The Retail Execution Quality Landscape*

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

March 13, 2023

Abstract. Market observers criticize the practice of retail brokers routing retail orders to wholesalers and argue that retail flow should execute on exchanges. Using comprehensive data, we show that both wholesalers and exchanges have characteristics beneficial for retail orders. Wholesalers provide substantial (significantly beyond *de minimis*) price improvement, while exchanges offer lower liquidity costs (realized spreads). On balance, price improvement dominates, and wholesaler intermediation saves retail investors close to a billion dollars per month. Four characteristics of the market for retail order flow are inconsistent with wholesaler market power. First, retail brokers reward wholesalers that offer lower liquidity costs with more order flow. Second, the largest two wholesalers charge the lowest liquidity costs. Third, neither a new wholesaler entry nor an increase in retail broker bargaining power reduces liquidity costs charged by wholesalers. Fourth, cross-sectional differences in liquidity costs are driven by proxies for inventory costs.

Key words: Retail Trading, Wholesalers, Execution Quality

JEL: G20; G24; G28

*We thank James Brugler, Tom Ernst, Peter Haynes, David Hecht, Peter Reiss, and seminar participants at Indiana University and Macquarie University 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, 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.

How retail orders are handled is currently actively debated. On the one hand, some observers argue that wholesalers wield market power and provide limited benefits to retail investors. They suggest that price improvement offered by wholesalers tends to be *de minimis*, or the smallest amount possible. 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.[1] 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 the majority of 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.

[1] 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).

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.[2] 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, 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 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 with a notably lower adverse selection cost. In our sample, order flow received by them generates 31% less adverse selection 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 called the *realized spread*, which is the difference between the effective spread (the market maker's liquidity provision revenue) and the price impact. We think of the realized spread as the cost of liquidity generation, because many of its components depend on market makers' strategic and technological choices. The data show that exchanges generate liquidity at a substantially lower cost than wholesalers; these costs are even negative for large parts of the cross-section of securities.

[2]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.

How does the difference in liquidity generation costs 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 liquidity generation costs should be smaller than those observed in a pure market maker setting. According to industry estimates, market making algorithms represent about 16% of all liquidity generation on modern exchanges.[3] With such a split, exchange liquidity should be notably cheaper to generate than wholesaler liquidity, as confirmed by the data.

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.[4] For wholesalers, the data cover predominantly retail flow, while for exchanges they cover predominantly institutional flow. For simplicity, in what follows we refer to wholesaler-executed trades as *retail* and to exchange-executed trades as *institutional*. We restrict our analysis to liquidity-demanding orders and drop ETFs.[5]

The data also 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 costs facing retail investors.

The above-mentioned differences between retail and institutional flows allow us to consider what would happen if retail flow were to move to exchanges. On the one hand, such a move would benefit retail investors via lower liquidity generation costs. On the other hand, retail investors

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

[4] 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.

[5] Results for ETFs are in the Appendix.

would lose price improvement provided by wholesalers. When we compute the net effect of such a move on retail traders, giving careful consideration to adverse selection costs and possible changes in exchange realized spreads, we obtain that retail investors would generally be worse off on exchanges. In the meantime, institutional traders would gain, since trading costs on exchanges would decline due to lower toxicity. Consequently, relocation of retail flow to exchanges would subsidize institutional flow at the expense of retail flow.

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.

What is the role of retail brokerages in obtaining lower execution costs for their clients? Do they preference wholesalers with lower liquidity generation costs? The data suggest that they indeed do. Wholesalers that offer lower liquidity generation costs today, are indeed rewarded with additional order flow in the future. We document large differences in wholesaler liquidity costs across securities, but find that these can largely be explained by proxies for inventory costs. Taken together, these results suggest that the market for retail order flow is competitive.

The nature of competition in the retail investor segment changes during our sample period both on the supply side because of entry by a new player (Jane Street) and the demand side because of a merger between two large retail brokers (Schwab and TD Ameritrade). If wholesalers had market power and thus were able to reap economic rents prior to these events, we expect competitive forces to increase the pressure on wholesalers to deliver better execution quality post entry/merger. In other words, we expect realized spreads to decrease. However, we find no ev-

idence that additional entry or increased retail broker bargaining power result in lower realized spreads earned by wholesalers, leading us to conjecture that wholesalers do not appear to have market power.

In summary, despite concerns expressed by some critics, the marketplace for retail order flow does not seem to be controlled by wholesalers. Rather, retail brokers appear to be controlling execution quality by routing to wholesalers who require lower compensation for liquidity generation.

Our final comments address the SEC's recent proposal to revamp retail trading practices by mandating that retail flow be directed to auctions for order-by-order competition. The proposal assumes that institutions would show significant interest in engaging with retail flow and would provide superior price improvement compared to wholesalers. However, institutional trading data suggest that this may only hold true for index stocks, and not for most stocks currently traded by retail investors. Based on our findings, it appears that many retail investors, particularly those dealing with less liquid stocks, would face slower transaction times and inferior execution quality if the proposal were put into effect.

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) and Kothari, So, and Johnson (2021) argue that wholesalers deliver retail trading costs that are lower than those offered by exchanges. Battalio and Jennings (2022) find that wholesalers provide substantial savings to investors and better prices relative to executing directly on 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.[6]

[6]Citadel, Virtu, and other wholesalers perform a dual function in the modern marketplace. They serve both as major on-exchange market makers and the largest 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 smaller 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 liquidity generation costs. Second, we show that even though the wholesale industry is indeed concentrated, it provides a significant net benefit to retail investors. Third, we show that retail brokerages act as monitors, as they base future routing decisions on current wholesaler performance. Finally, we report that the neither entry of new wholesalers, nor an increase in retail brokerage bargaining power reduces wholesaler spread capture, which is inconsistent with significant incumbent wholesaler market power.

Three concurrent studies report results 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 such data are potentially valuable, the focus on Robinhood may limit the external validity of inferences. Eaton, Green, Roseman, and Wu (2022) and Schwarz, Barber, Huang, Jorion, and Odean (2022) for instance show that Robinhood trader behaviour and execution quality substantially differ from those observed for the other retail brokerages. Battalio and Jennings (2022) also use proprietary data, but from one or more wholesaler(s) and

only for May of 2022. We differ from these studies by carefully analyzing a public Rule 605 dataset that has so far been only cursorily examined by the academic researchers. According to industry participants, this dataset allows for the cleanest identification of retail order flow that is possible without proprietary data. We believe that the quality of this dataset allows us to speak more clearly of internal validity of the results.

Compared to the retail trading literature for equities, the literature that examines retail trading in options is in relative consensus. Ernst and Spatt (2022) argue 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 Rule 605 reports filed by execution venues in the U.S. and generously makes the resulting data available to us. The 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.[7] As an example of a special instruction, a trader may ask that a limit order is not forwarded to venues

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

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 professional traders. As such, Rule 605 data cover virtually all retail orders, but 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 representatives, 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*.[8] 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 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 whole-

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

salers 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% of order flow, respectively. In total, the dataset contains information on execution quality for executable orders representing almost 2.9 trillion executed shares, which amounts to about 40% 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. In the U.S., 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 as 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 reported at the 5-minute horizon. We calculate the price impact from the 605 reports as the effective spread minus the realized spread.

Rule 605 data exclude odd lots, so our analysis is restricted to the orders of 100 shares or more.[9] 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 other metrics as discussed in the Appendix. We scale all metrics by the CRSP closing stock price and use share volume-weighted averages.

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 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]

[9] 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).

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. The difference in quoted spreads is expected given the clienteles served by wholesalers and exchanges. Many institutional trading 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 strategic 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 posit 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 as-

sertion 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 representatives 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 exchange-intermediated 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, a natural question is: What would happen if retail order flow were to be moved to the exchanges? Proposals to do so are occasionally heard in current market structure discussions. Such a move is likely to lead to three changes for retail flow. First, mixing retail flow with existing exchange flow will likely reduce average price impacts on exchanges. Second, retail flow will possibly incur smaller exchange realized spreads. Finally, price improvement currently provided by wholesalers will be

replaced by the smaller price improvement provided on exchanges. What would be the net effect of such a move on market participants?

To get a sense for the net effect, consider the following calculation based on Table 2. First, assume that retail volume relative to exchange 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. 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 on exchanges are likely to be censored – they 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.

We begin the discussion of possible consequences of moving retail flow to exchanges by assuming that current exchange realized spreads will remain unchanged at -0.34. In this case, the bold line in Panel A of Table 3 shows that while traders, who currently execute on exchanges, may benefit from the move, 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. Exchange flow would benefit from a substantial reduction in on-exchange adverse selection. Retail flow would experience a large reduction in realized spreads, but these gains would be offset by the loss of price improvements currently provided by wholesalers.

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 argues that greater trading volume results in 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, and the gains for RET LDs will therefore likely materialize.

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[10] 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.

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. Exchanges represent 68.13% of share volume for S&P 500 stocks, but they represent smaller and smaller share as size falls, reaching a low of 36.21% for Tercile 3 stocks. In other words, wholesalers play an out-sized role for less liquid stocks, a point we will return to below.

[10]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]

We explore whether the differences in market capitalization also affect execution quality in Table 5. We begin with the S&P 500 sub-sample. Wholesalers price-improve 75% of marketable orders, and they receive a price improvement corresponding to 44% of the quoted spread. By comparison, 12% of marketable orders receive price improvement on Exchanges, and traders get an average price improvement of 6%. As expected, the differences in price improvement lead to differences in trading costs and market maker gross revenues. The adverse selection in the order flow obtained by exchanges is 89% (= 6.45/3.41-1) higher than the order flow received by wholesalers. However, even though marketable orders on exchanges pay effective spreads that are 13% larger than those paid by retail investors whose orders are routed to wholesalers (5.22 vs. 4.62 bps), this is not enough to compensate for the differences in adverse selection. Consequently, wholesalers earn substantially larger realized spreads than liquidity providers on exchanges, 1.21 vs. -1.23 bps.

[Table 5]

We note that the negative value of exchange realized spreads should not necessarily be interpreted as evidence that exchange market makers lose money. First, recall that liquidity provision on exchanges is dominated by non-market maker limit orders that do not typically focus on spread revenue. As such, the negative realized spread averages may result from mixing negative realized spreads earned by non-professional liquidity providers and positive realized spreads earned by their professional counterparts.

Second, the negative realized spreads may be an artefact of the 5-minute horizon used to compute the realized spread metric. Such a horizon is mandated by Rule 605 and may be too long to capture the true profitability of modern high-speed liquidity provision strategies. In this regard, Conrad and Wahal (2020) argue that the longer horizons for realized spread calculations may understate true revenues of modern market makers, who are able to turn over their inventories

within sub-seconds. We therefore suggest that rather than focusing on the magnitude of realized spread figures, one could use them for comparing overall trading costs net of adverse selection between wholesalers and exchanges.

When it comes to Tercile 1, Tercile 2, and Tercile 3 stocks, the general pattern discussed for their S&P 500 counterparts is 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 than for those that are routed to exchanges 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 for Tercile 1, Tercile 2, and Tercile 3 stocks is substantially greater on exchanges, with price impacts 51%, 43%, and 114% higher for Tercile 1, Tercile 2, and Tercile 3 stocks, respectively. Finally, the exchange realized spreads are even more negative for Tercile 1 and Tercile 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.

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:

$$DepVar_{it} = \alpha + \beta_1 WHOL_{it} + \beta_2 price_{it} + \beta_3 volatility_{it} + \beta_4 volume_{it} + \varepsilon_{it}, \quad (1)$$

where $DepVar_{it}$ is one of the following execution quality variables for stock i 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 our overall sample in Panel A, and we are primarily interested in the coefficient on the $WHOL$ dummy, but note that the coefficients on the control variables are significant and of the expected signs. The results reported in columns [2] to [5] of Table 6 confirm that our earlier univariate findings hold also after controlling for price, volatility, and volume as well as for stock and month fixed effects. Wholesaler executions tend to occur when quoted spreads are relatively wide. For example, the quoted spreads (column [2]) 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 augment the regression by interacting the $WHOL$ dummy with dummies indicating whether a stock belongs to Tercile 1, Tercile 2, or Tercile 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 outcome variable for orders routed to wholesalers is 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 also confirm that wholesalers provide greater price improvement compared to exchanges (column [1]). 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 point 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 stocks we estimate that the price improvement is 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 order flow is lower than that of the exchange-bound order 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 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) higher 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. Table 1 above shows that these two wholesalers execute over 71% of the marketable order flow that is routed to wholesalers. The remaining six wholesalers are substantially smaller.

In Table 7, we use panel regressions to evaluate whether 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:

$$DepVar_{it} = \alpha + \beta_1 top2_{it} + \beta_2 price_{it} + \beta_3 volatility_{it} + \beta 4_v olume_{it} + \varepsilon_{it}, \tag{2}$$

where $DepVar_{it}$ is one of the following execution quality variables for stock i 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 significantly more toxic order flow (the difference is 3.02 bps) and earn significantly 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 dummy variables that take on a value of one for Tercile 1,

Tercile 2, and Tercile 3 stocks respectively. 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 higher adverse selection. The order flow routed to $top2$ is associated with a 1.00 bps higher price impact, and the realized spreads they earn are 0.55 bps lower than what their competitors earn from trading the same stocks. While the differences in price improvements shrink as we go from Tercile 1 to Tercile 3 stocks, 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) higher for $top2$ than for other wholesalers. Toxicity facing $top2$ in Tercile 3 stocks is 7.3 bps (= 0.999 + 6.285) higher and realized spreads are 9.5 bps (= -0.552 - 8.809) lower than for other wholesalers trading the same stocks.

Given that wholesalers tend to receive order flow of varying toxicity, price improvement may not be the most appropriate comparison metric for our analyses as it is determined in part by the effective spreads. We believe that the realized spreads is a better metric. Assuming that retail brokerages understand the toxicity of their own flow, they too should benchmark against a toxicity-adjusted performance metric. We use this reasoning in the subsequent analyses, in which 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 the wholesalers that they route to. Such evaluations typically occur on a monthly basis. 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 whether such relationship is observed

in the data, we estimate the following regression:

$$mkt.\ share_{j,t} = \alpha + \beta_1 realized\ spread_{j,t-1} + \beta_2 price_t + \beta_3 volatility_t + \beta_4 volume_t + \varepsilon_{j,t}, \quad (3)$$

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.

Table 8 shows that prior performance indeed appears to have an effect on order flow allocations. A lower realized spread earned on marketable orders by wholesaler j in month $t-1$ leads to larger market shares for this wholesaler in the subsequent month. The only exception is Tercile 2 where the coefficient on the lagged realized spread is negative but insignificant.

[Table 8]

These results show that wholesalers have an strong incentive to offer low toxicity adjusted spreads, doing so increases their market share. While this evidence does not prove an absence of wholesaler market power, it does suggest that there is at least some level of competition for retail order flow based on execution quality among wholesalers.

3.4 Competitive Shocks

The nature of competition in the retail investor segment changes during our sample period both on the supply side because of entry by a new player (Jane Street) and the demand side because of a merger between two large retail brokers (Schwab and TD Ameritrade). If wholesalers had market power and thus were able to reap economic rents prior to these events, we expect

competitive forces to increase the pressure on wholesalers to deliver better execution quality post entry/merger. In other words, we expect realized spreads to decrease.

The first competitive change we study is the entry of Jane Street into the business offering retail order execution services as a wholesaler. Jane Street entered in 2019, but throughout 2020 the firm still had a negligible market share based on Rule 605 data. This increased gradually in the spring of 2021, reaching a 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 diff-in-diff with the pre-period being the last three months of 2020, and the post-period being the last three months of 2021.

The second competitive event we study is the merger between Charles Schwab and TD Ameritrade. 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 9 reports the results from running the following regression:

$$\begin{aligned} realized\ spread_{it} = \alpha + \beta_1 WHOL_{it} + \beta_2 WHOL \times POST_{it} + \beta_3 price_{it} + \beta_4 volume_{it} \\ + \beta_5 volatility_{it} + \varepsilon_{it}, \end{aligned} \tag{4}$$

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 or after 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 9]

Based on the β_2 coefficients in Table 9, we do not find that realized spreads for wholesalers relative to exchanges decline following either event for any sub-sample. For index and Tercile 1 stocks, the entry/merger did not cause realized spreads to change suggesting that they were already at a competitive level. For Tercile 2 and Tercile 3 stocks, realized spreads actually increase.

To understand the results, consider first what happens when Jane Street enters in Panel A. Order flow is now divided among a larger number of different 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 Tercile 2 and Tercile 3 where wholesalers handle the majority of the volume (see Table 4). Our results are consistent with wholesalers reducing retail price improvements to cover the higher inventory costs for illiquid stocks they face following Jane Street's entry. By contrast, for S&P 500 and Tercile 1 securities, wholesalers handle a lower fraction of volume overall, and the liquidity of these stocks likely make it easier to manage inventory risk even with a lower market share, reducing the need to reduce retail price improvements to cover higher costs. We return to the topic of inventory risk in the next subsection.

Second, consider the Schwab/TD Ameritrade merger in Panel B. Here, again 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.

We tentatively conjecture that the wholesaler market is already competitive prior to these events since we see no evidence that additional entry or increased retail broker bargaining power result in lower spread capture by wholesalers. We also see tentative evidence consistent with inventory risk 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/merger events 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 on trading volume as a proxy for inventory costs. Specifically, a stock-month with lower volume is associated with higher inventory costs, especially when controlling for volatility. Share volume captures the ability for the wholesaler to lay off a position. To understand the role of inventory costs for wholesalers, we run the following panel regressions:

$$\begin{aligned} \textit{realized spread}_{it} = \alpha + \beta_1 T1 + \beta_2 T2 + \beta_3 T3 + \beta_4 price_{it} + \beta_5 volatility_{it} \\ + \beta_6 volume_{it} + \varepsilon_{it}, \end{aligned} \tag{5}$$

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; *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, Tercile 2, and Tercile 3 stocks have significantly higher realized spreads than S&P 500 stocks. When we control for price and volatility (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 cost, 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 earn significantly lower realized spreads for less liquid stocks than they do for S&P 500 stocks. What at face value appeared to be excessively large realized spreads of 39.17 bps for Tercile 3 stocks relatively to S&P 500 stocks in column [1] are insufficient to compensate for incremental inventory costs based on columns [3] which shows that realized spreads are 20.37 bps lower. Recall that S&P 500 stocks had a realized spread of 1.21 bps (see Table 5). In other words, realized spreads would likely be negative for Tercile 1, Tercile 2, and Tercile 3 stocks once we control for inventory costs.

An important caveat is that the realized spreads are measured over a five minute horizon in Rule 605 data. While this may be an appropriate horizon for less liquid securities, wholesalers clearly are often able to manage their position much faster than that, particularly for index stocks. Hence, it is possible that wholesalers capture more of the spread than we estimate based on the Rule 605 data for index stocks, perhaps even the entire effective spread of 4.62 bps (see Table 5). But, even if that were the case our results suggest that realized spreads may in practice be insufficient to cover inventory costs for less liquid securities. Therefore, we conclude that the

cross-sectional differences in wholesaler realized spreads that we observe between less liquid and index stocks appear to reflect differences in inventory costs.

3.6 Institutional Interest

In December 2022, the SEC proposed rules that would significantly change the equity markets.[11] To analyze these comprehensive rules is beyond the scope of the current study, but we believe our results can 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. [12] 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.[13] 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 and add to that changes in short interest (which are available bi-monthly). This gives us a proxy for institutional trading interest in a particular

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

stock. We then calculate the fraction of retail trading as reflected in Rule 605 data divided by our proxy for institutional trading interest, *rat*. We supplement this measure with the average across stocks of the fraction of quarters with no institutional trading interest based on our proxy, *nonerat*. 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] ([3]) shows that average (median) retail order flow represents 61% (20%) of institutional interest for S&P 500 stocks, so for index stocks there is significant institutional interest. Yet, even for index stocks 2.6% of stock-quarters have no institutional trading 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.

The conclusion we draw is that institutional trading interest may be very low or entirely absent for much 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 results in Table 10 suggest that realized spreads may be insufficient to cover inventory costs for less liquid stocks in the current environment. 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 much 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 (33% on average) 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 33% wider than when institutional orders are placed. Second, wholesalers mitigate these higher transaction costs by providing a substantial price improvement of 24% on average. 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.

There are several events that 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 there is a merger between two large retail brokers (Schwab and TD Ameritrade). If wholesalers had market power and thus were able to reap economic rents prior to these events, we expect competitive forces to increase the pressure on wholesalers to deliver better execution quality post entry/merger. However, we find no evidence that additional entry or increased retail broker bargaining power result in lower spread capture by wholesalers, leading us to tentatively conjecture that that the wholesaler market is already competitive prior to these events.

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 index stocks, it is unlikely to be true for the majority of 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 would be 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. 6, 7

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

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*. 7

Bartlett, R. P., J. McCrary, and M. O'Hara, 2022, "The Market Inside the Market: Odd-Lot Quotes," *Available at SSRN 4027099*. 10

Battalio, R., and R. Jennings, 2022, "Why do Brokers who do not Charge Payment for Order Flow Route Marketable Orders to Wholesalers?," *Available at SSRN 4304124*. 6, 7

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. 16

Boehmer, E., C. M. Jones, X. Zhang, and X. Zhang, 2021, "Tracking retail investor activity," *The Journal of Finance*, 76(5), 2249–2305. 7

Bogousslavsky, V., and P. Collin-Dufresne, 2022, "Liquidity, volume, and order imbalance volatility," *Journal of Finance, Forthcoming*. 16

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. 10

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. 8

Conrad, J., and S. Wahal, 2020, "The term structure of liquidity provision," *Journal of Financial Economics*, 136(1), 239–259. 17

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. 7

Ernst, T., and C. S. Spatt, 2022, "Payment for Order Flow and Asset Choice," working paper. 8

Ernst, T., C. S. Spatt, and J. Sun, 2022, "Would Order-by-Order Auctions Be Competitive?," *Available at SSRN*. 28

Hendershott, T., C. M. Jones, and A. J. Menkveld, 2011, "Does algorithmic trading improve liquidity?," *Journal of Finance*, 66, 1–33. 10

Hendershott, T., S. Khan, and R. Riordan, 2022, "Option Auctions," *Working paper*, University of California at Berkeley. 8

Hu, E., and D. Murphy, 2022, "Competition for Retail Order Flow and Market Quality," *Working paper*New York University. 7

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. 6

Kothari, S., E. So, and T. Johnson, 2021, "Commission Savings and Execution Quality for Retail Trades," *Working paper*, MIT Sloan School of Management. 6, 7

Lee, C. M., and M. J. Ready, 1991, "Inferring trade direction from intraday data," *Journal of Finance*, 46, 733–746. 10

Li, S., M. Ye, and M. Zheng, 2020, "Refusing the Best Price?," *Available at SSRN 3763455*. 9

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*. 7

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×*T1*	0.063***	1.520***	0.160	-1.446***	1.607***
	(0.00)	(0.33)	(0.17)	(0.23)	(0.27)
WHOL×*T2*	0.124***	11.396***	2.058***	-9.492***	11.550***
	(0.01)	(0.56)	(0.47)	(0.50)	(0.68)
WHOL×*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:

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

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. 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.
	[1]	[2]	[3]	[4]	[5]
Panel A: Base Specification					
top2	0.008	-1.180***	-0.417	3.022***	-3.439***
	(0.01)	(0.13)	(0.57)	(0.48)	(0.77)
price	-0.032***	-19.550***	-19.803***	-8.620***	-11.184***
	(0.00)	(1.32)	(1.33)	(0.86)	(0.92)
volatility	0.000***	0.314***	0.266***	0.168***	0.099***
	(0.00)	(0.03)	(0.02)	(0.02)	(0.02)
volume	-0.005***	-32.910***	-27.649***	-12.887***	-14.764***
	(0.00)	(1.60)	(1.66)	(0.99)	(0.82)
intercept	0.819***	481.222***	407.612***	193.495***	214.142***
	(0.02)	(20.46)	(21.43)	(12.66)	(10.99)
Adj. R^2	0.292	0.764	0.696	0.391	0.258
Panel B: Specification with Interaction Terms					
top2	0.055***	-0.332***	0.446***	0.999***	-0.552**
	(0.01)	(0.05)	(0.14)	(0.19)	(0.24)
top2×*T1*	-0.028***	-0.209***	-0.173*	0.086	-0.261**
	(0.00)	(0.05)	(0.10)	(0.12)	(0.11)
top2×*T2*	-0.066***	-1.138***	-0.619	1.510***	-2.130***
	(0.01)	(0.13)	(0.49)	(0.34)	(0.54)
top2×*T3*	-0.077***	-1.852***	-2.522*	6.285***	-8.809***
	(0.01)	(0.34)	(1.37)	(1.00)	(1.76)
price	-0.032***	-19.550***	-19.803***	-8.620***	-11.184***
	(0.00)	(1.32)	(1.33)	(0.86)	(0.92)
volatility	0.000***	0.314***	0.266***	0.167***	0.099***
	(0.00)	(0.03)	(0.02)	(0.02)	(0.02)
volume	-0.005***	-32.910***	-27.649***	-12.886***	-14.765***
	(0.00)	(1.60)	(1.66)	(0.99)	(0.82)
intercept	0.819***	481.229***	407.617***	193.482***	214.158***
	(0.02)	(20.46)	(21.43)	(12.66)	(10.99)
Adj. R^2	0.296	0.764	0.696	0.392	0.259

Table 8
Wholesaler Order Flow Determinants: Regression

we estimate the following regression:

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

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.

	S&P 500	Tercile 1	Tercile 2	Tercile 3
	[1]	[2]	[3]	[4]
lagged realized spread	-0.098**	-0.126***	-0.002	-0.110***
	(0.04)	(0.03)	(0.00)	(0.03)
price	0.000	0.000	0.001	-0.002***
	(0.00)	(0.00)	(0.00)	(0.00)
volatility	-0.000	0.000	0.027***	0.016***
	(0.01)	(0.00)	(0.00)	(0.00)
volume	-0.000	-0.002***	-0.009***	-0.006***
	(0.00)	(0.00)	(0.00)	(0.00)
intercept	0.134***	0.152***	0.231***	0.206***
	(0.01)	(0.00)	(0.01)	(0.00)
Adj. R^2	0.010	0.002	0.037	0.034

Table 9
Competitive Shocks

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) 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.

	S&P 500	Tercile 1	Tercile 2	Tercile 3
	[1]	[2]	[3]	[4]
Panel A: Jane Street entry				
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: 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

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.[14]

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.

[14]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 ($RS* is the share-weighted average spread in dollars using a five minute horizon.[15]
- *Average effective spread ($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.

[15] 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 ETF Tables

Table A1
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 A2
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 A3
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 A4
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 A5
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 A6
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 A7
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