The Retail Execution Quality Landscape*

Anne Haubo Dyhrberg[†] **Andriy Shkilko**[‡] **Ingrid M. Werner**[§]

July 19, 2023

Abstract. Using a comprehensive multi-year U.S. dataset, we show that off-exchange (wholesaler) executions tend to benefit retail investors by separating their flow from the more toxic non-retail flow. Although the wholesale industry is concentrated, three findings suggest that wholesalers may not abuse market power. Firstly, brokers reward wholesalers who offer lower liquidity costs with more order flow. Secondly, the largest wholesalers offer the lowest costs, due to economies of scale. Finally, the entry of a new large wholesaler does not result in a reduction of liquidity costs.

Key words: Retail Trading, Wholesalers, Execution Quality

JEL: G20; G24; G28

*We thank James Brugler, Sabrina Buti, Doug Clark, Laurence Daures, Tom Ernst, Marinela Finta, Corey Garriott, Carole Gresse, Peter Haynes, David Hecht, Bob Jennings, S.P. Kothari, Josh Mollner, Dmitriy Muravyev, Peter Reiss, Vincent Skiera, Chen Yao, Marius Zoican, and conference/seminar participants at the China International Conference in Finance, Conference on Financial Market Regulation, Case Western Reserve University, Indiana University, Macquarie University, McMaster University, NBER Big Data and High-Performance Computing for Financial Economics Conference, the Toronto Stock Exchange, the University of Toronto, and Université Paris Dauphine-PSL for valuable comments. The most recent version may be downloaded from: https://bit.ly/3ATa2v5.

[†]Wilfrid Laurier University, Canada, e-mail: adyhrberg@wlu.ca

[‡]Wilfrid Laurier University, Canada, e-mail: ashkilko@wlu.ca

[§]Ohio State University, United States of America, and CEPR, e-mail: werner.47@osu.edu

1. Introduction

In the United States, trading volume generated by retail investors represents close to 20% of total trading volume.[1] Retail brokers typically send customer orders to over-the-counter market making firms known as *wholesalers*. Wholesalers internalize liquidity demanding orders by buying from retail sellers and aiming to re-sell to retail buyers, capturing the bid-ask spread. A portion of the spread is retained by the wholesaler, another portion goes to the retail broker as payment for order flow (PFOF), and yet another portion is passed on to the retail trader as price improvement. While wholesalers internalize liquidity demanding orders, they send liquidity providing orders to exchanges since regulation requires such orders to be displayed.

Many markets around the globe fully or partially prohibit internalization of retail flow, so the U.S. landscape is rather unique. In the other markets that allow for internalization, the share of internalized retail orders is relatively low, and the majority of such orders still end up on exchanges.[2] In the U.S., the situation is reverse, and exchanges typically see only a minor portion of retail flow, usually representing the exhaust that wholesalers do not wish to internalize.

Which approach to handling retail flow brings more benefits to retail investors is an empirical question. On the one hand, separating retail and institutional investors may benefit the former as their orders cost less to intermediate, allowing for lower liquidity costs. On the other hand, the wholesale market segment may not be sufficiently competitive, and the liquidity cost savings may be captured by the wholesalers instead of being transferred to retail investors.[3]

How retail orders are handled is currently actively debated. On the one hand, some observers argue that wholesalers wield (and abuse) market power and provide limited benefits to retail in-

[1] "Retail Trading Just Hit An All-Time High," by D. Saul, Forbes, February 3, 2023 (https://bit.ly/3oSBvtB).

[2] For instance, in Canada, more than 90% of retail flow executes on exchanges. See "Exchange Q&A: Adam Inzirillo, Cboe Global Markets" (https://bit.ly/3NoArqa).

[3] In the U.S., two wholesalers – Citadel Securities and Virtu Financial – capture more than 70% of retail flow. Industry participants often express concerns with this level of market concentration. See, "IEX Supports SEC Equity Market Proposals," by A. Lyudvig, Traders Magazine, March 22, 2023 (https://bit.ly/3Hks0K2).

vestors. They suggest that price improvement offered by wholesalers tends to be *de minimis*, or the smallest amount possible. To enhance competition in this segment, the Securities and Exchange Commission (SEC) is considering implementing a system of auctions, in which wholesalers would compete for each individual retail order, obtaining execution rights only if they provide the largest amount of price improvement. On the other hand, wholesalers argue that it is retail brokerage firms that have market power, and that they route orders to wholesalers because it is in the best interest of their retail clients. Wholesalers claim that they offer significant price improvement, and that retail investors are well-served by the current system.

Reconciling these possibilities is an empirical task, and we do so using public data contained in the SEC Rule 605 reports.[4] Each order-handling venue must file such reports on a monthly basis to maintain a public record of execution quality. In a comprehensive sample of all U.S. equities traded from January 2019 through March 2022, we carefully compare the benefits of wholesaler and exchange executions and find that retail orders are better off being routed to wholesalers. We also show that abolishing the wholesaler system would cost retail investors close to a billion dollars per month in additional trading costs.

Notably, the data show that both wholesalers and exchanges offer unique execution benefits. On their part, wholesalers provide substantial price improvement, that is, they execute liquidity-demanding orders at prices that are better than those offered by exchanges. Wholesaler price improvement is far from *de minimis*. An average retail order in our sample receives price improvement of 24% of the quoted spread.[5] More strikingly, a retail order in an average S&P 500 stock receives price improvement of 44% of the quoted spread. By comparison, price improvement offered by exchanges is only 3% in the full sample and 6% in S&P 500 stocks. As such,

[4]A recent analysis by the SEC shows that these reports are highly consistent with the audit trail data available to the agency. See Release 34-96495 "Order Competition Rule" from December 14, 2022 (https://bit.ly/3v1Z96V).

[5]For example, if a stock trades at $50.00 on the bid and $50.10 on the offer, a retail trader would typically purchase it at $50.088 instead of $50.10 and sell at $50.012 instead of $50.00. The average quoted spread in our sample is $0.091.

along one dimension of execution quality, wholesalers offer a clear advantage.

Another important execution quality dimension is the cost of generating liquidity. Insofar as liquidity is supplied by professional market makers, including wholesalers, two considerations come into play. First, market makers incur three costs: the adverse selection cost, the cost of holding inventory, and the technology cost. Second, they aim to make profits. Researchers typically measure the adverse selection cost by estimating trade *price impacts*, the difference between the midquote at the time of a trade and the midquote at a future time. Adverse selection cost arises when midquotes increase after a buyer-initiated trade and decrease after a seller-initiated trade. Wholesalers are professional market makers facing notably lower adverse selection cost. In our sample, order flow received by them generates 31% less price impact than the exchange-bound flow.

Adverse selection costs aside, the inventory and technology costs as well as market making profits are captured by a conventional metric, the *realized spread*, which is the difference between the effective spread (the market maker's liquidity provision revenue) and the price impact. The data show that exchanges offer realized spreads that are substantially lower than those offered by wholesalers. How does the difference in realized spreads arise? On exchanges, limit orders submitted by market making algorithms compete with limit orders submitted by non-market making algorithms. The latter operate multiple strategies: from limit order legs of latency arbitrage (Aquilina, Budish, and O'Neill (2021)) to managing institutional investment positions (O'Hara (2015)). Covering market making costs and earning liquidity provision profits is not as important to non-market making algorithms as to their market making counterparts, if at all. Therefore, exchange realized spreads should be smaller than those observed in a pure market maker setting. According to industry estimates, pure market making algorithms represent about 16% of all liquidity provision on modern exchanges.[6] With such a split, exchange liquidity should be cheaper than wholesaler liquidity, as confirmed by our data.

[6]"Who is Trading on U.S. Markets?" by P. Mackintosh, January 28, 2021 (https://bit.ly/3za9W1k).

We examine Rule 605 reports from 14 U.S. exchanges and from the largest eight wholesalers for a sample of almost 2.9 trillion traded shares.[7] We focus on liquidity-demanding orders because wholesalers mainly internalize these orders. For wholesalers, the data primarily cover retail flow. For exchanges, the data predominantly cover institutional flow because wholesalers absorb most of the retail flow. For simplicity, in the following sections, we refer to trades executed by wholesalers as *retail* and trades executed by exchanges as *institutional*. The main analysis focuses on 8,165 equities other than ETFs. We report the results for a sample of ETFs in the Appendix.

The data show that while institutions tend to time their liquidity demand to periods of relatively low trading costs, retail investors engage in such timing substantially less. A typical liquidity-demanding retail order executes when the quoted spread is 33% greater than when a typical institutional order executes. In this light, price improvement offered by wholesalers is quite important to mitigate the trading costs facing retail investors.

The aforementioned distinctions between retail and institutional flows enable us to consider what would happen if retail flow were to shift to exchanges. Such proposals are often heard in current market structure discussions.[8] On the one hand, moving to exchanges would benefit retail investors by reducing realized spreads they pay. On the other hand, retail investors would lose price improvements provided by wholesalers and would face higher adverse selection costs resulting from being pooled with institutional flow. When we analyze the overall impact of such a move on retail traders, we find that they would generally be worse off on exchanges. Retail investor losses would be accompanied by gains to institutional traders, whose costs would decline due to lower on-exchange toxicity. Consequently, the transfer of retail flow to exchanges would essentially subsidize institutional flow at the expense of retail flow.

[7]The exchanges include BATS, BYXX, EDGA, EDGX, IEX, MEMX, Nasdaq, NSDQ Boston, NSDQ Philadelphia, NYSE, NYSE American, NYSE Arca, NYSE Chicago, and NYSE National. The wholesalers include Citadel, G1, Jane Street, Merrill Lynch, Morgan Stanley, Two Sigma, UBS, and Virtu.

[8]For example, the SEC has received numerous comment letters from investors, primarily retail, advocating for a significant reduction in off-exchange retail trading. See for instance the form letter promulgated by the We The Investors group: https://bit.ly/434ceeE.

Our analysis and conclusions are generally conservative, as we focus exclusively on execution costs. We, however, note that the current system of retail order flow handling is inextricably linked to commission-free trading for retail investors. The PFOF payments that brokers receive from wholesalers subsidize brokerage business, allowing the brokerages to operate without charging commissions. If the current system were to be dismantled, commissions may again be necessary, increasing the overall cost of retail market participation. Furthermore, if the return of commissions leads to a decline in retail volumes, moving retail flow to exchanges will be less effective in diluting the current toxicity levels making the move even less attractive.

The wholesale industry exhibits a relatively high level of concentration, with Citadel and Virtu capturing over 70% of retail flow. Furthermore, none of the stocks in the sample would be classified by the U.S. Department of Justice (DOJ) as having a competitive wholesale environment. The DOJ often relies on the Herfindahl-Hirschman Index (HHI) for such classification, and for each stock in the sample, the HHI indicates either moderate or high concentration among wholesalers. It is not surprising then that some market observers express concerns about the potential market power abuses by wholesalers. Although we cannot directly observe whether such abuses occur, the data offer several indicators that lead us to think otherwise.

First, the largest retail brokerage is larger than the largest wholesaler. This suggests that market power could potentially reside with retail brokerages rather than wholesalers. To examine this possibility, we ask if retail brokerages actively manage their relationships with wholesalers by favoring those that offer lower liquidity costs. The data indicate that they indeed do so. Wholesalers that provide lower costs today are rewarded with additional order flow in the future. Interestingly, brokerages appear to evaluate wholesalers not on a stock-by-stock basis but rather on a bundled basis. In other words, if Citadel offers the cheapest liquidity in AAPL, it will not necessarily receive more future AAPL flow. Instead, Citadel must outperform its competitors across the entire range of stocks to attract more order flow in AAPL or any other stock.

This finding highlights an intriguing aspect of the retail ecosystem, where the brokerages compel wholesalers to compete in stocks that have relatively low trading frequency and high inventory costs. Typically, small low-volume stocks are less attractive to intermediaries due to their lower profitability, and market regulators and exchanges often seek ways to improve liquidity in such stocks (Foley, Liu, Malinova, Park, and Shkilko (2023)). When we account for inventory costs, our analysis suggests that wholesalers tend to charge relatively low liquidity costs in small stocks compared to large stocks, pointing to cross-subsidization facilitated by bundling. The proposed SEC retail trading reform aims to introduce order-by-order competition in the retail segment, which would eliminate bundling. Our analysis suggests that this reform may result in poorer intermediary-supplied retail liquidity in small stocks.[9]

Second, considering that the top two wholesalers capture more than two-thirds of retail flow, their market power could potentially lead to higher costs for retail investors. Contrary to this expectation, we observe that Citadel and Virtu charge the lowest liquidity costs, even though they handle the most toxic retail flow. Therefore, if any market power abuses exist, they are not immediately evident from our data. Furthermore, when we account for the trading volume received by each wholesaler, we find that the relatively low liquidity costs charged by the top two are entirely explained by the scale of their operations. This suggests that economies of scale play a significant role in generating cost savings at the wholesaler level, and broker monitoring facilitates the transfer of these savings to retail customers.

Lastly, the dynamics of wholesaler competition undergo a transformation during our sample period with the entry of a new player, Jane Street. Within a few months, Jane Street gains a significant market share, capturing close to 15% of retail flow. If wholesalers were indeed exploiting their market power and enjoying economic rents prior to this entry, we would expect competitive pressures to intensify, leading to lower liquidity costs. The data however do not support this conjecture; we find no evidence of a decrease in liquidity costs. In fact, in low-volume stocks, the costs increase, likely due to the incumbents' loss of economies of scale.

[9] Ernst, Spatt, and Sun (2023) come to a similar conclusion in a theoretical model of order-by-order auctions.

In summary, despite concerns expressed by critics, it does not appear that the marketplace for retail order flow is controlled by wholesalers. Instead, retail brokers seem to exert control over execution quality by routing orders to wholesalers that require lower compensation for providing liquidity. The marketplace is also contestable as evidenced by a new entrant successfully capturing a sizable market share from the incumbents in a surprisingly short time.

In our final analysis, we revisit the SEC's proposal to overhaul retail trading practices, which involves directing retail flow to auctions for order-by-order competition. The proposal assumes that non-professional liquidity providers such as hedge funds, mutual funds, and pension funds would demonstrate significant interest in engaging with retail flow and offer superior price improvement compared to wholesalers. However, our analysis of institutional trading data indicates that this assumption may only hold true for large stocks and may not apply to many stocks currently traded by retail investors. Additionally, considering our earlier analysis that eliminating bundling could potentially reduce the incentives for intermediaries to engage with retail traders in small stocks, we caution that many retail investors would likely experience lower execution quality if the SEC proposal were to be implemented.

Related literature. We aim to contribute to a growing literature that examines retail execution quality and the effects of wholesaler internalization on lit market quality. Notably, this literature has not yet reached a consensus. On the one hand, Adams, Kasten, and Kelley (2021), Kothari, So, and Johnson (2021), and Battalio and Jennings (2023) argue that wholesalers deliver retail trading costs that are lower than those offered by exchanges. Also, Jain, Mishra, O'Donoghue, and Zhao (2022) suggest that internalization revenues may boost the ability of market makers such as Citadel and Virtu to compete on exchanges, thereby improving overall liquidity.[10] Similarly, Baldauf, Mollner, and Yueshen (2023) show theoretically that market makers may use retail flows to offset inventory imbalances.

[10]Citadel, Virtu, and several other wholesalers perform a dual function in the modern marketplace. They serve both as major on-exchange market makers and wholesalers.

On the other hand, two recent studies show that internalization of retail orders may negatively affect overall liquidity via the inventory and market power channels. Eaton, Green, Roseman, and Wu (2022) find that some retail investors may increase market maker costs by occasionally herding and thus creating inventory imbalances. Market makers respond to herding by increasing overall exchange trading costs. In turn, Hu and Murphy (2022) argue that the wholesale industry is highly concentrated and the resulting non-competitive behavior leads to wider exchange spreads and small price improvement for retail customers.

Our study complements this literature in several ways. First, we provide a comprehensive analysis of retail trading costs that accounts for the wholesaler-supplied benefit of price improvement and the exchange-supplied benefit of lower realized spreads. Second, we show that even though the wholesale industry is indeed concentrated, it provides a significant net benefit to retail investors due to economies of scale. Third, we show that retail brokerages act as monitors, as they base future routing decisions on current wholesaler performance. Finally, we report that the entry of a new wholesaler does not reduce wholesaler spread capture, which is inconsistent with significant incumbent wholesaler market power.

Three concurrent studies come to overall retail execution quality conclusions similar to ours. Adams, Kasten, and Kelley (2021) identify retail trades using the algorithm developed by Boehmer, Jones, Zhang, and Zhang (2021), which has been recently shown to have limitations. In particular, the algorithm tends to miss retail trades around the midquote and retail trades that do not receive price improvement. It also tends to mix institutional executions (e.g., VWAP trades) with retail executions (Barber, Huang, Jorion, Odean, and Schwarz (2022)).

Kothari, So, and Johnson (2021) use proprietary data from the Robinhood brokerage. While their study delivers valuable insights, a complementary comprehensive analysis across multiple brokerages may help shed light on the external validity of their inferences. Eaton, Green, Roseman, and Wu (2022) and Schwarz, Barber, Huang, Jorion, and Odean (2022) show that Robinhood trader behaviour and execution quality tend to differ from those observed for the other retail brokerages. Battalio and Jennings (2023) also use proprietary data, but from only one wholesaler and only for May of 2022.

We complement these studies by carefully analyzing a multi-year comprehensive public dataset that academic researchers have only cursorily examined. According to industry participants, this dataset allows for the cleanest identification of retail order flow that is possible without proprietary data. Not only does this dataset enable us to speak about the external validity of the results, but it also allows for an analysis of competitive forces by observing interactions between multiple wholesalers and retail brokerages.

Compared to the retail trading literature for equities, the literature that examines retail trading in options is in relative consensus. Ernst and Spatt (2022) suggest that options markets provide less price improvement compared to the equity markets and that retail brokerages have an incentive to nudge their customers into options trading, which is more profitable for the brokerages yet detrimental to customer investment returns. Along similar lines, Bryzgalova, Pavlova, and Sikorskaya (2022) argue that options market makers behave non-competitively and disproportionately benefit from the growth in retail trading. Finally, Hendershott, Khan, and Riordan (2022) show that options wholesalers engage in cream-skimming of less informed trades into auctions and suggest that eliminating the auction structure may result in lower liquidity costs overall.

2. Data and Sample

We obtain monthly order execution data from a service provider that focuses on compliance and trade analytics. The service provider compiles publicly available Rule 605 and Rule 606 reports filed by execution venues in the U.S. and generously makes the resulting data available to us. The Rule 605 data cover the period from January 2019 through March 2022 and are described in more detail in the Appendix.

SEC Rule 605 applies to market and limit orders that are executed during regular trading hours and contain no special handling instructions. We refer to them as basic orders.[11] As an example of a special instruction, a trader may ask that a limit order is not forwarded to venues other than the original receiving venue, an order known as *Do Not Ship*. Li, Ye, and Zheng (2020) show that orders that contain special instructions are typically submitted by relatively sophisticated traders such as high-frequency traders. As such, Rule 605 data cover virtually all retail orders and also include basic institutional orders. We focus exclusively on liquidity-demanding orders because wholesalers are required to forward liquidity-providing retail orders to exchanges. Furthermore, motivated by our discussions with industry participants, we group market orders and marketable limit orders together as marketable orders.

Rule 605 data include a wide range of securities (over 16,400 unique symbols). We restrict our sample to non-ETF ordinary and Class A, B, and C shares for a total of 8,165 symbols and refer to them as *stocks*.[12] We also create four sub-samples consisting of S&P 500 stocks and size-based terciles (T1-T3) of non-S&P 500 stocks (see the Appendix for details).

The data cover 70 execution venues, including all stock exchanges, all major wholesalers, many dark pools, crossing networks, etc. We focus on the first two venue categories, that is fourteen stock exchanges and the eight largest wholesalers. Table 1 reports trading volumes and market shares of all exchanges and wholesalers. Panel A shows that exchanges execute the majority, 58.76%, of Rule 605 orders with wholesalers capturing the remaining 41.24%. Our conversations with industry participants indicate that the flow routed to wholesalers consists predominantly of retail orders, while the flow routed to exchanges is mainly institutional orders. In later tests, we provide empirical support to this view.

[Table 1]

Panel B of Table 1 contains statistics for the individual exchanges and wholesalers. Among

[11] For details, see "Final Rule: Disclosure of Order Execution and Routing Practices," 17 CFR Part 240 (https://bit.ly/3zyrpB1).

[12] Summary statistics for 3,241 ETFs are provided in the Appendix.

exchanges, the leading roles are played by Nasdaq and the NYSE/NYSE Arca that respectively execute 17.27% and 17.85% (=10.44+7.41) of order flow. Among wholesalers, Citadel and Virtu stand out as the largest, capturing respectively 16.60% and 12.58% of order flow. Other wholesalers are considerably smaller, with the third largest, G1, processing 5.17%, and the next two, Two Sigma and UBS, processing 2.25% and 1.77%, respectively. Overall, the dataset contains information on execution quality for orders representing almost 2.9 trillion executed shares, which amounts to about 45% of trading volume reported by CRSP during the sample period.

Market structure studies typically rely on a set of execution quality metrics that consists of quoted, effective, and realized spreads as well as price impacts. The *quoted spread* is the difference between the national best offer (the offer quote that is the lowest across all lit markets) and the national best bid (the bid quote that is the highest across lit markets). It represents trading costs advertised by liquidity providers. Liquidity demanders do not always incur these costs exactly as advertised. Their orders may be price improved as is often done by wholesalers, or interact with better-priced non-displayed orders on exchanges (Bartlett, McCrary, and O'Hara (2022)). To assess trading costs actually incurred by liquidity demanders, Rule 605 data contain the *effective spread* computed as twice the signed difference between the traded price and the midquote (the average of the best offer and the best bid) at the time of the trade. Trade signs are observed by the filers and therefore do not need to be inferred using an algorithm such as Lee and Ready (1991).

Effective spreads are typically further divided into two components. The first component, the *price impact*, captures toxicity of a trade by computing the change in the midquote between the trade time and a future point in time. A buyer(seller)-initiated trade followed by a positive (negative) midquote change is considered informed and contributes to the adverse selection cost of market making. The second component, the *realized spread* is the difference between the effective spread and the price impact. The realized spread is a composite metric that captures (i) the costs of market making that are unrelated to adverse selection (i.e., inventory and fixed costs as well as trading fees); and (ii) market maker profits (Hendershott, Jones, and Menkveld (2011), Brogaard, Hagströmer, Nordén, and Riordan (2015)). Because of the composite nature of the metric, its interpretation is somewhat nuanced, and the upcoming discussions carefully take these nuances into account.

Rule 605 requires that price impacts and realized spreads are estimated at the 5-minute horizon. This horizon is an eternity in the modern marketplace, and we approach the interpretation of these statistics with due caution. This said, an analysis of the Trade and Quote data shows that adverse selection costs are incurred quickly, while market makers trade out of positions slowly. As a result, price impacts estimated at 5-minute horizons are rather effective at capturing adverse selection costs.

Rule 605 data exclude odd lots, so our analysis is restricted to the orders of 100 shares or more.[13] When working with the metrics, we remove outliers by trimming all variables at the 0.1 and 99.9 percentiles. Reporting of the quoted spread is not required by Rule 605, and we derive it from the other metrics as discussed in the Appendix. We scale all metrics by the CRSP closing stock price and use share volume-weighted averages.[14]

3. Execution Quality

3.1 Wholesalers vs. Exchanges

Table 2 reports summary execution quality metrics for our sample. During the sample period, wholesalers execute 146.36 million shares in an average sample stock, whereas exchanges execute 207.65 million shares. Rule 605 requires that venues report how their executions compare to

[13]Data from an industry initiative titled Financial Information Forum (FIF) include odd-lots and suggest that odd-lot market quality is similar to that reported for orders of other sizes, and especially the orders in the 100-499-share bin. See for example “Q1-2019 FIF Supplemental Retail Execution Quality Statistics Citadel Securities LLC” (https://bit.ly/3m2RC33).

[14]The SEC instead uses dollar-volume-weighted averages in their recent analysis of retail execution quality. See Release 34-96495 “Order Competition Rule" from December 14, 2022 (`https://bit.ly/3v1Z96V`).

the NBBO. Wholesalers price-improve a substantial portion, 65.71%, of order flow they receive, whereas exchanges only price-improve 9.49%. However, exchanges fare better than wholesalers with respect to their ability to match the existing NBBO, executing 98.34% of shares at the NBBO prices or better versus 92.98% by the wholesalers. Institutional traders typically split larger orders to avoid walking the book (slippage), and since their orders are predominantly routed to exchanges this may account for the difference in the proportion of flow that matches the NBBO.

[Table 2]

Notably, wholesalers tend to execute when the NBBOs are relatively wide, 64.92 bps vs. the exchange equivalent of 48.67 bps, a 33% difference. This difference cannot be attributed to wholesaler choices because commercial agreements with retail brokerages do not allow wholesalers to choose what orders to execute and when. Rather, wholesalers are required to execute all orders routed to them. As such, the difference in quoted spreads must be driven by trader decisions and is perhaps expected given the clienteles served by wholesalers and exchanges. Many institutional liquidity-taking algorithms time their activity to periods of narrow quoted spreads. When spreads are wide, they either switch from liquidity demand to liquidity supply or reduce trading altogether. Retail traders are much less likely to engage in such timing. Since the metrics in Table 2 are volume-weighted, it is not surprising that liquidity-demanding exchange trades (institutional flow) tend to occur when spreads are relatively narrow.

Even though retail trade executions occur when quoted spreads are relatively wide, the differential is reduced significantly once we account for the substantial price improvement wholesalers provide to retail flow. Consequently, effective spreads reported by wholesalers are much closer to those reported by their exchange counterparts, at 49.06 bps and 46.98 bps, respectively. With this in mind, we suggest that an execution quality metric appropriate for our setting should account for both quoted and effective spreads. We adopt a ratio of effective to quoted spreads as such a metric. In Table 2, this ratio is 0.76 for wholesalers, suggesting that orders executed by them pay

76% of the prevailing quoted spread, and 0.97 for exchanges. Within the existing market structure, wholesalers therefore appear to play a valuable role. They provide substantial, rather than *de minimis*, price improvement that may not be available from the exchanges when retail trades are executed.

As we discuss above, market structure studies typically distinguish between two components of the effective spread. One such component is price impact that captures the adverse selection cost associated with a trade. The other is the realized spread that reflects three important market making considerations: inventory costs, fixed costs, and profits. Table 2 confirms our earlier assertion that wholesalers obtain order flow that is considerably less toxic (price impact of 32.53 bps) than that routed to exchanges (price impact of 47.32 bps). These figures are consistent with the statements by the industry participants that retail order flow is predominantly routed to wholesalers, while exchanges end up receiving mainly institutional flow.

Given similar effective spreads and lower price impacts, wholesalers earn substantially larger realized spreads compared to those earned by exchanges, 16.53 vs. -0.34 bps. At first glance, this large difference may appear suggestive of excess profits earned by wholesalers; however, it is important not to over-interpret these figures. Liquidity on exchanges is only partially provided by professional market makers. For instance, Nasdaq attributes only 16% of liquidity provision to pure market making strategies. The remaining liquidity-providing orders are submitted by non-market markers, whose main goal is to manage positions rather than earn spread revenue. The realized spreads that non-market makers earn are therefore not reflective of market making costs and profits. Since non-market makers' share of exchange liquidity provision is significant, caution should be used when comparing exchange realized spreads to wholesaler realized spreads. Put differently, the 16.53 bps realized spread earned by wholesalers may represent either a substantial profit, or a combination of inventory and fixed costs that allows only for a zero profit, or anything in-between. We examine this issue in more detail later in the manuscript.

So far, we have identified two important differences between wholesaler- and exchangeintermediated executions. On the one hand, wholesalers provide sizeable price improvement, while exchange price improvement is noticeably smaller. On the other hand, exchange liquidity providers earn considerably lower realized spreads. With these differences in mind, what would happen if retail order flow were to be moved to the exchanges?

To grasp the potential impact of this move, it is helpful to examine its outcomes, both positive and negative. On the positive side, retail flow will pay considerably less in realized spreads. However, there are also two negative outcomes to consider. Firstly, retail flow will receive considerably less price improvement. Secondly, when combined with institutional flow, retail investors will bear the cost of the resulting mix's higher toxicity, which surpasses that of pure retail flow. These three outcomes, taken as a whole, sum to a negative net effect of shifting retail investors to exchanges. We elaborate below.

To get a sense of the net effect, consider the following calculation based on Table 2. First, assume that retail volume relative to institutional volume, w, remains the same post-migration. Rule 605 data capture close to 45% of all volume traded in the U.S., with the remaining 55% mainly representing sophisticated institutional volume at 531.02 million shares in an average stock during the sample period. So, retail volume represents w = 16.54% = $146.36/(146.36 + 207.65 + 531.02)$ of total volume.

Second, assume that the price impact of retail trades originally routed to wholesalers remains the same once routed to exchanges, so that the average price impact on exchanges after the routing change is the volume-weighted average of price impacts prior to migration or 44.87 bps (= $w * 32.53 + (1 - w) * 47.32$). Third, assume that the realized spread of -0.34 bps on exchanges before the routing change becomes the required compensation for all liquidity providers on exchanges post migration. We relax this assumption shortly. This means that the imputed average effective spread, which is the sum of price impact and realized spread, is 44.53 bps post migration. Fourth, assume that there will still be opportunities to interact with non-displayed liquidity, so that all orders on exchanges post migration enjoy the same price improvements observed before the change, which means the average quoted spread will become 45.91 bps (44.53/0.97) post migration.

As noted above, marketable orders that currently arrive to exchanges are timed to when quoted spreads are narrow. This needs to be taken into account when estimating the likely spreads facing retail traders if their orders were routed to exchanges. Assuming that retail and exchange traders maintain their pre-migration order submission patterns, the imputed spreads on retail (exchange) orders post move need to be adjusted to reflect the pre-migration quoted spreads faced by retail (exchange) traders relative to the volume-weighted average quoted spreads pre-migration. In other words, the quoted spreads facing retail traders would be the volume-weighted average post-migration quoted spread of 45.91 bps multiplied by $64.92/(w*64.92+(1-w)*48.67)$ or 58.04 bps. Similarly, the quoted spreads facing exchange traders would be 45.91 bps multiplied by $48.67/(w*64.92+(1-w)*48.67)$ or 43.51 bps. In each case, the imputed quote spread is multiplied by the pre-migration price improvement on exchanges, 0.97, to get the imputed effective spread.

So far, we have assumed that current exchange realized spreads will remain unchanged at -0.34 if retail and institutional flows are pooled on exchanges. For this case, the bold line in Panel A of Table 3 shows that while institutions may benefit from such pooling, retail traders are likely to lose. For the existing exchange liquidity demanders (EXCH LDs), effective spreads would decline by 10.17%. For retail liquidity demanders (RET LDs), effective spreads would increase by 14.75%. The reasons for such changes are straightforward. Institutions would benefit from a substantial reduction in on-exchange adverse selection. Retail traders would experience a large reduction in realized spreads, but these gains would be offset by the loss of price improvements currently provided by wholesalers and by having to pay for the additional adverse selection coming from the institutional flow.

To shed additional light on the economic magnitude of these effects, Panel B reports total gains and losses for four market participant categories: RET LDs, EXCH LDs, exchange liquidity providers (EXCH LPs), and wholesalers (WHOL LPs). The gains represent total dollar amounts across all sample stocks during our entire sample period. Commensurate with the above-mentioned increase in effective spreads, the loss for RET LDs would be \$28.12 billion, whereas the gain for EXCH LDs would be a more substantial \$93.70 billion. Unsurprisingly, WHOL LPs will loose from the switch – a \$64.25 billion loss. Given that realized spreads are negative on exchanges, EXCH LPs will also loose a modest \$1.32 billion.

One of the assumptions that goes into the gains calculations is that realized spreads would not change if retail volume moved to the exchanges. To shed light on the importance of this assumption, we examine the sensitivity of the results to possible changes in realized spreads. To do so, we vary the realized spread figure of -0.34 bps by 0.1-bps increments. While the majority of conclusions discussed above remain qualitatively the same, they change for EXCH LPs, as their losses turn into gains once the realized spread turns less negative than in the base case. Most importantly, our conclusions for retail traders appear to be relatively insensitive to the non-increasing realized spreads assumption.

Although this assumption may appear brave, we believe that it is in fact quite conservative. The existing literature generally associates greater trading volumes with lower realized spreads (e.g., Bogousslavsky and Collin-Dufresne (2022)). In addition, Bessembinder, Carrion, Tuttle, and Venkataraman (2016) find that anticipated arrivals of large uninformed volume are accompanied by additional liquidity coming off the sidelines and improving market quality. With these results in mind, we cautiously suggest that realized spreads are more likely to decline from the status quo upon the addition of retail volume; however this decline must be exceptionally large to make the move to exchanges worthwhile for retail traders.

Our simple calculations in Table 3 omit two additional possibilities. First, consider what would happen if institutional order flow patters do not change in response to the arrival of retail traders. Since institutions time their marketable order to periods when exchange spreads are narrow, and retail orders arrive more uniformly over time, the share of retail flow in overall volume tends to be higher when quoted spreads on exchanges are wide. In other words, retail flow as a fraction of volume is greater when adverse selection is high. From the perspective of exchange liquidity providers, this means that retail orders offer additional benefits because they help lower average adverse selection costs exactly when these costs would otherwise be high. With competitive liquidity provision on exchanges, quoted spreads may fall further. To the extent that this effect is significant, our calculation may underestimate the benefits to retail from moving to exchanges.[15]

Second, consider allowing for a strategic response by institutions to the arrival of retail flow. The fact that retail trades disproportionately reduce quoted spreads when adverse selection is high imply that institutions have less incentives to focus their marketable orders to periods when spreads are narrow.[16] When institutions smooth out their order flow over time, and therefore increase trading when adverse selection is high, the additional diversification benefits we just described will be muted and may even be reversed. If this effect is large, our calculations may even underestimate the cost to retail from moving to exchanges.

3.2 Cross-Sectional Differences

Because we have a large cross-section, including many illiquid securities, we separately examine four sub-samples, the S&P 500 and size-based terciles of non-S&P 500 stocks labeled Tercile 1, Tercile 2, and Tercile 3. During the sample period, there are 514 stocks in the S&P 500 sub-sample[17] and the Tercile 1, Tercile 2, and Tercile 3 sub-samples include 2,550, 2,550, and 2,551 stocks respectively. In this section, we investigate if wholesaler involvement and execution quality differs between the four sub-samples.

[15] We thank Josh Mollner for clarifying this point. See Battalio and Holden (2001) for a model laying out the intuition. Simulations suggest that our qualitative conclusions are robust to including this effect.

[16] See, e.g., Kyle (1985), where informed traders strategically increase their volume when uninformed order flow increases.

[17] There are 503 stocks in the S&P 500 index during our sample period, and the additional stocks account for turnover within the index.

Table 4 shows that the differences across sub-samples are noticeable. Wholesalers represent 31.87% of share volume for S&P 500 stocks, but their share increases monotonically in size reaching a high of 63.79% for Tercile 3 stocks. In other words, retail flow plays an out-sized role for less liquid stocks, a point we will return to below.

[Table 4]

Next, we ask if market capitalization affects execution quality (Table 5). We begin with the S&P 500 sub-sample. Wholesalers price-improve 75% of marketable orders and provide price improvement corresponding to 44% of the quoted spread. By comparison, only 12% of marketable orders receive price improvement on exchanges, and price improvement is a modest 6%. The adverse selection that accrues on exchanges is 89% (= 6.45/3.41-1) greater than that accruing to wholesalers. Wholesalers earn substantially larger realized spreads than liquidity providers on exchanges, 1.21 vs. -1.23 bps.

[Table 5]

When it comes to terciles 1 though 3, the pattern discussed for the S&P 500 stocks is generally preserved. First, wholesalers price improve a substantially larger portion of marketable orders than exchanges for each sub-sample (e.g., 64% vs. 9% for tercile 2). Note also that the fraction of price improved orders falls as we move from larger to smaller size firms both for wholesalers and exchanges. Second, the magnitude of price improvement continues to be significantly larger for marketable orders routed to wholesalers for all terciles (e.g., 27% vs. 5% of the quoted spread for tercile 2). This metric is generally declining as we move from larger to smaller size firms for orders routed to wholesalers, but is relatively constant for orders routed to exchanges.

Order flow toxicity is substantially greater on exchanges, with exchange price impacts 51%, 43%, and 114% greater than at wholesalers for terciles 1, 2, and 3, respectively. Finally, the exchange realized spreads are even more negative for tercile 1 and 2 stocks than for S&P 500 stocks, but turn positive for tercile 3 stocks. By contrast, wholesalers earn realized spreads that are positive and increase as we move from larger to smaller size firms. We note that although the realized spreads obtained by wholesalers appear quite large, reaching 33.09 bps for tercile 3, they may be representative of substantial inventory and fixed costs incurred in these relatively infrequently-traded stocks. We therefore refrain from linking these figures to excessive profits earned by wholesalers and return to inventory costs shortly.

Prior market structure literature has linked execution quality to several market characteristics. Among these are price, trading volume and volatility. A higher price is typically related to lower execution costs because of the fixed tick size in the U.S. Greater volatility is typically associated with greater fundamental information flows, and as such may negatively affect execution quality through the adverse selection channel. In turn, with volatility controlled for, greater volume is typically associated with lower adverse selection as it is thought to represent uninformed flow. In Table 6, we examine the robustness of our findings to controlling for these characteristics in the following regression model:

$$DepVar_{ijt} = \alpha + \beta_1 WHOL_j + \beta_2 price_{it} + \beta_3 volatility_{it} + \beta_4 volume_{it} + \varepsilon_{ijt}, \tag{1}$$

where $DepVar_{it}$ is one of the following execution quality variables for stock i intermediary type j (wholesaler or exchange) in month t: the ratio of effective to quoted spread, quoted spread, effective spread, price impact, and realized spread as defined previously; $WHOL$ is a dummy variable that has a value of 1 for orders executed by wholesalers and 0 for orders executed by exchanges; $price$ is the natural log of the stock price; $volatility$ is the difference between the high and low prices scaled by the high price, and $volume$ is the natural log of trading volume. The regression controls for stock and month fixed effects and uses double-clustered standard errors.

We estimate equation (1) for the full sample in Panel A of Table 6. We are primarily interested in the coefficient on the $WHOL$ dummy but note that the control coefficients are significant and of the expected signs. The univariate findings we discussed earlier hold. Wholesaler executions tend to occur when quoted spreads are relatively wide. For example, the quoted spreads that prevail during wholesaler executions are 15.01 bps wider than those that prevail during exchange executions. For comparison, the univariate results in Table 5 suggest that this difference is 16.21 bps. Price improvements offered by wholesalers are 27.6% larger, but the effective spreads facing retail investors are still slightly larger, by 1.74 bps. These results confirm our earlier assertion that due to differences in quoted spreads that prevail at the time of wholesaler and exchange executions, effective spreads are not the optimal execution quality comparison metric. With this in mind, we omit effective spreads from subsequent discussions. Finally, price impacts are 15.50 bps lower and realized spreads are 17.24 bps higher for orders routed to wholesalers.

[Table 6]

We next augment the regression by interacting the $WHOL$ dummy with dummies indicating whether a stock belongs to tercile 1, 2, or 3 in Panel B. The coefficient on the $WHOL$ dummy captures the difference between outcome variables for orders in S&P 500 stocks routed to wholesalers compared to exchanges. The interaction terms, e.g., $WHOL \times T3$, test whether the outcomes for orders routed to wholesalers are significantly different for tercile 3 stocks relative to S&P 500 stocks. To obtain the total difference in outcome variables between wholesalers and exchanges for T3 stocks, we add the coefficient on the $WHOL$ dummy to the coefficient on the $WHOL \times T3$ dummy.

In all sub-samples, the data confirm that wholesalers provide greater price improvement compared to exchanges. For S&P 500 stocks, Panel B shows that the difference between exchange and wholesaler effective-to-quoted spread ratios is 0.38, a 38 percentage points larger price improvement relative to the quoted spread. In the univariate results, this difference was similar in magnitude, at 0.39. As noted earlier, wholesaler price improvements decline as we move from large to smaller size firms. Still, even for tercile 3 we estimate that price improvements are 20%

larger (= -0.376 + 0.175 bps) for wholesalers than for exchanges.

Finally, we confirm for all four sub-samples that toxicity of wholesaler-bound flow is lower than that of the exchange-bound flow, and that wholesalers earn larger realized spreads. For instance, column [4] in Panel B shows that price impacts for wholesalers in S&P 500 stocks are 4.31 bps lower than for their exchange counterparts, whereas the realized spreads earned by wholesalers are 4.74 bps greater than those earned by exchange liquidity providers. The corresponding numbers for tercile 3 stocks are a 36.45 bps (= -4.307 - 32.145) lower price impact, and a 39.61 bps (= 4.741 + 34.867) greater realized spread.

So far, we have shown that retail order execution quality varies across the sub-samples of stocks. Yet the data allow for an even more detailed examination. Rule 605 reports are filed by individual venues, and therefore we are able to examine execution quality across wholesalers. To keep this analysis manageable, we divide wholesalers into two groups, the *top 2*, which includes Citadel and Virtu, and the *others*. Our group assignment is driven mainly by the market share, and therefore likely importance, of Citadel and Virtu. Recall that Table 1 shows that these two firms execute over 71% of marketable order flow that is routed to wholesalers.

In Table 7, we use panel regressions to ask if execution quality is systematically different for the $top2$ compared to the $other$ wholesalers overall (Panel A), and for the sub-samples (Panel B). The regressions are of the following form:

$$DepVar_{ijt} = \alpha + \beta_1 top2_j + \beta_2 price_{it} + \beta_3 volatility_{it} + \beta_4 volume_{it} + \varepsilon_{ijt}, \tag{2}$$

where $DepVar_{ijt}$ is one of the following execution quality variables for stock i wholesaler group j (top2 vs. the rest) in month t: the ratio of effective to quoted spread, price impact, and realized spread as defined previously; $top2$ 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; *price* is the natural log of the stock price; *volatility* is the difference between the high and low prices scaled by the high price, and *volume* is the natural log of trading volume. The regression controls for stock and month fixed effects and uses double-clustered standard errors. Note that we only use wholesaler data for these regressions.

The results show that price improvement is roughly the same for the two groups in the overall sample. However, $top2$ wholesalers face more toxic order flow (the difference is 3.02 bps) and earn lower realized spreads (a difference of -3.44 bps). We explore differences across sub-samples in Panel B, where we augment regression (2) by adding interaction variables between $top2$ and tercile dummies. The coefficient on $top2$ shows that Citadel and Virtu offer a 5.5 percentage point lower price improvement for S&P 500 stocks on average, but this does not suffice to compensate for the fact that they face significantly greater adverse selection. While the differences in price improvements shrink as we go from tercile 1 to 3, the differences in toxicity and realized spreads are magnified. Consider tercile 3 stocks, where the price improvements are 2.2 percentage points (= 0.055 - 0.077) greater for $top2$ than for other wholesalers. Toxicity facing $top2$ in tercile 3 stocks is 7.3 bps (= 0.999 + 6.285) greater and realized spreads are 9.5 bps (= -0.552 - 8.809) lower than for other wholesalers trading the same stocks.

In the final column in Panels A and B, we explore whether the differences in realized spreads between $top2$ and other wholesalers can be explained by the size of a wholesaler's operation. The idea is that a wholesaler that obtains more retail flow can more easily internalize the orders and therefore faces lower inventory costs as well as lower per-share fixed costs. To understand the inventory cost argument, let us assume that for every ten retail orders a brokerage receives, it sends four orders to Citadel. This assumption is consistent with Citadel's 40% share of retail flow. On average, retail flow is balanced meaning that the orders are likely to reconcile against each other, leaving a zero or a small inventory imbalance. Even with an imbalance however Citadel has the shortest wait time out of all wholesalers before the next batch of orders arrives. Such short wait times are essential for keeping inventory costs low.

A higher volume of retail flow is associated with significantly lower realized spreads in both panels. Including this variable as a control renders the coefficient on $top2$ insignificant in Panel A. This evidence is consistent with the top two wholesalers being better able to lower inventory and fixed costs due to the size of their operations, and as a result being able to offer lower realized spreads. However, Panel B reveals significant cross-sectional differences. Specifically, $top2$ wholesalers actually charge significantly higher realized spreads for S&P 500 (3.1 bps), tercile 1 (3.3 bps) and tercile 2 stocks (1.4 bps), but they charge significantly lower realized spreads for tercile 3 stocks (-5.0 bps), after controlling for operation size. This evidence suggests that Citadel and Virtu may be cross-subsidizing retail trading in tercile 3 stocks relative to the larger stocks more than other wholesalers. We return to the cross-subsidization issue in a later test.

Given that wholesalers tend to receive order flow of varying toxicity, price improvement may not be the most appropriate comparison metric for our analyses because it does not account for variation in toxicity across wholesalers. If wholesaler 1 provides slightly smaller price improvement than wholesaler 2, yet wholesaler 1 receives considerably more toxic flow, comparing price improvements will result in an unfair comparison between the two wholesalers. Meanwhile, the realized spread is a metric that takes both price improvement and order flow toxicity into account. Assuming that retail brokerages understand the toxicity of their own flow, they should also benchmark against a toxicity-adjusted performance metric. We apply this reasoning in the subsequent analyses where we ask if a wholesaler is able to increase its market share based on prior performance.

3.3 Wholesaler Past Performance

Industry participants suggest that retail brokerages regularly evaluate the performance of wholesalers that they route to.[18] Such evaluations typically occur on a monthly basis.[19] We propose that if the market for retail order flow is competitive, brokerages should adjust their routing to favor wholesalers with better past performance.

To examine if such a relationship is observed in the data, we use the econometric model of order routing proposed by Boehmer, Jennings, and Wei (2007). The model uses a combination of geometric and arithmetic means to allow predicted market shares to lie between zero and one for each wholesaler and to allow the sum of market shares across wholesalers to equal one. Specifically, we estimate the following regression:

$$\begin{aligned} \textit{mkt. share}_{ijt} = \alpha + \beta_1 \textit{abn. realized spread}_{ijt-1} + \beta_2 \textit{abn. realized spread}_{jt-1} \\ + \beta_3 price_{it} + \beta_4 volatility_{it} + \beta_5 volume_{it} + \varepsilon_{ijt}, \end{aligned} \tag{3}$$

where $\textit{mkt. share}_{ijt}$ is the market share of volume in stock i executed by wholesaler j in month t expressed as the deviation from the geometric mean across all wholesalers; $abn.\ realized\ spread_{ijt-1}$ is the average realized spread earned in stock i by wholesaler j in month $t-1$ expressed as the deviation from the arithmetic mean across all other wholesalers; $abn.\ realized\ spread_{jt-1}$ is the average realized spread earned by wholesaler j in all stocks routed to it in month $t-1$ expressed as a deviation from the arithmetic mean across all other wholesalers; *price* is the natural log of the stock price; $volatility$ is the difference between the high and low prices scaled by the high price, and *volume* is the natural log of trading volume. The realized spread variables are scaled, so the economic significance corresponds to basis points. We run these re-

[18] FINRA Rule 5310 requires brokers to conduct rigorous execution quality reviews on at least a quarterly basis (https://bit.ly/46GDy5B).

[19] See for instance item 24 in the SEC administrative proceedings against Robinhood Financial (https://bit.ly/3JUUs6J).

gressions for the full sample and then separately for each sub-sample using stock, wholesaler, and month fixed effects, and clustering standard errors by stock and month.

Table 8 shows that if wholesaler j offers a relatively low realized spread across all stocks, retail brokerages will reward the wholesaler with a greater market share next month. This result holds both for the full sample and for all sub-samples. The full-sample coefficient indicates that a one basis point reduction in a wholesaler's realized spread relative to the mean is associated with a 2.3% greater market share for the full sample, and between 2.0 and 2.5% greater market share for the sub-samples.

[Table 8]

In addition, a relatively low realized spread for wholesaler j in stock i is associated with a significantly greater market share next month for the full sample and for stocks in terciles 2 and 3. However, the economic magnitude of this effect is relatively small. Taken together, the evidence in Table 8 is consistent with wholesalers competing by offering lower liquidity costs across all stocks as opposed to on a security-by-security basis, and retail brokerages playing pivotal roles in ensuring the the best-performing wholesalers are recognized and awarded with more order flow.

3.4 A Competitive Shock

The nature of competition in the retail investor segment changes during our sample period because of entry by a new player, Jane Street.[20] If wholesalers had market power and thus were able to reap economic rents prior to this event, we expect competitive forces to increase the pressure on wholesalers to deliver better execution quality post entry. In other words, we expect realized spreads to decrease.

Jane Street entered into the wholesale business in 2019, but throughout 2020 the firm still had a very small market share. This increased gradually in the late summer of 2021, reaching a

[20] We study the effects on realized spreads of a merger between two large retail brokers, Schwab and TD Ameritrade, in the Appendix and find very similar results.

substantial level by October 2021. By the end of our sample period, Jane Street had a market share of 12.4% (13.9%) of market orders in S&P 500 (non S&P 500) stocks. To evaluate whether the entry of Jane Street results in lower realized spreads, we run a difference-in-differences regression of wholesalers against exchanges with the pre-period being April-June 2021, when Jane Street still has a small market share, and the post-period being the last three months of 2021, during which Jane Street has already established itself as a sizeable wholesaler.

Table 9 reports the results from running the following regression:

$$\begin{aligned} \textit{realized spread}_{ijt} = {} & \alpha + \beta_1 WHOL_j + \beta_2 WHOL \times POST_{jt} + \beta_3 price_{it} + \beta_4 volume_{it} \\ & + \beta_5 volatility_{it} + \varepsilon_{ijt}, \end{aligned} \tag{4}$$

where $\textit{realized spread}_{ijt}$ is the realized spread in stock i for intermediary type j 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; $POST$ is a dummy variable that has a value of 1 after the Jane Street market share capture and 0 otherwise, *price* is the natural log of the stock price; *volume* is the natural log of trading volume; $volatility$ is the difference between the high and low prices scaled by the high price. The models are estimated with stock and month fixed effects, which is why the standalone $POST$ variable is omitted. We run the regressions separately for each sub-sample.

[Table 9]

Based on the β_2 coefficients in Table 9 Panel A, we do not find that realized spreads for wholesalers relative to exchanges decline following Jane Street's entry for any sub-sample. For the index and tercile 1 stocks, the entry did not cause realized spreads to change consistent with the possibility that they were already at a competitive level. For tercile 2 and 3 stocks, realized spreads actually increase.

Panels B and C of Table 9 report regression results for the incumbent wholesalers and Jane Street separately. Recall that Jane Street was present but at a much lower market share in our pre-period. The results show that the incumbents did not change their realized spreads for the largest stocks, but increased the spreads for tercile 1-3 stocks significantly after Jane Street's entry. By contrast, Jane Street, who already offered lower costs in terciles 1 and 2 relative to the other wholesalers, reduced its costs significantly for tercile 3 stocks.

To understand this result, consider first what happens when Jane Street enters. Order flow is now divided among a larger number of wholesalers, and this likely results in lower order flow for each of the incumbent wholesalers. In other words, their inventory costs likely increase. This may be particularly true for less liquid securities such as those in terciles 1-3. By contrast, the S&P 500 securities are highly liquid and this likely makes it easier for the incumbents to manage inventory risk, reducing the need to increase liquidity costs. We return to the topic of inventory risk in the next subsection. Finally, note that Jane Street appears to be competing particularly fiercely in tercile 3 stocks, where the evidence shows that they lowered their realized spreads significantly, cutting them by almost fifty percent.

We cautiously conjecture that the wholesaler market is already rather competitive prior to Jane Street's entry since we see no evidence that additional entry results in lower spread capture by wholesalers. We also see tentative evidence consistent with inventory risk and the economies of scale playing a significant role for wholesalers, a topic we turn to next.

3.5 Inventory Costs

We find suggestive evidence that the wholesalers compete for order flow by offering low realized spreads, and we find no evidence of market power around the entry event discussed above. Yet, the wholesaler realized spreads we document may appear large, particularly for less liquid stocks. Are the realized spreads evidence of market power, or are they compensating for the inventory costs facing wholesalers in less liquid securities?

Inventory costs are of course difficult to measure, so we will rely, as we did earlier, on trading volume as a proxy for wholesalers ability to manage inventory. We conjecture that, controlling for volatility, a stock-month with lower volume is associated with greater inventory costs as outstanding positions take longer to lay off. To understand the role of inventory costs (for wholesalers only), we run the following panel regressions:

$$\begin{aligned} \textit{realized spread}_{it} = \alpha &+ \beta_1 T1_i + \beta_2 T2_i + \beta_3 T3_i + \beta_4 price_{it} + \beta_5 volatility_{it} \\ &+ \beta_6 volume_{it} + \varepsilon_{it}, \end{aligned} \tag{5}$$

where *realized spread*$_{jt}$ is the realized spread in stock i in month t; $T1$, $T2$, and $T3$ are dummies indicating whether a stock is in size-based tercile 1, 2, or 3 of non-S&P 500 stocks; *price* is the natural log of the stock price; $volatility$ is the difference between the high and low prices scaled by the high price; and *volume* is the natural log of trading volume (CRSP). The regressions control for month fixed effects, and we use two-way clustered standard errors.

[Table 10]

As expected, column [1] of Table 10 shows that tercile 1, 2, and 3 stocks have significantly greater realized spreads than S&P 500 stocks. When we control for price and volatility in column [2], there is no longer a significant difference between S&P 500 stocks and Tercile 1 stocks, but realized spreads for the remaining size terciles are still significantly higher. Column [3] includes volume, which is our proxy for inventory costs, and this makes all the coefficients on the size terciles turn negative and significant. Note also that the coefficient on volume is itself highly significant and negative as predicted. We conclude that after controlling for inventory costs, wholesalers seem to earn significantly lower realized spreads for less liquid stocks than they do for the S&P 500 stocks. This result once again points to a cross-subsidy from large stocks to small stocks that arises due to the portfolio approach used by retail brokers to evaluate wholesaler performance.

3.6 Institutional Interest

In December 2022, the SEC proposed rules that would significantly change the equity markets.[21] To analyze these comprehensive rules is beyond the scope of the current study, but we believe our results may shed some light on SEC's conjecture that retail traders would be better off if there was order-by-order competition for retail orders (labeled segmented orders in the rule) as envisioned in the proposed Order Competition Rule.[22] In a nutshell, the rule proposes a requirement that segmented orders be forwarded, either by the retail broker directly or by the wholesaler receiving retail order flow, to auctions run by exchanges and/or certain ATSs where institutions can interact with the order flow.[23] The SEC believes that since retail order flow has lower toxicity (as we document above), it should get larger price improvement than what is currently offered by wholesalers and that institutions would be willing to trade with retail at the NBBO midquote.

How realistic is this proposal, and how would it affect the cross-section of stocks currently handled by wholesalers? The answer depends on whether or not there is institutional interest to trade the stocks favored by retail investors. We document in Table 4 that wholesalers currently execute the bulk of retail share volume in less liquid stocks (Tercile 2 and Tercile 3). Is there sufficient institutional interest to do without the intermediation offered by wholesalers for these stocks?

To answer this question, we estimate institutional trading in the sample stocks based on changes in reported quarterly holdings from 13F reports and add to that changes in short interest (which are available bi-monthly). To account for intra-quarter trading, we gross-up institu-

[21] https://www.sec.gov/newsroom/market-structure-proposals-december-2022
[22] https://www.sec.gov/rules/proposed/2022/34-96495.pdf.
[23] See Ernst, Spatt, and Sun (2023) for a theoretical analysis of the auction proposal.

tional volume inferred from 13F reports by a factor of 1.17 based on Chakrabarty, Moulton, and Trzcinka (2017). This gives us a proxy for institutional trading interest in a particular stock. We then calculate the ratio of retail trading as reflected in Rule 605 data divided by our proxy for institutional trading interest. Table 11 reports the across stock means, medians, and quartiles for each sub-sample, that is S&P 500, Tercile 1, Tercile 2, and Tercile 3 stocks.

[Table 11]

Column [1] ([2]) shows that average (median) retail order flow represents 62% (20%) of institutional interest for S&P 500 stocks, so for index stocks there is significant institutional interest. Importantly, as we move to less liquid stocks, it becomes clear that retail order flow swamps institutional trading interest. Already for Tercile 1 stocks, the institutional trading interest starts to become insufficient on average as the ratio of retail to institutional interest exceeds one. For Tercile 2 stocks retail interest is more than double the institutional interest on average, and for Tercile 3 stocks, average retail order flow is more than seven times larger than institutional interest. The ratios are highly skewed, suggesting that retail interest tends to be focused in particular stocks, and that those stocks are not favored by institutions. If we switch our attention to the median values for a more conservative view, we observe that institutional interest is substantially below retail interest only for Tercile 3.

The conclusion we draw is that institutional trading interest may be low for some of the cross-section of securities traded by retail investors. At best, the effect of the proposed auctions for these securities would be to delay executions. However, the auctions could actually have even more detrimental consequences for retail investors in less liquid stocks. Our earlier results suggest that realized spreads may be insufficient to cover inventory costs for less liquid stocks in the current environment and that wholesalers may cross-subsidize small stocks with their large-stock revenues. If that is indeed the case, and wholesalers end up losing a significant fraction of order flow in liquid stocks through the proposed auctions, they may be unable to offer price improvements at the level we observe today for less liquid stocks. In other words, we could see execution quality deteriorate for some of the universe of securities retail investors currently trade.

4. Conclusion

In the United States, retail brokers typically route order flow to wholesalers rather than directly to exchanges. Wholesalers immediately fill the retail order from their inventory in hopes of receiving an offsetting order in the near future. We show that, contrary to public perception, a substantial portion of the spread is passed on to retail traders via price improvements. Yet another part of the spread is paid to the broker who routed the order, known as payment for order flow. The remaining part of the spread covers wholesaler inventory costs, technology costs, and wholesaler profits.

Using public SEC Rule 605 data, this paper suggests that retail orders are better off being routed to wholesalers than directly to exchanges. If wholesalers were to be removed, retail investors would pay billions in additional trading costs according to our estimates. The net effect consists of three components. First, unlike institutions, retail investors do not time the order submission to periods when spreads are narrow. As a result, retail orders are placed when the quoted spread is wider than when institutional orders are placed. Second, wholesalers mitigate these higher transaction costs by providing a substantial price improvement. Third, retail investors would benefit from lower liquidity generation costs (realized spreads) on exchanges. However, this benefit is not substantial enough to compensate for the loss of price improvements resulting in higher trading costs for retail investors. By contrast, institutional traders would gain because the lower toxicity of retail flow would help reduce the spreads needed on exchanges to compensate liquidity providers for adverse selection. Institutional order flow would therefore benefit at the expense of retail order flow if retail orders were to be relocated from wholesalers to exchanges.

Retail brokerages play an important role in this discussion, as they make routing decisions.

Our analysis suggests that retail brokerages base their routing decisions on the wholesaler liquidity generation costs. If the wholesaler offers low costs this month, the broker will route additional order flow in the future. This result indicates that brokers seek to enhance retail execution quality through their routing decisions.

Entry and exit of wholesalers may affect the nature of competition in the retail investor segment during our sample period. A new player, Jane Street, enters the retail wholesaler business and gains a significant market share. If wholesalers had market power and thus were able to reap economic rents prior to this event, we expect competitive forces to increase the pressure on wholesalers to deliver better execution quality post-entry. However, we find no evidence that additional entry results in lower spread capture by wholesalers, leading us to conjecture that the wholesaler market may be already competitive prior to Jane Street's entry.

We document large differences in wholesaler liquidity generation costs in the cross-section, with particularly large realized spreads for the least liquid securities. To examine whether the differences in wholesaler realized spreads can plausibly be explained by differences in inventory costs, we use trading volume to proxy for inventory costs. Once we control for volume, realized spreads for less liquid stocks are actually lower than for index stocks, suggesting that the realized spreads, while large, are not necessarily reflective of wholesaler market power.

We close by commenting on the recent SEC proposal to overhaul the retail trading landscape by requiring that retail flow be routed to auctions for order-by-order competition. The proposal rests on the assumption that there would be significant institutional trading interest that would like to interact with retail flow, and would offer better prices than those currently offered by wholesalers. Proxies for institutional interest suggest that, while this may be true for large stocks, it is unlikely to be true for small stocks currently traded by retail investors. Our results suggest that many retail investors, particularly those trading less liquid stocks, would be worse off if the proposal were implemented, as they would likely face both delays and lower execution quality.

References

Adams, S., C. Kasten, and E. K. Kelley, 2021, "Do Investors Save When Market Makers Pay? Retail Execution Costs Under Payment for Order Flow Models," *Working paper*, University of Tennessee, Knoxville. 8, 9

Aquilina, M., E. B. Budish, and P. O'Neill, 2021, "Quantifying the high-frequency trading "arms race": A simple new methodology and estimates," *Quarterly Journal of Economics,* forthcoming. 4

Baldauf, M., J. Mollner, and B. Z. Yueshen, 2023, "Siphoned Apart: A Portfolio Perspective on Order Flow Segmentation," *Available at SSRN 4173362*. 8

Barber, B. M., X. Huang, P. Jorion, T. Odean, and C. Schwarz, 2022, "A (Sub) penny For Your Thoughts: Tracking Retail Investor Activity in TAQ," *Available at SSRN 4202874*. 9

Bartlett, R. P., J. McCrary, and M. O'Hara, 2022, "The Market Inside the Market: Odd-Lot Quotes," *Available at SSRN 4027099*. 12

Battalio, R., and C. W. Holden, 2001, "A simple model of payment for order ow, internalization, and total trading cost," *Journal of Financial Markets*, 4, 33–71. 19

Battalio, R., and R. Jennings, 2023, "Why do Brokers who do not Charge Payment for Order Flow Route Marketable Orders to Wholesalers?," *Available at SSRN 4304124*. 8, 10

Bessembinder, H., A. Carrion, L. Tuttle, and K. Venkataraman, 2016, "Liquidity, resiliency and market quality around predictable trades: Theory and evidence," *Journal of Financial economics*, 121(1), 142–166. 18

Boehmer, E., R. Jennings, and L. Wei, 2007, "Public disclosure and private decisions: Equity market execution quality and order routing," *The Review of Financial Studies*, 20(2), 315–358. 26

Boehmer, E., C. M. Jones, X. Zhang, and X. Zhang, 2021, "Tracking retail investor activity," *The Journal of Finance*, 76(5), 2249–2305. 9

Bogousslavsky, V., and P. Collin-Dufresne, 2022, "Liquidity, volume, and order imbalance volatility," *Journal of Finance, Forthcoming*. 18

Brogaard, J., B. Hagströmer, L. Nordén, and R. Riordan, 2015, "Trading fast and slow: Colocation and liquidity," *Review of Financial Studies*, 28(12), 3407–3443. 13

Bryzgalova, S., A. Pavlova, and T. Sikorskaya, 2022, "Retail Trading in Options and the Rise of the Big Three Wholesalers," *Working paper*, London Business School. 10

Chakrabarty, B., P. C. Moulton, and C. Trzcinka, 2017, "The performance of short-term institutional trades," *Journal of Financial and Quantitative Analysis*, 52(4), 1403–1428. 32

Eaton, G. W., T. C. Green, B. Roseman, and Y. Wu, 2022, "Retail Trader Sophistication and Stock Market Quality: Evidence from Brokerage Outages," *Journal of Financial Economics*, forthcoming. 9

Ernst, T., and C. S. Spatt, 2022, "Payment for Order Flow and Asset Choice," working paper. 10

Ernst, T., C. S. Spatt, and J. Sun, 2023, "Would Order-by-Order Auctions Be Competitive?," *Available at SSRN*. 7, 31

Foley, S., A. Liu, K. Malinova, A. Park, and A. Shkilko, 2023, "Cross-subsidizing liquidity," *Working Paper*, Macquarie University. 7

Hendershott, T., C. M. Jones, and A. J. Menkveld, 2011, "Does algorithmic trading improve liquidity?," *Journal of Finance*, 66, 1–33. 13

Hendershott, T., S. Khan, and R. Riordan, 2022, “Option Auctions,” *Working paper*, University of California at Berkeley. 10

Hu, E., and D. Murphy, 2022, “Competition for Retail Order Flow and Market Quality,” *Working paper*New York University. 9

Jain, P. K., S. Mishra, S. O’Donoghue, and L. Zhao, 2022, “Trading Volume Shares and Market Quality: Pre-and Post-Zero Commissions,” *Working paper*, University of Memphis. 8

Kothari, S., E. So, and T. Johnson, 2021, “Commission Savings and Execution Quality for Retail Trades,” *Working paper*, MIT Sloan School of Management. 8, 9

Kyle, A. S., 1985, “Continuous Auctions and Insider Trading,” *Econometrica*, 53, 1315–1336. 19

Lee, C. M., and M. J. Ready, 1991, “Inferring trade direction from intraday data,” *Journal of Finance*, 46, 733–746. 12

Li, S., M. Ye, and M. Zheng, 2020, “Refusing the Best Price?,” *Available at SSRN 3763455*. 11

O’Hara, M., 2015, “High frequency market microstructure,” *Journal of Financial Economics*, 116(2), 257–270. 4

Schwarz, C., B. M. Barber, X. Huang, P. Jorion, and T. Odean, 2022, “The’Actual Retail Price’of Equity Trades,” *Available at SSRN 4189239*. 9

Table 1
Market Shares

The table contains the list of 22 trading venues that execute held liquidity-demanding orders during the sample period (2019-2022). The data are from the SEC Rule 605 reports. Wholesalers are highlighted in bold font. We report the total number of shares executed by each venue (in billions) and each venue's market share. Panel A aggregates by venue type, while Panel B contains the results by venue.

	venue type	shares executed, bil.	mkt. share, %
Panel A: by venue type			
	EXCH	1,695.50	58.76
	WHOL	1,190.05	41.24
Panel B: by venue			
Nasdaq	EXCH	498.29	17.27
Citadel	WHOL	479.14	16.60
Virtu	WHOL	363.00	12.58
NYSE	EXCH	301.21	10.44
NYSE Arca	EXCH	213.92	7.41
EDGX	EXCH	205.06	7.11
BATS	EXCH	173.77	6.02
G1	WHOL	149.05	5.17
BYXX	EXCH	72.51	2.51
Two Sigma	WHOL	65.06	2.25
EDGA	EXCH	62.92	2.18
IEX	EXCH	54.80	1.90
UBS	WHOL	50.98	1.77
Jane Street	WHOL	49.54	1.72
NYSE National	EXCH	45.51	1.58
NSDQ Boston	EXCH	29.44	1.02
Merrill Lynch	WHOL	22.99	0.80
NSDQ Philadelphia	EXCH	20.62	0.71
NYSE American	EXCH	15.44	0.54
Morgan Stanley	WHOL	10.31	0.36
NYSE Chicago	EXCH	1.03	0.04
MEMX	EXCH	0.95	0.03
Total		2,885.55	100.00

Table 2
Execution Quality

The table contains execution quality statistics for held liquidity-demanding orders. We compute the statistics separately for orders executed by wholesalers (WHOL) and exchanges (EXCH). We report the average number of shares executed and the average stock price in a sample stock during the sample period, followed by the percentage share of shares that are price improved or executed at or better the corresponding NBBO. Further, we report the quoted and effective spreads in basis points, and to better understand the magnitude of price improvement, we compute the ratio of the effective to the quoted spread. Finally, we compute the components of the effective spread: price impact and realized spread. All variables are volume-weighted. Asterisks *** in column [3] indicate statistical significance of differences between columns [1] and [2] at the 1% level.

	WHOL	EXCH	diff. [1]-[2]
	[1]	[2]	[3]
# shares, mil.	146.36	207.65	***
price, $.	32.06	32.62	
improved, %	65.71	9.49	***
at or better, %	92.98	98.34	***
quoted spread, bps	64.92	48.67	***
effective spread, bps	49.06	46.98	**
effective / quoted	0.76	0.97	***
price impact, bps	32.53	47.32	***
realized spread, bps	16.53	-0.34	***

Table 3
Moving Retail Flow to Exchanges

The table illustrates possible consequences of moving retail flow to exchanges. Among such consequences are an overall reduction in price impacts for all exchange trades, a reduction in realized spreads incurred by retail traders, and a reduction in price improvement obtained by retail traders. Panel A reports percentage changes in effective spreads for retail liquidity demanders (RET LDs) and liquidity demanders, whose orders are currently routed to exchanges (EXCH LDs). Panel B reports gains measured in terms of effective spreads for LDs and realized spreads for LPs from the move for four categories of market participants: RET LDs, EXCH LDs, exchange liquidity providers (EXCH LPs), and wholesalers (WHOL LPs). The line in bold font represents an assumption that the currently prevailing exchange realized spreads will not change if retail flow moves to exchanges. The remaining lines allow realized spreads to vary as a result of the move, in 0.1 bps increments.

realiz. spr.,	Panel A: Δ eff. spread, %		Panel B: gains, in $ bil.			
bps.	RET LDs	EXCH LDs	RET LDs	EXCH LDs	EXCH LPs	WHOL LPs
-0.84	13.46	-11.17	-25.67	102.99	-13.07	-64.25
-0.74	13.72	-10.97	-26.16	101.14	-10.72	-64.25
-0.64	13.97	-10.77	-26.65	99.28	-8.37	-64.25
-0.54	14.23	-10.57	-27.14	97.42	-6.02	-64.25
-0.44	14.49	-10.37	-27.63	95.56	-3.67	-64.25
-0.34	**14.75**	**-10.17**	**-28.12**	**93.70**	**-1.32**	**-64.25**
-0.24	15.00	-9.96	-28.61	91.84	1.03	-64.25
-0.14	15.26	-9.76	-29.11	89.98	3.38	-64.25
-0.04	15.52	-9.56	-29.60	88.12	5.73	-64.25
0.06	15.78	-9.36	-30.09	86.26	8.08	-64.25
0.16	16.04	-9.16	-30.58	84.40	10.43	-64.25

Table 4
Market Shares: Sub-samples

The table reports market shares in held liquidity-demanding orders for wholesalers and exchanges, with the sample divided into S&P 500 and size-based terciles of non-S&P 500 stocks labeled Tercile 1, Tercile 2, and Tercile 3.

	S&P 500	Tercile 1	Tercile 2	Tercile 3
	[1]	[2]	[3]	[4]
WHOL	31.87	34.30	51.02	63.79
EXCH	68.13	65.70	48.98	36.21
No. Stocks	514	2,550	2,550	2,551

Table 5
Execution Quality: Sub-samples

The table contains execution quality statistics for held liquidity-demanding orders. The sample is divided into S&P 500 and size terciles T1, T2, and T3 of non-S&P 500 stocks. We report the average number of shares executed and the average stock price in a sample stock during the sample period, followed by the percentage share of orders that are price improved or executed at or better the corresponding NBBO. Further, we report the quoted and effective spreads in basis points, and to better understand the magnitude of price improvement, we compute the ratio of the effective to the quoted spread. Finally, we compute the components of the effective spread: price impact and realized spread. All variables are volume-weighted. Asterisks *** (**) in columns [3] and [6] indicate statistical significance of differences between columns [1] and [2] and [4] and [5] at the 1% (5%) level.

	WHOL	EXCH	diff.	WHOL	EXCH	diff.
	[1]	[2]	[3]	[4]	[5]	[6]
	S&P 500			Tercile 1		
# shares, mil.	419.81	913.8	***	173.98	325.26	***
price, $	146.29	146.85		55.86	57.51	
improved, %	75.07	12.28	***	68.96	11.71	***
at or better, %	94.47	97.84	***	92.80	98.15	***
quoted spread, bps	8.28	5.53	***	24.16	15.53	***
effective spread, bps	4.62	5.22	***	15.96	14.70	***
effective / quoted	0.56	0.94	***	0.66	0.95	***
price impact, bps	3.41	6.45	***	11.77	17.81	***
realized spread, bps	1.21	-1.23	***	4.19	-3.10	***
	Tercile 2			Tercile 3		
# shares, mil.	91.77	88.63		116.92	66.15	***
price, $	15.34	15.52		7.82	7.96	
improved, %	63.94	9.06	***	62.68	7.26	***
at or better, %	92.97	98.59	***	92.99	98.33	***
quoted spread, bps	58.72	40.65	***	118.46	95.17	***
effective spread, bps	42.84	38.72	**	93.63	92.75	**
effective / quoted	0.73	0.95	***	0.79	0.97	***
price impact, bps	28.72	41.08	***	41.08	88.10	***
realized spread, bps	14.12	-2.36	***	33.09	4.65	***

Table 6
Execution Quality: Regression

Panel A of the table reports coefficient estimates from market quality regressions of the following form:

$$DepVar_{it} = \alpha + \beta_1 WHOL_{it} + \beta_2 price_{it} + \beta_3 volatility_{it} + \beta_4 volume_{it} + \varepsilon_{it},$$

where $DepVar_{it}$ is one of the following market quality variables for stock i in month t: the ratio of effective to quoted spreads, quoted spread, effective spread, price impact, and realized spread as defined previously; $WHOL$ is a dummy variable that has a value of 1 for orders executed by wholesalers and 0 for orders executed by exchanges; $price$ is the natural log of the stock price; $volatility$ is the difference between the high and low prices scaled by the high price, and $volume$ is the natural log of trading volume. Panel B augments the specification by including interaction terms between the $WHOL$ dummy and indicator variables for the size-based terciles of non-S&P 500 stocks; Tercile 1 ($T1$), Tercile 2 ($T2$), and Tercile 3 ($T3$). The models are estimated with stock and month fixed effects, and the standard errors are double-clustered across stocks and months. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	eff. spr. / quot. spr.	quoted spr.	effective spr.	price imp.	realized spr.
	[1]	[2]	[3]	[4]	[5]
Panel A: Base Specification					
WHOL	-0.276***	15.006***	1.739***	-15.503***	17.240***
	(0.01)	(0.44)	(0.48)	(0.62)	(0.94)
price	-0.017***	-19.896***	-19.777***	-15.701***	-4.073***
	(0.00)	(1.22)	(1.20)	(1.17)	(0.64)
volatility	0.000***	0.261***	0.237***	0.206***	0.031**
	(0.00)	(0.02)	(0.02)	(0.02)	(0.01)
volume	-0.002*	-29.172***	-26.294***	-17.114***	-9.185***
	(0.00)	(1.51)	(1.53)	(1.28)	(0.72)
intercept	1.026***	424.876***	390.658***	277.653***	113.059***
	(0.01)	(19.27)	(19.44)	(17.09)	(9.01)
Adj. R^2	0.660	0.757	0.734	0.520	0.182
Panel B: Specification with Interaction Terms					
WHOL	-0.376***	5.651***	0.435**	-4.307***	4.741***
	(0.01)	(0.40)	(0.17)	(0.31)	(0.36)
$WHOL \times T1$	0.063***	1.520***	0.160	-1.446***	1.607***
	(0.00)	(0.33)	(0.17)	(0.23)	(0.27)
$WHOL \times T2$	0.124***	11.396***	2.058***	-9.492***	11.550***
	(0.01)	(0.56)	(0.47)	(0.50)	(0.68)
$WHOL \times T3$	0.175***	22.692***	2.729**	-32.145***	34.867***
	(0.01)	(0.80)	(1.17)	(1.44)	(2.39)
price	-0.017***	-19.897***	-19.777***	-15.701***	-4.073***
	(0.00)	(1.22)	(1.20)	(1.17)	(0.64)
volatility	0.000***	0.261***	0.237***	0.206***	0.031**
	(0.00)	(0.02)	(0.02)	(0.02)	(0.01)
volume	-0.002**	-29.175***	-26.294***	-17.111***	-9.188***
	(0.00)	(1.51)	(1.53)	(1.28)	(0.72)
intercept	1.026***	424.907***	390.663***	277.619***	113.098***
	(0.01)	(19.26)	(19.44)	(17.09)	(9.01)
Adj. R^2	0.685	0.761	0.734	0.529	0.203

Table 7
Execution Quality Across Wholesalers: Regressions

Panel A of the table reports coefficient estimates from wholesaler market quality regressions of the following form:

$$DepVar_{ijt} = \alpha + \beta_1 top2_{ijt} + \beta_2 price_{it} + \beta_3 volatility_{it} + \beta_4 volume_{it} + \varepsilon_{it},$$

where $DepVar_{ijt}$ is one of the following market quality variables for stock i wholesaler group j in month t: the ratio of effective to quoted spreads, quoted spread, effective spread, price impact, and realized spread as defined previously; $top2$ 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; *price* is the natural log of the stock price; $volatility$ is the difference between the high and low prices scaled by the high price, and *volume* is the natural log of trading volume. In the last specification, we add the natural log of retail volume executed by a wholesaler to control for the wholesaler's ability to manage inventory internally. Panel B augments the specification by including interaction terms between the $top2$ dummy and indicator variables for the size-based terciles of non-S&P 500 stocks; Tercile 1 ($T1$), Tercile 2 ($T2$), and Tercile 3 ($T3$). The models are estimated with stock and month fixed effects, and the standard errors are double-clustered across stocks and months. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	eff. spr. / quot. spr.	quoted spr.	effective spr.	price imp.	realized spr.	realized spr.
	[1]	[2]	[3]	[4]	[5]	[6]
Panel A: Base Specification						
top2	0.008	-1.180***	-0.417	3.022***	-3.439***	0.866
	(0.01)	(0.13)	(0.57)	(0.48)	(0.77)	(0.74)
price	-0.032***	-19.550***	-19.803***	-8.620***	-11.184***	-11.921***
	(0.00)	(1.32)	(1.33)	(0.86)	(0.92)	(0.94)
volatility	0.000***	0.314***	0.266***	0.168***	0.099***	0.104***
	(0.00)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
volume	-0.005***	-32.910***	-27.649***	-12.887***	-14.764***	-9.599***
	(0.00)	(1.60)	(1.66)	(0.99)	(0.82)	(0.78)
retail volume						-5.280***
						(0.35)
intercept	0.819***	481.222***	407.612***	193.495***	214.142***	224.141***
	(0.02)	(20.46)	(21.43)	(12.66)	(10.99)	(11.46)
Adj. R^2	0.292	0.764	0.696	0.391	0.258	0.260
Panel B: Specification with Interaction Terms						
top2	0.055***	-0.332***	0.446***	0.999***	-0.552**	3.111***
	(0.01)	(0.05)	(0.14)	(0.19)	(0.24)	(0.32)
top2×T1	-0.028***	-0.209***	-0.173*	0.086	-0.261**	0.156
	(0.00)	(0.05)	(0.10)	(0.12)	(0.11)	(0.17)
top2×T2	-0.066***	-1.138***	-0.619	1.510***	-2.130***	-1.761***
	(0.01)	(0.13)	(0.49)	(0.34)	(0.54)	(0.62)
top2×T3	-0.077***	-1.852***	-2.522*	6.285***	-8.809***	-8.133***
	(0.01)	(0.34)	(1.37)	(1.00)	(1.76)	(1.86)
price	-0.032***	-19.550***	-19.803***	-8.620***	-11.184***	-11.898***
	(0.00)	(1.32)	(1.33)	(0.86)	(0.92)	(0.94)
volatility	0.000***	0.314***	0.266***	0.167***	0.099***	0.104***
	(0.00)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
volume	-0.005***	-32.910***	-27.649***	-12.886***	-14.765***	-9.761***
	(0.00)	(1.60)	(1.66)	(0.99)	(0.82)	(0.79)
retail volume						-5.116***
						(0.36)
intercept	0.819***	481.229***	407.617***	193.482***	214.158***	223.847***
	(0.02)	(20.46)	(21.43)	(12.66)	(10.99)	(11.44)
Adj. R^2	0.296	0.764	0.696	0.392	0.259	0.261

Table 8
Wholesaler Order Flow Determinants: Regression

We estimate the following regression:

$$\begin{aligned} \textit{mkt. share}_{ijt} &= \alpha + \beta_1 \textit{abn. realized spread}_{ijt-1} + \beta_2 \textit{abn. realized spread}_{jt-1} \\ &+ \beta_3 \textit{price}_{it} + \beta_4 \textit{volatility}_{it} + \beta_5 \textit{volume}_{it} + \varepsilon_{ijt}, \end{aligned}$$

where *mkt. share*$_{ijt}$ is the market share of volume in stock i executed by wholesaler j in month t expressed as the deviation from the geometric mean across market centers; *abn. realized spread*$_{ijt-1}$ is the average realized spread earned in stock i by wholesaler j in month $t-1$ expressed as a deviation from the arithmetic mean across market centers; *abn. realized spread*$_{jt-1}$ is the average realized spread earned by wholesaler j in all stocks routed to it in month $t-1$ expressed as a deviation from the arithmetic mean across market centers; *price* is the natural log of the stock price; *volatility* is the difference between the high and low prices scaled by the high price, and *volume* is the natural log of trading volume. The realized spread variables are scaled, so the economic significance corresponds to basis points. We run these regressions for the full sample and then separately for each sub-sample, use stock, wholesaler, and month fixed effects, and cluster standard errors by stock and month. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	Full sample	S&P 500	Tercile 1	Tercile 2	Tercile 3
	[1]	[2]	[3]	[4]	[5]
abn. realized spread$_{ij}$	-0.000***	-0.000	-0.000	-0.000***	-0.000***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
abn. realized spread$_j$	-0.023***	-0.023***	-0.025***	-0.024***	-0.020***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
price	0.010***	0.052	0.001	0.006***	0.030***
	(0.00)	(0.04)	(0.00)	(0.00)	(0.00)
volatility	-0.065***	-0.480*	-0.140**	-0.047***	-0.046***
	(0.02)	(0.25)	(0.06)	(0.01)	(0.01)
volume	0.019***	0.124	0.025**	0.017***	0.022***
	(0.01)	(0.08)	(0.01)	(0.00)	(0.00)
Adj. R^2	0.690	0.729	0.729	0.681	0.714

Table 9
Jane Street Entry

The table reports coefficient estimates from the following regression:

$$realized\ spread_{it} = \alpha + \beta_1 WHOL_{it} + \beta_2 WHOL \times POST_{it} + \beta_3 price_{it} + \beta_4 volatility_{it} + \beta_5 volume_{it} + \varepsilon_{it},$$

where *realized spread*$_{jt}$ is the realized spread in stock i 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; $POST$ is a dummy variable that has a value of 1 after the Jane Street entry (Panel A), *price* is the natural log of the stock price; $volatility$ is the difference between the high and low prices scaled by the high price, and *volume* is the natural log of trading volume. 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 with stock and month fixed effects, which is why the standalone $POST$ variable is omitted. The standard errors are double-clustered across stocks and months. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	S&P 500	Tercile 1	Tercile 2	Tercile 3
	[1]	[2]	[3]	[4]
Panel A: All Wholesalers				
WHOL	1.174***	5.218***	11.228***	20.167***
	(0.18)	(0.29)	(0.61)	(0.84)
WHOL×POST	0.375	1.535	5.578**	14.003***
	(0.19)	(0.69)	(1.66)	(3.33)
price	-0.338	-0.019	1.912	-3.682
	(0.44)	(0.51)	(1.08)	(2.03)
volatility	0.013	-0.013	0.029	-0.015
	(0.01)	(0.01)	(0.02)	(0.03)
volume	-0.417	-1.655***	-6.122***	-2.704**
	(0.23)	(0.27)	(1.36)	(0.80)
intercept	6.369	18.484***	59.526**	37.316***
	(4.72)	(3.67)	(15.90)	(8.15)
Adj. R^2	0.350	0.328	0.235	0.220
Panel B: Incumbents				
WHOL	1.171***	5.689***	11.574***	20.106***
	(0.16)	(0.30)	(0.57)	(1.02)
WHOL×POST	0.327	1.932**	6.525**	16.863***
	(0.17)	(0.75)	(1.82)	(3.79)
Panel C: Jane Street				
WHOL	1.183**	4.563***	8.752***	26.885***
	(0.34)	(0.16)	(0.40)	(2.15)
WHOL×POST	0.421	0.617	0.206	-12.680***
	(0.35)	(0.59)	(0.96)	(2.18)

Table 10
Inventory Costs

The table reports coefficient estimates from the following regression:

$$\begin{aligned} \textit{realized spread}_{it} = \alpha + \beta_1 T1 + \beta_2 T2 + \beta_3 T3 + \beta_4 \textit{price}_{it} + \beta_5 \textit{volatility}_{it} \\ + \beta_6 \textit{volume}_{it} + \varepsilon_{it}, \end{aligned} \tag{6}$$

where $\textit{realized spread}_{jt}$ is the realized spread in stock i in month t; $T1$, $T2$, and $T3$ are dummies indicating whether a stock is in size-based Tercile 1, Tercile 2, or Tercile 3 of non-S&P 500 stocks; *price* is the natural log of the stock price; *volatility* is the difference between the high and low prices scaled by the high price; and *volume* is the natural log of trading volume. The regressions control for month fixed effects, and we use two-way clustered standard errors. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	[1]	[2]	[3]
T1	3.514***	1.335	-18.203***
	(0.23)	(0.83)	(1.75)
T2	13.035***	8.627***	-32.457***
	(0.51)	(1.59)	(3.27)
T3	41.307***	34.427***	-20.374***
	(2.55)	(1.76)	(3.15)
price		-1.653**	-8.843***
		(0.72)	(0.93)
volatility		0.124***	0.197***
		(0.02)	(0.03)
volume			-9.928***
			(0.52)
intercept	0.798	6.869**	169.756***
	(0.66)	(2.79)	(10.33)
Adj. R^2	0.110	0.114	0.190

Table 11
Institutional Interest

The table reports descriptive statistics for stock-quarter ratios of Rule 605 volume of liquidity-demanding orders to institutional volume, *rat*. Institutional volume is proxied for as changes in institutional holdings from quarterly 13F filings plus changes in short interest.

	avg. rat [1]	*med. rat* [2]	*std. rat* [3]	*p25 rat* [4]	*p75 rat* [5]
S&P 500	0.620	0.198	1.617	0.128	0.394
Tercile 1	1.215	0.218	2.926	0.112	0.681
Tercile 2	2.316	0.571	4.056	0.155	2.364
Tercile 3	7.399	4.209	7.497	0.802	13.795

Appendix

A.1 Data details

We obtain our data from a service provider that specializes in compliance and trade analytics. The Rule 605 data we have access to cover 70 market centers for January 2016 - March 2022. While our service providers' Rule 605 data coverage is extensive it is not complete. To patch the missing data, we download Rule 605 data directly in order to add NYSE National (XCIS) and missing months.[24]

We define our S&P 500 sample based on all stocks indicated as being part of the index between January 2019 - March 2022. We merge the S&P 500 stocks with CRSP data and are able to find data for 514 unique symbols. It is more than 503 stocks because our sample includes later additions of stocks due to increasing market capitalization and large spin-offs, and deletions due to decreasing market capitalization and M&A activity.

Data for Other (non-S&P 500) are also available from our service provider and this sample consists of a broad range of securities. Market centers use a number of different ways to indicate that a security is of a particular type, e.g., a series A preferred, and extensive re-coding of symbols is necessary. The result is a sample that includes 15,888 unique symbols (365,065 symbol-months), where 11,475 are ordinary shares, 169 are Class A shares, 117 Class B shares, and 2 Class C shares, for a total of 11,763 symbols – we call these securities stocks. The remainder are warrants, preferred stocks, units, rights issues, convertible bonds, etc. Stocks represent 99% of share volume (and 84.5% of symbol-months). We drop the other security types (warrants etc.) for the remainder of the analysis. We merge the Other stocks with CRSP, and are able to match 93.2% of the symbols and 95.8% of the symbol-months. Finally, we merge with TAQ data, and end up with a sample of 11,406 stocks, 8,165 ordinary stocks and 3,241 ETFs.

[24] There are individual missing months for some market centers, but the data is more uniformly missing for September 2020 (when only four market centers are covered).

To study cross-sectional differences, we divide non-S&P 500 stocks into terciles based on average market capitalization (defined as CRSP number of shares outstanding multiplied by the closing monthly price) during our sample period. Terciles 1 and 2 have 2,550 securities, and Tercile 3 includes 2,551.

The 605 reports provide a selection of variables for each stock, market center, month, order type (market, marketable, and limit order), and order size (100-499, 500-1999, 2000-4999, and 5000-9999 shares). For this analysis we use a subset of the variables which are defined as follows:

- *Executed shares (EXshs)* are the cumulative number of shares executed at the receiving market center.
- *Away executed shares (AWshs)* are the cumulative number of shares executed at another venue.
- *Average realized spread ($RS* is the share-weighted average spread in dollars using a five minute horizon.[25]
- *Average effective spread ($ES)* is the share weighted average in dollars.
- *Price improved shares (PIshs)* is the cumulative number of shares executed with a price improvement.
- *Price improved average amount ($PI)* is the per share share-weighted average dollar amount that prices were improved.
- *At the quote shares (AQshs)* is the cumulative number of shares executed at the quote.
- *Outside the quote shares (OQshs)* is the cumulative number of shares executed outside the quote.
- *Outside the quote average amount ($OQ)* is the per share share-weighted average dollar amount that prices were outside the quote.

[25] If the order is executed less than five minutes before the close of regular trading hours, the midpoint used is the final midpoint of regular trading hours.

The service provider uses these variables to compute a series of market quality metrics which are defined as:

$$SHS = EXshs + AWshs \tag{7}$$

$$quoted\ spread \equiv \$QS = \$ES + 2 \cdot \frac{1}{SHS} \cdot (\$PI \cdot PIshs + 0 \cdot AQshs - \$OQ \cdot OQshs) \tag{8}$$

$$price\ impact = \$ES - \$RS \tag{9}$$

$$effective\ /\ quoted = \frac{\$ES}{\$QS} \cdot 100 \tag{10}$$

$$at\ or\ better = \frac{AQshs + PIshs}{SHS} \cdot 100 \tag{11}$$

$$price\ improved = \frac{PIshs}{SHS} \cdot 100 \tag{12}$$

After data cleaning to correct for inconsistent coding of missing vs 0 in share volume fields across market centers, we re-calculate the quoted spread and truncate this variable to be at least $0.01. We also re-calculate the effective / quoted metric.

We merge the patched Rule 605 dataset with CRSP monthly data to obtain information on closing monthly price (prc), volume (vol), shares outstanding so we can calculate size (prc*shrout), and askhi and bidlo so we can calculate monthly price range ((askhi-bidlo)/askhi). We trim the following variables at 0.1 and 99.9% separately for market and marketable limit orders: quoted spread (before setting it to be minimum $0.01); effective spread; realized spread; price impact; and CRSP closing price. Finally, we calculate the quoted, effective, realized spreads and price impact in basis poins relative to the monthly price from CRSP.

A.2 Robustness Checks

The merger between Charles Schwab and TD Ameritrade provides another opportunity to examine the effects of a competitive shock. The merger of the corporate entities closed in October 2020, however, the merger of retail trading operations is not expected to be completed until the third quarter of 2023. For our purposes, we are interested in when the contracts between Schwab/TD Ameritrade and wholesalers were renegotiated, and the combined entity was able to use its potentially larger bargaining power to obtain better execution quality. We cannot observe these negotiations, or the nature of the contracts. This said, based on Rule 606 disclosures, we are able to observe when the payments for order flow charged by Schwab and TD Ameritrade were homogenized. This occurred in July 2021. Therefore, to examine whether the higher bargaining power resulted in lower realized spreads, we run a diff-in-diff with the pre-period being April, May, and June 2021, and the post period being August, September, and October of 2021.

Table A1 reports the results from running the following regression:

$$\begin{aligned} \textit{realized spread}_{it} = \alpha + \beta_1 WHOL_{it} + \beta_2 WHOL \times POST_{it} + \beta_3 price_{it} + \beta_4 volume_{it} \\ + \beta_5 volatility_{it} + \varepsilon_{it}, \end{aligned} \tag{13}$$

where *realized spread*$_{jt}$ is the realized spread in stock i 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; $POST$ is a dummy variable that has a value of 1 after the Schwab/TD Ameritrade fee unification and 0 otherwise, *price* is the natural log of the stock price; *volume* is the natural log of trading volume; $volatility$ is the difference between the high and low prices scaled by the high price. The models are estimated with stock and month fixed effects, which is why the standalone $POST$ variable is omitted. We run the regressions separately for each sub-sample.

[Table A1]

Based on the β_2 coeficients in Table A1, we find no effect on realized spreads for S&P 500, Tercile 1 or Tercile 2 stocks, but wholesaler realized spreads relative to exchanges increase following the event for Tercile 3 stocks. It is difficult to reconcile this evidence with a bargaining story, but since our events overlap it is possible that what we see in Panel B is again the results of Jane Street's entry into the retail wholesale business.

Table A1
Competitive Shocks: Schwab/TD Ameritrade Fee Unification

The table reports coefficient estimates from the following regression:

$$\begin{aligned} \textit{realized spread}_{it} = &\ \alpha + \beta_1 WHOL_{it} + \beta_2 WHOL \times POST_{it} + \beta_3 price_{it} + \beta_4 volatility_{it} \\ &+ \beta_5 volume_{it} + \varepsilon_{it}, \end{aligned}$$

where *realized spread*$_{jt}$ is the realized spread in stock i 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; *POST* is a dummy variable that has a value of 1 after the Schwab/TD Ameritrade fee unification (Panel B), *price* is the natural log of the stock price; $volatility$ is the difference between the high and low prices scaled by the high price, and *volume* is the natural log of trading volume. The models are estimated with stock and month fixed effects, which is why the standalone *POST* variable is omitted. The standard errors are double-clustered across stocks and months. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	S&P 500	Tercile 1	Tercile 2	Tercile 3
	[1]	[2]	[3]	[4]
Panel A: Schwab-TD Ameritrade fee unification				
WHOL	1.174***	5.216***	11.228***	20.158***
	(0.18)	(0.29)	(0.61)	(0.84)
WHOL×POST	0.298	0.214	1.783	5.316***
	(0.22)	(0.33)	(0.98)	(1.00)
price	-0.901**	-1.332	1.670	-2.596
	(0.28)	(0.74)	(1.40)	(1.95)
volatility	0.004	-0.047**	0.034	-0.054
	(0.01)	(0.02)	(0.02)	(0.04)
volume	-0.606	-1.101	-5.642***	-1.174
	(0.24)	(0.47)	(1.16)	(1.02)
intercept	11.488**	17.160**	55.324***	19.302
	(4.15)	(6.07)	(13.18)	(9.05)
Adj. R^2	0.226	0.325	0.216	0.221

A.3 ETF Tables

Table A2
Market Shares for ETFs

The table contains the list of 22 trading venues that execute held liquidity-demanding orders in ETFs during the sample period (2019-2022). The data are from the SEC Rule 605 reports. Wholesalers are highlighted in bold font. We report the total number of shares executed by each venue (in billions) and each venue's market share. Panel A aggregates by venue type, while Panel B contains the results by venue.

	venue type	shares executed, bil.	mkt. share, %
Panel A: by venue type			
	EXCH	405.38	66.67
	WHOL	202.70	33.33
Panel B: by venue			
NASDAQ	EXCH	102.82	16.91
NYSE ARCA	EXCH	97.26	15.99
Citadel	WHOL	77.68	12.77
Virtu	WHOL	56.61	9.31
BATS	EXCH	50.37	8.28
EDGX	EXCH	38.16	6.28
G1	WHOL	29.64	4.88
BYXX	EXCH	24.70	4.06
NYSE	EXCH	22.50	3.70
EDGA	EXCH	19.40	3.19
NYSE NAT	EXCH	16.81	2.76
NSDQ PHIL	EXCH	13.47	2.22
UBS	WHOL	12.03	1.98
Two Sigma	WHOL	10.21	1.68
NSDQ BOS	EXCH	10.18	1.67
Jane Street	WHOL	9.23	1.52
IEX	EXCH	5.85	0.96
Merrill Lynch	WHOL	3.82	0.63
Morgan Stanley	WHOL	3.47	0.57
NYSE AMER	EXCH	3.26	0.54
NYSE CHI	EXCH	0.40	0.07
MEMX	EXCH	0.21	0.03
Total		608.08	100.00

Table A3
Execution Quality for ETFs

The table contains execution quality statistics for held liquidity-demanding orders in ETFs. We compute the statistics separately for orders executed by wholesalers (WHOL) and exchanges (EXCH). We report the average number of shares executed and the average stock price in a sample stock during the sample period, followed by the percentage share of shares that are price improved or executed at or better the corresponding NBBO. Further, we report the quoted and effective spreads in basis points, and to better understand the magnitude of price improvement, we compute the ratio of the effective to the quoted spread. Finally, we compute the components of the effective spread: price impact and realized spread. All variables are volume-weighted. Asterisks *** in column [3] indicate statistical significance of differences between columns [1] and [2] at the 1% level.

	WHOL	EXCH	diff. [1]-[2]
	[1]	[2]	[3]
# shares, mil.	62.66	125.16	***
price, $.	42.46	42.15	
improved, %	75.34	11.02	***
at or better, %	95.35	98.88	***
quoted spread, bps	27.95	24.34	***
effective spread, bps	17.97	23.23	***
effective / quoted	0.76	0.97	***
price impact, bps	3.70	3.70	***
realized spread, bps	14.28	6.11	***

Table A4
Moving ETF Retail Flow to Exchanges

The table illustrates possible consequences of moving ETF retail flow to exchanges. Among such consequences are an overall reduction in price impacts for all exchange trades, a reduction in realized spreads incurred by retail traders, and a reduction in price improvement obtained by retail traders. Panel A reports percentage changes in effective spreads for retail liquidity demanders (RET LDs) and liquidity demanders, whose orders are currently routed to exchanges (EXCH LDs). Panel B reports gains measured in terms of effective spreads for LDs and realized spreads for LPs from the move for four categories of market participants: RET LDs, EXCH LDs, exchange liquidity providers (EXCH LPs), and wholesalers (WHOL LPs). The line in bold font represents an assumption that the currently prevailing exchange realized spreads will not change if retail flow moves to exchanges. The remaining lines allow realized spreads to vary as a result of the move, in 0.1 bps increments.

realiz. spr.,	Panel A: Δ eff. spread, %		Panel B: gains, in $ bil.			
bps.	RET LDs	EXCH LDs	RET LDs	EXCH LDs	EXCH LPs	WHOL LPs
5.61	31.21	-11.61	-4.81	15.01	2.02	-12.24
5.71	31.83	-10.77	-5.00	13.92	3.31	-12.24
5.81	32.46	-10.77	-5.00	13.92	3.31	-12.24
5.91	33.09	-10.34	-5.09	13.37	3.95	-12.24
6.01	33.71	-9.92	-5.19	12.82	4.59	-12.24
6.11	**34.34**	**-9.50**	**-5.29**	**12.28**	**5.24**	**-12.24**
6.21	34.97	-9.08	-5.38	11.73	5.88	-12.24
6.31	35.59	-8.66	-5.48	11.19	6.52	-12.24
6.41	36.22	-8.23	-5.58	10.64	7.16	-12.24
6.51	36.85	-7.81	-5.67	10.10	7.80	-12.24
6.61	37.47	-7.39	-5.77	9.55	8.45	-12.24

Table A5
ETF Execution Quality: Regression

The table reports coefficient estimates from market quality regressions for ETFs of the following form:

$$DepVar_{it} = \alpha + \beta_1 WHOL_{it} + \beta_2 price_{it} + \beta_3 volatility_{it} + \beta_4 volume_{it} + \varepsilon_{it},$$

where $DepVar_{it}$ is one of the following market quality variables for stock i in month t: the ratio of effective to quoted spreads, quoted spread, effective spread, price impact, and realized spread as defined previously; $WHOL$ is a dummy variable that has a value of 1 for orders executed by wholesalers and 0 for orders executed by exchanges; *price* is the natural log of the stock price; $volatility$ is the difference between the high and low prices scaled by the high price; and *volume* is the natural log of trading volume. The models are estimated with stock and month fixed effects, and the standard errors are double-clustered across stocks and months. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	eff. spr. / quot. spr.	quoted spr.	effective spr.	price imp.	realized spr.
	[1]	[2]	[3]	[4]	[5]
WHOL	-0.415***	3.646***	-4.039***	-9.638***	5.599***
	(0.01)	(0.37)	(0.15)	(0.65)	(0.60)
price	-0.010***	-5.052***	-4.293***	-4.077***	-0.215
	(0.00)	(0.66)	(0.56)	(0.85)	(0.49)
volatility	0.000	-0.002	-0.001	0.001	-0.003
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
volume	-0.005***	-2.131***	-1.603***	-0.853***	-0.749***
	(0.00)	(0.20)	(0.16)	(0.17)	(0.12)
intercept	1.031***	55.713***	46.872***	34.845***	12.022***
	(0.02)	(3.56)	(3.13)	(4.64)	(2.56)
Adj. R^2	0.549	0.665	0.600	0.300	0.210

Table A6
ETF Execution Quality Across Wholesalers: Regressions

The table reports coefficient estimates from wholesaler ETF market quality regressions of the following form:

$$DepVar_{it} = \alpha + \beta_1 top2_{it} + \beta_2 price_{it} + \beta_3 volatility_{it} + \beta_4 volume_{it} + \varepsilon_{it},$$

where $DepVar_{it}$ is one of the following market quality variables for stock i in month t: the ratio of effective to quoted spreads, quoted spread, effective spread, price impact, and realized spread as defined previously; $top2$ 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; *price* is the natural log of the stock price; $volatility$ is the difference between the high and low prices scaled by the high price; and *volume* is the natural log of trading volume. The models are estimated with stock and month fixed effects, and the standard errors are double-clustered across stocks and months. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	eff. spr. / quot. spr.	quoted spr.	effective spr.	price imp.	realized spr.
	[1]	[2]	[3]	[4]	[5]
top2	0.070***	-0.080	1.654***	0.842**	0.812**
	(0.01)	(0.05)	(0.14)	(0.35)	(0.37)
price	-0.023***	-5.189***	-3.641***	-1.075***	-2.566***
	(0.00)	(0.76)	(0.55)	(0.33)	(0.46)
volatility	0.000	0.002	0.001	-0.003	0.005
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
volume	-0.004	-2.537***	-1.404***	0.149	-1.552***
	(0.00)	(0.23)	(0.15)	(0.10)	(0.14)
intercept	0.594***	63.700***	37.257***	4.199**	33.052***
	(0.03)	(4.26)	(3.02)	(1.61)	(2.65)
Adj. R^2	0.143	0.692	0.523	0.151	0.311

Table A7
Wholesaler ETF Order Flow Determinants: Regression

The table reports the results from estimating the following regression on ETF data:

$$mkt.\ share_{j,t} = \alpha + \beta_1 realized\ spread_{j,t-1} + \beta_2 price_t + \beta_3 volatility_t + \beta_4 volume_t + \varepsilon_{j,t},$$

where *mkt. share*$_{jt}$ is the market share of volume executed by wholesaler j in month t; *realized spread*$_{j,t-1}$ is average realized spread earned by wholesaler j in month $t-1$ from marketable orders; *price* is the natural log of the stock price; *volatility* is the difference between the high and low prices scaled by the high price, and *volume* is the natural log of trading volume. We run these regressions separately for each sub-sample, use stock and month fixed effects, and cluster standard errors by stock and month. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	mkt. share$_{j,t}$
lagged realized spread	5.881
	(3.240)
price	-0.005***
	(0.00)
volatility	0.023***
	(0.00)
volume	-0.013***
	(0.00)
intercept	0.295***
	(0.01)
Adj. R^2	0.096

Table A8
Competitive Shocks: ETFs

The table reports coefficient estimates from the following regression for ETFs:

$$realized\ spread_{it} = \alpha + \beta_1 WHOL_{it} + \beta_2 WHOL \times POST_{it} + \beta_3 price_{it} + \beta_4 volatility_{it} + \beta_5 volume_{it} + \varepsilon_{it},$$

where $realized\ spread_{jt}$ is the realized spread in stock i 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; $POST$ is a dummy variable that has a value of 1 after the Jane Street entry (Panel A) and after Schwab/TD Ameritrade fee unification (Panel B), *price* is the natural log of the stock price; $volatility$ is the difference between the high and low prices scaled by the high price, and *volume* is the natural log of trading volume. The models are estimated with stock and month fixed effects, which is why the standalone $POST$ variable is omitted. The standard errors are double-clustered across stocks and months. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

Panel A: Jane Street entry	
WHOL	3.865***
	(0.13)
WHOL×*POST*	0.528
	(0.25)
price	-1.445
	(0.96)
volatility	0.005
	(0.00)
volume	-0.168
	(0.31)
intercept	11.991**
	(3.69)
Adj. R^2	0.334
Panel B: Schwab-TD Ameritrade fee unification	
WHOL	3.861***
	(0.13)
WHOL×*POST*	0.656
	(0.27)
price	0.565
	(2.04)
volatility	0.005
	(0.01)
volume	-0.093
	(0.22)
intercept	3.906
	(8.44)
Adj. R^2	0.358