Journal of Financial Economics

The retail execution quality landscape☆

Anne Haubo Dyhrberg [a], Andriy Shkilko [a,b], Ingrid M. Werner [c,d,*]

[a] Wilfrid Laurier University, Canada
[b] Macquarie University, Australia
[c] The Ohio State University, United States of America
[d] CEPR, United Kingdom

ARTICLE INFO

JEL classification:
G20
G24
G28

Keywords:
Retail trading
Wholesalers
Execution quality

ABSTRACT

We demonstrate that off-exchange (wholesaler) executions provide significant cost savings to retail investors. Wholesaler concentration has raised regulatory concerns; however, we show that the largest wholesalers offer the lowest costs due to economies of scale. The entry of a new large wholesaler reduces incumbent scale economies, resulting in higher execution costs. Most retail brokers route to multiple wholesalers and actively monitor their performance, rewarding those offering lower execution costs with more volume. While retail investors benefit from the current landscape across all stocks, those trading small stocks benefit the most.

1. Introduction

Academics (e.g., Autor et al., 2017; Grullon et al., 2019; Autor et al., 2020), journalists (Economist, 2016), and policy makers (CEA, 2016) point to increasing levels of market concentration throughout the economy, raising concerns that consumers pay too much for goods and services. These concerns are warranted if large firms have pricing power, but less so if concentration results from a competitive process where success depends heavily on efficiency and innovation. As a result of this process, firms that better serve their customers may gain larger market shares, leading to long-run efficiency gains and more favorable consumer prices (Demsetz, 1973; Peltzman, 1977; Focarelli and Panetta, 2003).

In this paper, we provide the first comprehensive empirical analysis of intermediation in the U.S. market for retail order flow. The market is concentrated by conventional measures, but our evidence suggests that concentration allows intermediaries to benefit from economies of scale that are, in turn, passed through to retail customers in the form of lower execution costs. The transfer is facilitated by retail brokers, who continuously monitor execution quality and reallocate flows to better-performing intermediaries (wholesalers). Entry by a sizable new wholesaler reduces the economies of scale for the incumbents, and execution costs increase.

In the United States, the trading volume generated by retail investors represents close to 20% of equity trading volume (Saul, 2023). In the 1990s, many U.S. retail brokers executed client orders in-house through their own market-making businesses. However, market developments and regulatory initiatives of the early 2000s that spurred increased competition for liquidity provision prompted them to exit these businesses. Instead, brokers began outsourcing order execution

☆ We thank Dimitris Papanikolaou (the co-editor), the anonymous referee, Robert Battalio, James Brugler, Sabrina Buti, Doug Clark, Carole Comerton-Forde, Laurence Daures, Amy Edwards, Greg Eaton, Tom Ernst, Marinela Finta, Corey Garriott, Carole Gresse, Peter Haynes, David Hecht, Bob Jennings, Travis Johnson, S. P. Kothari, Phil Mackintosh, Thomas Marta, Josh Mollner, Pam Moulton, Dmitriy Muravyev, Fabricio Perez, Peter Reiss, Julia Reynolds, Chris Schwarz, Vincent Skiera, Wendi Wu, Pradeep Yadav, Chen Yao, Marius Zoican, and conference/seminar participants at American Finance Association, Brock University, Case Western Reserve University, Central Bank Conference on the Microstructure of Financial Markets, Chicago Quantitative Alliance, China International Conference in Finance, Columbia University Workshop on Handling of Retail Orders, European Finance Association, Financial Management Association, Indiana University, Macquarie University, McMaster University, NBER Big Data and High-Performance Computing for Financial Economics Conference, Northern Finance Association, SEC Conference on Financial Market Regulation, Toronto Stock Exchange, University of Georgia, University of Graz, University of Memphis, University of Mississippi, University of Toronto, Université Paris Dauphine-PSL, Wilfrid Laurier University, and Women in Microstructure Meeting for valuable comments.

Received 7 June 2024; Received in revised form 19 March 2025; Accepted 20 March 2025

Available online 3 April 2025

to vertically integrated over-the-counter market making firms known as wholesalers, whom they believe to have the capital, scale, and expertise necessary to most effectively serve their retail clients (Schwab, 2022). Fast forward to today, the majority of retail brokers send their customer flow to wholesalers who internalize most liquidity demanding orders by buying from retail sellers and aiming to re-sell to retail buyers, capturing the bid–ask spread. The wholesaler retains a portion of the spread, another portion is passed on to the retail trader as price improvement, and in some cases yet another portion goes to the retail broker as payment for order flow (PFOF).

How retail orders are currently handled has been actively debated.[1] Many observers, including top-ranking Securities and Exchange Commission (SEC) officials, argue that the wholesaler business is too concentrated and therefore provides limited benefits to retail investors. These observers suggest that price improvement offered by wholesalers tends to be *de minimis*, or the smallest amount possible. Others (often retail investors) are opposed to brokers receiving PFOF, arguing that it distorts best execution incentives. To enhance competition for retail orders, the SEC has proposed to implement a system of auctions in which liquidity providers would compete for each individual retail order, obtaining execution rights only if they provide the largest amount of price improvement (SEC, 2022). In comments to the SEC, both brokers and wholesalers assert that wholesalers already compete vigorously for the opportunity to execute retail order flow, and brokers route orders to wholesalers in such a way as to serve the best interest of their customers.

We undertake an empirical task of shedding light on these views, doing so using four years of SEC Rule 605 reports by all major wholesalers covering more than 12,000 U.S. National Market System (NMS) securities from 2019 through 2022. Each order-handling venue, wholesalers and exchanges alike, must file such reports on a monthly basis to maintain a public record of execution quality. Our analyses show that wholesalers provide substantial price improvement, executing liquidity-demanding orders at prices better than those quoted on exchanges. Wholesaler price improvement is far from *de minimis*, amounting to 27% of the quoted spread in the full sample. Even more remarkable is that retail traders in an average S&P 500 stock receive price improvement amounting to 51% of the quoted spread.

By comparison, based on disclosures by all major U.S. retail brokers pursuant to the SEC Rule 606, we estimate that PFOF paid by wholesalers to brokers that accept such payments represents only 1% of the quoted spread for the average security. Moreover, brokers that accept PFOF charge all wholesalers the same PFOF rate, presumably to address concerns about conflicts of interest in their routing decisions. In other words, a broker's PFOF revenue does not depend on how it allocates order flow across wholesalers. It is also notable that several large brokers such as Fidelity and Vanguard do not accept PFOF, but still route extensively to wholesalers to satisfy their best execution obligations.

Brokers often evaluate wholesaler performance based on price improvement relative to the quoted spread. When a broker compares execution quality across wholesalers, any broker-specific differences in order flow characteristics, such as order flow toxicity generated by the broker's average customer, are inherently held constant.[2] Therefore, a toxicity-unadjusted metric, such as price improvement, is appropriate for a broker to use when comparing wholesalers. By contrast, price improvement does not adequately capture the economics of intermediation in a dataset like ours, which consists of filings by individual wholesalers. The reason is that, as we show, the association between order flow and future price moves differs significantly across brokers. Legacy brokers such as E*TRADE, Schwab, and TD Ameritrade generate more toxic flow, while newer brokers such as Robinhood and Webull generate less toxic flow. Rule 606 reports show that each wholesaler receives a different mix of broker flows, resulting in significant differences in toxicity across wholesalers. For example, Rule 605 data reveal that the two largest wholesalers, Citadel Securities and Virtu Financial, receive the most toxic flow. With this in mind, we argue that the appropriate measure to compare performance across wholesalers in our data is a toxicity-adjusted spread (realized spread).[3]

While the sizable price improvement and the effectively *de minimis* PFOF rates support the view that retail investors benefit from the current market structure, the data also reveal that the wholesale business is concentrated, with the two largest wholesalers capturing close to 70% of retail flow. With this level of concentration, it is not surprising that some express the concern that wholesalers may leverage their size in negotiations with brokers, resulting in high costs for retail investors. While we cannot directly refute the occurrence of such behavior, the data provide several indicators that help alleviate this concern.

To start with, contrary to the notion that large wholesalers might charge high execution costs, we find that Citadel and Virtu actually charge less than their smaller competitors. Furthermore, we find that the scale of operations explains the entirety of the difference in execution cost between the top two and the other wholesalers. This suggests that economies of scale generate cost savings at the wholesaler level, and these savings tend to be passed through to retail customers.

Why do the top two wholesalers pass the savings through instead of using them to boost their revenue? We show that the broker side of the market is also rather concentrated. For instance, TD Ameritrade Holding Corporation, even prior to its recent acquisition by Charles Schwab, represents over 47% of retail order flow that we can estimate from Rule 606 disclosures. Other brokers also hold sizeable market shares leading us to believe that they may very much be in the driver's seat. Discussions with industry representatives also reveal that brokers expect wholesalers to execute all orders they receive, and that they are evaluated based on past performance. Indeed, the data are consistent with brokers carefully monitoring execution costs and rewarding superior wholesaler performance with more order flow.

We find that brokers often and significantly adjust order flow allocations across wholesalers, with those offering lower execution costs generally receiving a larger share. Notably, a typical broker appears to evaluate wholesalers not on a security-by-security basis but rather on a bundled basis. For instance, if Citadel offers the lowest execution costs in Apple, it will not necessarily receive more future Apple flow. Instead, Citadel must outperform its competitors across the entire range of securities, including smaller securities, to attract more order flow.

This finding highlights an intriguing aspect of the retail ecosystem, where the brokers compel wholesalers to compete in micro and small capitalization securities that have relatively low trading frequency and high inventory costs. Typically, low-volume securities are less attractive to intermediaries, and market regulators and exchanges often seek ways to improve liquidity in such securities. When we account for inventory costs, our analysis suggests that wholesalers tend to charge relatively low execution costs in micro and small capitalization securities compared to large securities, underscoring the role of broker-enforced bundling in boosting liquidity for smaller securities.

To end with, the dynamics of wholesaler competition undergo a transformation during our sample period with the entry of a new player, Jane Street Capital. Within a few months, Jane Street gains a sizable market share, capturing over 12% of retail flow. It is reasonable to expect that competitive pressures among wholesalers would intensify after this entry, leading to lower execution costs. Our difference-in-differences analyses, however, do not confirm this expectation. In

[1] See the Internet Appendix A.1 for sources.

[2] The term *toxicity* refers to the link between liquidity-demanding retail orders and the adverse price moves faced by a wholesaler after supplying liquidity to these orders.

[3] This measure is also the focus of the economic analysis of the SEC's proposed Order Competition Rule.

fact, execution costs increase, consistent with the incumbents' loss of economies of scale.

In summary, our data depict a landscape where brokers actively control execution quality for their clients' orders by routing to wholesalers that offer lower execution costs, resulting in cost savings for retail investors. The threat of entry creates additional pressure on wholesaler pricing, and a new entrant successfully captures a sizable market share in a surprisingly short time. Wholesale business appears to be characterized by economies of scale, with the largest wholesalers offering the lowest execution costs. The latter characteristic raises the question of why the market has not naturally gravitated toward an equilibrium with just one wholesaler. We posit that such an equilibrium would prove sub-optimal for brokers, as a monopolistic wholesaler would be challenging to control. Needless to say, having just one wholesaler would also entail significant operational risk. Consequently, the current state with several competing wholesalers resembles monopolistic competition, as we explain shortly, and maintains an intriguing balance allowing certain players to become large while retaining a competitive fringe to serve as a credible ongoing threat, discouraging any rogue behavior.

Our results highlight two considerations regarding the SEC's proposal to require retail orders to be sent to auctions for order-by-order competition. First, the proposal expects that non-professional liquidity providers would demonstrate significant interest in engaging with retail flow and offer superior price improvement compared to wholesalers. While our data indicate that this assumption might hold true for large securities, where the retail to non-retail ratio is 1:13, it is less likely to hold in the thousands of micro and small capitalization securities currently actively traded by retail investors, where trading is sparse and the ratio is only 1:2. Second, order-by-order auctions would eliminate bundling and are therefore likely to reduce the incentives for intermediaries to engage with retail traders in micro and small capitalization securities. We caution that many retail investors might experience lower execution quality if the proposal were to be implemented. Ernst et al. (2024) come to a similar conclusion based on a theoretical model.

2. Background

2.1. Market structure

To aid in the interpretation of our empirical results, we offer some background from the industrial organization literature. The wholesale market structure has several salient characteristics. First, the business is concentrated among relatively few firms, with two market leaders and a competitive fringe of remaining firms. This concentration suggests that fixed costs are likely significant, indicating the presence of economies of scale, as we will demonstrate shortly. Second, while we focus on one aspect of wholesaler services – execution costs – their services include other features we are unable to measure, such as size improvement, as shown by Battalio and Jennings (2023), reliability, capital commitment, etc. This implies that wholesaler services are differentiated. Third, while brokers actively monitor wholesalers and shift demand based on execution costs, they continue using multiple wholesalers, presumably to mitigate operational risk and enhance bargaining power. Lastly, entry to and exit from the wholesale business occur with some regularity suggesting that entry (sunk) costs are not prohibitive.[4]

Taken together, these observations suggest that the wholesale market exhibits many characteristics of monopolistic competition (Chamberlin, 1933; Dixit and Stiglitz, 1977; Asplund and Nocke, 2006). In such a setting, each wholesaler faces a downward-sloping demand curve and sets the price of its services to maximize profit. An increase in demand or innovations that reduce production costs could spur competitive entry, leading to broker demand being spread across more wholesalers, each with a smaller market share. In the absence of significant sunk costs, entry (or the threat of entry) and exit should eventually drive the price of wholesale services to equal average cost.

2.2. Inventory management

The following observations about the wholesale business are also useful to consider. Wholesalers assume the responsibility of achieving best execution for retail orders routed to them, which means they are evaluated based on the National Best Bid and Offer (NBBO). However, unlike in classic models, such as those by Amihud and Mendelson (1980) and Ho and Stoll (1981), where market makers manage inventory by adjusting their quotes, wholesalers have limited influence over the NBBO.[5] Since arrangements with brokers require wholesalers to accept all marketable orders routed to them, wholesalers tend to manage inventory imbalances by trading on other venues. This often involves forgoing the expected spread revenue from the order and incurring venue fees. Therefore, the profit maximization problem for a risk-averse wholesaler becomes an inventory management problem with costly control.

This problem is modeled formally by Huang et al. (2012) for the case of one security and by Song (2010) for multiple-securities.[6] The authors show that the optimal inventory policy is a threshold policy, where the upper and lower thresholds determine when to route out orders. A wholesaler sets these thresholds further apart, and hence spends less resources on controlling inventory, if her risk aversion is low and/or if the volatility of the security is low. The former helps explain why wholesaler scale is beneficial as risk aversion declines in wealth (e.g., Arrow, 1971; Paravisini et al., 2017). The latter aligns with the fact that wholesalers offer execution services in thousands of securities — managing inventory for a diversified portfolio is less costly than doing so at the individual security level.

2.3. Related literature

We analyze a four-year comprehensive public dataset at the monthly frequency that academic researchers have not examined under the current retail market structure.[7] According to industry participants, this dataset allows for the clearest view into retail execution quality that is possible without proprietary data. Since our paper was first made public, three new related working papers that use either proprietary or self-generated data have been circulated. Battalio and Jennings (2023) use proprietary data from one or more anonymous wholesalers in May 2022 to examine execution quality, while Ernst et al. (2023) and Huang et al. (2024) study broker routing and wholesaler performance. To do so, Ernst et al. (2023) use proprietary data from three retail brokers, while Huang et al. (2024) conduct a field experiment with self-generated odd-lot orders. We discuss our contributions relative to these working papers below. For a more in-depth discussion of the literature, we refer interested readers to the Internet Appendix A.2.

[4] Jane Street entered the equity wholesale business in July 2019 and in 2.5 years established itself as a sizable player. Wolverine Trading exited the business in February 2021. Hudson River Trading entered in July 2022 and receives orders from one retail broker as of the end of 2022.

[5] While Citadel and Virtu have substantial presence on all exchanges, industry estimates suggest that they provide less than 20% of exchange liquidity (Mackintosh, 2023). Internet Appendix A.7 contains results consistent with this notion, suggesting that much of exchange liquidity is supplied by non-professional liquidity providers trading via limit orders.

[6] A similar idea is in the two-period model by Ho and Stoll (1983) where dealers lay off inventory by crossing the spread to hit competing dealers' quotes, but in their model the dealer has influence over the NBBO.

[7] Several studies were written after the SEC mandated in 2001 that market centers publicly disclose execution quality metrics, known as Dash 5 reports (e.g., Bessembinder, 2003; Lipson, 2003; Boehmer, 2005; Boehmer et al., 2007). More recently, O'Hara and Ye (2011) study the impact of fragmentation on market quality using Rule 605 data.

Battalio and Jennings (2023) examine order-level data for a comprehensive cross-section of securities and demonstrate that their wholesaler(s) provided retail investors with substantial price and size improvements in May 2022. Therefore, our finding that wholesalers offer retail investors significant price improvements is reinforced by their more granular, though narrower in scope, dataset. Our primary contribution relative to Battalio and Jennings (2023) is the analysis of execution quality data for all eight major wholesalers over a four-year period. As we demonstrate, execution quality varies among wholesalers, with larger firms delivering superior outcomes, due to economies of scale. Such an analysis is only feasible with a comprehensive sample of wholesalers. Furthermore, our extended time frame allows us to explore the dynamics of competition in the provision of execution services and examine patterns in broker routing. A limitation of our study is that our data are monthly and do not include odd lots or short sales. Battalio and Jennings (2023) show that these account for 7.34% of retail share volume (2.01% in odd lots and 5.33% in short sales) and that including them amplifies the value of price improvements.[8] This suggests that we may be understating the value of the services wholesalers provide to retail investors.

Ernst et al. (2023) examine unique data on the metrics and methods used by three unnamed retail brokers in their order routing decisions. Data availability differs across brokers, spanning various months in the 2019–2023 period. Corroborating our findings, the data indicate that order routing responds to performance, with wholesalers that offer better execution quality receiving more future order flow. Beyond the contributions discussed in relation to Battalio and Jennings (2023), our contribution compared to Ernst et al. (2023) is that our data encompass all retail brokers, and therefore 100% of the retail flow routed to the eight major wholesalers, allowing us to draw more general inferences about how broker routing decisions are influenced by wholesaler performance.

Huang et al. (2024) conduct a field experiment by routing self-generated odd-lot orders for a sample of fewer than 150 common stocks to six retail brokers (one with two account types). Like us and Ernst et al. (2023), they find that retail brokers on average base their routing decisions on execution quality, routing orders to better performing wholesalers. However, when they focus on the four largest wholesalers, only two brokers route more of the authors' odd-lot orders to the wholesaler that offers superior performance, while two appear to reward poor performance. Our contribution relative to Huang et al. (2024) is to provide evidence from the full spectrum of retail brokers, showing that the average broker, routing orders in over 12,000 securities traded by retail investors, directs order flow based on the prior performance of wholesalers. In other words, the finding in Huang et al. (2024) that certain brokers route more of the authors' odd lots to underperforming wholesalers does not generalize to the typical broker's routing of non-odd-lot order sizes in our sample.

Due to the nature of their broker data and field experiment, neither Ernst et al. (2023) nor Huang et al. (2024) can measure toxicity, preventing them from comparing toxicity of order flow across brokers. Battalio and Jennings (2023) focus on one or more wholesaler(s) and also do not discuss order flow toxicity across brokers. Another contribution of our paper relative to these three working papers is therefore to demonstrate that the average toxicity of order flow varies across retail brokers and that the composition of broker flow differs among wholesalers, with the largest wholesalers receiving more toxic order flow. We also contribute to the literature by demonstrating that the brokers' routing strategies allow small caps and micro caps to benefit from relatively lower execution costs.

Table 1
Wholesaler market shares.

	Full sample [1]	S&P 500 [2]	Tercile 1 [3]	Tercile 2 [4]	Tercile 3 [5]
Citadel	39.22	38.20	38.92	40.43	39.71
Virtu	29.55	28.45	28.97	29.87	31.75
G1	13.36	14.30	13.89	12.46	12.04
Jane Street	5.89	5.16	5.78	6.10	6.65
Two Sigma	4.99	4.54	4.70	5.25	5.89
UBS	4.20	5.36	4.54	3.60	2.81
Merrill Lynch	1.81	2.41	1.96	1.66	1.01
Morgan Stanley	0.98	1.57	1.23	0.63	0.14

The table contains the list of 8 wholesalers that execute Rule 605 liquidity-demanding orders from 2019 through 2022. We report each wholesaler's market share for the full sample and for the four subsamples: S&P 500 and terciles 1 through 3.

3. Data

We obtain monthly order execution quality, routing, and payments data for all NMS securities traded in the U.S. from publicly available Rule 605 and Rule 606 reports. Each U.S. market center, including wholesalers, is required to file a monthly Rule 605 report detailing its execution quality on a security-by-security basis. Similarly, each brokerage must submit a monthly Rule 606 report outlining the payments it sends to and receives from market centers where it routes customer orders as well as the proportion of orders it routes to each market center. These reports are filed for two groups of securities, the S&P 500 stocks and all others.

The data are self-reported, but many of the wholesalers and brokers in our sample use independent third-party analytics providers to calculate the required statistics. Notably, the raw Rule 605 dataset includes over 4,000 securities that are not NMS securities and therefore not subject to Rule 605 reporting obligations (e.g., warrants and convertible bonds). This suggests that reporting entities likely submit their entire raw execution data in bulk to the analytics providers, rather than selectively altering the data before submission.

Most importantly, the economic analysis in the proposed Order Competition Rule states that Rule 605 reports are highly consistent with regulatory Consolidated Audit Trail (CAT) data, which is accessible to the SEC. Therefore, we believe it is unlikely that Rule 605 data suffer from self-reporting biases. Although we are unaware of comparable sources to verify the accuracy of Rule 606 data, two points are worth noting. First, many brokers use third-party analytics providers to construct their Rule 606 reports, similarly to how wholesalers handle Rule 605 reports. Second, FINRA regularly audits compliance with both Rule 605 and Rule 606 filings, and to our knowledge, none of the entities in our sample have been cited by FINRA for reporting inconsistencies.

3.1. Rule 605: Order execution quality

We collect Rule 605 execution quality data for a four-year period from January 2019 through December 2022.[9] We restrict our main sample to NMS securities that merge with monthly data from the Center for Research in Securities Prices (CRSP) for a total of 12,012 securities, including Exchange Traded Products. For simplicity, we use the terms *securities* and *stocks* interchangeably throughout the remainder of the manuscript and Internet Appendix. We then divide these into four subsamples. The first subsample includes the S&P 500 stocks, and the remaining three subsamples are formed from size-based terciles of non-S&P 500 stocks. The sizes (market capitalizations) are $600 million and above for tercile 1 firms, between $110 million and $600 million for tercile 2 firms, and below $110 million for tercile 3 firms. Based on FINRA's definition of market capitalization groups, tercile 1 includes

[8] We detail the calculation of these figures in Internet Appendix A.2.3.

[9] Internet Appendix A.3 contains additional details of this dataset.

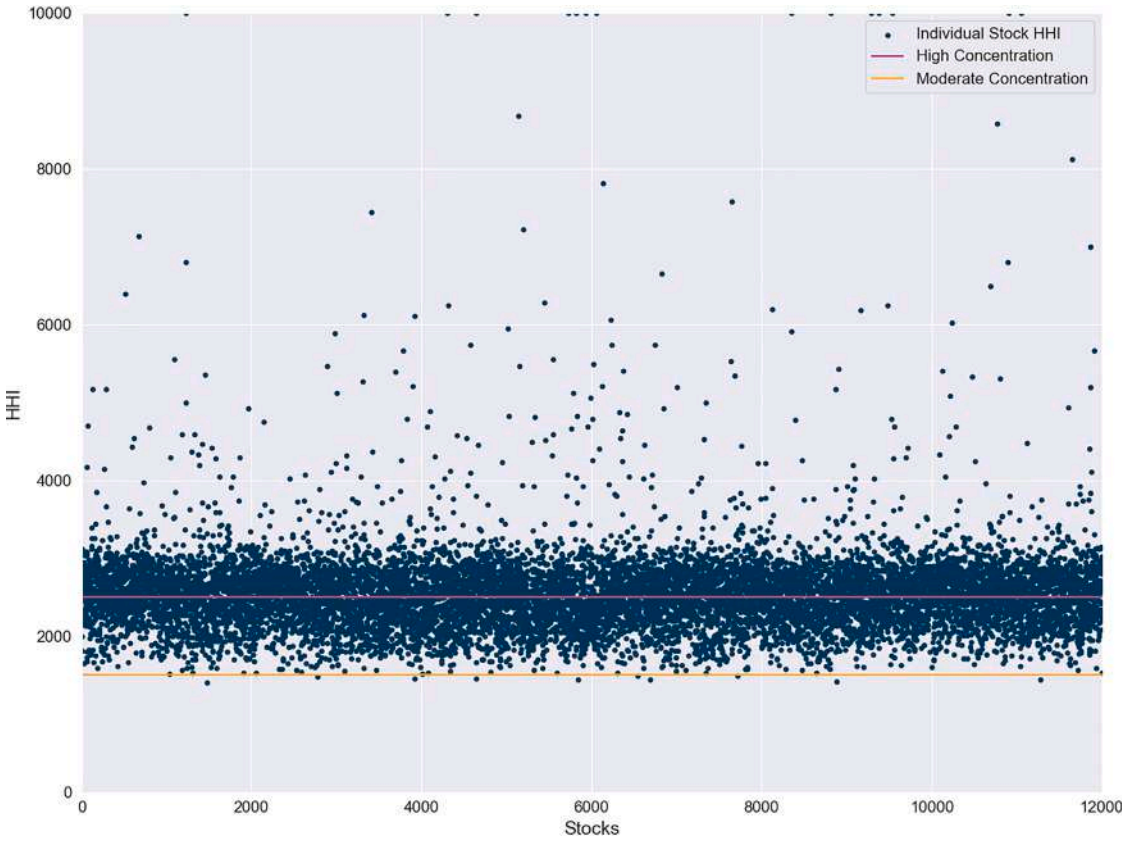

Fig. 1. Herfindahl–Hirschman Indices for Individual Stocks.
The figure plots as blue dots the Herfindahl–Hirschman Indices (HHIs) computed for each sample stock as

$$HHI = \sum_{j=1}^{8} s_j^2,$$

where s_j is the market share percentage of wholesaler j expressed as a whole number, not a decimal. The stocks are sorted alphabetically along the x-axis. The U.S. Department of Justice considers a marketplace (i) competitive if its HHI is less than 1,500 (the space below the yellow line), (ii) moderately concentrated if the HHI is between 1,500 and 2,500 (the space between the yellow and purple lines), and (iii) highly concentrated if the HHI is 2,500 or greater (the space above the purple line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(non-S&P 500) mega, large, medium, and the upper end of small caps, tercile 2 consists of the remaining small and the upper end of micro caps, and tercile 3 is populated by the remaining micro caps (FINRA, 2022).

We concentrate on the eight largest wholesalers. According to industry consensus, Rule 605 reports submitted by wholesalers contain predominantly retail orders, covering virtually all such orders, and are therefore the most comprehensive information source about retail execution quality. The reports cover execution quality for all orders wholesalers receive, not only the orders they internalize. As a result, if a wholesaler routes all or part of a retail order for execution elsewhere, the execution quality of the entire order will still be reflected in the wholesaler's Rule 605 report.

We focus on liquidity-demanding (market and marketable) orders because retail investors submit liquidity-providing (limit) orders relatively infrequently, with such orders comprising only about 12% of retail share volume. Limit orders are also handled quite differently by wholesalers; they are either routed directly to exchanges or executed by wholesalers under the *no knowledge* exemption (FINRA Rule 5320.02) or the *riskless principal* exemption (FINRA Rule 5320.03). In contrast, wholesalers are expected to promptly execute liquidity-demanding orders, and their execution quality is benchmarked against the NBBO at the time the order is received.

Rule 605 applies to orders ranging from 100 to 9,999 shares that are executed during regular trading hours (SEC, 2000). As a result, the data exclude odd lots. Some brokers and wholesalers, as part of the Financial Information Forum (FIF) initiative, have periodically reported quarterly odd-lot execution quality statistics on their websites (e.g., Citadel Securities (2019)). These reports suggest that retail odd-lot price improvements exceed those reported for orders in the 100 to 499 share range, which is the smallest category required for Rule 605 reporting.[10] Therefore, Rule 605 data somewhat underestimate the market quality delivered by wholesalers.

Table 1 reports the market shares of the eight wholesalers. Citadel and Virtu dominate, together accounting for almost 70% of all retail flow. The remaining wholesalers capture smaller market shares, ranging from 13% held by G1 to less than 1% by Morgan Stanley. The wholesaler market shares remain relatively stable across the four subsamples. This result is consistent with industry practice, where wholesalers do not selectively choose groups of stocks for execution.

[10] Using the Wayback Machine (https://web.archive.org), we retrieved the entire 2019–2022 time series of FIF reports for Charles Schwab's market orders in S&P 500 stocks. The data reveal that the average per-share price improvement for odd lots is 48% greater than for orders of 100–499 shares. Wholesalers report data both for S&P 500 and non-S&P 500 stocks separately, but the time series of FIF data are sparser. For Two Sigma, the average price improvement for odd-lot market orders across 10 available quarterly reports is 45% greater than for the 100–499 share bin. Virtu's Q1–Q3 2019 data show 14% greater price improvement, while for Citadel, the only overlapping report from Q1 2019 shows 9% greater price improvement.

Table 2
Broker market shares and routing.

		Routed to							
	% all flow	Citadel	Virtu	G1	Jane Street	Two Sigma	UBS	Merrill Lynch	Morgan Stanley
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
TD Clearing	34.5	32.9	32.4	25.2	6.8	0.7	2.0	0.0	0.0
TD Ameritrade	13.3	52.8	42.2	0.0	0.0	4.9	0.0	0.0	0.0
Robinhood	12.8	40.2	24.5	14.5	5.4	15.4	0.0	0.0	0.0
Schwab	12.5	30.9	28.0	18.8	7.6	3.0	11.6	0.0	0.0
E*TRADE	12.3	37.7	30.4	17.8	6.4	4.2	3.5	0.0	0.0
ViewTrade	6.0	23.0	32.6	0.0	0.0	28.7	15.7	0.0	0.0
Webull	3.5	35.0	29.4	1.4	30.4	3.8	0.0	0.0	0.0
TradeStation	1.9	31.1	41.2	0.0	18.4	4.2	5.1	0.0	0.0
Merrill Lynch	1.6	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0
Morgan Stanley	1.6	9.1	0.7	0.0	0.0	3.5	0.0	0.0	86.8

The table reports each broker's market share of Rule 606 share volume, along with the proportion of volume routed by each broker to each wholesaler during the 2020–2022 sample period. To obtain routed volumes, we use two variables available from Rule 606 data: the total PFOF dollar amounts received by retail brokers and the PFOF amounts in cents per one hundred shares. Dividing the former by the latter allows us to estimate the share amounts sent by the brokers to the wholesalers. Interactive Brokers routes some of their IBKR Lite flow to wholesalers during our sample period, but since the flow from their more sophisticated IBKR Pro platform appears not to be routed to wholesalers, our method would vastly underestimate Interactive Broker's overall volume. We therefore exclude them. Since Fidelity and Vanguard do not receive PFOF, we are unable to estimate their share volume.

Instead, retail brokerages expect wholesalers to accept all orders sent to them.

The statistics in Table 1 suggest that the wholesale environment is rather concentrated. To provide additional detail, we examine concentration in the cross-section. Specifically, we calculate the Herfindahl–Hirschman Index (HHI) for each sample stock as

$$HHI = \sum_{j=1}^{8} s_j^2, \tag{1}$$

where s_j is wholesaler j's market share percentage expressed as a whole number, not a decimal. The U.S. Department of Justice (DOJ) considers a marketplace (i) competitive if its HHI is less than 1,500, (ii) moderately concentrated if the HHI is between 1,500 and 2,500, and (iii) highly concentrated if the HHI is 2,500 or greater. We report individual stock HHIs, sorted alphabetically, in Fig. 1. The wholesale environment qualifies as competitive for only 10 out of 12,012 sample securities. The majority of the remaining stocks have either a moderately concentrated (48% of firms) or a highly concentrated (52% of firms) environment.

Rule 605 data allow us to observe four metrics that are commonly used in market structure research: quoted, effective, and realized spreads, as well as price impacts. Quoted spreads measure liquidity costs displayed by liquidity providers, while effective spreads reflect liquidity costs incurred by liquidity demanders. Effective spreads are typically further divided into two components. The first component, the price impact, captures adverse selection (toxicity) generated by a trade, while the second component, the realized spread, captures toxicity-unrelated costs of market making such as inventory and fixed costs, along with market making profits.[11]

When working with the metrics, we remove outliers by trimming all variables at the 1st and 99th percentiles. Reporting of the quoted spreads is not required by Rule 605, and we derive them as discussed in Internet Appendix A.3. We scale all execution quality metrics by the CRSP closing stock price and calculate stock-level statistics using share-volume weights. When further aggregating across stocks, we use a simple average, which enables us to assess execution quality in the average stock and allows for a view of the entire equity landscape. As we demonstrate shortly, retail traders are quite active across the landscape, particularly in smaller stocks.[12]

3.2. Rule 606: Broker routing

Rule 606 disclosure of PFOF received by brokers from wholesalers enables us to estimate share volumes routed to each wholesaler by each broker that accepts PFOF.[13] For each such broker–wholesaler pair, Rule 606 filings contain two variables that are of interest to us. First, they report the dollar amount of PFOF. Second, they report the same amount in cents per 100 shares. Dividing the former by the latter, we estimate the number of shares sent by each broker to each wholesaler in each sample month.[14]

While brokers like Fidelity and Vanguard use wholesalers extensively, they do not accept PFOF and therefore we are unable to reliably reconstruct the volumes they route to wholesalers. Interactive Brokers routes some of their IBKR Lite flow to wholesalers during our sample period, but since the flow from their more sophisticated IBKR Pro platform appears not to be routed to wholesalers, our method would vastly underestimate Interactive Broker's overall volume. Hence, we exclude Fidelity, Vanguard, and Interactive Brokers from Rule 606 analyses.

We focus on the nine remaining large retail brokerages: TD Ameritrade, Schwab, E*TRADE, Robinhood, ViewTrade, Webull, TradeStation, Merrill Lynch, and Morgan Stanley. TD Ameritrade Holding Corporation has two separate wholly owned subsidiaries that each file Rule 606 reports: TD Ameritrade and TD Ameritrade Clearing (TD Clearing). According to industry representatives, TD Ameritrade flow generally originates from TD's thinkorswim® platform, which caters to relatively sophisticated retail investors, offering customization of strategies, advanced access to stock screens, news, and analytics. The remaining TD Ameritrade flow is reported by TD Clearing.[15]

[11] Rule 605 requires that price impacts are estimated over five-minute horizons. This timeframe may seem long to some readers, as they may believe that horizons relevant to modern market makers are considerably shorter, measured in seconds or even sub-seconds. In Internet Appendix A.6 we use intraday Trade and Quote (TAQ) data to demonstrate that the five-minute price impacts serve as a suitable proxy for adverse selection costs.

[12] In its analysis of retail execution quality for the proposed Order Competition Rule, the SEC uses dollar-volume weights. This approach skews the execution quality metrics towards high-price high-volume stocks. Our results align with those of the Commission when we apply the same weighting technique, as we discuss shortly.

[13] Rule 606 routing and payments data are not available in 2019, so for this analysis, we use a shorter three-year period from January 2020 through December 2022.

[14] Additional details about Rule 606 analysis are in Internet Appendix A.4.

[15] Several smaller retail brokers and Retail Investment Advisors rely on clearing and custody services provided by TD Clearing and Schwab. Therefore, the flows routed by TD Clearing and Schwab include a small portion of orders originating from these entities.

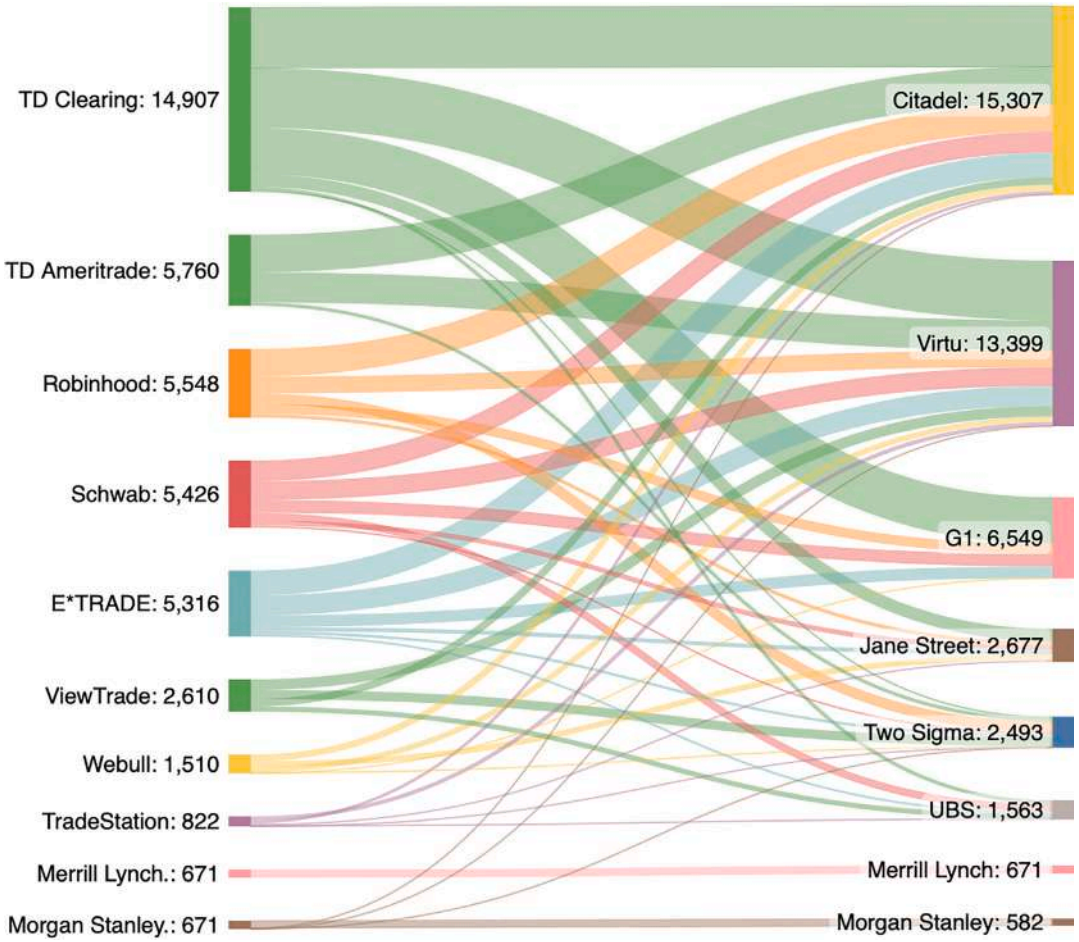

Fig. 2. Retail Broker Routing.
The figure reports order routing patterns (in millions of shares per month) by retail brokers to wholesalers using Rule 606 data from 2020–2022. To obtain routed volumes, we use two variables available from Rule 606 data: the total PFOF dollar amounts received by retail brokerages and the PFOF amounts in cents per one hundred shares. Dividing the former by the latter allows us to estimate the share amounts sent by the brokerages to the wholesalers. For brokerages such as Fidelity and Vanguard that do not accept PFOF, we are unable to compute the share amounts, so these brokerages are not included in the figure. Interactive Brokers routes some of their IBKR Lite flow to wholesalers during our sample period, but since the flow from their more sophisticated IBKR Pro platform appears not to be routed to wholesalers, our method would vastly underestimate Interactive Broker's overall volume. We therefore exclude them.

Fig. 2 and Table 2 illustrate the flows of shares between brokers and wholesalers. TD Clearing routes the largest flows, while Citadel receives the largest flows. Notably, while Citadel and Virtu are the largest receiving wholesalers for most brokers, the proportion of flow they receive varies substantially across brokers. For instance, TD Ameritrade directs 52.8% of its flow to Citadel, whereas ViewTrade only directs 23.0% of its flow to Citadel. Meanwhile, Two Sigma, the fifth largest wholesaler, receives 4.9% of TD Ameritrade's flow, yet its allocation from ViewTrade is 28.7%, larger than ViewTrade's allocation to Citadel.[16]

Because Rule 606 data focus on self-directed retail activity, the flows that we discuss do not reflect the size of each brokerage's entire book of business. For instance, Merrill Lynch and Morgan Stanley may seem small relative to the size of their assets under management. This is because a significant portion of their client base consists of affluent individuals who use wealth management or investment advisory services. Most orders generated by such services are considered not-held, while Rule 606 only captures orders that are held (self-directed investment activity).[17,18] The two smallest brokerages have another distinguishing feature. Morgan Stanley primarily routes to its own wholesale facility, and Merrill Lynch routes to its own facility exclusively. No other brokerage routes to these two facilities. Meanwhile, the larger brokerages (representing more than 96% of retail flow) send orders to multiple wholesalers, a strategy whose costs and benefits we discuss next.

The costs of the multi-wholesaler strategy include setting up and regularly fine-tuning the connections, monitoring execution quality, and periodically renegotiating the terms of engagement. Meanwhile, a major benefit of this strategy is increased resilience should one wholesaler go out of business, have technical difficulties, or run into capacity constraints. Moreover, the literature on customer–supplier relationships suggests that the threat of a broker switching wholesalers may provide incentives for wholesalers to invest in technology and to deliver good execution quality (Denski et al., 1987; Wagner and Friedl, 2007).

Table 3 shows that brokers switch their routing frequently and by large amounts month-to-month. For example, while TD Clearing routes

[16] Jane Street enters the wholesale business in mid-2019, and its market share increases during our sample period as more brokerages begin routing to it. The statistics for Jane Street in Table 2 represent the averages computed during this growth period. In Section 4.6, we examine Jane Street's entry in detail.

[17] A not-held order gives the executing broker time and price discretion to secure the best possible outcome for the client. The vast majority of self-directed retail orders are considered held and require immediate execution.

[18] Morgan Stanley directs its retail customers wishing to open self-directed accounts to E*TRADE, its subsidiary.

Table 3
Broker routing changes.

	Citadel [1]	Virtu [2]	G1 [3]	Jane Street [4]	Two Sigma [5]	UBS [6]	Merrill Lynch [7]	Morgan Stanley [8]
TD Clearing	3.13	3.45	3.49	2.25	0.14	1.12	0.00	0.00
TD Ameritrade	2.90	2.94	0.00	0.00	1.44	0.00	0.00	0.00
Robinhood	4.92	2.56	3.84	1.55	1.99	0.00	0.00	0.00
Schwab	0.61	0.76	0.89	0.70	0.37	0.72	0.00	0.00
E*TRADE	0.99	1.74	1.55	0.60	0.54	0.49	0.00	0.00
ViewTrade	1.50	1.59	0.00	0.00	1.56	1.67	0.00	0.00
Webull	6.06	4.39	0.40	4.27	2.40	0.00	0.00	0.00
TradeStation	6.72	5.00	0.00	4.76	1.09	1.41	0.00	0.00
Merrill Lynch	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Morgan Stanley	0.72	0.04	0.00	0.00	0.54	0.00	0.00	1.01

This table reports month-to-month percentage point changes in order flow routing by each broker to each wholesaler. The share volume is estimated using two variables available from Rule 606 data: the total PFOF dollar amounts received by retail brokerages and the PFOF amounts in cents per one hundred shares. Dividing the former by the latter allows us to estimate the share amounts sent by the brokerages to the wholesalers. Interactive Brokers routes some of their IBKR Lite flow to wholesalers during our sample period, but since the flow from their more sophisticated IBKR Pro platform appears not to be routed to wholesalers, our method would vastly underestimate Interactive Broker's overall volume. We therefore exclude them. Since Fidelity and Vanguard do not receive PFOF, we are unable to estimate their share volume.

Table 4
Execution quality.

	Full sample [1]	S&P 500 [2]	Tercile 1 [3]	Tercile 2 [4]	Tercile 3 [5]
% of total volume	22.37	7.08	13.70	22.49	32.74
Panel A: Equally-weighted					
% price-improved	68.95	76.11	71.92	65.88	68.13
% at NBBO or better	93.75	94.71	93.62	94.09	93.40
Quoted spread, bps	58.35	8.16	25.92	62.87	89.52
Effective spread, bps	42.59	4.36	17.26	46.23	70.02
Price impact, bps	26.84	3.13	12.68	30.67	40.78
Realized spread, bps	15.75	1.22	4.58	15.56	29.24
Panel B: Dollar-volume-weighted					
% price-improved	81.39	83.06	80.61	76.59	71.46
% at NBBO or better	95.37	95.12	95.66	95.12	93.45
Quoted spread, bps	7.88	4.07	7.93	28.69	61.73
Effective spread, bps	4.56	1.94	4.45	19.11	45.48
Price impact, bps	3.34	1.51	3.22	13.38	33.99
Realized spread, bps	1.22	0.42	1.24	5.72	11.48

The table contains execution quality statistics for liquidity-demanding orders executed by wholesalers. We report the results for the full sample and for the four subsamples: S&P 500 and terciles 1 through 3. We report the percentage share of total volume executed by wholesalers, the percentage of shares that are price improved or executed at or better than the corresponding NBBO, the quoted, effective, and realized spreads, and the price impacts. In Panel A, all variables are share-volume-weighted up to the cross-section and then equal-weighted across stocks. In Panel B, all variables are dollar-volume-weighted throughout.

an average of 32.9% and 32.4% of its flow to Citadel and Virtu (Table 2), it changes routing by, respectively, 3.13 and 3.45 percentage points monthly. In relative terms, these changes are 9.5% (= 3.13 ÷ 32.9) and 10.6% (= 3.45 ÷ 32.4). In Internet Appendix Table A.1, we show that TD Clearing routes as much as 49% and as little as 18% of its order flow to Citadel during the sample period. For Virtu, the corresponding figures are 43% and 23%. These variations are not driven entirely by a time trend, as TD Clearing routes 31% (33%) to Citadel (Virtu) at the start of the sample period and 30% (32%) at the end.[19] Other broker–wholesaler pairs, with the exception of Merrill Lynch, also display significant fluctuations in share allocations over time, some of which do exhibit a time trend component.[20] It is possible that these fluctuations are responses to the execution quality provided by the wholesalers. We formally test this hypothesis in the following section.

4. Empirical results

4.1. Execution quality

Table 4 reports summary share volume and execution quality statistics based on Rule 605 data. Note that even though we are unable to compute share volumes for all brokers from Rule 606 data, all brokers are included in Rule 605 reports and are therefore reflected in these statistics. Retail volume is 22.37% of the total volume in an average stock, with retail share significantly greater in smaller stocks than in larger stocks. For instance, while retail traders generate only 7.08% of volume in S&P 500 stocks, they contribute 32.74% of volume in tercile 3 stocks.

Panel A of Table 4 shows that of the volume they receive, wholesalers price-improve a substantial portion, 68.95%. The share of price improvement is the highest for the S&P 500 stocks, at 76.11%, while it is 68.13% for tercile 3 stocks. When retail orders arrive, the prevailing quoted spread is 58.35 bps. Meanwhile, the effective spread they pay is 42.59 bps for a price improvement of 27%. Thus, the data show that wholesalers provide substantial, rather than *de minimis*, price improvement.

Two additional factors should be considered when evaluating the magnitude of price improvement. First, Rule 605 requires market centers to report execution quality relative to the NBBO, likely inflating the reported price improvement, since the NBBO did not account for odd lots during our sample period (Bartlett et al., 2025). Internet Appendix Table A.2 shows that price improvement on exchanges, which includes odd lots and hidden/reserve orders, averages around 3% (1.4 bps). Therefore, the net benefit of wholesaler executions is 24% (= 27%−3%).

Second, while retail investors often time their trades poorly relative to the NBBO spread (see Internet Appendix A.7), it is possible that they time their trades relative to depth imbalances — selling when the limit order book has a negative imbalance and buying when it has a positive imbalance. If so, effective spreads measured against the NBBO midpoint could be overstated (Muravyev and Pearson, 2020; Hagströmer, 2021). Hagströmer (2021) finds that this bias averages 7% for a subset of stocks comparable in size to the average firm in our sample. Therefore, retail investors may gain not only from price improvement but also from actual effective spreads that are 7% lower on average than what is reflected in Rule 605 data.

Some market structure commentators contend that the PFOF payments wholesalers make to retail brokerages are substantial compared to the price improvement they offer. Combining data from Rule 605 and 606 reports for 2020–2022 (the years for which we have Rule 606 data) suggests that the opposite is true. Using the estimated share volumes routed by brokers to wholesalers and each broker's PFOF per 100 shares, we calculate the share volume-weighted average PFOF per share. This allows us to evaluate how the quoted spread is divided between price improvement, PFOF, and wholesaler revenue. In comparison to price improvement, which accounts for 28% of the quoted spread in the 2020–2022 sample, PFOF is only 1%. Thus, PFOF as reported in Rule 606 data accounts for less than 4% of the price improvement reported in Rule 605 data.

It is worth noting that if the option to route to wholesalers were not available, brokers would instead need to pay a taker fee to execute marketable retail orders on exchanges. From Rule 606 disclosures, we calculate that if brokers in our sample had sent their entire retail flow to exchanges instead of wholesalers, they would have had to pay over half

[19] While the two largest wholesalers retain a substantial portion of TD Clearing's flow over the sample period, they do lose a significant share to Jane Street. Their share increases in the early months but declines as TD Clearing reallocates some of its flow to Jane Street.

[20] We discuss these figures in more detail in Internet Appendix A.5.

a billion dollars per year in trading fees. In the current zero commission environment, such a strategy would significantly impact their bottom lines and potentially the cost and quality of services they provide to retail investors.[21]

Next, we turn to the components of the effective spread: price impact and realized spread. In Panel A of Table 4, the average price impact is 26.84 bps, varying between 3.13 bps for the S&P 500 stocks and 40.78 bps for tercile 3 stocks, reflecting larger informational asymmetries in smaller stocks. Wholesalers collect realized spreads of 15.75 bps in an average stock, and again there is substantial variation across the four stock groups. For instance, the realized spread is 1.22 bps in the S&P 500 stocks and 29.24 bps in tercile 3 stocks. These figures are consistent with the greater inventory costs incurred by intermediaries in small stocks. While relatively frequent trading in large stocks is conducive to quick rebalancing of inventory positions, small stocks trade less frequently, resulting in longer inventory holding periods.

The statistics in Panel A are share-volume weighted across order types, order sizes, wholesalers, and months, and then equally-weighted across stocks. We rely on this weighting scheme to provide a comprehensive picture of the full cross-section of securities traded by retail investors. In Panel B, we use dollar-volume weighting throughout, consistent with the SEC's analysis accompanying the Order Competition Rule. Using this weighting scheme, both the fraction of orders receiving price improvement and the magnitude of the improvement are greater for the full sample and each subsample, consistent with the tendency for high-volume and high-price stocks to receive more price improvement. In the full sample, for instance, 81.39% of orders are price-improved, with an average improvement of 42.13%.

Finally, we observe that dollar-volume weighting brings the full-sample average metrics closer to those seen for S&P 500 stocks in Panel A. This result is expected, given that a significant portion of trading volume, including retail volume, concentrates in the largest stocks. Moving forward, we continue to report results for all four subsamples to capture the full range of cross-sectional effects. The results for the S&P 500 subsample should be viewed as representative of the average dollar invested by retail investors.

4.2. Differences across wholesalers

In an earlier section, we suggest that toxicity-adjusted trading costs, as measured by realized spreads, are the most appropriate metric for comparing execution quality across wholesalers in our dataset. Recall that the execution quality we observe is at the stock-wholesaler level, and it is the volume-weighted average execution quality provided to the mix of brokers that the wholesaler serves. If different brokers serve retail clienteles that produce varying levels of toxicity (a concept we confirm shortly), and since the mix of broker flows differs across wholesalers (Fig. 2 and Table 2), it is likely that order flow toxicity also differs across wholesalers. While a metric such as price improvement does not account for these differences, a realized spread metric does. In this section, we examine these possibilities in detail.

We begin by demonstrating that differences in toxicity among brokers indeed exist. To do so, we merge Rule 605 price impact statistics with Rule 606 data on order flows between specific brokers and wholesalers and estimate how changes in these flows affect the price impacts incurred by wholesalers. As an illustration, suppose that in the current month a wholesaler receives most of its flows from Broker A, whose retail customers' orders are relatively toxic. The wholesaler's flow mix changes however in the following month when Broker B, catering to investors with less toxic flow, chooses to route to the wholesaler. After this adjustment, the price impact incurred by the wholesaler will decrease. By regressing the magnitude of price impacts faced by each wholesaler on the share of the wholesaler's flow received from each broker, we can approximate the toxicity generated by different brokers.

A caveat associated with such an analysis, known in statistics as *compositional analysis*, is that any change in the share of one broker will necessarily impact the shares of other brokers, as they are constrained to add up to 1. Consequently, a traditional regression analysis, which assumes independence between regressors, is not appropriate. To address this issue, we follow Greenacre (2021) and use log-ratios of the shares of each broker relative to the share of a reference broker. We choose Robinhood as the reference broker; the results are robust to choosing other brokers for this role. To facilitate the interpretation of the coefficients, we use log base 2 (log2) ratios as suggested by Coenders and Pawlowsky-Glahn (2020). A unit increase in the log2 ratio captures a doubling of the flow share of one broker relative to the shares of each of the other brokers. To avoid applying a log to zeros, we substitute, as suggested by Greenacre (2021), a near zero value computed as $\frac{2}{3}$ of the smallest non-zero broker share observed in the entire dataset.

The regression model is estimated as follows:

$$price\ impact_{jt} = \alpha_t + \sum_{\gamma=1}^{9} \beta_\gamma share_{\gamma jt} + \varepsilon_{jt}, \tag{2}$$

where $price\ impact_{jt}$ is the log2 price impact incurred by wholesaler j in month t, and $share_{\gamma jt}$ is the log2 ratio of the share of broker γ (TD Clearing, TD Ameritrade, Schwab, etc.) in wholesaler j's total flow in month t relative to Robinhood's share for that wholesaler in the same month. The model is estimated with month fixed effects, and the standard errors are clustered by month. As is standard in log-ratio analyses, to obtain the coefficient for Robinhood, we compute the sum of the estimated $\hat{\beta}_\gamma$s and multiply it by -1. Finally, we transform the coefficients for interpretation using $(2^{\hat{\beta}_\gamma} - 1) \times 100$.

Aitchison (1986) points out that compositional regression models are both scale and shift invariant, meaning the interpretation of coefficients should focus on the relative relationships between them, rather than their absolute values. With this in mind, the coefficient estimates plotted in Fig. 3 indicate that TD Ameritrade, TradeStation, and Schwab have the most toxic orders while Merrill Lynch, ViewTrade, and Morgan Stanley have the least toxic orders. Overall, the results are in line with our expectations of trader sophistication across different brokers. As we mention earlier, TD Ameritrade reports order flow from the thinkorswim® platform, and we expect its users to exhibit higher sophistication compared to those whose orders are reported by TD Clearing. The results in Fig. 3 validate this expectation. Furthermore, Eaton et al. (2022) explain that Robinhood users are less sophisticated compared to the clients of legacy brokers such as TD, Schwab, and E*TRADE. Our data support this assertion, offering the added benefit of a comprehensive view of the varying levels of toxicity among several brokers.

Having shown that flow toxicity varies from broker to broker, we now examine wholesaler-level toxicity, which we also expect to vary. To simplify the exposition going forward, we categorize wholesalers into two groups: the *top two*, which includes Citadel and Virtu, and the *others*. Recall that the top two are significantly larger than their peers, accounting for nearly 70% of retail flow, which is the primary concern of the SEC when it comes to market concentration and pricing power. We first examine price impacts, or the toxicity of order flow received by each group, and then delve into realized spreads, which, as we have argued, is the most suitable metric of execution quality in our wholesaler-level data.

In Table 5, we use the following panel regressions to ask if price impacts systematically differ for the top two compared to the other wholesalers:

$$price\ impact_{ijt} = \alpha_{it} + \beta_1 top2_j + \varepsilon_{ijt}, \tag{3}$$

[21] An alternative of routing to inverted exchanges and therefore receiving payments for marketable orders is not always feasible as such exchanges are rarely at the NBBO (Mackintosh, 2020).

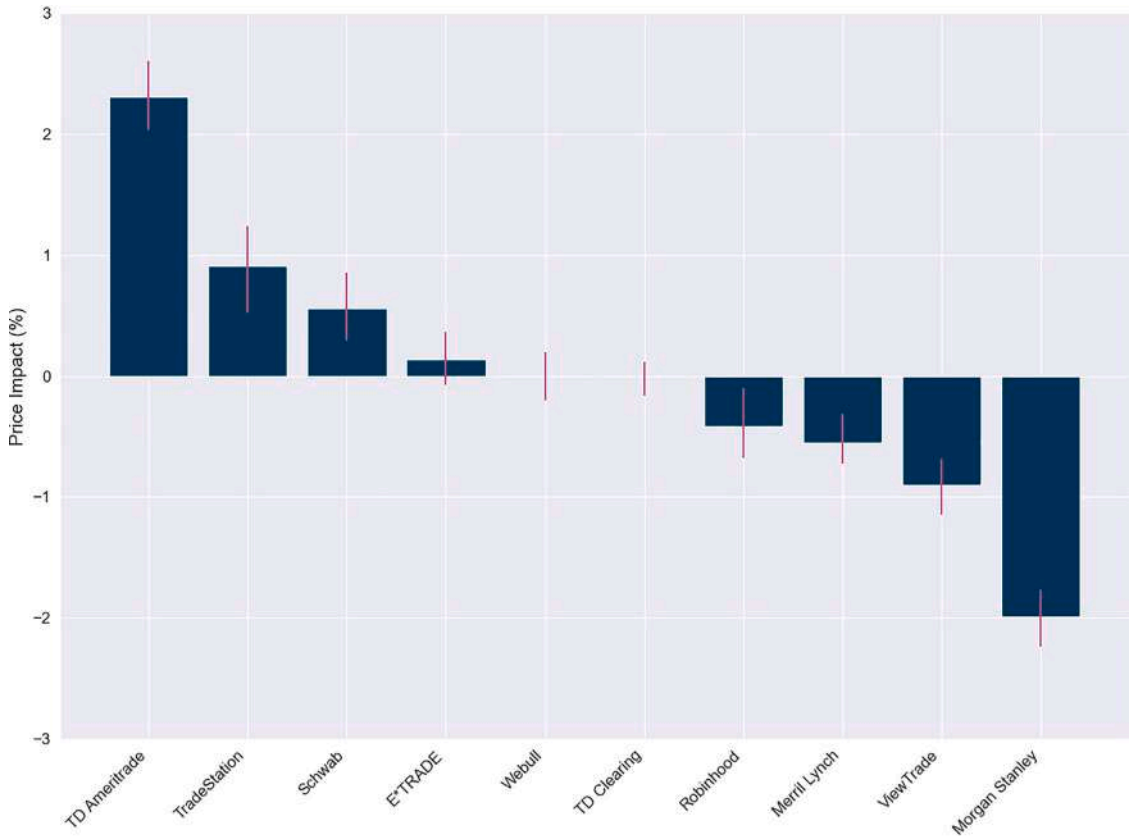

Fig. 3. Broker Toxicity.
The figure depicts the regression coefficients from estimating:

$$price\ impact_{jt} = \alpha_t + \sum_{\gamma=1}^{9} \beta_\gamma share_{\gamma jt} + \epsilon_{jt},$$

where $price\ impact_{jt}$ is the log 2 price impact incurred by wholesaler j in month t, $share_{\gamma jt}$ is the log 2 ratio of the share of broker γ (TD Ameritrade, TradeStation, Schwab, E*TRADE, Webull, TD Clearing, Merril Lynch, ViewTrade, or Morgan Stanley) in wholesaler j's total flow in month t relative to Robinhood's share for that wholesaler in the same month. The model is estimated with month fixed effects, and the standard errors are clustered by month. We implement this modeling to account for the compositional nature of the data, in which any change in the share of one broker will necessarily impact the shares of other brokers, as they are constrained to add up to 1. To address this issue, we follow Greenacre (2021) and use log-ratios of the shares of each broker relative to the share of a reference broker. We choose Robinhood as the reference broker; the results are robust to choosing other brokers for this role. To facilitate the interpretation of the coefficients, we use log base 2 (log2) ratios as suggested by Coenders and Pawlowsky-Glahn (2020). A unit increase in the log2 ratio captures a doubling of the flow share of one broker relative to the shares of each of the other brokers. To avoid applying a log to zeros, we substitute, as suggested by Greenacre (2021), a near zero value computed as $\frac{2}{3}$ of the smallest non-zero broker share observed in the entire dataset. As is standard in log-ratio analyses, to obtain the coefficient for Robinhood, we compute the sum of the estimated $\hat{\beta}_\gamma$s and multiply it by -1. Finally, we transform the coefficients for interpretation using $(2^{\hat{\beta}_\gamma} - 1) \times 100$.

where $price\ impact_{ijt}$ is the share-volume-weighted metric for stock i wholesaler j in month t, and $top2$ is a dummy variable that has a value of 1 for executions by Citadel and Virtu and 0 for executions by the other wholesalers. The models control for stock-by-month fixed effects and use double-clustered standard errors. The fixed effects absorb the controls commonly used in market structure research, such as stock price, trading volume, and price volatility, since these variables are fixed for each stock-month. In Internet Appendix A.8, we estimate similar regression models using stock and month fixed effects along with the aforementioned controls, and our results remain unchanged.

Panel A of Table 5 shows that the top two wholesalers receive more toxic flow compared to the others, a pattern consistent across all four stock groups. In the full sample for instance, Citadel and Virtu face price impacts that are 4.757 bps greater than the price impacts faced by the other wholesalers, a difference of 17.7% relative to the average price impact of 26.84 bps in Panel A of Table 4. Most brokers use a routing wheel, which rotates to wholesaler 2 after a preset quantity has been routed to wholesaler 1 and so on. In a hypothetical scenario with two wholesalers assigned 60% and 40% of the flow, the broker would send 6 orders to the first wholesaler, 4 orders to the second wholesaler, and then revert to the first wholesaler for the next 6 orders. By design, the wheel aims to achieve a random allocation of orders. Therefore, the variation in flow toxicity across wholesalers is driven by the mix of brokers they serve, rather than by individual brokers selectively routing more toxic flow to the top two wholesalers.

As we discuss previously, the composition of order flow varies across wholesalers and also over time. Recall for instance that TD Ameritrade, the broker with the most toxic flow, sends almost all of it to Citadel and Virtu, increasing adverse selection for these wholesalers compared to others. Additionally, many brokers significantly vary their allocations across wholesalers throughout the sample period and therefore cause variation in adverse selection costs over time. Given that adverse selection is a major cost of liquidity provision, any metric aiming to compare execution costs across wholesalers in a wholesaler-level dataset must account for the differences in adverse selection wholesalers face. We believe that the realized spread is such a metric and use it to measure retail execution costs offered by wholesalers in the remainder of the manuscript.[22]

In Panel B of Table 5, we replace the dependent variable in Eq. (3) with realized spreads to examine if realized spreads systematically

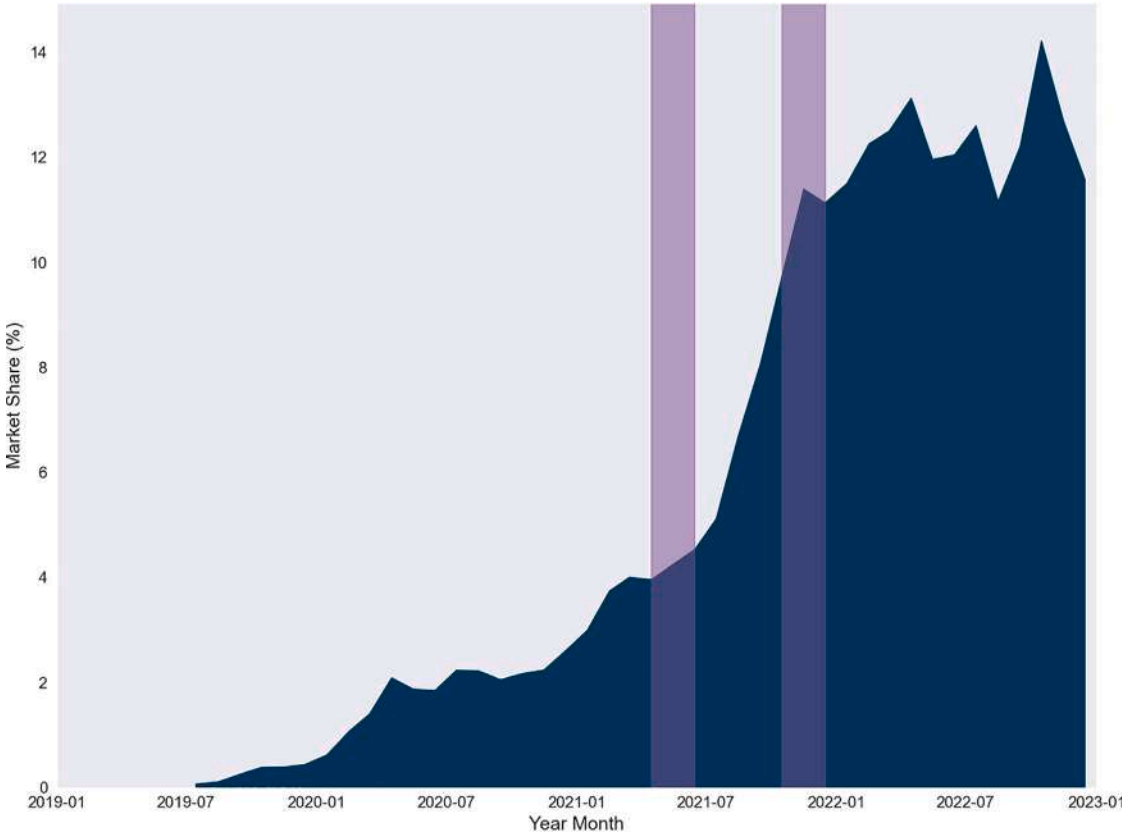

Fig. 4. Jane Street Entry.
The figure reports the market share of Jane Street. The sample covers all stocks and ETFs during the 2019–2022 period, and the data source is Rule 605 reports. We highlight in shaded purple the pre-period (April–June, 2021) and post-period (October–December, 2021) used in the event study in Section 4.6.

differ between the top two and the other wholesalers. Despite the high adverse selection costs they face, the top two charge relatively low realized spreads. In the full sample, their realized spreads are 2.806 bps lower than those charged by the others, a difference of 17.8% compared to the sample mean. This pattern persists in the cross-section. Overall, the data indicate that the top two wholesalers provide execution cost savings to retail investors. In the next section, we explore the sources of these savings.

4.3. Economies of scale

Are the differences in execution costs between the top two wholesalers and their competitors attributable to economies of scale? One reason wholesalers may achieve economies of scale is their substantial fixed costs from large investments in technology and human capital, which are necessary to support sophisticated order routing and inventory management. As an example, from 2020 to 2022, Virtu's annual revenue from market making declined from $1.782 billion to $1.428 billion and then further to $1.058 billion, a total reduction of 40.6%. In the meantime, its operating costs remained rather stable at $690, $677, and $675 million, declining only by 2.2%.[23] With much of the operating costs being fixed, a wholesaler handling more flow can allocate these costs across a larger volume of business, resulting in lower production costs per unit of liquidity.

In addition to its effect on fixed costs per share, wholesaler size may have both positive and negative effects on inventory management. On the one hand, a wholesaler handling more retail flow may be able to internalize orders more efficiently, leading to lower inventory costs. Recall that retail brokerages distribute orders among wholesalers using a routing wheel. Suppose that for every ten orders a retail brokerage receives, it routes four to Citadel. Given the typical balance of retail flow, with buys arriving as frequently as sells, the orders will tend to reconcile against each other, and inventory will tend to zero. Even in the case of a leftover imbalance, Citadel benefits from the shortest waiting time (only six orders) before the wheel rotates to it again. In contrast, a wholesaler with a 10% market share faces a waiting time of nine orders before having an opportunity to undo the imbalance created by the first order it receives. The ability to balance the flow quickly and the relatively short wait between rotations enjoyed by large wholesalers may therefore help keep inventory costs low.

On the other hand, as customer buys detract from a wholesaler's inventory and customer sells add to it, inventory imbalance may be thought of as a random walk process. Consequently, while the unconditional expectation of inventory may be zero, its variance may increase in volume, potentially leading to greater rebalancing costs.

[22] Other studies measure execution quality using price improvement, which is calculated as the difference between quoted and effective spreads. This metric is suitable when applied to broker-level datasets, such as for instance the data used by Ernst et al. (2023), since broker flow toxicity is fixed at the broker level.

[23] We use the figures from Virtu because it is one of the few publicly traded firms primarily engaged in trade intermediation. Although the parent companies of three other wholesalers – UBS, Merrill Lynch, and Morgan Stanley – are also public, their cost and revenue streams are less informative given the diversified nature of their businesses. Meanwhile, in 2020 (2022), 78% (72%) of Virtu's adjusted net trading revenues came from market making, with much of the remainder deriving from execution services (https://ir.virtu.com). Virtu acquired ITG in Q1 2019, and we therefore exclude 2019 from the comparisons.

Table 5
Execution quality across wholesalers.

	Full sample [1]	S&P 500 [2]	Tercile 1 [3]	Tercile 2 [4]	Tercile 3 [5]
Panel A: Price impacts					
top2	4.757***	0.823***	1.779***	5.285***	11.666***
	(0.26)	(0.10)	(0.13)	(0.33)	(0.73)
R^2	0.209	0.053	0.190	0.259	0.145
Panel B: Realized spreads					
top2	−2.806***	−0.191**	−0.733***	−2.889**	−7.979***
	(0.47)	(0.09)	(0.16)	(0.58)	(1.24)
R^2	0.173	0.056	0.121	0.201	0.116
Panel C: Realized spreads, controlling for economies of scale					
top2	0.834	−0.102	−0.256	1.696	3.738
	(1.01)	(0.11)	(0.34)	(1.16)	(2.77)
operation size	−1.717***	−0.052	−0.231**	−2.184***	−5.736***
	(0.35)	(0.04)	(0.10)	(0.39)	(1.17)
dev. operation size	−2.285***	−0.137*	−0.940***	−1.770***	−3.683***
	(0.24)	(0.07)	(0.11)	(0.19)	(0.72)
R^2	0.173	0.056	0.121	0.202	0.118
Panel D: Effective spreads					
top2	1.952***	0.632***	1.046***	2.395***	3.692***
	(0.39)	(0.05)	(0.13)	(0.52)	(0.94)
R^2	0.734	0.613	0.687	0.668	0.679

Panels A, B, and D report coefficient estimates from the following regression:

$$depvar_{ijt} = \alpha_{it} + \beta_1 top2_j + \epsilon_{ijt},$$

where $depvar_{ijt}$ is the price impact, or realized spread, or effective spread for stock i wholesaler j in month t, and $top2$ is a dummy variable that has a value of 1 for orders executed by Citadel and Virtu and 0 for orders executed by other wholesalers. We estimate the model using the full sample and the four subsamples: S&P 500 and terciles 1 through 3. Panel C, estimated for realized spreads, reports coefficient estimates from the following regression:

$$realized\ spread_{ijt} = \alpha_{it} + \beta_1 top2_j + \beta_2 operationsize_{jt} + \beta_3 dev.operationsize_{ijt} + \epsilon_{ijt},$$

where $operation\ size_{jt}$ is the average $operation\ size_{ijt}$ across all stocks handled by wholesaler j in month t, with $operation\ size_{ijt}$ defined as the log of the ratio of retail volume captured by wholesaler j in stock i during month t to the total CRSP trading volume for that stock-month. In turn, $dev.\ operation\ size_{ijt}$ ($= operation\ size_{ijt} - operation\ size_{jt}$) is the stock-specific deviation from the average operation size. The models are estimated using stock volume weights and include stock-by-month fixed effects, with standard errors double-clustered by stock and month. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

Additionally, Eaton et al. (2022) show that the typically balanced nature of retail flow may occasionally change if retail investors follow each other in a herding-like manner, leading to unwanted inventory accumulations. Hoffmann and Jank (2024) and industry participants suggest that instead of holding onto such accumulations, wholesalers offload them to other wholesalers or to exchanges. Huang et al. (2012) provide a theoretical model detailing the economics of such actions. Offloading is costly; therefore, greater volume may have a negative effect on wholesalers' bottom lines.

Similar to much of microstructure data, our dataset lacks the granularity to precisely differentiate between wholesalers' fixed costs and inventory costs. However, the mere existence of large wholesalers suggests that the benefits of scale outweigh the drawbacks. In the following analysis, we aim to distinguish between these costs and evaluate the overall economies of scale by estimating the size of a wholesaler's operation both in aggregate and for specific stocks.

To approximate the size of a wholesaler's operation, we construct an *operation size* variable for each wholesaler j in each month t as follows. We start by measuring the retail volume processed by the wholesaler in each stock-month. To ensure comparability across securities, we normalize this retail volume by the total CRSP volume for that stock-month. We then average the resulting values, $operation\ size_{ijt}$, for each wholesaler in each month to determine the size of a wholesaler's operation in an average stock, $operation\ size_{jt}$. If economies of scale allow large wholesalers to generate liquidity at a lower cost, we would expect $operation\ size_{jt}$ to be negatively related to realized spreads. This effect would capture both the fixed cost considerations and the overall inventory cost implications.

Further, to approximate the additional benefits or costs potentially associated with inventory management, we compute the deviation from the average operation size for each wholesaler in each stock as $operation\ size_{ijt} - operation\ size_{jt}$. If having a relatively large operation size in a particular stock enables a wholesaler to manage inventories more efficiently, this stock-specific deviation should exhibit a negative relationship with realized spreads.

Finally, if economies of scale enable the top two wholesalers to provide liquidity at a lower cost, controlling for operation size should diminish or even eliminate the observed differences between them and other wholesalers. To test this hypothesis and examine the relationships between operation size and realized spreads discussed above, we estimate the following regression:

$$realized\ spread_{ijt} = \alpha_{it} + \beta_1 top2_j + \beta_2 operation\ size_{jt} + \beta_3 dev.\ operation\ size_{ijt} + \epsilon_{ijt}, \quad (4)$$

with all variables defined as previously discussed. As before, the regression includes stock-by-month fixed effects and uses double-clustered standard errors.

Recall that Panel B of Table 5 shows that the top two wholesalers charge lower realized spreads than their competitors. If economies of scale explain some (or all) of this difference, we expect the *top2* coefficient to become smaller (or statistically indistinguishable from zero) once operation size is accounted for. Alternatively, if the coefficient turns positive, one could conclude that the top two charge too much given the economies of scale they enjoy. Panel C shows that when *operation size* and *dev. operation size* are included as controls, the *top2* coefficient is insignificant in all subsamples. Taken together with the negative *operation size* and *dev. operation size* coefficients, this result suggests that the top two benefit from lower inventory and fixed costs per share due to their operational scale, which explains the low execution costs they offer.

For comparison with studies that focus on effective spreads as the primary trading cost metric, in Panel D, we report the results for effective spreads as the dependent variable. Recall that Panels A and B show that the top two wholesalers incur 4.757 basis points more in adverse selection than their competitors and charge 2.806 basis points less in realized spreads. The approximate net of these figures – a positive 1.952 basis points – is the effective spread difference reported in Panel D. Although this suggests that the top two charge relatively high effective spreads, it seems they do so to partially offset their significantly larger adverse selection costs.

4.4. Broker routing and wholesaler performance

Why do the top two wholesalers transfer their economies of scale savings to retail customers instead of using them to boost their bottom lines? Do brokers play a role in encouraging such transfers? FINRA Rule 5310 mandates that brokers conduct thorough reviews of execution quality received by their customers on at least a quarterly basis.[24] In reality, such evaluations typically occur on a monthly basis (Ernst et al., 2023). In their own disclosure statements, individual brokers claim to adjust routing to favor wholesalers with superior past performance.[25] If the brokers abide by these statements, a wholesaler's market share should increase if execution costs it charges are lower than those charged by the competitors.

As we show in Table 1, retail customers are noticeably active in micro and small caps, where liquidity is naturally limited. In the

[24] Rule 5310: Best Execution and Interpositioning (https://bit.ly/46GDy5B).
[25] See, for instance, Schwab's Order Routing Process (https://bit.ly/3UMECB1), Robinhood's Stock, ETF, and options order routing (https://bit.ly/3wgfbhV), and Webull's Execution Quality and Order Routing (https://bit.ly/3JQfTFz).

Table 6
Wholesaler order flow determinants.

	$\tau = 3$					$\tau = 1$
	Full sample [1]	S&P 500 [2]	Tercile 1 [3]	Tercile 2 [4]	Tercile 3 [5]	Full sample [6]
Panel A: Stock-wholesaler-month panels						
$realized\ spread_{ijt-\tau}$	−0.000***	−0.001	−0.000**	−0.000***	−0.000**	−0.000***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
$realized\ spread_{jt-\tau}$	−0.045***	−0.043***	−0.046***	−0.046***	−0.041***	−0.039***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
R^2	0.676	0.782	0.720	0.667	0.642	0.638
Panel B: Wholesaler-month panels						
$realized\ spread_{jt-\tau}$	−0.039***	−0.043***	−0.040***	−0.039***	−0.034***	−0.030***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
R^2	0.802	0.729	0.781	0.800	0.846	0.774
Panel C: Wholesaler-month panels						
$effective\ spread_{jt-\tau}$	−0.026**	−0.036***	−0.030***	−0.025**	−0.017**	−0.030**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
R^2	0.793	0.726	0.774	0.791	0.836	0.776
Panel D: Wholesaler-month panels						
$realized\ spread_{jt-\tau}$	−0.037***	−0.044***	−0.040***	−0.037***	−0.031***	−0.033***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
$price\ impact_{jt-\tau}$	0.015	−0.008	0.005	0.016	0.031	−0.021
	(0.03)	(0.04)	(0.03)	(0.03)	(0.02)	(0.03)
R^2	0.802	0.728	0.780	0.800	0.849	0.776

To infer if the share of retail order flow received by a wholesaler depends on the wholesaler's prior performance, in Panel A, we estimate the following regression using stock-wholesaler-month panels:

$$market\ share_{ijt} = \alpha_{it} + \theta_j + \beta_1 realized\ spread_{ijt-\tau} + \beta_2 realized\ spread_{jt-\tau} + \epsilon_{ijt},$$

where $market\ share_{ijt}$ is the share of retail volume in stock i executed by wholesaler j in month t expressed as the deviation from the geometric mean across all wholesalers, $realized\ spread_{ijt-\tau}$ is the average realized spread charged by wholesaler j in stock i over the previous τ months expressed as the deviation from the arithmetic mean across all other wholesalers, and $realized\ spread_{jt-\tau}$ is the average realized spread charged by wholesaler j in all stocks routed to it over the previous τ months expressed as a deviation from the arithmetic mean across all other wholesalers. The model is estimated for the full sample and four subsamples using stock volume weights, incorporating stock-by-month fixed effects, with standard errors double-clustered by stock and month. In specifications [1] through [5], $\tau = 3$ to account for relatively long lookback windows, and in specification [6], $\tau = 1$ to capture shorter lookback windows. In Panels B through D, we collapse the stock dimension and use wholesaler-month panels to estimate:

$$market\ share_{jt} = \alpha_t + \theta_j + \beta_1 indepvar_{jt-\tau} + \epsilon_{jt},$$

where $indepvar_{jt}$ is computed similarly to $realized\ spread_{jt-\tau}$ above for realized spreads, effective spreads, and price impacts. These regressions are estimated using wholesaler and month fixed effects and standard errors clustered by month. The independent variables are scaled so that the economic significance corresponds to basis points. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

meantime, industry participants indicate that retail brokerages aim to obtain the highest execution quality in all stocks, regardless of size, and therefore expect wholesalers to offer low execution costs across the entire spectrum of securities, rather than focusing on specific securities that wholesalers may prefer to intermediate.

To investigate whether brokerages actively monitor and reward high-performing wholesalers, and whether they incentivize wholesalers to deliver superior execution quality across all stocks, we use the econometric framework for order routing proposed by Boehmer et al. (2007). The framework, designed for data similar to ours, uses a combination of geometric and arithmetic means to ensure that predicted wholesaler market shares fall between zero and one, and that the total market share across wholesalers sums to one. Using this framework, we estimate:

$$market\ share_{ijt} = \alpha_{it} + \theta_j + \beta_1 realized\ spread_{ijt-\tau} + \beta_2 realized\ spread_{jt-\tau} + \epsilon_{ijt}, \quad (5)$$

where $market\ share_{ijt}$ is the share of retail volume in stock i executed by wholesaler j in month t expressed as the deviation from the geometric mean across all wholesalers, $realized\ spread_{ijt-\tau}$ is the average realized spread charged by wholesaler j in stock i over the previous τ months expressed as the deviation from the arithmetic mean across all other wholesalers, and $realized\ spread_{jt-\tau}$ is the average realized spread charged by wholesaler j in all stocks routed to it over the previous τ months expressed as a deviation from the arithmetic mean across all other wholesalers. We use $\tau = 3$ and $\tau = 1$ to capture relatively long and short lookback windows, respectively. The realized spread variables are scaled so the economic significance corresponds to basis points. We run these regressions for the full sample and then for each subsample using stock-by-month and wholesaler fixed effects and clustered standard errors. Since Morgan Stanley (Merrill Lynch) primarily (exclusively) routes to its own facility, we exclude both from this analysis.

Panel A of Table 6 shows that if a wholesaler charges a relatively low realized spread across all stocks, retail brokerages respond by granting the wholesaler a larger market share. Conversely, wholesalers charging relatively large spreads face a reduction in their allocations. This result holds for the full sample and for all subsamples. A one basis point reduction in a wholesaler's realized spread relative to the average across wholesalers is associated with a 4.5% greater future market share for the full sample and between 4.1% and 4.6% greater market shares for the subsamples.

The regression results also generally suggest that a low spread charged by a wholesaler in one particular stock is associated with a greater future market share. However, the economic significance of this effect is small; wholesalers seem to compete by offering lower liquidity costs across all stocks, rather than on a stock-by-stock basis. In light of this, we simplify the analysis and repeat the market share evaluation using a wholesaler-month panel:

$$market\ share_{jt} = \alpha_t + \theta_j + \beta_1 indepvar_{jt-\tau} + \epsilon_{jt}, \quad (6)$$

where $indepvar_{jt-\tau}$ is computed similarly to $realized\ spread_{jt-\tau}$ in Eq. (5). In addition to realized spreads, we also consider effective spreads and price impacts as independent variables.

The results in Panel B reinforce the earlier findings: better wholesaler performance, as measured by realized spreads, is associated with

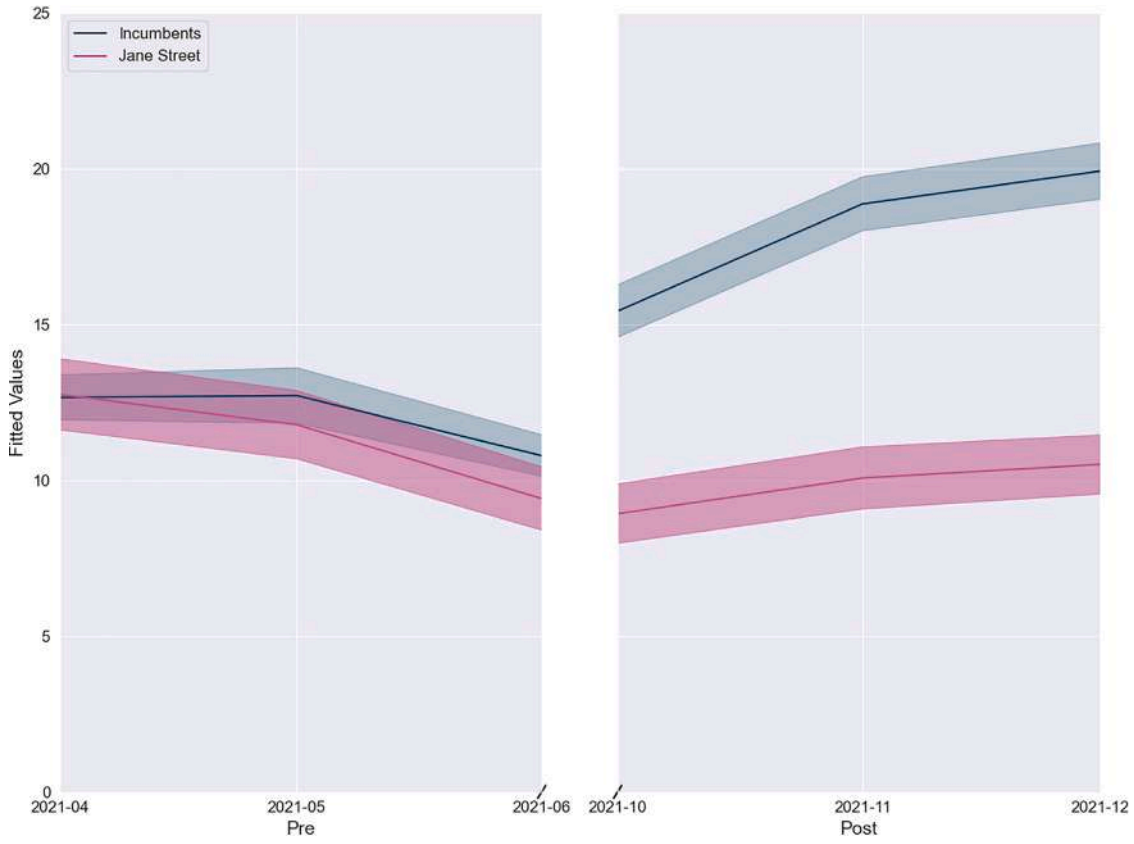

Fig. 5. Execution Costs Around Jane Street Entry.
The figure presents the realized spread fitted values for the incumbents and Jane Street before and after Jane Street gained substantial market share. The values are estimated from Eq. (8) and are based on the full sample regressions from Table 8. Each line is accompanied by 10% and 90% confidence intervals.

greater subsequent order flow to the wholesaler. This relationship holds for both the full sample and all four subsamples.[26] Notably, Panel C shows similar results when using effective spreads instead of realized spreads. To unpack this finding, Panel D decomposes effective spreads into realized spreads and price impacts. In this breakdown, realized spreads continue to have negative coefficients, while price impacts are statistically insignificant, indicating that the effect observed for effective spreads is driven by the realized spread component.

4.5. The cross-section

Our earlier discussion relied on inventory costs to explain why realized spreads in small stocks were greater than those in large stocks. Although measuring inventory costs is notoriously challenging, we propose using processed retail volume as a proxy for a wholesaler's ability to manage them. We suggest that in lower-volume stocks, outstanding inventory positions may take longer to offload[27] and run the following regression to examine this possibility:

$$realized\ spread_{it} = \alpha_t + \beta_1 Tercile1_i + \beta_2 Tercile2_i + \beta_3 Tercile3_i + \beta_4 price_{it} + \beta_5 volatility_{it} + \beta_6 volume_{it} + \varepsilon_{it}, \quad (7)$$

where $realized\ spread_{it}$ is the realized spread in stock i in month t, $Tercile1$ through $Tercile3$ are dummy variables indicating the tercile a stock belongs to, with the intercept capturing S&P 500 stocks, $price$ is the natural log of the stock price, and $volatility$ is the difference between the high and low prices scaled by the high price. We use two versions of the *volume* variable. First, we use *retail volume*, calculated as the natural log of total retail volume across all wholesalers in stock i in month t, as reported in Rule 605 data. Second, we estimate an alternative specification using *total volume*, defined as the natural log of CRSP trading volume, in place of retail volume. The first specification allows us to examine the retail routing wheel argument, while the second specification accounts for the possibility that vertically integrated wholesalers may also use non-retail flow to manage inventory positions. Since the tercile dummies are unique to each security, we control only for month fixed effects, yet use double-clustered standard errors.

Column [1] of Table 7 reports the base specification that controls only for the month fixed effects, confirming that tercile 1, 2, and 3 stocks have significantly greater realized spreads than S&P 500 stocks. Introducing controls for price and volatility in column [2], followed by the retail volume control in column [3], turns the difference between S&P 500 stocks and stocks in terciles 1 and 2 negative. In other words, when controlling for the difficulty of managing inventory in volatile stocks and stocks with relatively low retail volume, it turns out that wholesalers undercharge for liquidity in tercile 1 and tercile 2 stocks relative to S&P 500 stocks. Finally, column [4] considers the possibility that wholesalers may use non-retail volume to manage inventory. In

[26] In Internet Appendix A.9, we show that wholesaler performance across all stock subsamples, from mega- to micro-caps, influences broker routing decisions across the entire range of stock groups. In other words, while performance in S&P 500 stocks affects the routing of S&P 500 stocks, it also impacts routing decisions for all other stock subsamples. Similarly, wholesaler performance in tercile 3 influences the routing of both tercile 3 stocks and all other stocks, though this effect is slightly less economically significant.

[27] Returning to our earlier example, where Citadel waits for six orders for the routing wheel to cycle back to it, the speed of the wheel's rotation varies with trading volume – turning faster during periods of high volume and slower when volume is low – directly affecting Citadel's inventory holding costs.

this specification, all coefficients on the size terciles are negative and significant consistent with the notion that wholesalers undercharge for liquidity in all non-S&P 500 stocks. This result suggests that retail brokers' emphasis on superior execution quality across the entire portfolio of securities routed to wholesalers yields especially significant benefits for smaller stocks.

4.6. A competitive shock

The dynamics of wholesaler competition change during our sample period due to the entry of a new wholesaler, Jane Street. On the one hand, it is possible that competitive forces among wholesalers intensify post-entry, leading to a decrease in execution costs. On the other hand, any loss of market share in an economies of scale business is likely to result in a redistribution of fixed costs across a smaller number of executed shares, potentially leading to greater execution costs. In what follows, we aim to understand which of these effects dominates.

Fig. 4 illustrates Jane Street's entry and market share growth over time. The firm enters the wholesale business in the middle of 2019, but throughout 2020 it still has a very small market share. Its market share begins to increase more rapidly in the late summer of 2021, reaching a substantial level by October 2021.[28] By the end of 2021, all major brokers but one route to Jane Street, and by the end of our sample period (the end of 2022), it has a market share around 12%. All incumbent wholesalers, large and small, experience a market share loss of 11% or greater to Jane Street.

While we cannot be certain why Jane Street chose to enter the equity wholesale business in mid-2019, its overall approach to growth is to scale the business by reaching a wider client base and expanding into new markets to achieve cost efficiencies (Wigglesworth, 2021; Seligson and Doherty, 2024). Consequently, two factors may have contributed to its decision to become a wholesaler. First is Jane Street's significant experience operating as an ETF market maker/authorized participant and running a single-dealer platform in the U.S. as well as operating a systematic internalizer in Europe. Second is the increase in retail trading spurred by the emergence of zero-commission brokers such as Robinhood and the anticipated switch to zero commissions industry-wide.

To assess how Jane Street's entry impacts execution costs charged by wholesalers, we run a difference-in-differences regression. The pre-period is defined as April–June 2021, when Jane Street holds a small market share, and the post-period covers the last three months of 2021, following Jane Street's rapid growth. The pre and post periods are highlighted in Fig. 4. As a control sample, we use the spreads charged by liquidity providers on exchanges. The exchange data are sourced from Rule 605 reports, and we discuss the descriptive statistics and parallel trends in Internet Appendix A.7. We use exchange spreads solely to control for potential market-wide confounding events or trends and not for making direct comparisons between two platform types. Table 8 reports the results from the following regression:

$$depvar_{ijt} = \alpha_{it} + \beta_1 whol_j + \beta_2 whol \times post_{jt} + \varepsilon_{ijt}, \quad (8)$$

where $depvar_{ijt}$ is either the realized or effective spread in stock i for intermediary j (wholesaler or exchange) in month t, $whol$ is a dummy variable that has a value of 1 for orders executed by wholesalers and 0 for orders executed by exchanges, and *post* is a dummy variable that has a value of 1 after Jane Street market share capture and 0 otherwise. We use stock-by-month fixed effects and double-cluster the standard errors by stock and month.

[28] Rule 606 data indicate that upon entry, Jane Street first enters agreements with smaller brokerages such as TradeStation and Webull. In 2021, having established itself as a reliable wholesaler, it begins collaborating with large brokerages; first with E*TRADE, then Schwab, followed by TD Clearing, and finally Robinhood.

Table 7
The cross-section.

	[1]	[2]	[3]	[4]
Tercile 1	3.778***	−2.411***	−10.647***	−15.043***
	(0.15)	(0.71)	(0.93)	(1.16)
Tercile 2	17.487***	5.442***	−11.389***	−21.942***
	(0.63)	(0.99)	(1.28)	(1.82)
Tercile 3	39.042***	24.044***	3.023***	−12.077***
	(2.00)	(1.25)	(1.07)	(1.53)
price		−5.686***	−9.487***	−9.994***
		(0.57)	(0.62)	(0.01)
volatility		0.049***	0.039***	0.069***
		(0.01)	(0.01)	(0.01)
retail volume			−4.495***	
			(0.19)	
total volume				−5.867***
				(0.28)
R^2	0.138	0.170	0.214	0.222

The table reports coefficient estimates from the following regression:

$$realized\ spread_{it} = \alpha_t + \beta_1 Tercile1_i + \beta_2 Tercile2_i + \beta_3 Tercile3_i + \beta_4 price_{it} + \beta_5 volatility_{it} + \beta_6 volume_{it} + \varepsilon_{it},$$

where $realized\ spread_{it}$ is the realized spread in stock i in month t, $Tercile1$ through $Tercile3$ are dummy variables indicating the tercile a stock belongs to, with the intercept capturing S&P 500 stocks, $price$ is the natural log of the stock price, and $volatility$ is the difference between the high and low prices scaled by the high price. We use two versions of the *volume* variable. First, we use *retail volume*, calculated as the natural log of total retail volume across all wholesalers in stock i in month t, as reported in Rule 605 data. Second, we estimate an alternative specification using *total volume*, defined as the natural log of CRSP trading volume, in place of retail volume. Since the tercile dummies are unique to each security, the regressions include only month fixed effects while using stock volume weights and double-clustered standard errors. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

Panel A of Table 8 shows that wholesaler realized spreads increase following Jane Street's market share capture in the full sample. This result holds in all subsamples, although its statistical significance is marginal for the S&P 500 stocks.[29] Panels B and C report the results for the incumbents and Jane Street separately. Recall that Jane Street was present in the pre-period, albeit at a much lower market share, making it possible to run a difference-in-differences analysis for it alone. The incumbents increase their realized spreads in all subsamples. In contrast, Jane Street keeps the S&P 500 and tercile 1 and 2 spreads unchanged while reducing the spreads for tercile 3. Fig. 5 illustrates the differences between the incumbents and Jane Street. It reports the fitted values of the realized spread for the pre- and post-periods, along with the 10% and 90% confidence intervals. Consistent with the regression results, in the pre-period, the spreads are similar, but in the post-period, the incumbents increase their spreads relative to Jane Street.

To understand the result for the incumbents, let us consider what happens when Jane Street enters. Order flow is now divided among a larger number of wholesalers resulting in less flow for each incumbent. For instance, Citadel and Virtu each lose more than 11% of their flow to Jane Street, while many smaller wholesalers lose even more. With less flow, the incumbents likely face increased per-share costs and pass them onto the price of liquidity.

It is also useful to consider pricing behavior of Jane Street itself. Having captured a substantial market share, it should begin benefiting from economies of scale, transferring the associated cost savings to retail customers. We only observe such behavior in tercile 3 stocks, while the execution costs it charges in other stocks remain the same. While we cannot definitively pinpoint the cause of this behavior, Jane Street may have initially undercharged for liquidity outside of tercile

[29] The full-sample results for effective spreads, as reported in specification 6, are consistent with the findings for realized spreads.

Table 8
Jane Street entry.

	Realized spread					Effective spread
	Full sample [1]	S&P 500 [2]	Tercile 1 [3]	Tercile 2 [4]	Tercile 3 [5]	Full sample [6]
Panel A: All wholesalers						
whol	8.968***	1.280***	5.145***	11.208***	16.268***	−2.122***
	(0.32)	(0.16)	(0.25)	(0.63)	(0.71)	(0.25)
whol × *post*	4.575**	0.387*	1.789**	6.071**	8.653**	2.364***
	(1.28)	(0.16)	(0.69)	(1.62)	(2.34)	(0.32)
R^2	0.337	0.246	0.303	0.385	0.304	0.923
Panel B: Incumbents						
whol	8.955***	1.281***	5.220***	11.212***	16.042***	−2.271***
	(0.33)	(0.15)	(0.25)	(0.64)	(0.80)	(0.29)
whol × *post*	5.299**	0.380*	1.974**	7.113***	10.230**	2.917***
	(1.38)	(0.15)	(0.71)	(1.72)	(2.65)	(0.40)
R^2	0.323	0.238	0.295	0.377	0.287	0.920
Panel C: Jane Street						
whol	9.045***	1.318**	4.067***	9.712***	22.727***	−1.878***
	(0.51)	(0.34)	(0.17)	(0.90)	(1.44)	(0.31)
whol × *post*	−1.998**	0.451	0.599	−1.300	−10.939***	−2.862**
	(0.75)	(0.34)	(0.57)	(1.18)	(1.39)	(0.84)
R^2	0.239	0.043	0.095	0.229	0.257	0.832

The table examines changes in execution costs offered by wholesalers from April–June 2021, when Jane Street has a small market share, to the last three months of 2021, when Jane Street has established itself as a sizeable wholesaler. It reports coefficient estimates from the following difference-in-differences regression:

$$depvar_{ijt} = \alpha_{it} + \beta_1 whol_j + \beta_2 whol \times post_{jt} + \varepsilon_{ijt},$$

where $depvar_{ijt}$ is either the realized spread in stock i for intermediary j (wholesaler or exchange) in month t or the effective spread, $whol$ is a dummy variable that has a value of 1 for orders executed by wholesalers and 0 for orders executed by exchanges, and $post$ is a dummy variable that has a value of 1 after Jane Street market share capture and 0 otherwise. We run regressions separately for the full sample and each subsample in specifications 1 through 5 for realized spreads, and for the full sample in specification 6 for effective spreads. In Panel A, the model is estimated for all wholesalers, while in Panels B and C, it is estimated separately for the incumbents and Jane Street. The models are estimated using stock volume weights and include stock-by-month fixed effects, with standard errors double-clustered by stock and month. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

3 to break into the market, and achieving scale could have made these prices sustainable.

In theories of monopolistic competition, new entry occurs if prices are above the entrant's projected average cost. Therefore, a wholesaler may consider entering if demand for the product has recently expanded or is expected to expand. A wholesaler may also enter if it has a cost advantage, such as access to better technology or risk management practices, or if it is willing to cross-subsidize the new market from its other businesses for a period of time with the expectation of taking market share from the incumbents. In our case, Jane Street's entry appears to have taken market share from the incumbent wholesalers, triggering losses of economies of scale and causing them to raise prices. This result is consistent with markets that are monopolistically competitive, where a reduction in demand per firm is associated with higher average production costs per unit.

We find that incumbent wholesalers increase execution costs after Jane Street captures a significant market share, which we attribute to a loss of economies of scale. By contrast, Ernst et al. (2023) and Huang et al. (2024) find that execution quality improves after the broker Robinhood begins routing to Jane Street, suggesting that increased competition among wholesalers helps lower execution costs. Two observations reconcile these seemingly conflicting findings. First, by the time Robinhood began routing to Jane Street in December 2021, Jane Street had been operating for over 28 months and had reached a market share plateau of approximately 12%. As a result, most of the market-wide effects of Jane Street's entry, which is what we focus on, had already occurred before Robinhood changed its routing. Second, prior to adding Jane Street, Robinhood relied on four wholesalers. Huang et al. (2024) document that only one of these four significantly reduced execution costs for their odd-lot orders when Robinhood added Jane Street, yet it still lost 20% of its Robinhood flow. The other incumbents did not significantly reduce execution costs; however, one saw no loss of order flow, while another lost a quarter of its flow. This outcome is consistent with what one would expect in a monopolistically competitive market with differentiated products. In this environment, brokers can leverage their ability to reroute order flow across wholesalers to lower execution costs for retail investors, though the success of these efforts ultimately depends on the wholesalers' cost structure and product characteristics.

5. Conclusion

The U.S. retail trading volume, which constitutes nearly 20% of total volume, is primarily executed off-exchange by intermediaries known as wholesalers. This practice has sparked a debate, mainly due to the concentration of the wholesale business, prompting the SEC to contemplate introducing new rules to encourage additional competition for retail executions. Conversely, brokers and wholesalers state that wholesalers compete vigorously for retail flow, and that retail brokers choose to execute through wholesalers in the best interest of their clients.

Our data tend to support the latter claims. Retail investors receive a price improvement of roughly one-quarter of the spread. The evidence also suggests that it is the retail brokers who are in control. Brokers are large, with the largest one surpassing the largest wholesaler. They closely monitor wholesaler performance, rewarding the best performers with more order flow and reducing allocations to those that perform poorly. Furthermore, the wholesale environment is characterized by economies of scale. The largest wholesalers are able to provide liquidity at lower cost, and broker oversight ensures that these savings transfer to retail customers. The wholesale market is also open to entry as evidenced by a new wholesaler gaining a substantial market share during our sample period. With the arrival of this wholesaler, the incumbents lose a substantial portion of their economies of scale, leading to an increase in retail customer trading costs.

CRediT authorship contribution statement

Anne Haubo Dyhrberg: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources,

Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Andriy Shkilko:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Ingrid M. Werner:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Andriy Shkilko and Anne Haubo Dyhrberg report financial support was provided by Government of Canada Social Sciences and Humanities Research Council. Ingrid M. Werner reports board memberships with Dimensional Mutual Funds and ETFs, Royal Swedish Academy of Sciences, and Fourth Swedish Pension Fund (AP4). Ingrid M. Werner reports consulting or advisory relationships with Danish Finance Institute, Swiss Finance Institute Geneva, and Swedish House of Finance.

References

Aitchison, J., 1986. The statistical analysis of compositional data. Monogr. Statist. Appl. Probab.

Amihud, Y., Mendelson, H., 1980. Dealership market: Market-making with inventory. J. Financ. Econ. 8 (1), 31–53.

Arrow, K.J., 1971. The theory of risk aversion. In: Individual Choice under Certainty and Uncertainty, collected papers of Kenneth J. Arrow, 1984. Harvard University Press, pp. 147–171.

Asplund, M., Nocke, V., 2006. Firm turnover in imperfectly competitive markets. Rev. Econ. Stud. 73, 295–327.

Autor, D., Dorn, D., Katz, L.F., Patterson, C., Van Reenen, J., 2017. Concentrating on the fall of the labor share. Am. Econ. Rev. (Pap. & Proceedings) 107, 180–185.

Autor, D., Dorn, D., Katz, L.F., Patterson, C., Van Reenen, J., 2020. The fall of the labor share and the rise of superstar firms. Q. J. Econ. 135, 645–709.

Bartlett, R.P., McCrary, J., O'Hara, M., 2025. The Market Inside the Market: Odd-Lot Quotes. Rev. Financ. Studies 38 (3), 661–711.

Battalio, R.H., Jennings, R.H., 2023. Wholesaler execution quality. Working paper, University of Notre Dame.

Bessembinder, H., 2003. Selection biases and cross-market trading cost comparisons. Working paper, University of Utah.

Boehmer, E., 2005. Dimensions of execution quality: Recent evidence for US equity markets. J. Financ. Econ. 78 (3), 553–582.

Boehmer, E., Jennings, R., Wei, L., 2007. Public disclosure and private decisions: Equity market execution quality and order routing. Rev. Financ. Stud. 20 (2), 315–358.

CEA, 2016. Benefits of competition and indicators of market power. Counc. Econ. Advis. https://bit.ly/3wCNKic. (Accessed 17 January 2025).

Chamberlin, E.H., 1933. The Theory of Monopolistic Competition. Harvard University Press.

Citadel Securities, 2019. Q1-2019 FIF supplemental retail execution quality statistics. https://bit.ly/3WnhsSi. (Accessed 17 January 2025).

Coenders, G., Pawlowsky-Glahn, V., 2020. On interpretations of tests and effect sizes in regression models with a compositional predictor. SORT-Stat. Oper. Res. Trans. 44 (1), 201–220.

Demsetz, H., 1973. Industry structure, market rivalry, and public policy. J. Law Econ. 16, 1–9.

Denski, J., Sappington, D., Spiller, P., 1987. Managing supplier switching. RAND J. Econ. Spring, 77–97.

Dixit, A.K., Stiglitz, J.E., 1977. Monopolistic competition and optimum product diversity. Am. Econ. Rev. 67 (3), 297–308.

Eaton, G.W., Green, T.C., Roseman, B.S., Wu, Y., 2022. Retail trader sophistication and stock market quality: Evidence from brokerage outages. J. Financ. Econ. 146 (2), 502–528.

Economist, 2016. Too much of a good thing. The Economist https://econ.st/4bb8yN2. (Accessed 17 January 2025).

Ernst, T., Malenko, A., Spatt, C., Sun, J., 2023. What does best execution look like? Working paper, University of Maryland.

Ernst, T., Spatt, C.S., Sun, J., 2024. Would order-by-order auctions be competitive? J. Finance (Forthcoming).

FINRA, 2022. Market cap explained. https://bit.ly/44TgcJw. (Accessed 17 January 2025).

Focarelli, D., Panetta, F., 2003. Are mergers beneficial to consumers? Evidence from the market for bank deposits. Am. Econ. Rev. 93, 1152–1172.

Greenacre, M., 2021. Compositional data analysis. Annu. Rev. Stat. Appl. 8, 271–299.

Grullon, G., Larkin, Y., Michaely, R., 2019. Are US industries becoming more concentrated? Rev. Financ. 23 (4), 697–743.

Hagströmer, B., 2021. Bias in the effective bid-ask spread. J. Financ. Econ. 142 (1), 314–337.

Ho, T., Stoll, H.R., 1981. Optimal dealer pricing under transactions and return uncertainty. J. Financ. Econ. 9 (1), 47–73.

Ho, T.S., Stoll, H.R., 1983. The dynamics of dealer markets under competition. J. Financ. 38 (4), 1053–1074.

Hoffmann, P., Jank, S., 2024. What is the value of retail order flow? Deutsche Bundesbank Discussion Paper No. 33/2024.

Huang, X., Jorion, P., Lee, J., Schwarz, C., 2024. Who is minding the store? Order routing and competition in retail trade execution. Working paper, Washington University in St. Louis.

Huang, K., Simchi-Levi, D., Song, M., 2012. Optimal market-making with risk aversion. Oper. Res. 60 (3), 541–556.

Lipson, M., 2003. Competition among market centers. Working paper, University of Virginia.

Mackintosh, P., 2020. Quantifying the cost of maker-taker markets. https://bit.ly/3UGS7lA. (Accessed 17 January 2025).

Mackintosh, P., 2023. The 2023 intern's guide to the market structure galaxy. https://bit.ly/3vmqp3L. (Accessed 17 January 2025).

Muravyev, D., Pearson, N.D., 2020. Option trading costs are lower than you think. Rev. Financ. Stud. 33 (11), 4973–5014.

O'Hara, M., Ye, M., 2011. Is market fragmentation harming market quality? J. Financ. Econ. 100 (3), 459–474.

Paravisini, D., Rappoport, V., Ravina, E., 2017. Risk aversion and wealth: Evidence from person-to-person lending portfolios. Manag. Sci. 63 (2), 279–297.

Peltzman, S., 1977. The gains and losses from industrial concentration. J. Law Econ. 20, 229–263.

Saul, D., 2023. Retail trading just hit an all-time high. Forbes https://bit.ly/3oSBvtB. (Accessed 17 January 2025).

Schwab, 2022. U.S. equity market structure: Order routing practices, considerations, and opportunities. https://bit.ly/3RhOpNI. (Accessed 17 January 2025).

SEC, 2000. Final rule: Disclosure of order execution and routing practices. https://bit.ly/3zyrpB1. (Accessed 17 January 2025).

SEC, 2022. Order competition rule. https://bit.ly/3v1Z96V. (Accessed 17 January 2025).

Seligson, P., Doherty, K., 2024. Jane street scores $10.6 billion trading haul. Bloomberg News, https://bloom.bg/4fBMm0i. (Accessed 17 January 2025).

Song, M., 2010. Applications of Stochastic Inventory Control in Market-Making and Robust Supply Chains (Dissertation). Department of Civil and Environmental Engineering, Massachusetts Institute of Technology.

Wagner, S.M., Friedl, G., 2007. Supplier switching decisions. European J. Oper. Res. 183 (2), 700–717.

Wigglesworth, R., 2021. Jane street: the top wall street firm 'no one's heard of'. Financial Times, https://on.ft.com/4c4BHJK. (Accessed 17 January 2025).