of research teams about dispersion across research teams matches realized dispersion. More precisely, the test statistic is defined as the difference between the average belief on the SD across teams, minus the realized SD, where this difference is divided by the realized SD. The *p*-values in parentheses are obtained by bootstrapping 1000 times. $^{*}/^{**}$ correspond to significance at the 5/0.5% level, respectively.

	RT-H1 Efficiency	RT-H2 RSpread	RT-H3 Client Volume	RT-H4 Client RSpread	RT-H5 Client MOrders	RT-H6 Client GTR	All
Estimate	-99.5%** (0.00)	-95.4%** (0.00)	-9.0% (0.74)	-97.5%** (0.00)	-45.3% (0.63)	-83.3%** (0.00)	-71.7%** (0.00)

Table OA.7: Statistical significance under multiple testing

This table tests how many stage-4 estimates are statistically significant, where significance levels for the individual tests are adjusted to account for multiple testing. The number of significantly negative tests and significantly positive tests is reported in brackets. The reported family *p*-value is the probability that out of all test statistics, at least one is larger than the reported value, under the null of a multivariate normal with zero means, and a covariance matrix with squared SEs (reported by the RTs) on the diagonal and off-diagonals that are either zero (Bonferroni) or based on the multiverse analysis (Section 2).

Panel (a): Multiple tests (Bonferroni)

	Reject at 0.5%?	*p*-value of family test	Mean (SD) correlation test statistics	Effective number of tests
RT-H1	Yes (31, 4)	< 0.0001	0.00 (0.00)	164
RT-H2	Yes (38, 3)	< 0.0001	0.00 (0.00)	164
RT-H3	Yes (123, 2)	< 0.0001	0.00 (0.00)	164
RT-H4	Yes (15, 8)	< 0.0001	0.00 (0.00)	164
RT-H5	Yes (13, 9)	< 0.0001	0.00 (0.00)	164
RT-H6	Yes (3, 3)	< 0.0001	0.00 (0.00)	164

Panel (b): Multiple tests (based on multiverse analysis)

	Reject at 0.5%?	*p*-value of family test	Mean (SD) correlation test statistics	Effective number of tests
RT-H1	Yes (32, 5)	< 0.0001	0.01 (0.24)	73
RT-H2	Yes (47, 3)	< 0.0001	0.03 (0.28)	41
RT-H3	Yes (125, 2)	< 0.0001	0.14 (0.53)	13
RT-H4	Yes (15, 8)	< 0.0001	0.03 (0.28)	91
RT-H5	Yes (16, 11)	< 0.0001	0.10 (0.46)	21
RT-H6	Yes (3, 3)	< 0.0001	0.01 (0.25)	51

Figure OA.1: Distribution of the median/mean estimate

This figure plots the empirical probability density functions (PDFs) of the median estimate and the mean estimate across 164 researchers. They are plotted based on 1000 draws from a multivariate normal with means equal to the realized #fincap medians, and a covariance matrix with squared SEs (reported by the RTs) on the diagonal and off-diagonals that are based on the multiverse analysis (Section 2). This distribution is also used to generate panel (b) of Table 2.3.

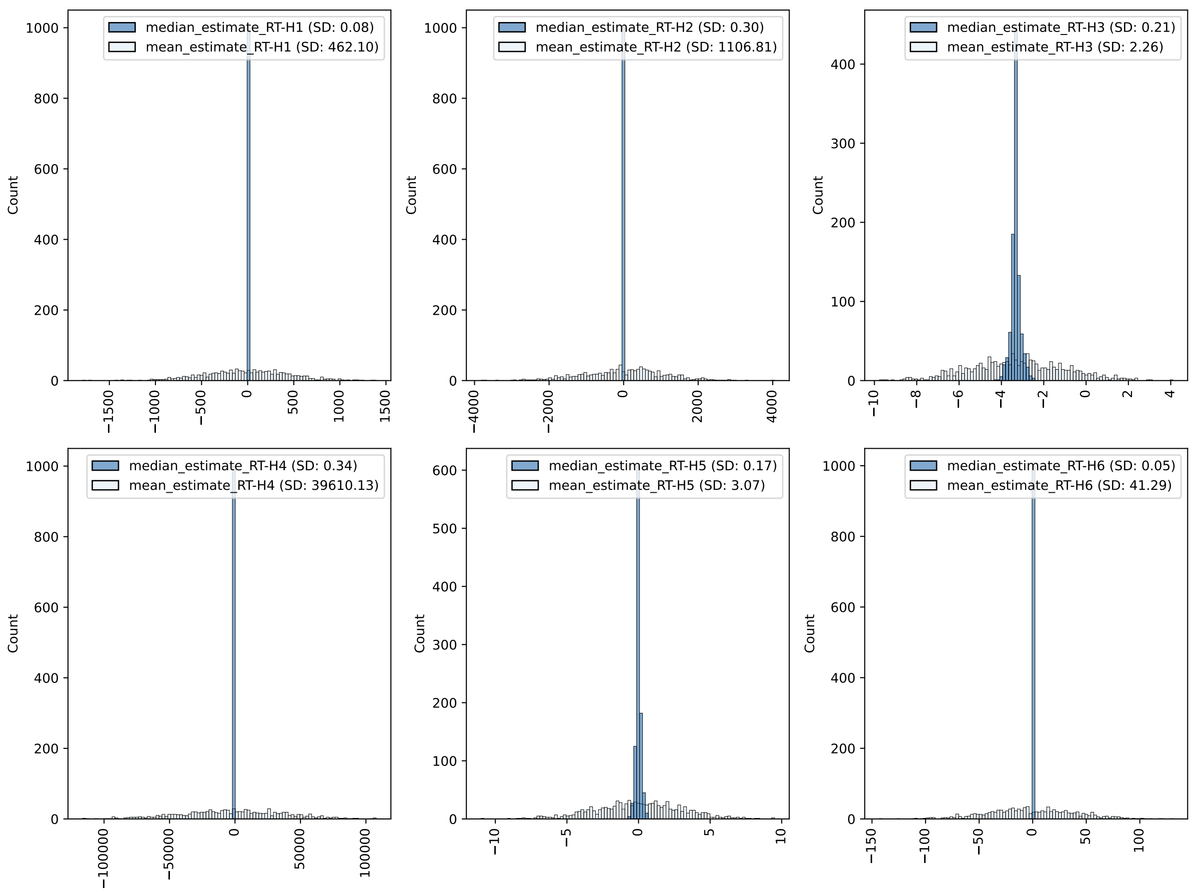


Figure OA.2: Countries of #fincap community

This plot illustrates how the #fincap community is dispersed around the globe. The top plot depicts how the members of the research teams are dispersed across countries. The bottom plot does the same for the peer evaluators.

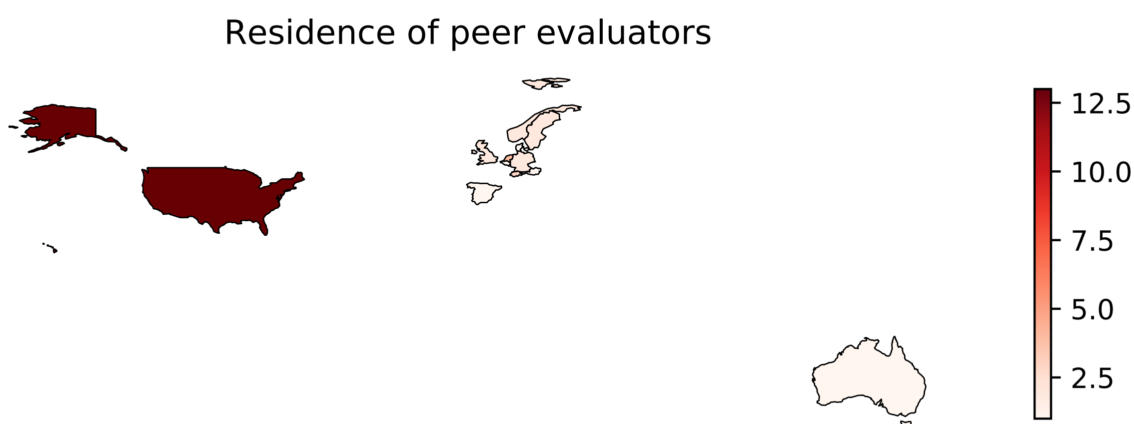


Figure OA.3: Dispersion estimates across all feedback stages

This plot illustrates the dispersion in RT estimates across all feedback stages. The boxes depict the first and third quartile. The horizontal line in the box corresponds to the median. The whiskers depict the 2.5% and 97.5% quantile.

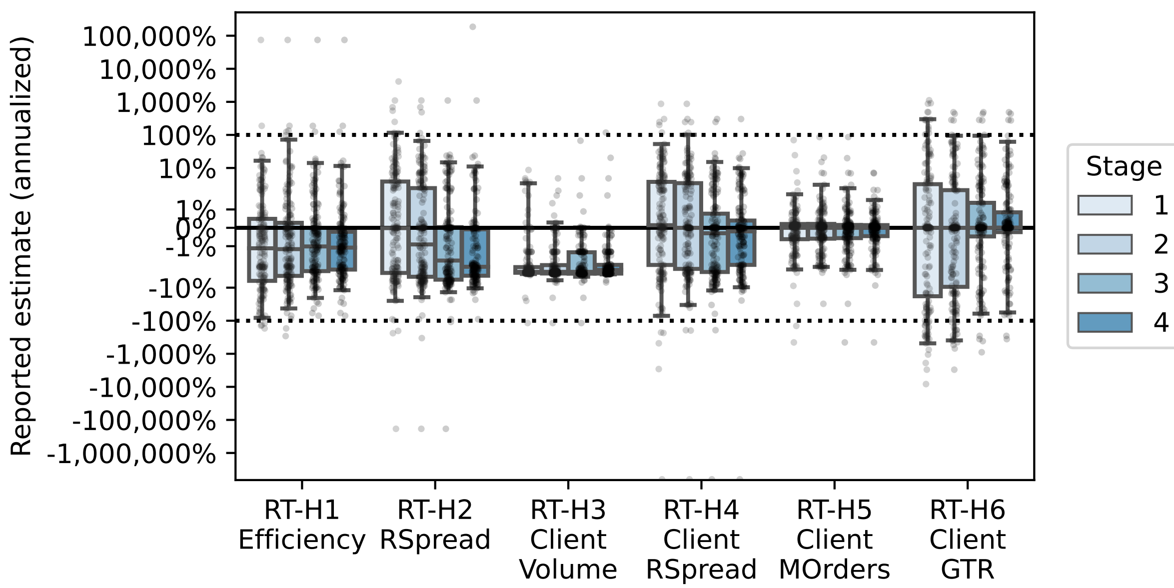


Figure OA.4: Multiverse fit

This figure plots the quality of the fit between the paths in the multiverse analysis and the paths that RTs followed when conducting their analyis. More specifically, for the multiverse analysis, each RT was asked to select among a set of pre-defined alternatives for a number of key forks on the analyis path (see Table 2.6). After selecting an alternative, they were they were asked to rate how close the selected alternative was to what they actually did on a scale from 1 ("Far from what we did.") to 5 ("Very close to what we did."). The graph plots the overall average across all RTs and all forks.

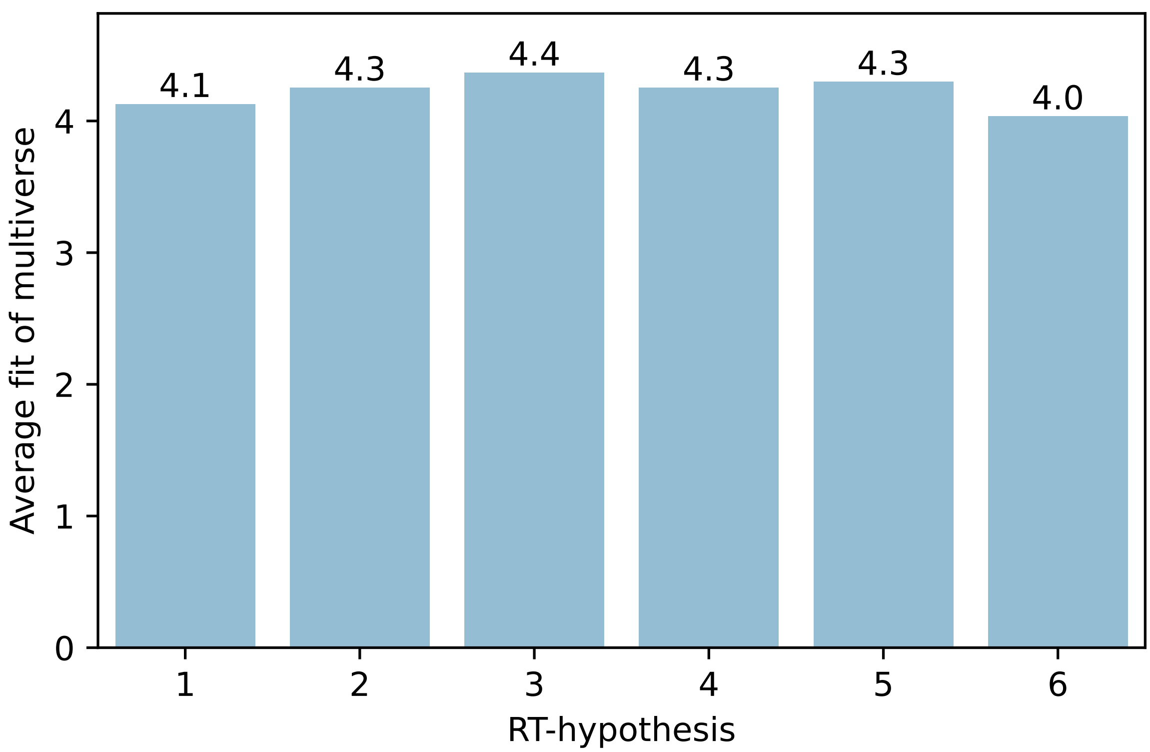


Figure OA.5: Distribution decisions at the forks

This figure plots the distribution of decisions for all forks in the multiverse. It depicts the information that is also available in Table 2.6.

Figure OA.6: Details on the data sample

about the data

The data pertain to 17 years (2002-2018) of trading of EuroStoxx 50 futures, which are among the world's most actively traded index derivatives. They give investors exposure to "Europe," or, more precisely, to a basket of euro-area blue-chip equities. All trading is done through an electronic limit-order book (see, e.g., Parlour and Seppi, 2008). Please find more background information on the futures in this factsheet.

The data consist of 720 million trade records and will be made available in monthly gzipped semicolon separated text files ("csv"). Each zipped monthly file is no larger than 50 MB. The data is clean in the sense that for all files the format is identical. Please find below the first ten lines of the December 2018 file as an example.

```
DATETIME; EXPIRATION; BUY_SELL_ID; TRADE_SIZE; MATCH_PRICE; AGGRESSOR_FLAG;ACCOUNT_ROLE; EXEC_TYPE_ID
2018-12-03 08:00:06.400; 201812; S; 2; 3229; N; A; F
2018-12-03 08:00:06.410; 201812; S; 1; 3229; N; A; F
2018-12-03 08:00:06.410; 201812; S; 1; 3229; N; A; F
2018-12-03 08:00:06.410; 201812; B; 4; 3229; Y; A; F
2018-12-03 08:00:06.540; 201812; S; 1; 3229; N; A; F
2018-12-03 08:00:06.550; 201812; B; 2; 3229; Y; A; F
2018-12-03 08:00:06.550; 201812; S; 1; 3229; N; A; F
2018-12-03 08:00:06.630; 201812; B; 1; 3229; Y; A; F
2018-12-03 08:00:06.630; 201812; S; 1; 3229; N; A; F
```

The variables are defined as follows (the characterizations are short and therefore imprecise, please refer to any standard textbook on futures to get a detailed description of what futures are and how they are traded):

- **DATETIME:** Time stamp of the trade denoted as YYYY-MM-DD hh:mm:ss.sss where ss.sss denotes seconds up to a third decimal (i.e., the precision is tens of milliseconds as the last digit is always zero).
- **EXPIRATION:** The expiration date of the futures contract being traded. All data pertain to Eurex trading in EuroStoxx 50 (SX5E) futures contracts. Expiration months are: March, June, September, and December. Contracts expire on the third Friday of the expiration month. The notation of expiration is YYYYMM (where MM is in [03, 06, 09, 12]).
- **BUY_SELL:** This indicator shows if the trade record is for a buyer "B" (who goes long the index) or for a seller "S" (who goes short the index).
- **TRADE_SIZE:** This is the size of the trade expressed in number of contracts. The contract value per index point is EUR 10 (e.g., per contract traded, the long side is entitled to receive 10 euro from the short side of the trade each time the index increases by one point).
- **MATCH_PRICE:** The price at which the trade between buyer and the seller is concluded (i.e., the long and the short side of the trade, respectively).
- **AGGRESSOR_FLAG:** If the trade record pertains to a market order (or marketable limit order) that is executed against a standing limit order, this flag takes the value "Y". If the record pertains to a limit order, resting in the book before being matched with an incoming market order, or to an order in an auction (e.g., the opening and closing auction), then this flag takes the value "N". This flag became available as of November 2009.
- **ACCOUNT_ROLE:** This variable is either:
  - A: Agency trade (i.e., a trade an exchange member does for a client).
  - M: Market-maker principal trade (i.e., a trade an exchange member does for his own account in his role as market maker).
  - P: Non-market-maker principal trade (i.e., a trade an exchange member does for his own account).
  - P.S.: The distinction between M and P is not an economically meaningful one for the purpose of this project.
- **EXEC_TYPE_ID:** This variable is:
  - F if the full order was executed in the trade.
  - P if the order was only partially executed in the trade.
  - N if not assigned.

Figure OA.7: Sign-up sheet research team

sign-up as a research team

Please fill in the following form if you wish to participate as a research team (RT). You can either participate **alone** or **in a team of two** (in case you want to join the project as a team of two researchers, the form below only needs to be submitted once). Upon submission of the sign-up form, a link to the entry survey will be forwarded to the e-mail address(es) you enter below. Please make sure that the entry survey is filled out by each team member within three days. Looking forward to collaborating with you in this project!

team member #1:

full name: *

first name, last name

e-mail address: *

e-mail address the entry survey will be sent to

team member #2 (optional):

full name:

first name, last name

e-mail address:

e-mail address the entry survey will be sent to

informed consent: *

I understand that the project requires research expertise and experience in empirical finance (with a particular focus on market liquidity) and the analysis of large datasets. Also, at least one of the team members has to hold a PhD in finance or economics. I consent to elaborate on why our team fulfills the requirements for participating in the study in the form field below.

☐ I agree

fulfillment of RT requirements: *

Please provide a brief explanation why you (and your potential team member) are sufficiently skilled to participate as a research team in #fincap. In particular, briefly sketch your background in empirical finance research, research methodology, etc. Please note that two to three sentences are sufficient.

Our team fulfills the participation requirements because ...

Figure OA.8: Non-disclosure agreement research teams

non-disclosure agreement for research team members

Note: The following agreement was signed by all nine project coordinators (Albert J. Menkveld, Anna Dreber, Jürgen Huber, Felix Holzmeister, Magnus Johannesson, Michael Kirchler, Michael Razen, and Utz Weitzel) - hereinafter referred to as "Data Supplier" - and each research team member - herinafter referred to as "Data Recipient" - on an individual basis.

Whereas,

Data Supplier intends to provide records of executions of the EuroStoxx® future by means of a temporary electronic access via a shared drive to Data Recipient ("Data").

The Data comprises the following semicolon separated fields for executions in the simple instruments of the EuroStoxx® Future between 1st January 2002 and 31st December 2018:

FieldName	*Description*	*Format*	*Example*
DATETIME	Date and Time of the execution	String: yyyy-mm-dd hh:mm:ss.000	2018-10-15 08:00:04.840
EXPIRATION	Expiration months of the contract	Integer: yyyymm	201809
BUY_SELL_ID	Buy/Sell Indicator	Char: 1 = Buy, 2 = Sell	1
TRADE_SIZE	Number of contracts executed	Integer	2
MATCH_PRICE	Price of execution in points	Integer	3176
AGGRESSOR_FLAG	Flag indicating if an order was aggressive or passive	Char: Y = aggressive, N = not aggressive	Y
ACCOUNT_ROLE	Account Role	Char: M = MarketMaker, P = Proprietary, A = Agency	M
EXEC_TYPE_ID	Full/partial execution indicator	Char: F: Full execution, P: Partial Execution	F

Therefore,

Data Recipient and Data Supplier agree as follows:

1. Data Recipient shall only use the Data for empirical analyses of the Data and/or for evaluations of other data recipients' analyses within a project where multiple decentral teams report their results and/or evaluations to the core research team (represented by Data Recipient) ("Project"). Data Recipient shall only use the Data for academic purposes and not make any commercial use of the Data in whatsoever form. The Data Recipient shall indicate the origin of the Data (i.e. Deutsche Börse AG, Mergenthalerallee 61, 65760 Eschorn) in any publication resulting from or in connection with the Project by a proof of source and shall provide Data Supplier with a voucher copy by e-mail (pdf) or mail.

2. Data Recipient shall not disclose or reveal any Data to any third party without the prior written consent of the Data Supplier.

3. Data Recipient shall use all reasonable efforts to keep the Data in confidence and to safeguard the Data. In so doing, the Data Recipient shall take at least the same precautions which it would take to safeguard its own similarly valued proprietary and confidential data, but shall in no event take less than commercially reasonable precautions.

4. Data Recipient acknowledges that Data Supplier does not make any representation, warranty or undertaking, express or implied, as to the accuracy, reliability, completeness or reasonableness of the Data.

5. Except as set forth in Section 1, this Agreement does not transfer to the Data Recipient or any other person any ownership, license, intellectual property rights, or other rights including, but not limited to, patent and copyright rights, in or to the Data. All Data shall remain the exclusive property of the Data Supplier or the third party which owns it.

6. The Data Recipient shall be allowed to access and use the Data for research purposes within the scope of the Project for a period of 18 months. **After the period of 18 months the Data Recipient shall promptly permanently destroy, erase or delete all Data in its possession or control**, in particular, but not limited to, from any computer, word processor, mobile telecommunication device or similar device into which it was stored or programmed, and provide to the Data Supplier written confirmation of such destruction, erasure or deletion.

7. No provision of this Agreement creates a partnership between the Parties or makes a Party the agent of the other Party for any purpose.

8. This Agreement, including this Section 8, may only be supplemented, amended, varied or modified in writing. In case any provision in or obligation under this Agreement shall be held invalid, illegal or unenforceable, the validity, legality and enforceability of the remaining provisions or obligations of this Agreement shall not be affected. The invalid, illegal or

unenforceable provision shall be replaced by a valid, legal and enforceable provision which reflects as close as possible the original intention of the Parties. The foregoing shall apply mutatis mutandis in case of any gaps of this Agreement.

9. This Agreement, and any non-contractual obligations arising out of or in connection with it, shall be governed by, construed and enforced in accordance with the laws of the Netherlands without reference to or inclusion of the principles of choice of law or conflicts of law of that jurisdiction and the Parties hereby submit to the jurisdiction of the courts of Amsterdam.

Figure OA.9: Non-disclosure agreement peer evaluators

width=!,height=!,pages=-

Figure OA.10: Registration as a team member

registration as a team member

To register as a research team member, please fill out the form below. Fields indicated with an asterisk are required. For your contact information, please make sure to provide the details in the way they should be added to the final paper.

first name: *

first name

middle name(s):

middle name(s)

last name: *

last name

e-mail address: *

e-mail address

webpage:

webpage url (incl. 'http://' or 'https://' prefix)

your ORCID:

if available (see https://orcid.org/)

full affiliation: *

department, organization, country

years since PhD: *

please enter 0 if you do not have obtained your PhD yet

gender: *

please select...

country of residence: *

country you currently work in

highest degree: *

please select...

Other:

highest degree: if other, please specify field of education: *

please select...

Other:

field of education: if other, please specify

current position: *

please select...

Other:

current position: if other, please specify

What is your number of Google Scholar citations? (If you don't have a Google Scholar profile, please provide an approximate number of citations.) *

number of citations

Have you published (or had a paper accepted for publication) in at least one of the following journals: Journal of Finance, Journal of Financial Economics, Review of Financial Studies, American Economic Review, Econometrica, Journal of Political Economy, Review of Economic Studies, or Quarterly Journal of Economics? *

please select...

What is the largest dataset you have worked with so far (in terms of observations)? *

please select...

Do you consider the data to be analyzed in this project (652,000,000 observations) "big data"? *

please select...

expertise in empirical finance: *

please select...

expertise in market liquidity: *

please select...

any other comment:

anything you would like to tell us?

Figure OA.11: Instruction sheet for research teams

instruction sheet for research teams

This three-page instruction sheet clarifies what is expected of you as a research team in the #fincap project. It first provides some context for the hypotheses you are expected to test, then presents the assignment, and finally lists the hypotheses you are asked to test with only the Deutsche Börse data that is made available to you by the #fincap team. These data contain trade information on the EuroStoxx 50 futures.

Context

Electronic order matching systems (automated exchanges) and electronic order generation systems (algorithms) have changed financial markets over time. Investors used to trade through broker-dealers by paying the dealers' quoted ask prices when buying, and accepting their bid prices when selling. The wedge between dealer bid and ask prices, the bid-ask spread, was a useful measure of trading cost, and often still is.

Now, investors more commonly trade in electronic limit-order markets (as is the case for EuroStoxx 50 futures). They still trade at bid and ask prices. They do so by submitting so-called market orders and marketable limit orders. However, investors now also can quote bid and ask prices themselves by submitting (non-marketable) standing limit orders. Increasingly, investors now also use agency algorithms to automate their trades. Concurrently, exchanges have been continuously upgrading their systems to better serve their clients. Has market quality improved, in particular when taking the viewpoint of non-exchange members: (end-user) clients?

Assignment

You are expected to write an academic paper that is maximum five pages long. To make that feasible you can skip many parts of a typical academic paper. You only need to do the following for all hypotheses listed below:

1. Propose a statistical measure, briefly motivate it, and present the formula to calculate it.
2. For this measure, estimate the average per-year change in percentage terms, based on the full sample (or at least the longest possible period because some series are not available yet at the beginning of the sample). Test it against the null of no change.
3. Report this estimate along with its standard error in four decimals (e.g., "measure Z declined by 1.251% with a standard error 0.241%").
4. Briefly discuss your result.

For example, an appropriate outcome statement for testing hypothesis X which states that Y has not changed is:

> "We propose measure Z to test hypothesis X because [...]. It is calculated as Z = f(DATA). Implementing it leads to the following result: We reject the null of no change. We find that Y declined as our measure Z declined by 1.251% on average per year where the standard error of this change is 0.421% and the resulting t-statistic is 2.971. This result shows [...]"

We emphasize that you are asked to report your results in a brief manner. *If the paper is longer than five pages we will not consider the paper and we will have to exclude you as co-authors from the project.*

Hypotheses

1. Assuming that informationally-efficient prices follow a random walk, did market efficiency change over time?

   Null hypothesis 1: Market efficiency has not changed over time.

2. Did the (realized) bid-ask spread paid on market orders change over time? The realized spread could be thought of as the gross-profit component of the spread as earned by the limit-order submitter.

   Null hypothesis 2: The realized spread on market orders has not changed over time.

The remaining hypotheses focus on client trades only (i.e., trades implemented by exchange members on behalf of their clients).

3. Did the share of client volume in total volume change over time?

   Null hypothesis 3: Client share volume as a fraction of total volume has not changed over time.

4. On their market orders and marketable limit orders, did the realized bid-ask spread that clients paid, change over time?

   Null hypothesis 4: Client realized spreads have not changed over time.

5. Realized spread is a standard cost measure for market orders, but to what extent do investors continue to use market and marketable limit orders (as opposed to non-marketable limit orders)?

Null hypothesis 5: The fraction of client trades executed via market orders and marketable limit orders has not changed over time.

6. A measure that does not rely on the classic limit- or market-order distinction is *gross trading revenue* (GTR). Investor GTR for a particular trading day can be computed by assuming a zero position at the start of the day and evaluating an end-of-day position at an appropriate reference price. Relative investor GTR can then be defined as this GTR divided by the investor's total (euro) volume for that trading day. This relative GTR is, in a sense, a realized spread. It reveals what various groups of market participants pay in aggregate for (or earn on) their trading. It transcends market structure as it can be meaningfully computed for any type of trading in any type of market (be it trading through limit-orders only, through market-orders only, through a mix of both, or in a completely different market structure).

   Null hypothesis 6: Relative gross trading revenue (GTR) for clients has not changed over time.

Figure OA.12: Form to submit stage 1 results

submit results for stage 1

Please read the following instructions for submitting your results carefully.

- Please enter your team ID **before** uploading your short paper and/or your analysis scripts. You will find your team ID in the e-mail notifications sent by the project collaborators. Each research team is supposed to submit their results **only once**. Please coordinate with your team member to avoid double submissions.
- Please upload your short paper in **.PDF format**. Make sure that the file does not exceed the page limit of **five pages**. Note that if your paper exceeds the page limit, we will not consider it and we will have to exclude you as co-authors from the project. **Click on the "upload file" button on the right to submit the PDF.**
- To ensure anonymity throughout the project, please make sure that **no identifying information** is contained in any of the submitted files. Please note that we will lift your anonymity only for the peer evaluation process (see peer evaluation for details). That is, neither the short paper nor the analysis scripts should allow drawing any inferences on the identity of the team members.
- Please bundle all analysis and data processing scripts in a single .ZIP, .RAR, or .7ZIP file. **Please upload only the script files, not the data.** Please also include a "readme" file (in .TXT format) with precise instructions on how to reproduce your estimates from the raw data (e.g., which files to put in what folder, the order in which to run the scripts, etc.). From experience, the best guidance we can give you is to literally follow your own instructions to see if the workflow you describe yields the desired results. **Click on the "upload file" button on the right to submit the file.**
- For each of the six null hypotheses, please provide the average per-year change in percentage terms (effect size) and the corresponding standard error. Please report all estimates with **three digits after the comma** (e.g., effect size = 1.234%, standard error = 0.567%).

team ID: *

ABCD

team ID not yet provided

short paper: *

browse | | upload file

analysis scripts: *

browse | | upload file

upload status:

Please note that your results can be submitted only if all three bullet point below are indicated with a check mark. To do so, please provide your correct team ID first, select the two files AND press the "upload file" buttons on the right (for both files separately).

- *team ID not yet provided*
- *short paper not yet uploaded*
- *analysis scripts not yet uploaded*

null hypothesis 1 (H1): Market efficiency has not changed over time.

H1: effect size: *

%

H1: standard error: *

%

null hypothesis 2 (H2): The realized spread on market orders has not changed over time.

H2: effect size: *

%

H2: standard error: *

%

null hypothesis 3 (H3): Client share volume as a fraction of total volume has not changed over time.

H3: effect size: *

%

H3: standard error: *

%

null hypothesis 4 (H4): Client realized spreads have not changed over time.

H4: effect size: *

%

H4: standard error: *

%

null hypothesis 5 (H5): The fraction of client trades executed via market orders and marketable limit orders has not changed over time.

H5: effect size: *

%

H5: standard error: *

%

null hypothesis 6 (H6): Relative gross trading revenue (GTR) for clients has not changed over time.

H6: effect size: *

%

H6: standard error: *

%

Figure OA.13: Form to submit stage 2 results

submit results for stage 2

Please read the following instructions for submitting your results carefully.

- Please enter your team ID **before** uploading your short paper and/or your analysis scripts. You will find your team ID in the e-mail notifications sent by the project collaborators. Each research team is supposed to submit their results **only once**. Please coordinate with your team member to avoid double submissions.
- Please upload your revised short paper in **.PDF format**. Make sure that the file does not exceed the page limit of **five pages**. **Click on the "upload file" button on the right to submit the PDF.**
- To ensure anonymity throughout the project, please make sure that **no identifying information** is contained in any of the submitted files. That is, neither the short paper nor the analysis scripts should allow drawing any inferences on the identity of the team members.
- Please bundle all analysis and data processing scripts in a single .ZIP, .RAR, or .7ZIP file. **Please upload only the script files, not the data.** Please also include a "readme" file (in .TXT format) with precise instructions on how to reproduce your estimates from the raw data (e.g., which files to put in what folder, the order in which to run the scripts, etc.). From experience, the best guidance we can give you is to literally follow your own instructions to see if the workflow you describe yields the desired results. **Click on the "upload file" button on the right to submit the file.**
- For each of the six null hypotheses, please provide the **average per-year change in percentage terms (the annualized effect size)** and the corresponding **standard error**. Please report all estimates with **at least three digits after the decimal point**. In cases where this would yield 0.000 or 0.00x (e.g. 0.003), please report your result with at least two more digits after the first non-zero digit (e.g., effect size = 0.000123%, standard error = 0.00345%). To aggregate the data, it is important for us to correctly align the signs of the reported effect sizes. We thus ask you to also **indicate the direction of your results** with respect to the corresponding standardized statements in the form.
- Please be aware that you can leave comments and clarifications on the revision in the open text field at the bottom of the form.

team ID: *

ABCD

team ID not yet provided

short paper: *

browse | upload file

analysis scripts: *

browse | upload file

upload status:

Please note that your results can be submitted only if all three bullet points below are indicated with a check mark. To do so, please provide your correct team ID first, select the two files AND press the "upload file" buttons on the right (for both files separately).

- ⊗ *team ID not yet provided*
- ⊗ *short paper not yet uploaded*
- ⊗ *analysis scripts not yet uploaded*

null hypothesis 1 (H1): Market efficiency has not changed over time.

H1: effect size: *

%

H1: standard error: *

%

H1: Direction of the effect: *

---

null hypothesis 2 (H2): The realized spread on market orders has not changed over time.

H2: effect size: *

%

H2: standard error: *

%

H2: Direction of the effect: *

---

null hypothesis 3 (H3): Client share volume as a fraction of total volume has not changed over time.

H3: effect size: *

%

H3: standard error: *

%

H3: Direction of the effect: *

---

null hypothesis 4 (H4): Client realized spreads have not changed over time.

H4: effect size: *

%

H4: standard error: *

%

H4: Direction of the effect: *

---

null hypothesis 5 (H5): The fraction of client trades executed via market orders and marketable limit orders has not changed over time.

H5: effect size: *

%

H5: standard error: *

%

H5: Direction of the effect: *

---

null hypothesis 6 (H6): Relative gross trading revenue (GTR) for clients has not changed over time.

H6: effect size: *

%

H6: standard error: *

%

H6: Direction of the effect: *

Comments and clarifications on the revision:

Figure OA.14: Form to submit stage 3 results

submit results for stage 3

Please read the following instructions for submitting your results carefully.

- Please enter your team ID **before** uploading your short paper and/or your analysis scripts. You will find your team ID in the e-mail notifications sent by the project collaborators. Each research team is supposed to submit their results **only once**. Please coordinate with your team member to avoid double submissions.
- Please upload your revised short paper in **.PDF format**. Make sure that the file does not exceed the page limit of **five pages**. **Click on the "upload file" button on the right to submit the PDF.**
- To ensure anonymity throughout the project, please make sure that **no identifying information** is contained in any of the submitted files. That is, neither the short paper nor the analysis scripts should allow drawing any inferences on the identity of the team members.
- Please bundle all analysis and data processing scripts in a single .ZIP, .RAR, or .7ZIP file. **Please upload only the script files, not the data.** Please also include a "readme" file (in .TXT format) with precise instructions on how to reproduce your estimates from the raw data (e.g., which files to put in what folder, the order in which to run the scripts, etc.). From experience, the best guidance we can give you is to literally follow your own instructions to see if the workflow you describe yields the desired results. **Click on the "upload file" button on the right to submit the file.**
- For each of the six null hypotheses, please provide the **average per-year change in percentage terms (the annualized effect size)** and the corresponding **standard error**. Please report all estimates with **at least three digits after the decimal point**. In cases where this would yield 0.000 or 0.00x (e.g. 0.003), please report your result with at least two more digits after the first non-zero digit (e.g., effect size = 0.000123%, standard error = 0.00345%). To aggregate the data, it is important for us to correctly align the signs of the reported effect sizes. We thus ask you to also **indicate the direction of your results** with respect to the corresponding standardized statements in the form.
- Please be aware that you can leave comments and clarifications on the revision in the open text field at the bottom of the form.

team ID: *

ABCD

team ID not yet provided

short paper: *

browse | upload file

analysis scripts: *

browse | upload file

upload status:

Please note that your results can be submitted only if all three bullet points below are indicated with a check mark. To do so, please provide your correct team ID first, select the two files AND press the "upload file" buttons on the right (for both files separately).

- ⊗ *team ID not yet provided*
- ⊗ *short paper not yet uploaded*
- ⊗ *analysis scripts not yet uploaded*

null hypothesis 1 (H1): Market efficiency has not changed over time.

H1: effect size: *

%

H1: standard error: *

%

H1: Direction of the effect: *

---

null hypothesis 2 (H2): The realized spread on market orders has not changed over time.

H2: effect size: *

%

H2: standard error: *

%

H2: Direction of the effect: *

---

null hypothesis 3 (H3): Client share volume as a fraction of total volume has not changed over time.

H3: effect size: *

%

H3: standard error: *

%

H3: Direction of the effect: *

---

null hypothesis 4 (H4): Client realized spreads have not changed over time.

H4: effect size: *

%

H4: standard error: *

%

H4: Direction of the effect: *

---

null hypothesis 5 (H5): The fraction of client trades executed via market orders and marketable limit orders has not changed over time.

H5: effect size: *

%

H5: standard error: *

%

H5: Direction of the effect: *

---

null hypothesis 6 (H6): Relative gross trading revenue (GTR) for clients has not changed over time.

H6: effect size: *

%

H6: standard error: *

%

H6: Direction of the effect: *

Comments and clarifications on the revision:

Figure OA.15: Belief elicitation sheet

belief elicitation

Please read the following instructions for submitting your beliefs carefully.

- Please enter your team ID. You will find your team ID in the e-mail notifications sent by the project collaborators. Each research team is supposed to submit their results **only once**. Please coordinate with your team member to avoid double submissions.

- Task:

  Imagine that, at the end of the project, you receive the short papers of all teams who completed all stages in the project (more than 100 research teams registered for #fincap and were given access to the data on January 11). **What do you predict the dispersion in results to be across these teams in Stage 1?**

  More specifically, for each hypothesis, we ask you to predict...

  1. the **standard deviation of reported effect size estimates** across teams in Stage 1 (i.e., imagine that you collect the effect size estimates y_i of all teams in a single data series; then, what is the standard deviation of this series?)
  2. the **standard deviation of t-statistics** across teams in Stage 1 (the t-statistic is defined as the reported effect size estimate divided by the reported standard error).

  Recall that estimates of the **effect sizes were in terms of per-year percentage changes**. Thus, your prediction of the standard deviation across teams should be in the same units. Likewise, the prediction for the standard deviation in t-statistics should be provided in the equivalent units (i.e., in t units).

  In total you will make 12 predictions (the above two questions on each of the 6 hypotheses).

- Monetary rewards:

  One out of five (20%) RTs that complete all stages in the project will be randomly drawn for a monetary reward. If your RT is drawn for a monetary reward, **one out of the 12 prediction questions** will be randomly drawn to determine the monetary reward. If your prediction about the standard deviation in results across the teams is within +/- 50% of the observed value you will receive a **monetary reward of $300** and otherwise you will receive a reward of $0. If, for example, the observed standard deviation is X in the prediction question randomly drawn for payment, you will be paid a $300 reward if your predicted standard deviation is between 0.5X and 1.5X and you will get paid zero otherwise.

  The observed standard deviation will be computed based on the Stage 1 results of those RTs who completed all stages in the project. The monetary reward will be paid out as an Amazon gift card (if you are two persons in the RT, the reward will be split equally between you).

team ID: *

ABCD

team ID not yet provided

team ID not yet provided

Null hypothesis 1 (H1): Market efficiency has not changed over time.

H1: Standard deviation of effect size estimates: *

%

H1: Standard deviation of t-statistics: *

"t"

Null hypothesis 2 (H2): The realized spread on market orders has not changed over time.

H2: Standard deviation of effect size estimates: *

%

H2: Standard deviation of t-statistics: *

"t"

Null hypothesis 3 (H3): Client share volume as a fraction of total volume has not changed over time.

H3: Standard deviation of effect size estimates: *

%

H3: Standard deviation of t-statistics: *

"t"

Null hypothesis 4 (H4): Client realized spreads have not changed over time.

H4: Standard deviation of effect size estimates: *

%

H4: Standard deviation of t-statistics: *

"t"

Null hypothesis 5 (H5): The fraction of client trades executed via market orders and marketable limit orders has not changed over time.

H5: Standard deviation of effect size estimates: *

%

H5: Standard deviation of t-statistics: *

"t"

Null hypothesis 6 (H6): Relative gross trading revenue (GTR) for clients has not changed over time.

H6: Standard deviation of effect size estimates: *

%

H6: Standard deviation of t-statistics: *

"t"

Figure OA.16: Instruction sheet peer evaluators

instruction sheet for peer evaluators

For each hypothesis and for the short paper, please rate the quality of the analysis, briefly clarify and provide suggestions for possible improvement (a paragraph is sufficient). For your feedback please follow the guidelines of the Journal of Finance: "(P)lease focus on what you see as central weaknesses in terms of (...) the chosen econometric strategy (...). (Y)ou should provide clear and concise reasons why you see your proposed revision as material (...)."

Please read the following information carefully:

- the instruction sheet that was given to the research teams (RT), click [here], and
- the FAQs that the RTs were alerted to, click [here].

Importantly, RTs could use any frequency of the data for their analyses (e.g., daily data, monthly data, or annual data), but they were asked to report effect sizes in per-year/annualized terms (to make them comparable across teams).

The following list contains detailed instructions:

- At the end of this e-mail, please find the list of links to the 10 short papers we kindly ask you to evaluate. The papers are named after the team IDs (e.g., ``ABCD.pdf´´).
- As outlined in the project description, you as a peer evaluator will remain anonymous to the research teams, but not the other way around. For each paper, there are two anonymous evaluators. The list below also contains, for each of your papers, the team ID, the name(s) of the team members, and your blinded evaluator ID (1 or 2).
- For your evaluations, please use the following file as a template: Evaluation Template.
- For each RT, please open a new version of the provided template.
- On the first page, enter the team ID and your evaluator ID for the paper as outlined in the attached list.
- Please use the designated cells to provide your ratings and your suggestions.
- Save the file as "TeamID_EvaluatorID.xlsx" according to the attached list (e.g., ``ABCD_1.xlsx´´).
- Please double-check that you have entered all six hypothesis ratings as well as the overall rating for each paper.
- Please send all 10 files to info@fincap.academy until before Friday, May 7, 2021, 23:59 GMT-12.

In case you have any questions, please do not reach out to the research teams or other potential peer evaluators. We are happy to answer any questions you might have on info@fincap.academy.

Thank you very much for your time and for your contribution to the #fincap project.

Figure OA.17: The form that RTs completed for multiverse

#fincap RT decisions overview

This form is used to take stock of the key decisions that research teams (RTs) made in the FIRST STAGE of the #fincap project.

* Required

1. Please provide the four-letter code of your research team. *

2. Who is/are the member(s) of the research team? Please report first and last name. *

Stage-1 decisions for all hypotheses

The following five questions refer to decisions that affect all hypotheses. Please tick six boxes, one for each hypothesis. Please tick the alternative that you picked for your Stage-1 result (i.e., the result that you filled in on the reporting sheet).

3. QUESTION 1: Did you remove parts of the sample to get rid of idiosyncrasies associated with the open and close of trading? PLEASE TICK SIX BOXES, ONE FOR EACH HYPOTHESIS! *

Mark only one oval per row.

	No.	Yes, we removed the first and the last 30 minutes of trading.
H1	◯	◯
H2	◯	◯
H3	◯	◯
H4	◯	◯
H5	◯	◯
H6	◯	◯

4. How close were the alternatives that you picked in the previous question (Q1), to what you actually did? *

Mark only one oval.

	1	2	3	4	5	
Far from what we did.	◯	◯	◯	◯	◯	Very close to what we did.

5. Comment (optional)

6. QUESTION 2: Did you exclude periods from the sample? (Please tick six boxes, one for each hypothesis!) *

Mark only one oval per row.

	No.	Yes, settlement weeks were excluded.
H1	◯	◯
H2	◯	◯
H3	◯	◯
H4	◯	◯
H5	◯	◯
H6	◯	◯

7. How close were the alternatives that you picked in the previous question (Q2), to what you actually did? *

Mark only one oval.

	1	2	3	4	5	
Far from what we did.	◯	◯	◯	◯	◯	Very close to what we did.

8. Comment (optional)

9. QUESTION 3: Did you deal with potential outliers in your final sample (i.e., the sample you used for your final analysis in Stage 1)? (Please tick six boxes, one for each hypothesis!) *

Mark only one oval per row.

	No.	Yes, we winsorized at 2.5 and 97.5 percentile.	Yes, we trimmed at 2.5 and 97.5 percentile (trimming=removing).
H1	◯	◯	◯
H2	◯	◯	◯
H3	◯	◯	◯
H4	◯	◯	◯
H5	◯	◯	◯
H6	◯	◯	◯

10. How close were the alternatives that you picked in the previous question (Q3), to what you actually did? *

Mark only one oval.

	1	2	3	4	5	
Far from what we did.	◯	◯	◯	◯	◯	Very close to what we did.

11. Comment (optional)

12. QUESTION 4: What is the frequency of your final sample (i.e., at what frequency did you conduct your final analysis in Stage 1)? (Please tick six boxes, one for each hypothesis!) *

Mark only one oval per row.

	Daily frequency.	Weekly frequency.	Monthly frequency.	Annual frequency.
H1	◯	◯	◯	◯
H2	◯	◯	◯	◯
H3	◯	◯	◯	◯
H4	◯	◯	◯	◯
H5	◯	◯	◯	◯
H6	◯	◯	◯	◯

13. How close were the alternatives that you picked in the previous question (Q4), to what you actually did? *

Mark only one oval.

	1	2	3	4	5	
Far from what we did.	◯	◯	◯	◯	◯	Very close to what we did.

14. Comment (optional)

15. QUESTION 5: What is the statistical model/method that you used to estimate change? (Please tick six boxes, one for each hypothesis!) *

Mark only one oval per row.

	Assume trend stationarity for the log measure, and estimate a linear model with trend.	"Log difference" the measure, and estimate its mean.	Compute percentage change of the measure, and estimate its mean.
H1	◯	◯	◯
H2	◯	◯	◯
H3	◯	◯	◯
H4	◯	◯	◯
H5	◯	◯	◯
H6	◯	◯	◯

16. How close were the alternatives that you picked in the previous question (Q5), to what you actually did? *

Mark only one oval.

	1	2	3	4	5	
Far from what we did.	◯	◯	◯	◯	◯	Very close to what we did.

17. Comment (optional)

Stage-1 hypothesis-specific decisions

The remaining questions pertain to decisions for a subset of hypotheses. Please tick one box per hypothesis.

18. QUESTION 6: For H1, which measure did you use for market efficiency? *

Mark only one oval.

◯ R-squared of an autoregressive model for returns.
◯ Variance ratio for returns.

19. How close were the alternatives that you picked in the previous question (Q6), to what you actually did? *

Mark only one oval.

	1	2	3	4	5	
Far from what we did.	◯	◯	◯	◯	◯	Very close to what we did.

20. Comment (optional)

21. QUESTION 7: For H1, which frequencies did you use for computing your market efficiency statistic? (If you used the R-squared statistic, then the highest frequency is the frequency at which you ran the regression, and the lowest frequency drives the number of lags, e.g., second and minute frequencies imply you sampled in seconds and used 60 lags.) *

Mark only one oval.

◯ Highest frequency second; lowest frequency minute.
◯ Highest frequency minute; lowest frequency 5-minutes.
◯ Highest frequency 5-minutes; lowest frequency 30-minutes.
◯ Highest frequency day; lowest frequency week.
◯ Highest frequency day; lowest frequency month.

22. How close were the alternatives that you picked in the previous question (Q7), to what you actually did? *

Mark only one oval.

	1	2	3	4	5	
Far from what we did.	◯	◯	◯	◯	◯	Very close to what we did.

23. Comment (optional)

24. QUESTION 8: For H2, H4, and H5, did you use the "tick test" (e.g., Lee and Ready, 1991) to determine the aggressor side of the trade, or did you use the aggressor flag and, therefore, accepted that you can only use part of the sample? (Lee and Ready, 1991, "Inferring Trade Direction from Intraday Data") *

Mark only one oval per row.

	We used the tick test.	We used the aggressor flag.
H2	◯	◯
H4	◯	◯
H5	◯	◯

25. How close were the alternatives that you picked in the previous question (Q8), to what you actually did? *

Mark only one oval.

	1	2	3	4	5	
Far from what we did.	◯	◯	◯	◯	◯	Very close to what we did.

26. Comment (optional)

27. QUESTION 9: For H2 and H4, which post-trade reference price did you use to compute the realized spread? *

Mark only one oval per row.

	Price 5 minutes after the trade.	Price 10 minutes after the trade.	Price 30 minutes after the trade.
H2	⬭	⬭	⬭
H4	⬭	⬭	⬭

28. How close were the alternatives that you picked in the previous question (Q9), to what you actually did? *

Mark only one oval.

	1	2	3	4	5	
Far from what we did.	⬭	⬭	⬭	⬭	⬭	Very close to what we did.

29. Comment (optional)

30. QUESTION 10: For H2 and H4, how did you aggregate across trades? *

Mark only one oval per row.

	Trade-size-weighted average.	Equal-weighted average.
H2	⬭	⬭
H4	⬭	⬭

31. How close were the alternatives that you picked in the previous question (Q10), to what you actually did? *

Mark only one oval.

	1	2	3	4	5	
Far from what we did.	⬭	⬭	⬭	⬭	⬭	Very close to what we did.

32. Comment (optional)

33. QUESTION 11: For H3, which units did you use to compute volume shares? *

Mark only one oval.

- ⬭ Volume expressed in number of contracts.
- ⬭ Volume expressed in euro.

34. How close were the alternatives that you picked in the previous question (Q11), to what you actually did? *

Mark only one oval.

	1	2	3	4	5	
Far from what we did.	⬭	⬭	⬭	⬭	⬭	Very close to what we did.

35. Comment (optional)

36. QUESTION 12: For H6, which reference price did you pick when computing "gross trading revenue" (GTR)? *

Mark only one oval.

- ⬭ Last trade price in the day.
- ⬭ Last trade price one day later.
- ⬭ Volume-weighted average-price (VWAP) full-day.
- ⬭ Volume-weighted average-price (VWAP) based on last five trades in the day.

37. How close were the alternatives that you picked in the previous question (Q12), to what you actually did? *

Mark only one oval.

	1	2	3	4	5	
Far from what we did.	⬭	⬭	⬭	⬭	⬭	Very close to what we did.

38. Comment (optional)

39. QUESTION 13: For H6, to attenuate any complications arising from extreme outliers for daily GTRs, you might have aggregated using medians instead of means. Which aggregation approach did you use? *

Mark only one oval.

- ⬭ Mean.
- ⬭ Median.

40. How close were the alternatives that you picked in the previous question (Q13), to what you actually did? *

Mark only one oval.

	1	2	3	4	5	
Far from what we did.	⬭	⬭	⬭	⬭	⬭	Very close to what we did.

41. Comment (optional)

42. QUESTION 14: For H6, how did you deal with non-negative values for daily GTRs? (The idea is that client GTRs are expected to be negative to make intermediary GTRs positive, because they sum to zero. To make a series of mostly negative numbers, strictly negative, one option is to subtract the maximum of the series, and a little more to avoid one observation being zero. Another option is to set non-negative values to missing values.) *

Mark only one oval.

- If needed, we created a non-negative time series by applying f(X) = f(X) - max[f(X)] - 0.001.
- If needed, we created a non-negative time series by applying f(X) = f(X) - max[f(X)] - 1.
- If needed, we set non-negative values to missing values.

43. How close were the alternatives that you picked in the previous question (Q14), to what you actually did? *

Mark only one oval.

	1	2	3	4	5	
Far from what we did.	◯	◯	◯	◯	◯	Very close to what we did.

44. Comment (optional)

45. QUESTION 15: For H6, when reporting average per-year change in percentages, did you "retain the sign" of a negative trend? (For example, did you report a change from -1 to -2 as a negative number?) *

Mark only one oval.

- Yes, we did retain the sign (in the example we would have reported a negative number).
- No, we did not retain the sign (in the example we would have reported a positive number).

46. How close were the alternatives that you picked in the previous question (Q15), to what you actually did? *

Mark only one oval.

	1	2	3	4	5	
Far from what we did.	◯	◯	◯	◯	◯	Very close to what we did.

47. Comment (optional)

Annex 1

Non-Standard Errors*

Albert J. Menkveld[200,121], *Anna Dreber*[113], *Felix Holzmeister*[149], *Juergen Huber*[149], *Magnus Johannesson*[113], *Michael Kirchler*[149], *Sebastian Neusüss*[1], *Michael Razen*[149], *Utz Weitzel*[200,100,121], David Abad-Díaz[133], Menachem Abudy[12], Tobias Adrian[58], Yacine Ait-Sahalia[95], Olivier Akmansoy[17,21], Jamie T. Alcock[169], Vitali Alexeev[179], Arash Aloosh[81], Livia Amato[139], Diego Amaya[204], James J. Angel[44], Alejandro T. Avetikian[93], Amadeus Bach[154], Edwin Baidoo[118], Gaetan Bakalli[6], Li Bao[122], Andrea Barbon[174], Oksana Bashchenko[104], Parampreet C. Bindra[149], Geir H. Bjønnes[8], Jeffrey R. Black[158], Bernard S. Black[85], Dimitar Bogoev[32], Santiago Bohorquez Correa[129], Oleg Bondarenko[147], Charles S. Bos[200], Ciril Bosch-Rosa[116], Elie Bouri[69], Christian Brownlees[130], Anna Calamia[115], Viet Nga Cao[78], Gunther Capelle-Blancard[131], Laura M. Capera Romero[200], Massimiliano Caporin[170], Allen Carrion[158], Tolga Caskurlu[134], Bidisha Chakrabarty[107], Jian Chen[97], Mikhail Chernov[124], William Cheung[202], Ludwig B. Chincarini[172], Tarun Chordia[37], Sheung Chi Chow[7], Benjamin Clapham[46], Jean-Edouard Colliard[49], Carole Comerton-Forde[157], Edward Curran[74], Thong Dao[87], Wale Dare[47], Ryan J. Davies[9], Riccardo De Blasis[67], Gianluca F. De Nard[193], Fany Declerck[122], Oleg Deev[75], Hans Degryse[61], Solomon Y. Deku[87], Christophe Desagre[125], Mathijs A. van Dijk[38], Chukwuma Dim[43], Thomas Dimpfl[146], Yun Jiang Dong[97], Philip A. Drummond[78], Tom Dudda[117], Teodor Duevski[49], Ariadna Dumitrescu[36], Teodor Dyakov[33], Anne Haubo Dyhrberg[178], Michał Dzieliński[114], Asli Eksi[108], Izidin El Kalak[20], Saskia ter Ellen[22], Nicolas Eugster[171], Martin D. D. Evans[44], Michael Farrell[189], Ester Felez-Vinas[179], Gerardo Ferrara[11], El Mehdi Ferrouhi[56], Andrea Flori[92], Jonathan T. Fluharty[203], Sean D. V. Foley[74], Kingsley Y. L. Fong[165], Thierry Foucault[49], Tatiana Franus[14], Francesco Franzoni[195], Bart Frijns[89], Michael Frömmel[45], Servanna M. Fu[141], Sascha C. Füllbrunn[100], Baoqing Gan[179], Ge Gao[135], Thomas P. Gehrig[188], Roland Gemayel[65], Dirk Gerritsen[197], Javier Gil-Bazo[130,13], Dudley Gilder[20], Lawrence R. Glosten[25], Thomas Gomez[197], Arseny Gorbenko[78], Joachim Grammig[184], Vincent Grégoire[48], Ufuk Güçbilmez[143], Björn

Hagströmer[114], Julien Hambuckers[47], Erik Hapnes[1], Jeffrey H. Harris[3], Lawrence Harris[173], Simon Hartmann[199], Jean-Baptiste Hasse[2], Nikolaus Hautsch[188], Xue-Zhong (Tony) He[205], Davidson Heath[185], Simon Hediger[193], Terrence Hendershott[123], Ann Marie Hibbert[203], Erik Hjalmarsson[144], Seth Hoelscher[77], Peter Hoffmann[39], Craig W. Holden[57], Alex R. Horenstein[159], Wenqian Huang[10], Da Huang[185], Christophe Hurlin[168,21], Konrad Ilczuk[1], Alexey Ivashchenko[200], Subramanian R. Iyer[164], Hossein Jahanshahloo[20], Naji P. Jalkh[106], Charles M. Jones[24], Simon Jurkatis[11], Petri Jylhä[1], Andreas T. Kaeck[177], Gabriel Kaiser[152], Arzé Karam[30], Egle Karmaziene[200], Bernhard Kassner[162], Markku Kaustia[1], Ekaterina Kazak[153], Fearghal Kearney[98], Vincent van Kervel[94], Saad A. Khan[48], Marta K. Khomyn[179], Tony Klein[98], Olga Klein[190], Alexander Klos[63], Michael Koetter[50], Aleksey Kolokolov[153], Robert A. Korajczyk[85], Roman Kozhan[190], Jan P. Krahnen[46], Paul Kuhle[127], Amy Kwan[178], Quentin Lajaunie[91], F. Y. Eric C. Lam[51], Marie Lambert[47], Hugues Langlois[49], Jens Lausen[46], Tobias Lauter[71], Markus Leippold[193], Vladimir Levin[152], Yijie Li[109], Hui Li[68], Chee Yoong Liew[126], Thomas Lindner[201], Oliver Linton[138], Jiacheng Liu[96], Anqi Liu[178], Guillermo Llorente[127], Matthijs Lof[1], Ariel Lohr[4], Francis Longstaff[124], Alejandro Lopez-Lira[142], Shawn Mankad[27], Nicola Mano[102], Alexis Marchal[35], Charles Martineau[182], Francesco Mazzola[38], Debrah Meloso[115], Michael G. Mi[178], Roxana Mihet[101], Vijay Mohan[99], Sophie Moinas[122], David Moore[72], Liangyi Mu[120], Dmitriy Muravyev[76], Dermot Murphy[147], Gabor Neszveda[59], Christian Neumeier[60], Ulf Nielsson[26], Mahendrarajah Nimalendran[142], Sven Nolte[100], Lars L. Norden[114], Peter W. O'Neill[42], Khaled Obaid[18], Bernt A. Ødegaard[175], Per Östberg[193], Emiliano Pagnotta[110], Marcus Painter[107], Stefan Palan[145], Imon J. Palit[99], Andreas Park[183], Roberto Pascual[194], Paolo Pasquariello[160], Lubos Pastor[139], Vinay Patel[179], Andrew J. Patton[29], Neil D. Pearson[148,19], Loriana Pelizzon[46], Michele Pelli[105], Matthias Pelster[90], Christophe Pérignon[49,21], Cameron Pfiffer[167], Richard Philip[178], Tomáš Plíhal[75], Puneet Prakash[77], Oliver-Alexander Press[26], Tina Prodromou[192], Marcel Prokopczuk[71], Talis Putnins[179], Ya Qian[1], Gaurav Raizada[53], David Rakowski[180], Angelo Ranaldo[174], Luca Regis[181], Stefan Reitz[64], Thomas Renault[196], Rex W. Renjie[200], Roberto Reno[186], Steven J. Riddiough[182], Kalle Rinne[152], Paul J. Rintamäki[1], Ryan Riordan[97], Thomas Rittmannsberger[149], Iñaki Rodríguez Longarela[114], Dominik Roesch[112], Lavinia Rognone[153], Brian Roseman[88], Ioanid Rosu[49], Saurabh Roy[156], Nicolas Rudolf[151], Stephen R. Rush[15], Khaladdin Rzayev[140,66], Aleksandra A. Rzeźnik[206], Anthony Sanford[155], Harikumar Sankaran[82], Asani Sarkar[41], Lucio Sarno[138], Olivier Scaillet[103], Stefan Scharnowski[154], Klaus R. Schenk-Hoppé[153], Andrea Schertler[145], Michael Schneider[28,70], Florian Schroeder[74], Norman Schürhoff[104], Philipp Schuster[176], Marco A. Schwarz[31,16], Mark S. Seasholes[4], Norman J. Seeger[200], Or Shachar[41], Andriy Shkilko[204], Jessica Shui[40], Mario Sikic[193], Giorgia Simion[201], Lee A. Smales[191], Paul Söderlind[174], Elvira Sojli[165], Konstantin Sokolov[158], Jantje Sönksen[184], Laima Spokeviciute[20], Denitsa Stefanova[152], Marti G. Subrahmanyam[80,79], Barnabas Szaszi[34], Oleksandr Talavera[135], Yuehua Tang[142], Nick Taylor[137], Wing Wah Tham[165], Erik Theissen[154], Julian Thimme[62], Ian Tonks[137], Hai Tran[72], Luca Trapin[136], Anders B. Trolle[26], M. Andreea Vaduva[128], Giorgio Valente[52], Robert A. Van Ness[161], Aurelio Vasquez[55], Thanos Verousis[141], Patrick Verwijmeren[38], Anders Vilhelmsson[73], Grigory Vilkov[43], Vladimir Vladimirov[134], Sebastian Vogel[38], Stefan Voigt[150], Wolf Wagner[38], Thomas Walther[197], Patrick Weiss[198], Michel van der Wel[38], Ingrid M. Werner[119], Joakim Westerholm[178], Christian Westheide[188], Hans C. Wika[84], Evert Wipplinger[200], Michael Wolf[193], Christian C. P. Wolff[152], Leonard Wolk[200], Wing-Keung Wong[5], Jan Wrampelmeyer[200], Zhen-Xing Wu[1], Shuo Xia[50], Dacheng Xiu[139], Ke Xu[187], Caihong Xu[114], Pradeep K. Yadav[166], José Yagüe[163], Cheng Yan[141], Antti Yang[38], Woongsun Yoo[23], Wenjia Yu[1], Yihe Yu[132], Shihao Yu[200], Bart Z. Yueshen[54], Darya Yuferova[86], Marcin Zamojski[144], Abalfazl Zareei[114], Stefan M. Zeisberger[100], Lu Zhang[152], S. Sarah Zhang[153], Xiaoyu Zhang[200], Lu Zhao[111], Zhuo Zhong[157], Zeyang (Ivy) Zhou[192], Chen Zhou[38], Xingyu S. Zhu[113], Marius Zoican[183], and Remco Zwinkels[200]

[1]Aalto University, [2]Aix-Marseille University, [3]American University, [4]Arizona State University, [5]Asia University, [6]Auburn University, [7]Australian National University, [8]BI Norwegian Business School, [9]Babson College, [10]Bank for International Settlements, [11]Bank of England , [12]Bar-Ilan University, [13]Barcelona School of Economics, [14]Bayes Business School, [15]Bowling Green State University, [16]CESifo, [17]CNRS, [18]California State University - East Bay, [19]Canadian Derivatives Institute, [20]Cardiff University, [21]Cascad, [22]Central Bank of Norway, [23]Central Michigan University , [24]Columbia Business School, [25]Columbia University, [26]Copenhagen Business School, [27]Cornell University, [28]Deutsche Bundesbank, [29]Duke University, [30]Durham University, [31]Düsseldorf Institute for Competition Economics, [32]EDF Energy London, [33]EDHEC Business School, [34]ELTE, Eotvos Lorand University, [35]EPFL, [36]ESADE Business School, Univ. Ramon Llull, [37]Emory University, [38]Erasmus University Rotterdam, [39]European Central Bank, [40]Federal Housing Finance Agency , [41]Federal Reserve Bank of New York, [42]Financial Conduct Authority, [43]Frankfurt School of Finance and Management, [44]Georgetown University, [45]Ghent University, [46]Goethe University Frankfurt, [47]HEC Liège - University of Liège, [48]HEC Montréal, [49]HEC Paris, [50]Halle Institute for Economic Research, [51]Hong Kong Institute for Monetary and Financial Research, [52]Hong Kong Monetary Authority, [53]IIM Ahmedabad, [54]INSEAD, [55]ITAM, [56]Ibn Tofail University, [57]Indiana University, [58]International Monetary

Fund, [59]John von Neumann University, [60]Justus-Liebig University, [61]KU Leuven, [62]Karlsruhe Institute of Technology, [63]Kiel University, [64]Kiel university, [65]King's College London, [66]Koç University, [67]LUM University, [68]La Trobe University, [69]Lebanese American University, [70]Leibniz Institute for Financial Research SAFE, [71]Leibniz University Hannover, [72]Loyola Marymount University, [73]Lund University, [74]Macquarie University, [75]Masaryk University, [76]Michigan State University, [77]Missouri State University, [78]Monash University, [79]NYU Shanghai, [80]NYU Stern, [81]Neoma Business School, [82]New Mexico state University, [83]None, [84]Norges Bank, [85]Northwestern University, [86]Norwegian School of Economics (NHH), [87]Nottingham Trent University, [88]Oklahoma State University, [89]Open Universiteit, [90]Paderborn University, [91]Paris Dauphine University, [92]Politecnico di Milano, [93]Pontificia Universidad Católica de Chile, [94]Pontificial University of Chile, [95]Princeton University, [96]Purdue University, [97]Queen's University, [98]Queen's University Belfast, [99]RMIT University, [100]Radboud University, [101]SFI at HEC Lausanne, [102]SFI at USI Lugano, [103]SFI at University of Geneva, [104]SFI at University of Lausanne, [105]SFI at University of Zurich, [106]Saint Joseph University, [107]Saint Louis University, [108]Salisbury University, [109]SandP Global Ratings, [110]Singapore Management University, [111]Southwestern University of Finance and Economics, [112]State University of New York at Buffalo, [113]Stockholm School of Economics, [114]Stockholm University, [115]TBS Education, [116]Technische Universität Berlin, [117]Technische Universität Dresden, [118]Tennessee Technological University, [119]The Ohio State University, [120]The University of Manchester, [121]Tinbergen Institute, [122]Toulouse 1 Capitole University, [123]UC Berkeley, [124]UCLA, [125]UCLouvain, [126]UCSI University, [127]Universidad Autónoma de Madrid, [128]Universidad Carlos III de Madrid, [129]Universidad EAFIT, [130]Universitat Pompeu Fabra, [131]University Paris 1 Pantheon-Sorbonne, [132]University at Buffalo, [133]University of Alicante, [134]University of Amsterdam, [135]University of Birmingham, [136]University of Bologna, [137]University of Bristol, [138]University of Cambridge, [139]University of Chicago Booth School of Business, [140]University of Edinburgh, [141]University of Essex, [142]University of Florida, [143]University of Glasgow , [144]University of Gothenburg, [145]University of Graz, [146]University of Hohenheim, [147]University of Illinois at Chicago, [148]University of Illinois at Urbana-Champaign, [149]University of Innsbruck, [150]University of København, [151]University of Lausanne, [152]University of Luxembourg, [153]University of Manchester, [154]University of Mannheim, [155]University of Maryland, [156]University of Massachusetts, Amherst, [157]University of Melbourne, [158]University of Memphis, [159]University of Miami, [160]University of Michigan, [161]University of Mississippi, [162]University of Munich (LMU), [163]University of Murcia, [164]University of New Mexico, [165]University of New South Wales, [166]University of Oklahoma, [167]University of Oregon, [168]University of Orléans, [169]University of Oxford, [170]University of Padova, [171]University of Queensland, [172]University of San Francisco, [173]University of Southern California, [174]University of St. Gallen, [175]University of Stavanger, [176]University of Stuttgart, [177]University of Sussex, [178]University of Sydney, [179]University of Technology Sydney, [180]University of Texas at Arlington, [181]University of Torino, [182]University of Toronto, [183]University of Toronto Mississauga, [184]University of Tübingen, [185]University of Utah, [186]University of Verona, [187]University of Victoria, [188]University of Vienna, [189]University of Virginia, [190]University of Warwick, [191]University of Western Australia, [192]University of Wollongong, [193]University of Zurich, [194]University of the Balearic Islands, [195]Università della Svizzera italiana, [196]Université Paris 1 Panthéon-Sorbonne, [197]Utrecht University, [198]Vienna Graduate School of Finance, [199]Vienna University of Economics and Business, [200]Vrije

Universiteit Amsterdam, [201]WU Vienna University of Economics and Business, [202]Waseda University, [203]West Virginia University, [204]Wilfrid Laurier University, [205]Xi'an Jiaotong-Liverpool University, [206]York University, [207]Zhongnan University of Economics and Law

*The first nine authors in italics are the project coordinators. They conceptualized and designed the project, managed it, conducted the meta-analyses, and wrote the manuscript. Any errors are therefore their sole responsibility. The other authors all significantly contributed to the project by participating either as a member of a research team, or as a peer evaluator. The views expressed here are the authors' and do not represent the views of the Federal Reserve Bank of New York or the Federal Reserve System, or any other of the institutions that the authors are affiliated with or receive financing from. The coordinators thank Andrew Chen, Amit Goyal, Campbell Harvey, Lucas Saru, Eric Uhlmann, and participants at the Microstructure Exchange 2021, Derivatives Forum Frankfurt 2022, Financial Intermediation Research Society (FIRS) 2022, Research in Behavioral Finance Conference (RBFC) 2022, Society for Experimental Finance (SEF) 2022, Society for Financial Econometrics (SoFiE) 2022 where the paper was runner-up for the best-paper prize, Vienna-Copenhagen Conference on Financial Econometrics 2022, and the Western Finance Assocation (WFA) 2022 for valuable comments. They further thank Adam Gill, Eugénie de Jong, Ingrid Löfman, and Elmar Nijkamp for research assistance. The coordinators are grateful for financial support from (Dreber) the Knut and Alice Wallenberg Foundation, the Marianne, Marcus Wallenberg Foundation, the Jan Wallander, Tom Hedelius Foundation, (Huber) an FWF grant P29362, (Huber and Kirchler) FWF SFB F63, (Johannesson) Riksbankens Jubileumsfond grant P21-0168, and (Menkveld) NWO-Vici.

Chapter 3

Informed Alpha

3.1 Introduction

Most individuals do not invest in security markets directly but rather via investment vehicles such as mutual, pension, and hedge funds. Evaluating the performance of these vehicles is important for the well-being of individuals, for the efficient allocation of capital in the economy, and for market efficiency. In this paper we test some of the implications of the Rational Expectations Equilibrium paradigm regarding the evaluation of the performance of investment vehicles and derive a new measure of performance, which we call *Informed Alpha*, that beats the standard (single or multifactor) alpha. The superiority of *Informed Alpha* is illustrated in the following figure where we plot future fund performance as a function of past Jensen alphas (Figure 3.1 (a)) versus *Informed Alphas* (Figure 3.1 (b)). As we can observe, while past Jensen's alphas hardly predict future performance (low slope coefficient and R2), *Informed Alpha* are strongly predictive of future performance (slope and R2 is higher). This means that *Informed Alpha* represents a major improvement in the identification of funds managed by truly talented managers.

Insert Figure 3.1: *Informed Alpha*

The asset pricing literature has provided academia and the financial industry with alpha as perhaps the most important metric for performance evaluation.[1] But alpha is a very noisy measure of the quality in asset management. First, since all asset pricing models are incomplete, in the sense of they do not account for all sources of risk, alphas may reflect compensation for not accounted sources of risk rather than outstanding performance due to talent in asset management. Second, as pointed out in Karapandza and Marin (2014) the addition of factors within an incomplete factor model generates spurious alphas that just reflect the omitted variables bias prevailing in all incomplete asset pricing models. This means that many of the funds' alphas have nothing to do with outstanding performance, but are just the reflection of the omitted variable bias, which has no economic bearing at all. Finally, we cannot ignore that some of the alphas are the result of luck, rather than talent. These problems are not intellectual curiosities, but truly relevant observations that are easily verified in the relatively low predictability of alphas.

[1]There are many other metrics for performance evaluation such as the Sharpe Ratio, M2 of Modigliani and Modigliani, the Information Ratio, Treynor Index, etc.

At the heart of all the previous problems is the absence of a true microfoundation behind alpha as a measure of talent. Alpha is just a statistic that measures the deviation from a theoretical and wrongful (as it fails to account for all the existing risk factors) pricing model of observed data, and as such it may obey to many different explanations, of which talent is just one of them. For this reason, it is key to explicitly account for the actual behaviour expected from talented traders to identify them in the data. That is, if the goal is to identify and measure true talent in asset management, we need to incorporate the microeconomics of talent in performance evaluation.

The natural place to analyse this microfoundation is the Noisy Rational Expectations Equilibrium paradigm. Starting with Radner (1979) and several papers by S. Grossman (for instance, Grossman (1976); Grossman and Stiglitz (1980); Grossman (1981)), and complemented with the so-called market microestructure literature (for instance, Kyle (1985); Grundy and McNichols (1989); Kim and Verrecchia (1991), to name just a few of the leading contributions), the rational expectations paradigm emerged as a setting where agents trade based on different information and learn from prices. This setting explicitly characterizes the behaviour of better informed (or, more talented) and worse informed (or, less talented) traders in both competitive and strategic environments. In these models, prices change due to two reasons: the arrival of (private and public) information about fundamentals –informational price change– and the trading done by noise or liquidity traders –not informational price change–. In equilibrium, the arrival of private information generates momentum in asset prices and liquidity trading generates price reversals. In this setting, the better informed traders exhibit outstanding performance because they have better information about fundamentals and also about noise, as they learn from prices when a price change is due to noise. For instance, in the famous Grossman and Stiglitz (1980) setting of pure asymmetric information (as opposed to disperse information)[2] where market participants are split between informed and uniformed traders, informed traders exhibit outstanding performance because they buy cheap (sell high) when they receive good (bad) news, as prices still do not reflect that private information, and they buy cheap (sell high) after a large price drop (jump) generated by liquidity trading; uninformed traders exhibit poorer performance as they are on

[2]The logic is quite similar in models of disperse information (as opposed to asymmetric information)

the other side of the trades initiated by informed traders, and cannot figure out noise trading as precisely as informed traders.

Testing the performance evaluation predictions of this theory is not easy because traders private information is not an observable variable. But we can proxy for that. In this paper we proxy private information with the trading done by insiders. Its been recognized for a long time that the trading done by insiders (specially insiders purchases) is informative in general. Further, if we restrict our attention to insiders purchases with positive returns after the transaction, we will be filtering trading consistent with the informed trading assumed in the NREE. Hence, fund managers that buy stock around the dates in which these insiders purchases take place will be behaving as informed traders and should exhibit outstanding performance.

Based on this insight we define the variable *Insiders Purchases Coincidence* (*%INS PC*) as the fraction (percentage) of times the fund agrees with insiders when the latter make informative purchase of stocks in the fund's portfolio, during a given period of time. In other words, *%INS PC* is a metric that measures funds' abilities to spot trading opportunities similar to those of insiders that buy shares on private information. Notice that this proxy for informed trading is very incomplete as it only accounts for some information events that generate informed trading. First, informed traders in actual markets may even know more than insiders (specially about market wide fundamental information); second, insiders may decide not to trade on private information to avoid prosecution or just because of company self-imposed restrictions; finally, insiders are not professional asset managers and may not trade at all based on non-fundamental information (liquidity needs). So our proxy does not capture all events relevant for informed traders in the market place and consequently our test may fail even when the theory were true. In particular, our conjecture is that traders who systematically behave as insiders in the informational events related to insiders trading do so because they have a talent or investment philosophy very close to informed traders and consequently will also behave as informed traders in other informational events not accounted for by our limited proxy for informed trading. This is indeed the hypothesis we test in this paper.

The REE theory pass with honors the test. Indeed, our results as astonishing given the previous limitations of the analysis. The portfolio of funds in the top decile of the distribution of

%INS PC exhibit positive and significant alphas of 0.17%; while the portfolios of funds in the bottom decile exhibit alphas 0.00%. Furthermore, we have performed a battery of robustness tests and the results remain true. In conclusion, our analysis is conclusive in that the REE paradigm provides a powerful metric associated with outstanding performance.

Notice that the previous performance evaluation analysis is set in the context of Jensen's alpha, besides our previous strong concerns about this traditional performance metric as a measure of talent in asset management. But the analysis itself gives as the clue as to how to improve the traditional metric. In particular, the restriction of the set of Jensen's alphas to those in the top percentiles of the distributions of *%INS PC* will result in an improved pool of funds more concentrated in those with true talent in asset management. That is, the intersection of the top deciles of the distributions of *%INS PC* and unconditional alphas should yield a pool of funds whose outstanding performance is related to talent rather than luck or errors in the multi-factor asset pricing model used in the estimation of unconditional alphas. To test this hypothesis, we define *Informed Alpha* as the set of funds of unconditional alphas (positive and significant alphas at the 10% level of significance) in intersection with those funds in the 10th percentile of the distribution of *%INS PC*. We show that, unlike unconditional alpha, *Informed Alpha* exhibits strong persistence: past REE alphas are very predictive of future performance, as illustrated in Figure 1 and further corroborated using rigorous empirical methods.

To summarize, in this paper we test performance evaluation in the context of the REE Paradigm. Consistent with the theory, we provide robust evidence that funds that implement trading strategies more similar to insiders exhibit outstanding performance. Based on these results we introduce a new metric called *Informed Alpha* which represents a major improvement over alpha in the identification of truly talented managers.

The rest of the chapter is structured as follows. The next section presents the literature review. Section 3.3. is devoted to describe the data and the construction of the variables of interest. Section 3.4. presents the results of testing the performance of mutual funds in the REE Paradigm. Section 3.5. has the aim to go a step further and show the superiority of Informed Alpha over Jensen Alpha. Finally, in section 3.6., we conclude.

3.2 Related literature

The analysis in this paper relates to three branches of the literature. First, it relates to the CAPM and more broadly to the multifactor asset pricing literature that has produced alpha as a measure of performance; second, it relates to the Noisy Rational Expectations literature, where equilibrium is obtained under the assumption of traders in possession of disperse (or asymmetric) information and learn from prices; finally, it relates to the insider trading literature.

Jensen (1968) was the first to use alpha as a measure of performance and did it in the context of the one factor, market model. The one factor model was extended to a multi-factor model approach: first adding two factors (Fama and French (1993)), and afterwards adding more factors such as the momentum factor (Carhart (1997); Pástor and Stambaugh (2002)) liquidity factor or the Whited and Wu (2006) credit constraints factor and many other multifactor specifications, perhaps too large to cite here. As a measure of performance, alpha has been criticized for many reasons, but essentially because of its arbitrariness. First, Roll (1977) showed that the measure is highly sensitive to the choice of the benchmark portfolio; Cremers and Petajisto (2009) argue that some of the factors used are not directly tradable and they would be very costly to trade, hence estimated alphas are not in the actual asset span of the existing tradable securities; finally, Karapandza and Marin (2014) show that due to model incompleteness, it is not just that some alphas are meaningless because they may just reflect risk not accounted for, but also that due to the omitted variable bias, the addition or subtraction of factors in the model generates spurious alphas, that is alphas not related to outstanding performance. So this literature makes very clear that many of the alphas obtained in the context of any particular multifactor model are just pure statistical artifacts not related to quality in asset management; and, that some true talented managers are not spotted by the alpha metric.

This failure of alpha to identify true talent is manifested in a new wave of papers that provide alternatives to alpha. In this direction we have Cremers and Petajisto (2009) that suggests to look at performance by focusing on the share of portfolio holdings that differ from the benchmark index (self reported by the fund) holdings; also Berk and van Binsbergen (2015) that proposes measuring performance in dollars terms, instead of (alpha) returns. In this direction, our paper contributes by providing another alternative for alpha, but unlike the previous developments, imposing microeconomic principles.

In our view, at the bottom of these failures of alpha as a measure of skills in asset management is the absence of a true microfoundation behind the metric. So, this paper contributes to this literature developing a measure that improves alpha on microeconomic grounds.

The REE paradigm explicitly models and characterizes skill in asset management. Starting with Radner (1979) and later on with more finance-like settings such as those in Grossman (1976); Grossman and Stiglitz (1980); Kyle (1985), etc. this literature models the behaviour or better versus worse informed traders. So the paradigm has clear predictions for the measurement of performance. These predictions are summarized in Colla and Marin (2010). To the best of our knowledge, nobody has tested these empirical predictions before.[3] So our paper contributes to this literature as the first paper that tests performance evaluation in the context of REE models.

Finally, our work also relates to the literature on insider trading as we use the trading done by insiders as the proxy for private information in financial markets. It's been long recognized that insider trading is informative. There is plenty of evidence on insiders' ability to predict future price changes in their own firm's stocks and that in general insider purchases are more predictive than insiders sales, as the later can also be driven by no informational considerations such as diversification Seyhun (1992); Lakonishok and Lee (2001); Huddart et al. (2007); Cohen et al. (2012). Building on this literature we proxy for informed trading by looking at insiders purchases rather than sales and restricting these to those that are followed by positive returns, which are the most likely to be driven by private information. Hence our work also contributes to this literature by highlighting a new role of insider trading as a tool for performance evaluation.

3.3 Data and variable definition

3.3.1 Data and sample

Our sample consists of active US. mutual funds during the period 1990 to 2016. Fund characteristics have been downloaded from the CRSP and Thomson Reuters mutual fund database.

[3]The only exception to note is Kacperczyk and Seru (2007), who introduces a measure of performance based on the REE but, as explained in Colla and Marin (2010), the model prediction the authors test is faulty for several reasons. See details in Colla and Marin (2010).

Following, Kacperczyk et al. (2008), we discard those funds with investment objective: international, municipal bonds, bond and preferred, balanced, metal and unclassified funds as well as those funds that on average invest less than 80% or more than 105% in common equity.[4] As we are interested mainly on active mutual funds, following Hoberg et al. (2018) using the CRSP index funds, we exclude passive mutual funds and also funds whose name contain words such as: "INDEX", "S&P", DOW", etc. In addition, we exclude funds if the total net asset (TNA) falls below $5 million at one point in time and we keep just those funds which have at least 10 holdings in their portfolio. In addition, by following Coval and Stafford (2007), we discard those funds for which the discrepancies in terms of the TNA is high between the CRSP database and Thomson Reuters database.[5] Finally, we exclude funds which do not report returns for at least 36 months, thus, we are left with 2089 unique active mutual funds during the period: 1996 to 2016.

Additionally, we make use of the Thomson Reuters Institutional Holdings database. We complement this information, with the information on prices and outstanding shares provided by the CRSP database.

Since, we are interested in insider trading, we make use of the Thomson Reuter Insider Trading Dataset. Following Akin et al. (2020), we keep just transactions in Form 4, which reflects any change in insider's ownership positions, which could be any type of transactions: purchases, sales, grants and awards, exercises of derivatives, etc. We focus just on market purchases and sales. Following, Lakonishok and Lee (2001), we exclude: amended transactions, transactions for which no price or share is available, transactions with less than 100 shares, and transactions for which the number of shares exceeded 20% of the number of outstanding shares. In addition, we keep just common shares (CRSP shares code should be 10 or 11).

Insert Table 3.1 Here

In Table 3.1 we report the summary statistics for our final database on mutual funds. The summary statistics are similar to those reported in the literature. Such that the average age is

[4]By using the Thomson Reuters classification: discard those funds with Objective Code: international funds (IOC = 1), municipal bonds (IOC = 5), bond and preferred (IOC = 6), balanced (IOC = 7), metals (IOC = 8) and unclassified funds (IOC = 9). In addition, by using the CRSP database, we calculate the average percentage of stock holdings, and keep just those investing between 80% and 105% in common equity.

[5]Keep funds when the following condition holds: 1/1.3 < TNA TR/TNA CRSP <1.

of 17 years, the size of the fund is on average of 1,753 millions. The average expense ratio and turnover ratio is of 0.01 and 0.82, receptively. The monthly net and gross returns are returns and monthly returns are on average of 0.64% and 0.74%, respectively.

3.3.2 Informed funds measure

We refer to informed funds as funds that trade akin informed traders in the REE paradigm. And we identify them as those that spot trading opportunities similar to those of insiders that buy shares on private information. To this purpose we define the variable insider purchases coincidence, $\%INSPC$ that measures this similarity of fund trading with insider trading. $\%INSPC$ is defined as the fraction (percentage) of times the fund agrees with insiders when the latter purchase stocks in the fund portfolio, during a given period of time. For example, assume that a given fund holds 5 stocks: 1, 2, 3, 4, and 5. Assume now that during a given period insiders buy stocks 2 and 3, and that during that same period the fund increased the holding of stocks 1, 2 and 4. In this example, $\%INSPC_{f,t-k,t}$ takes the value 1/2.

Formally, for a given period of time t-k to t:

- Let $H_{j,f,t-k,t}$ denote the set of stocks, j, that fund f *holds* at the beginning of the period, t-k.
- Now consider the following subset of $H_{j,f,t-k,t}$: $FPH_{j,f,t-k,t}$ is the subset of stocks in $H_{j,f,t-k,t}$ for which *fund f* has increased holdings (between t-k and t).
- Let $I_{j,f,t-k,t}$ denote the set of stocks, that insiders purchased or sold, between t-k and t.
- Now consider the following subset of: $I_{j,f,t-k,t}$: $IPH_{j,f,t-k,t}$ is the set of stocks for which insiders have increased holdings (between t-k and t). We consider insiders have increased holding of stock j during a given period t-k to t if NET BUYS during the period –defined as the sum of value purchased by insiders normalized by market value at the transaction date, during the period, minus the sum of the value sold by insiders normalized by market value at the transaction date, during the period– is

positive and the stock has positive returns afterwards (after 18 months). Formally, denoting by τ the date in which the transaction took place (aggregating the values if more than one transaction took place the same day), NET BUYS are defined as:

$$NETBUYS_{j,t-k,t} = \sum_{\tau\in(t-k,t)} \frac{BUY_{j,\tau}}{MV_{j,\tau}} - \sum_{\tau\in(t-k,t)} \frac{SELL_{j,\tau}}{MV_{j,\tau}} \quad (3.1)$$

where $BUY_{j,\tau}$ is the total dollar value of all buy transactions in security j done by all type of insiders in day τ of period t-k to t; $SELL_{j,\tau}$ is he total dollar value of all sell transactions in security j done by all type of insiders in day τ of period t-k to t; and, $MV_{j,\tau}$ is the market value of security j at the day of the transaction. To qualify as a period of informative buying in stock j by insiders we require both that $NETBUYS_{j,t-k,t}$ is positive and that it has a positive return during the following 18 months from the date of the last insider transaction release date.[6]

- Furthermore, define the following indicator functions:

$$IFP_{j,f,t-k,t} = \begin{cases} 1 & if\, j\in FPH_{j,f,t-k,t} \subseteq H_{j,f,t-k,t}; \\ 0 & \text{otherwise.} \end{cases} \quad (3.2)$$

$$IIP_{j,f,t,t+h} = \begin{cases} 1 & if\, j\in IPH_{j,f,t-k,t} \subseteq I_{j,f,t-k,t}; \\ 0 & \text{otherwise.} \end{cases} \quad (3.3)$$

$$IFIP_{j,f,t,t+h} = \begin{cases} 1 & if\, j\in FPH_{j,f,t-k,t} \cap IPH_{j,f,t-k,t}; \\ 0 & \text{otherwise.} \end{cases} \quad (3.4)$$

[6]Notice that the transaction by an insider can take place in k, but may be released up to two days later (https://www.sec.gov/files/forms-3-4-5.pdf)

- Finally, we define

$$\%INSPC_{f,t-k,t} = \frac{\sum_{j=1} IFIP_{j,f,t-k,t}}{\sum_{j=1} IIP_{j,f,t-k,t}} \tag{3.5}$$

In the main body of the paper we assume that k is 4 (four quarters). Now we characterize $\%INSPC_{f,t-k,t}$. At the end of each quarter, t, we compute $\%INSPC_{f,t-k,t}$ over the previous 4 quarters and sort it in 10 deciles. In Table 3.2 we report the means of each decile over the period 1996Q1 to 2016Q4. Mutual funds in decile 10 have on average 3.04% of stocks in which the funds increases their holdings when informed insiders traders buy. While the ones in decile 1 have a proportion of 0.01%.

Insert Table 3.2

3.4 Testing performance evaluation in the REE Paradigm

3.4.1 Portfolio based analyses

According to the REE hypothesis, funds that trade like informed traders will exhibit outstanding performance. Thus according to this, funds in the top decile of the $\%INSPC_{f,t-k,t}$ will outperform those in the bottom decile. With the aim to test this hypothesis, we use portfolio-based analyses. Here and in the rest of the paper, in order not to contaminate the analysis with the return of the stocks traded by the fund in coincidence with the insiders, we start running portfolio returns 18 months after the sorting of funds based on *INSPC*. Formally, at the end of each quarter, we sort mutual funds into 10 portfolios based on our measure of informed mutual funds computed 6 quarters before: $\%INSPC_{f,t-10,t-6}$. Where decile 10 (D10) contains mutual funds with the highest $\%INSPC_{f,t-10,t-6}$, while decile 1 (D1) contains mutual funds with the lowest $\%INSPC_{f,t-10,t-6}$. We assume that we hold the portfolio for 3 months, thus for each portfolio in each decile, we calculate equally weighted gross and net returns over each of the following three months. Thus our sample with monthly returns spans 1996Q1 to 2016Q4. For each decile, we report i) the average return of equally weighted gross and net returns, and ii) the risk adjusted returns obtained by estimating the Capital Asset Pricing Model (CAPM). Table 3.3 reports the results. Panel (a) reports the results by using gross returns, while panel (b) reports the results by using net returns. The results point out to higher raw returns on decile 10 and lower ones in decile 1. If we were to make a 0 investment by going long in decile 10 and short in decile 1, we would earn (net) on Average 0.21% monthly. In addition, in terms of alphas, mutual funds in decile 10 (D10) earn 0.19% more than those funds in decile 1 (D1). Similar results are found by using the gross returns instead of net returns.

Insert Table 3.3 Here

3.4.2 Predictive regressions

By using predictive regressions, we examine if funds that trade like informed traders will exhibit outstanding performance. To this end, we estimate the following regression:

$$RAlpha_{f,q+1} = \beta_0 + \beta_1 \%INSPC_{f,t-10,t-6} + \beta_k' X_{k,t} + \epsilon_t \quad (3.6)$$

Where our dependent variable is $RAlpha_{f,q}$. $RAlpha_{f,q}$ is the average of each of the following realized alpha over the following 3 months for fund f. Realized alphas are calculated, following the methodology proposed by Carhart (1997), as the difference between the realized return in excess of the risk-free rate and the expected excess return estimated from one factor market model. With the aim to calculate the expected excess return estimated from one factor model, we estimate rolling-window time series regression over the previous three years. The independent variable of interest is $\%INSPC_{f,t-10,t-6}$. We control for different variables, which have been acknowledged to play a role on the future performance of mutual funds: Age, Total Net Assets, Expense Ratios, Turnover Ratios, Flows and Volatility of the Flow. In addition, we include time fixed effects to control for potential changes which could have affected mutual funds performance, such as regulatory changes, macroeconomic events or any other events which could affect systematically funds performance. In addition, we cluster standard errors at the fund level.

Insert Table 3.4 Here

The results are reported in Table 3.4. Panel A reports the results using the gross returns, while Panel B reports the results by using net returns. Model 1 reports the results without any controls. Model 2 reports the results by including past alpha as a control. Finally, model 3 includes controls which have been acknowledged in the literature to play a key role on future mutual funds performance. In Panel A, by using gross returns, in all three models, the variable of interest $\%INSPC$ is positive and statistically significant at the 1% level of significance, which suggests that the proxy gauges the informativeness of a fund. An increase of 1% in the variable of interest is associated with an increase of 1.5 in basis points in monthly realized alphas over the following 3 months (See: Table 4: Panel A: Realized Gross Alpha(t+1)). When we consider net returns instead of gross returns, the results are very similar (See Panel B). It is worth noting that the sign of the coefficients accompanying the control variables is consistent with the ones in the literature.

3.4.3 Persistence

3.4.3.1 Portfolio based analyses

In this subsection, we examine if funds with a higher $\%INSPC$ are relevant for performance persistence. With this aim, we proceed as follows. At the end of each quarter, we sort mutual funds into 10 portfolios based on our measure of informed mutual funds: $\%INSPC_{f,t-10,t-6}$. Where decile 10 (D10) contains mutual funds with the highest $\%INSPC_{f,t-10,t-6}$, while decile 1 (D1) contains mutual funds with the lowest $\%INSPC_{f,t-10,t-6}$. For each decile, we i) calculate the mean of the equally weighted gross and net returns, and ii) we estimate the risk adjusted returns obtained by estimating the Capital Asset Pricing Model (CAPM). We assume that we hold the portfolio for different time horizons h: 6, 9, 12, 15, 18, 24, and 36 months. The long-term differences in performance between more informed mutual funds and less informed ones are reported in Table 3.5. Panel A reports the results by using gross returns, while panel B reports the results by using net returns. The results point out to higher raw returns on decile 10 and lower ones in decile 1 for all the horizons assumed. The difference in returns dissipates in time.

However, even after 36 months, the difference in returns is positive and statistically significant at the 1% level of significance, in both cases for the Average and for the CAPM Alpha. If we were to make a 0 investment by going long in decile 10 and short in decile 1, and hold this portfolio for 24 months, we would earn (net) on average 0.15% monthly. In addition, in terms of alphas, mutual funds in decile 10 (D10) earn 0.11% more than those funds in decile 1 (D1). Similar results are found by using the gross returns instead of net returns.

Insert Table 3.5 here

4.3.2 Persistence Regressions

Following Hoberg et al. (2018), we run a regression, in which we want to assess to what extent increases in past performance are higher for mutual funds with a higher $\%INSPC_{f,t-10,t-6}$.[7]

$$\begin{aligned} RAlpha_{f,q+1} = \beta_0 + \beta_1 \%INSPC_{f,t-10,t-6} + \beta_2 High_{f,q} + \beta_3 Medium-High_{f,q} + \beta_4 Medium-Low_{f,q} \\ + \beta_5 \%INSPC_{f,t-10,t-6} * High_{f,q} + \beta_6 \%INSPC_{f,t-10,t-6} * Medium-High_{f,q} \\ + \beta_7 \%INSPC_{f,t-10,t-6} * Medium-Low_{f,q} + \beta_k' X_{k,t} + \epsilon_t \end{aligned} \tag{3.7}$$

Our dependent variable is $RAlpha_{f,q+1}$ which is the average of each realized alpha over the following 12 months for fund f. We calculate the realized alphas following the methodology proposed by Carhart (1997). In order to capture when a fund is characterized by higher $\%INSPC_{f,t-10,t-6}$, we sort the variable of interest: $\%INSPC_{f,t-10,t-6}$ in four groups. Based on the four groups, we construct 4 dummy variables: High is a dummy variable that takes the value of one if the fund falls in group 4 and 0 otherwise; Medium-High is a dummy variables that takes the value of 1 if the fund falls in group 3 and 0 otherwise; Medium-Low is a dummy variable that takes the value of 1 if the fund falls in group 2 and 0 otherwise; and Low is a dummy variable

[7] Hoberg et al. (2018) show that funds acting in a low competition environment have persistent performance. While we are using as a dependent variable the realized alphas obtained from running a rolling one factor model, Hoberg et al. (2018) calculates the Characteristics- Selectivity Alpha by using the methodology proposed in Daniel et al. (1997)

that takes the value of 1 if the funds falls in group 1 and 0 otherwise. Alpha stands for the alpha estimated from the CAPM regressions estimated over the previous 36 months. As we want to assess if increases in past alphas are associated with a higher persistence for funds with a high $\%INS\,PC$, we include interaction effect between past performance experienced by funds and one of the dummy variables constructed based on $\%INS\,PC$.

The results are reported in Table 3.6. Panel A reports the results using the gross returns, while Panel B reports the results by using net returns. In Panel A, in all three models, the dummy variables High is statistically significant and exhibit a positive sign. The results suggest that those funds falling under the category High exhibit higher returns than those falling under the category of Low. For instance, those funds characterized by High $\%INS\,PCT$ have on average a 0.1% (coefficient in model 3) higher gross future performance ($RAlpha_{f,q}$) than those funds with low $\%INS\,PC$ on a monthly basis. The coefficient of interest that accompanies the interaction effect: $Alpha * High$ is negative and not statistically significant at the 1% of significance. Similar results are obtained, by using Net Returns instead of Gross Returns. The previous results lead us to the following section: 3.5. Informed Alpha.

Insert Table 3.6 Here

3.5 Informed Alpha

Based in the previous analysis, we conjecture that restricting the set of Jensen's alphas to those funds that exhibits a large coincidence in trading with insiders will results in a subset of funds whose alpha is more likely to be the result of skill than luck or asset pricing model miss specification. We denote these funds that have a statistically significant alpha and that also trade consistently with the trading of informed traders in RRE models as funds with *Informed Alpha*.

To illustrate the superiority of the Informed Alpha over Jensen Alpha, we start by a simple regression and graphical analysis. Jensen Alpha is obtained as follows: at the end of each year we estimate fund alphas using 36 months of past gross monthly returns. We restrict the sample to those funds with a positive and statistical significant Jensen Alpha at the 10% level of significance (we call this the set of Jensen Alpha). Informed Alpha include the intersection of Jensen Alpha with the set of funds in the 10th percentile of the distribution of $\%INSPC$. Now, we regress the average of each of the realized alpha over the following 12 months over the alphas previously defined. The results are plotted in Figure 3.1. As we can observe, the predictive and explanatory power of *Informed Alphas* (panel (b)) is much larger than the one of Jensen‘s alphas (panel (a)): the slope in the case of Informed Alpha is larger in the case of Jensen‘s alphas (0.54 vs 0.36) and the R2 in the case of Informed Alpha is more than twice the R2 in the case of Jensen‘s alphas (0.178 vs 0.075). This simple analysis already points at a much better predictive power of Informed Alpha over Jensen's Alpha.

With the aim to show formally that Informed Alpha sorts truly talented managers who exhibit strong performance, we estimate the following fund-quarter regression:

$$RAlpha_{f,q+1} = \beta_0 + \beta_1 Alpha_{f,q} + \beta_2 Top_{f,q} + \beta_3 Alpha_{f,q} * Top_{f,q} + \beta_k' X_{k,t} + \epsilon_t \quad (3.8)$$

Where, $RAlpha_{f,q}$ is the average of each of the realized alpha over the following 12 months for fund f. We calculate the realized alphas following the methodology proposed by Carhart (1997). The realized alpha is calculated as the difference between the realized return in excess of the risk-free rate and the expected excess return estimated from one factor market model. With the aim to calculate the expected excess return estimated from one factor model, we estimate rolling-window time series regression over the previous three years. $Alpha_{f,q}$ is the alpha estimated over the previous 36 months by estimating the one factor model. $Top_{f,q}$ is a dummy variable that takes the value of one if the fund f in quarter q belongs to the top decile of the distribution of $\%INSPC$ and 0 otherwise. $Alpha_{f,q} * Top_{f,q}$ is the interaction effect which shed lights on the ability of our measure to predict if increases in past alpha is associated with higher stronger future performance for more informed funds relative to less informed ones. The remaining variables control for other factors, which has been acknowledged to play a role on the future performance of mutual funds: the Age, Total Net Assets, Expense Ratios, Turnover Ratios, Flows and Volatility of the Flow. We include time fixed effects to control for potential changes which could have affected mutual funds performance, such as regulatory changes, macroeconomic events or any other events which could affect systematically funds performance. In addition, we cluster standard errors at the fund level. Since the objective is to test if Informed Alpha beats the standard Jensen's alpha, we restrict the sample to the set of Jensen Alpha defined as just to those funds with positive and significant alpha ($Alpha_{f,q}$) at the 10% level of significance.

Table 3.7 reports the results by estimating the regression in Eq 3.8, where the dependent variable is the future average of each of the following Realized Alpha over the following 12 months. Panel A reports the results using gross returns, while Panel B reports the results by using net returns. Model 1 reports the results without any controls. Model 2 reports the results by including: $Top_{f,q}$ and the interaction effect: $Alpha_{f,q} * Top_{f,q}$. Finally, model 3 includes controls which have been acknowledged in the literature to play a key role on future mutual funds performance. In Panel A, by using gross returns, in all three models, the variable $Alpha_{f,q}$ is positive and statistically significant at the 1% level of significance, which suggests that past significant alpha can predict future performance. In particular, 1% increase in past alpha predicts a 16.6 in basis points increase in future monthly realized alpha monthly. Furthermore, if we take a look at the coefficient that accompanies the variable of interest in model 3: $Alpha_{f,q} * Top_{f,q}$, the coefficient is statistically significant at the 1% level of significance, which indicates that increases in past alpha is associated with stronger gross future performance for those informed funds located in the top decile of $\%INSPC$ than the other funds. In particular, an increase of 1% in gross Alpha will lead funds in the top decile of $\%INSPC$ to have higher future performance by 19.0 in basis points on a monthly basis in comparison to the rest of the funds. In economic terms, one standard deviation increase in gross Alpha will lead funds in the top decile of $\%INSPC$ to have higher future performance by 0.089 standard deviation. Similar results are obtained when we consider net returns instead of gross returns. This suggest that Informed Alpha predicts future performance, and informed funds do not just perform better than the rest of the funds, but their outstanding performance is persistent.

Insert Table 3.7 here

3.6 Conclusions

In this paper we test performance evaluation in the context of the REE Paradigm. Consistent with the theory, we provide robust evidence that funds that implement trading strategies more similar to insiders exhibit outstanding performance. In particular, we proxy for private information with the trading done by insiders. In line with this, we define the variable *Insiders Purchases Coincidence* ($\%INS\,PC$) as the fraction (percentage) of times the fund agrees with insiders when the latter make informative purchase of stocks in the fund's portfolio, during a given period of time. We show that the portfolio of funds in the top decile of the distribution of $\%INS\,PC$ perform much better than those in the bottom decile. By estimating predictive regressions, we show that an increase of 1% in $\%INS\,PC$ is associated with an increase in future performance of 1.5 in basis points on a monthly basis.

Based on the previous results we introduce a new metric called *Informed Alpha* which represents a major improvement over alpha in the identification of truly talented managers. We define *Informed Alpha* as the set of funds of unconditional alphas (those with positive and significant alphas) in intersection with those funds in the 10th percentile of the distribution of $\%INS\,PC$. We show that, unlike unconditional alpha, *Informed Alpha* exhibits strong persistence.

This is the first attempt to test the performance evaluation implications of REE models; in future research we will extend the set of informational or miss pricing events beyond insiders interventions to capture the model implications more comprehensively.

References

O. Akin, J. M. Marín, and a. J.-L. Peydró. Anticipating the financial crisis: Evidence from insider trading in banks. *Economic Policy*, 35(102):213–267, Apr. 2020. ISSN 0266-4658. doi: 10.1093/epolic/eiaa012.

J. B. Berk and J. H. van Binsbergen. Measuring skill in the mutual fund industry. *Journal of Financial Economics*, 118(1):1–20, Oct. 2015. ISSN 0304-405X. doi: 10.1016/j.jfineco.2015.05.002.

M. M. Carhart. On Persistence in Mutual Fund Performance. *The Journal of Finance*, 52(1): 57–82, 1997. ISSN 1540-6261. doi: 10.1111/j.1540-6261.1997.tb03808.x.

L. Cohen, C. Malloy, and L. Pomorski. Decoding Inside Information. *The Journal of Finance*, 67(3):1009–1043, 2012. ISSN 1540-6261. doi: 10.1111/j.1540-6261.2012.01740.x.

P. Colla and J. M. Marin. Performance Evaluation in Competitive REE Models. *SSRN Electronic Journal*, 2010. ISSN 1556-5068. doi: 10.2139/ssrn.1693216.

J. Coval and E. Stafford. Asset fire sales (and purchases) in equity markets. *Journal of Financial Economics*, 86(2):479–512, Nov. 2007. ISSN 0304-405X. doi: 10.1016/j.jfineco.2006.09.007.

K. J. M. Cremers and A. Petajisto. How Active Is Your Fund Manager? A New Measure That Predicts Performance. *Review of Financial Studies*, 22(9):3329–3365, Sept. 2009. ISSN 0893-9454, 1465-7368. doi: 10.1093/rfs/hhp057.

K. Daniel, M. Grinblatt, S. Titman, and R. Wermers. Measuring Mutual Fund Performance with Characteristic-Based Benchmarks. *The Journal of Finance*, 52(3):1035–1058, July 1997. ISSN 00221082. doi: 10.1111/j.1540-6261.1997.tb02724.x.

E. F. Fama and K. R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, Feb. 1993. ISSN 0304405X. doi: 10.1016/0304-405X(93)90023-5.

S. Grossman. On the Efficiency of Competitive Stock Markets Where Trades Have Diverse Information. *The Journal of Finance*, 31(2):573–585, 1976. ISSN 0022-1082. doi: 10.2307/2326627.

S. J. Grossman. An Introduction to the Theory of Rational Expectations Under Asymmetric Information. *The Review of Economic Studies*, 48(4):541–559, 1981. ISSN 0034-6527. doi: 10.2307/2297195.

S. J. Grossman and J. E. Stiglitz. On the Impossibility of Informationally Efficient Markets. *The American Economic Review*, 70(3):393–408, 1980. ISSN 0002-8282.

B. D. Grundy and M. McNichols. Trade and the Revelation of Information through Prices and Direct Disclosure. *The Review of Financial Studies*, 2(4):495–526, 1989. ISSN 0893-9454.

G. Hoberg, N. Kumar, and N. Prabhala. Mutual Fund Competition, Managerial Skill, and Alpha Persistence. *The Review of Financial Studies*, 31(5):1896–1929, May 2018. ISSN 0893-9454, 1465-7368. doi: 10.1093/rfs/hhx127.

S. Huddart, B. Ke, and C. Shi. Jeopardy, non-public information, and insider trading around SEC 10-K and 10-Q filings. *Journal of Accounting and Economics*, 43(1):3–36, Mar. 2007. ISSN 0165-4101. doi: 10.1016/j.jacceco.2006.06.003.

M. C. Jensen. The Performance of Mutual Funds in the Period 1945-1964. *The Journal of Finance*, 23(2):389–416, 1968. ISSN 0022-1082. doi: 10.2307/2325404.

M. Kacperczyk and A. Seru. Fund Manager Use of Public Information: New Evidence on Managerial Skills. *The Journal of Finance*, 62(2):485–528, Apr. 2007. ISSN 00221082. doi: 10.1111/j.1540-6261.2007.01215.x.

M. Kacperczyk, C. Sialm, and L. Zheng. Unobserved Actions of Mutual Funds. *Review of Financial Studies*, 21(6):2379–2416, Nov. 2008. ISSN 0893-9454, 1465-7368. doi: 10.1093/rfs/hhl041.

R. Karapandza and J. M. Marin. The Rate of Market Efficiency. 2014.

O. Kim and R. E. Verrecchia. Trading Volume and Price Reactions to Public Announcements. *Journal of Accounting Research*, 29(2):302–321, 1991. ISSN 0021-8456. doi: 10.2307/2491051.

A. S. Kyle. Continuous Auctions and Insider Trading. *Econometrica*, 53(6):1315–1335, 1985. ISSN 0012-9682. doi: 10.2307/1913210.

J. Lakonishok and I. Lee. Are Insider Trades Informative? *Review of Financial Studies*, 14(1): 79–111, Jan. 2001. ISSN 0893-9454, 1465-7368. doi: 10.1093/rfs/14.1.79.

L'. Pástor and R. F. Stambaugh. Mutual fund performance and seemingly unrelated assets. *Journal of Financial Economics*, 63(3):315–349, Mar. 2002. ISSN 0304405X. doi: 10.1016/S0304-405X(02)00064-8.

R. Radner. Rational Expectations Equilibrium: Generic Existence and the Information Revealed by Prices. *Econometrica*, 47(3):655–678, 1979. ISSN 0012-9682. doi: 10.2307/1910413.

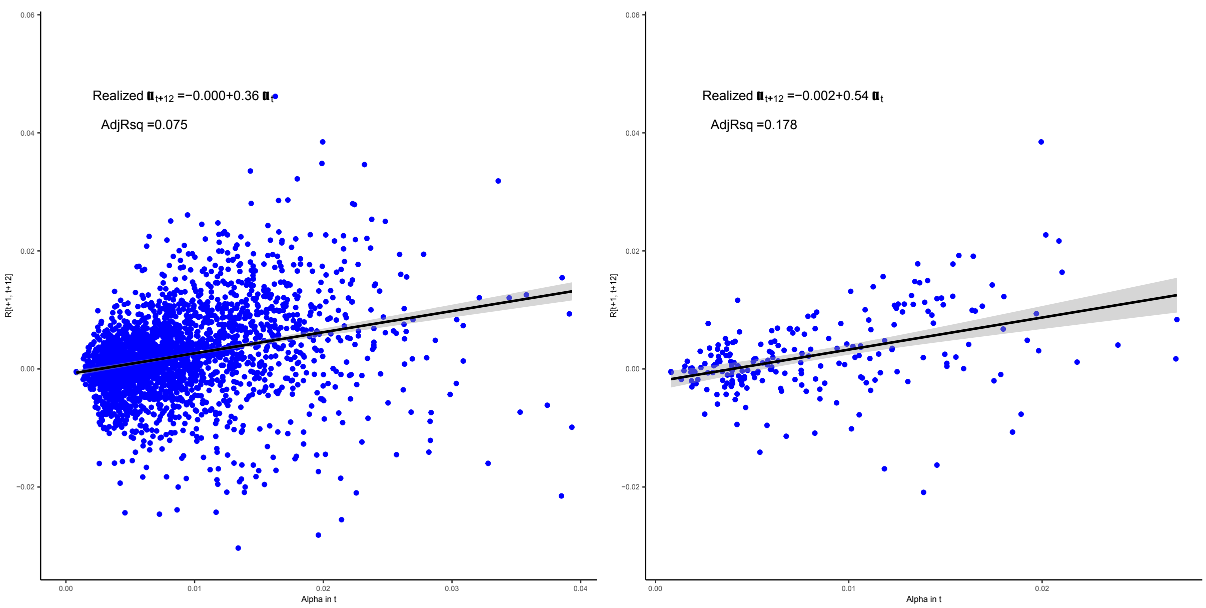
H. N. Seyhun. The Effectiveness of the Insider-Trading Sanctions. *Journal of Law & Economics*, 35(1):149–182, 1992.

T. M. Whited and G. Wu. Financial Constraints Risk. *The Review of Financial Studies*, 19(2): 531–559, July 2006. ISSN 0893-9454. doi: 10.1093/rfs/hhj012.

Figures and Tables

Figure 3.1: Informed Alpha

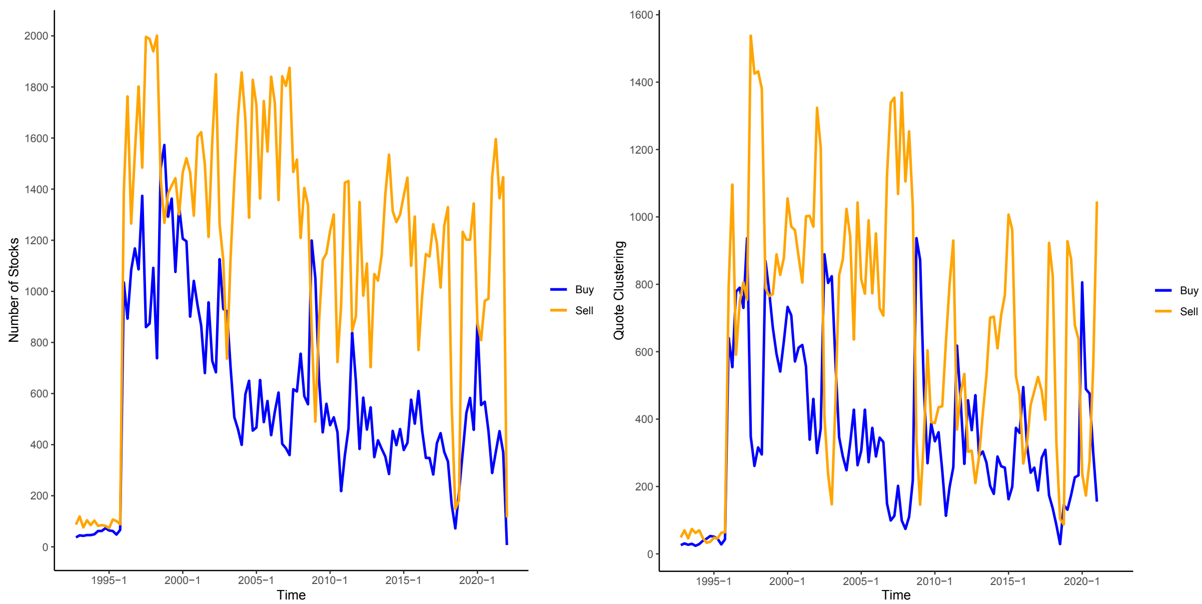
Panel (a) Jensen Alpha plots mutual funds past signficant Jensen alphas (x-axis) and the future realized alpha over the following 12 months (y-axis). Panel (b) plots mutual funds past significant Informed Alpha (x-axis) and the future realized alpha over the following 12 months (y-axis).



(a) Jensen Alpha **(b) Informed Alpha**

Figure 3.2: Number of stocks traded by Insiders quarterly

Figure 1 plots the number of stocks traded by insider traders regardless of their nature. Panel A, plots all the stocks traded by insider traders regardless of their nature. Panel B, plots just the number of stocks traded by informed insider traders.



(a) Stocks traded by all insider traders **(b) Stocks traded by Informed Insider Traders**

Table 3.1: Summary Statistics

	mean	Std.Dev.	Q0.05	Q0.25	Q0.5	Q0.75	Q.095
Size	1753.07	6277.14	22.90	108.30	361.00	1214.40	6706.36
Age	16.78	14.00	4.28	7.80	12.75	20.00	46.75
Expense Ratio	0.01	0.00	0.01	0.01	0.01	0.02	0.02
Turnover Ratio	0.82	0.84	0.12	0.35	0.63	1.05	2.07
Flow	0.00	0.26	-0.13	-0.05	-0.02	0.02	0.16
Net Returns	0.64	5.40	-8.52	-2.16	1.03	3.85	8.38
Gross Returns	0.74	5.40	-8.41	-2.06	1.13	3.95	8.48

This table reports the summary statistics for the sample of mutual funds (number of unique funds: 2089). Fund Size is the total net asset (TNA) reported at the end of each quarter. Age is the number of years since the inception of the mutual fund. Expense Ratio (Turnover Ratio) is the expense ratio (turnover ratio) reported in the CRSP database. For those cases in which the expense ratio (turnover ratio) is missing, we follow Fama and French (2010) and we attribute the expense ratio (turnover ratio) of other active mutual funds with similar total net assets. Flow are the mutual funds flows, calculated using the methodology used by Coval and Stafford (2007). Net Returns are the monthly returns reported by the CRSP database, while Gross Returns is calculated as the Net Return plus 1/12*Expense Ratios. For each variable, we report the mean (Mean), the standard deviation (Std. Dev.), the percentile 5% (Q.0.05), the percentile 25% (Q0.25), the median (Q0.5), the percentile 75% (Q0.75) and the percentile 95% (Q0.95). The time span is 1996 to 2016.

Table 3.2: Table: Proxy for informed mutual funds

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
INSPC	0.01	0.07	0.14	0.2	0.29	0.39	0.52	0.72	1.05	3.04

This table reports the mean of %INSPC obtained by forming portfolios based on the measure: % INSPC. Using all the funds in the final sample (number of unique funds: 2089): at the end of every quarter, form deciles based on the measure % INSPC. %INSPC is defined as the fraction (percentage) of times the fund agrees with informed insiders when the latter purchase stocks in the fund portfolio, during a given period of time. Informed insider traders are those characterized by a positive return after 18 months. Each decile is a portfolio and it is updated on a quarterly basis. We report the mean of the %INSPC for each decile.

Table 3.3: Portfolio Based Analyses

(a) Panel A: Gross Returns

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D10-D1
Average	0.65	0.71	0.67	0.72	0.70	0.71	0.77	0.80	0.80	0.85	0.2**
CAPM Alpha	0.00	0.05	0.01	0.05	0.04	0.05	0.10	0.13	0.13	0.17	0.17**

(b) Panel B:Net Returns

Average	0.55	0.60	0.57	0.62	0.60	0.61	0.67	0.70	0.70	0.76	0.21***
CAPM Alpha	-0.11	-0.06	-0.10	-0.05	-0.06	-0.06	0.00	0.03	0.04	0.08	0.19**

This table reports the returns obtained by forming portfolios based on the measure: % INSPC. Using all the funds in the final sample (number of unique funds: 2089): at the end of every quarter, form deciles based on the last 4 quarters of the measure %INSPC. %INSPC is defined as the fraction (percentage) of times the fund agrees with informed insiders when the latter purchase stocks in the fund portfolio, during a given period of time. Informed insider traders are those characterized by a positive return after 18 months. Each decile is a portfolio and it is updated on a quarterly basis. For each portfolio, in each decile we calculate the monthly equally weighted return over the next 3 months. We report: the Average, which stands for the average return of equally weighted gross and net returns, and CAPM alpha: which stands for the risk adjusted returns obtained by estimating the Capital Asset Pricing Model (CAPM). The results are expressed in percentages. We consider both gross returns and net returns. The difference between decile 1 and decile 10 is the return obtained by a zero investment in which we would short sell decile 1 and buy decile 10. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 3.4: Predictive Regressions

	Panel A: Realized Gross Alpha (t+1)			Panel B: Realized Net Alpha (t+1)		
	1	2	3	1	2	3
% INSPC	0.012***	0.012***	0.015***	0.012***	0.009***	0.015***
	[0.003]	[0.003]	[0.003]	[0.003]	[0.003]	[0.003]
Alpha		0.201***	0.212***		0.199***	0.212***
		[0.014]	[0.015]		[0.014]	[0.015]
SIZE			-0.000***			-0.000***
			[0.000]			[0.000]
log(Age)			0.000***			0.000***
			[0.000]			[0.000]
Expense Ratio			-0.084***			-0.006
			[0.015]			[0.015]
Turnover Ratio			-0.000			-0.000
			[0.000]			[0.000]
Flow			-0.000			-0.000
			[0.000]			[0.000]
sd Flow			-0.000***			-0.000***
			[0.000]			[0.000]
Constant	-0.009***	-0.010***	-0.007***	-0.009***	-0.009***	-0.007***
	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
NoObs	85,168	85,168	85,168	85,168	85,168	85,168
Adj. R-squared	0.0883	0.0918	0.0928	0.0883	0.0925	0.0934

Table 3.4: Predictive Regressions *(continued)*

	1	2	3	1	2	3

This table reports the coefficients obtained by estimating predictive regressiones. Realized alphas over the following quarter has beeen estimated on %INSPC and on a set of controls. The realized alphas are calculated as the difference between the realized return in excess of the risk-free rate and the expected excess return estimated from one factor model: the market. The expected excess return are estimated from one factor model, by running rolling-window time series regression over the previous 36 months. % INSPC is defined as the fraction (percentage) of times the fund agrees with informed insiders when the latter purchase stocks in the fund portfolio, during a given period of time. Informed insider traders are those characterized by a positive return after 18 months. Alpha is estimated using the one factor model: market over the previous 36 months. In addition, we include: Size: Total Net Asset (TNA), Age: number of years since the inception of the fund, Expense Ratios, Turnover Ratios. For those cases in which the Expense Ratios and Turnover Ratio are missing, we attribute the expense ratios and turnover ratio, respectively of other active mutual funds with similar total net assets. Furthemore, we control for: flow which are the mutual funds flows: calculated using the methodology used by Coval and Stafford (2007), and sd Flow: calculated as the standard deviation of the Flows over the previous 36 months. Standard errors are reported below the estimated coefficients and are clustered at the mutual funds level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 3.5: Portfolio Based Analyses Persistency

(a) Panel A: Gross Returns

	6	9	12	15	18	21	24	36
Average	0.2***	0.19***	0.17***	0.16***	0.15***	0.14***	0.13***	0.11***
CAPM Alpha	0.17***	0.16***	0.14***	0.13***	0.12***	0.11***	0.099***	0.08***

(a) Panel B:Net Returns

Average	0.22***	0.2***	0.19***	0.17***	0.16***	0.15***	0.15***	0.13***
CAPM Alpha	0.19***	0.18***	0.16***	0.14***	0.13***	0.12***	0.11***	0.095***

This table presents long-term differences in performance between more informed funds relative to less informed funds. Using all the funds in the final sample (number of unique funds: 2089): at the end of every quarter, we form deciles based on the last 4 quarters of the measure % INSPC. % INSPC is defined as the fraction (percentage) of times the fund agrees with informed insiders when the latter purchase stocks in the fund portfolio, during a given period of time. Informed insider traders are those characterized by a positive return after 18 months. Each decile is a portfolio and it is updated on a quarterly basis. The holding period for each portfolio varies from 3 months to 36 months, indicated by: 3, 6, 9, 12, 15, 18, 21, 24, 36. For each portfolio, in each decile we calculate the monthly equally weighted return over the different time horizons assumed. In the table we report the difference between decile 10 and decile 1, using either the Average or the CAPM alpha. Where, Average stands for the average of monhtly equally weighted gross and net returns, and CAPM Alpha stands for the risk adjusted returns obtained by estimating the Capital Asset Pricing Model (CAPM). The results are expressed in percentages. We consider both gross (Panel A: Gross Returns) and net (Panel B: Net Returns) returns. The difference between decile 1 and decile 10 is the return obtained by a zero investment in which we would short sell decile 1 and buy decile 10. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 3.6: Persistance

	Panel A: Realized Gross Alpha (t+1)			Panel B: Realized Net Alpha (t+1)		
	1	2	3	1	2	3
High	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Medium-High	0.000	0.000	0.000**	0.000*	0.000*	0.000**
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Medium-Low	0.000	-0.000	0.000	0.000	0.000	0.000
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Alpha		0.095***	0.108***		0.097***	0.108***
		[0.034]	[0.034]		[0.034]	[0.035]
Alpha*High		-0.046	-0.048		-0.050	-0.048
		[0.045]	[0.045]		[0.045]	[0.045]
Alpha*Medium-High		-0.049	-0.053		-0.051	-0.053
		[0.043]	[0.043]		[0.043]	[0.043]
Alpha* Medium-Low		-0.027	-0.028		-0.028	-0.027
		[0.038]	[0.038]		[0.038]	[0.038]
SIZE			-0.000***			-0.000***
			[0.000]			[0.000]
log(Age)			0.000***			0.000***
			[0.000]			[0.000]
Expense Ratio			0.014			-0.063***
			[0.018]			[0.018]
Turnover Ratio			-0.000			-0.000
			[0.000]			[0.000]
Flow			-0.000			-0.000
			[0.000]			[0.000]
sd Flow			-0.000***			-0.000***
			[0.000]			[0.000]
Constant	-0.007***	-0.007***	-0.006***	-0.008***	-0.008***	-0.006***
	[0.000]	[0.000]	[0.001]	[0.000]	[0.000]	[0.001]

Table 3.6: Persistance *(continued)*

	1	2	3	1	2	3
NoObs	79,423	79,423	79,423	79,423	79,423	79,423
R-squared	0.110	0.112	0.115	0.108	0.110	0.112

This table reports the coefficients obtained by estimating predictive regressiones. Realized alphas over the following quarter have beeen estimated on dummy variabls which reflects the level of %INSPC: High, Medium-High, Medium-Low and interaction effects between: Alpha*High, Alpha*Medium-High, Alpha*Medium-Low and on a set of controls. The realized alphas are calculated as the difference between the realized return in excess of the risk-free rate and the expected excess return estimated from one factor model: the market. The expected excess return are estimated from one factor model, by running rolling-window time series regression over the previous 36 months. Thus the dependent variable is the average of each of the following three months of the realized alphas. The dummy variables reflect the level of %INSPC, such that we sort the variable of interest: % INSPC in four groups. Based on the four groups, we construct four dummy variables. Where, % INSPC is defined as the fraction (percentage) of times the fund agrees with informed insiders when the latter purchase stocks in the fund portfolio, during a given period of time. Informed insider traders are those characterized by a positive return after 18 months. Alpha is estimated using the one factor model: market over the previous 36 months. In addition, we include: Size: Total Net Asset (TNA), Age: number of years since the inception of the fund, Expense Ratios, Turnover Ratios. For those cases in which the Expense Ratios and Turnover Ratio are missing, we attribute the expense ratios and turnover ratio, respectively of other active mutual funds with similar total net assets. Furthermore, we control for: flow are the mutual funds flows: calculated using the methodology used by Coval and Stafford (2007), and sd Flow: calculated as the standard deviation of the Flows over the previous 36 months. Standard errors are reported below the estimated coefficients and are clustered at the mutual funds level. *** p<0.01, ** p<0.05, * p<0.1.

Table 3.7: Informed Alpha

	Panel A: Realized Gross Alpha (t+1)			Panel B: Realized Net Alpha (t+1)		
	1	2	3	1	2	3
Alpha	0.165***	0.139***	0.166***	0.155***	0.130***	0.165***
	[0.041]	[0.044]	[0.043]	[0.041]	[0.044]	[0.043]
TopINSPC		-0.002***	-0.001***		-0.002***	-0.001***
		[0.000]	[0.000]		[0.000]	[0.000]
Alpha*TopINSPC		0.206***	0.190***		0.204***	0.191***
		[0.052]	[0.052]		[0.053]	[0.052]
SIZE			-0.000***			-0.000***
			[0.000]			[0.000]
log(Age)			0.001***			0.001***
			[0.000]			[0.000]
Expense Ratio			-0.019			-0.096***
			[0.036]			[0.036]
Turnover Ratio			-0.000*			-0.000*
			[0.000]			[0.000]
Flow			-0.000			-0.000
			[0.000]			[0.000]
sd Flow			-0.001***			-0.001***
			[0.000]			[0.000]
Constant	-0.005***	-0.005***	-0.003**	-0.006***	-0.006***	-0.003**
	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
NoObs	9,054	9,054	9,054	9,054	9,054	9,054
Adj. R-squared	0.295	0.297	0.303	0.291	0.293	0.300

Table 3.7: Informed Alpha *(continued)*

	1	2	3	1	2	3

This table reports the coefficients obtained by estimating predictive regressiones. Realized alphas over the following 12 months has beeen estimated on %INSPC and on a set of controls. The realized alphas are calculated as the difference between the realized return in excess of the risk-free rate and the expected excess return estimated from one factor model: the market. The expected excess return are estimated from one factor model, by running rolling-window time series regression over the previous 36 months.% INSPC is defined as the fraction (percentage) of times the fund agrees with informed insiders when the latter purchase stocks in the fund portfolio, during a given period of time. Informed insider traders are those characterized by a positive return after 18 months. $Top_{f,q}$ is a dummy variable that takes the value of one if the fund f in quarter q belongs to the top decile of the distribution of %INSPC and 0 otherwise. Alpha is estimated using the one factor model: market over the previous 36 months. $Top_{f,q}$*Alpha standard for the interaction effect between: $Top_{f,q}$ and Alpha. In addition, we include: Size: Total Net Asset (TNA), Age: number of years since the inception of the fund, Expense Ratios, Turnover Ratios. For those cases in which the Expense Ratios and Turnover Ratio are missing, we attribute the expense ratios and turnover ratio, respectively of other active mutual funds with similar total net assets. Furthermore, we control for: flow are the mutual funds flows: calculated using the methodology used by Coval and Stafford (2007), and sd Flow: calculated as the standard deviation of the Flows over the previous 36 months.. Standard errors are reported below the estimated coefficients and are clustered at the mutual funds level. *** p<0.01, ** p<0.05, * p<0.1.