The Retail Execution Quality Landscape*

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

December 18, 2023

Abstract. We show that off-exchange (wholesaler) executions provide significant trading cost savings to retail investors. Despite industry concentration, three findings suggest that wholesalers do not abuse market power. First, brokers closely monitor and reward wholesalers offering low liquidity costs with more order flow. Second, the largest wholesalers offer the lowest costs due to economies of scale. Finally, the entry of a new large wholesaler does not reduce liquidity costs. Drawing from these insights, we discuss the implications of two proposed alternatives to the status quo: (i) pooling retail and institutional flows on exchanges and (ii) sending retail flow to order-by-order auctions.

Key words: Retail Trading, Wholesalers, Execution Quality

JEL: G20; G24; G28

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

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

1. Introduction

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

Many U.S. retail brokerages historically operated their own market-making businesses, executing their clients' orders in-house and, in some cases, also orders from other smaller brokers. However, market developments and regulatory initiatives in the early 2000s that spurred increased competition for liquidity provision prompted them to exit these businesses. Instead, brokers began outsourcing order execution to vertically integrated wholesalers, whom they believe to have the capital, scale, and expertise necessary to most effectively serve their retail clients.[2]

How retail orders are currently handled is actively debated. Some observers argue that wholesalers wield (and abuse) market power and provide limited benefits to retail investors.[3] These observers suggest that price improvement offered by wholesalers tends to be *de minimis*, or the smallest amount possible.[4] To enhance competition in this segment, the Securities and Exchange Commission (SEC) is considering implementing a system of auctions in which market participants would compete for each individual retail order, obtaining execution rights only if they

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

[2] For example, consider statements by Charles Schwab, which sold its capital markets unit to UBS in 2004. "U.S. Equity Market Structure: Order Routing Practices, Considerations, and Opportunities," Charles Schwab, Order-Routing White Paper, Q2 2022 (https://bit.ly/3RhOpNI).

[3] "IEX Supports SEC Equity Market Proposals," by A. Lyudvig, Traders Magazine, March 22, 2023 (https://bit.ly/3Hks0K2).

[4] This perception seems to originate from a July 2012 study by Nanex (https://bit.ly/3Ooijgp).

provide the largest amount of price improvement. Wholesalers instead argue that retail investors are well-served by the current system, contending that market power lies with retail brokerages, who route orders to wholesalers in the best interest of their customers.

Reconciling these conflicting views is an empirical task, and we do so using four years of SEC Rule 605 reports for all U.S. equities from 2019 to 2022.[5] Each order-handling venue must file such reports on a monthly basis to maintain a public record of execution quality. Our analyses show that wholesalers provide substantial price improvement, executing liquidity-demanding orders at prices better than those quoted on exchanges. Wholesaler price improvement is far from *de minimis*, amounting to 24% of the quoted spread in the full sample, meaning that the retail buyer and seller each receive a liquidity cost discount of 12%, or 1.7 cents per share.[6] Even more noteworthy is that retail traders in an average S&P 500 stock receive price improvement amounting to 47% of the quoted spread. In comparison, exchanges offer only 3% price improvement in the full sample and 5% in S&P 500 stocks. Therefore, when it comes to one aspect of execution quality, wholesalers offer a significant advantage.[7]

While the magnitude of price improvement is encouraging, the data also reveal that the wholesale industry is concentrated, with the two largest wholesalers – Citadel Securities and Virtu Financial – capturing over 70% of retail flow. Given this context, the concerns regarding possible market power abuses by wholesalers are not surprising. Although we cannot directly observe whether such abuses occur, the data offer several indicators that lead us to think otherwise.

To begin with, and contrary to the expectation that the larger wholesalers abuse market power, we observe that Citadel and Virtu charge the lowest liquidity costs, even though they handle the

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

[6] It is commonly assumed that quoted spreads are close to 1 cent, especially for large firms. In reality, the average U.S. stock has a quoted spread of 9.1 cents, and the average S&P 500 stock has a quoted spread of 11.7 cents.

[7] One often hears that exchanges are unable to provide price improvement on the same scale as wholesalers because regulations do not permit sub-penny executions on exchanges. The data however show that price improvement by wholesalers in an average stock amounts to more than 1 cent, and in S&P 500 stocks it exceeds 2 cents. Therefore, the inability to trade in sub-pennies is likely not the primary impediment to exchange price improvement.

most informed retail flow. Furthermore, we find that the scale of their operations explains the relatively low liquidity costs charged by the top two wholesalers. This suggests that economies of scale generate cost savings at the wholesaler level, and these savings tend to be transferred to retail customers.

Why do the top two wholesalers choose to pass on the savings instead of keeping them to boost their profits? The data suggest that a system of broker monitoring and rewards facilitates these transfers. We observe that the largest retail brokerage is larger than the largest wholesaler. This suggests that market power could potentially reside with the brokerages. To examine this possibility, we ask if brokerages actively manage their relationships with wholesalers by favoring those that offer lower liquidity costs. The data indicate that they indeed do so. Wholesalers that provide lower costs today are rewarded with additional order flow in the future. Notably, brokerages appear to evaluate wholesalers not on a stock-by-stock basis but rather on a bundled basis. For instance, if Citadel offers the cheapest liquidity in Apple stock, it will not necessarily receive more future Apple flow. Instead, Citadel must outperform its competitors across the entire range of stocks to attract more order flow.

This finding highlights an intriguing aspect of the retail ecosystem, where the brokerages compel wholesalers to compete in small stocks that have relatively low trading frequency and high inventory costs. Typically, small low-volume stocks are less attractive to intermediaries, and market regulators and exchanges often seek ways to improve liquidity in such stocks. When we account for inventory costs, our analysis suggests that wholesalers tend to charge relatively low liquidity costs in small stocks compared to large stocks, pointing to cross-subsidization facilitated by bundling.

To end with, the dynamics of wholesaler competition undergo a transformation during our sample period with the entry of a new player, Jane Street. Within a few months, Jane Street gains a significant market share, capturing over 12% of retail flow. If wholesalers were abusing their market power before this entry, we would expect competitive pressures to intensify, leading to

lower liquidity costs. Our difference-in-differences analyses, however, do not support this conjecture; we find no evidence of a decrease in liquidity costs. In fact, in medium and small stocks the costs increase, likely due to the incumbents' loss of economies of scale.

In summary, despite the concerns voiced by critics, wholesalers do not seem to abuse the marketplace for retail executions. Instead, retail brokers exert control over execution quality by routing to wholesalers that offer lower costs. The marketplace is also contestable, and a new entrant successfully captures a sizable market share in a surprisingly short time. Wholesale business appears to be characterised by economies of scale, with the largest wholesalers offering the lowest costs. The latter characteristic raises the question of why the market has not naturally gravitated toward an equilibrium with just one wholesaler. We posit that such an equilibrium would prove sub-optimal for brokerages, as a monopolistic wholesaler would be challenging to control. Consequently, the current state with several competing wholesalers appears to maintain an intriguing balance where certain players are allowed to become large while smaller players are retained to serve as an ongoing credible threat, discouraging any rogue behavior.

Our concluding analyses discuss alternatives to the U.S. status quo for retail executions. First, we contemplate what would happen if retail orders were routed to exchanges instead of wholesalers, a practice that prevails in many countries around the globe.[8] The data indicate that moving to exchanges would benefit retail investors by reducing realized spreads that they pay. However, it would also disadvantage them with increased adverse selection costs stemming from being pooled with institutional flow. When we analyze the overall impact of such a move on retail traders, we find that they would most likely be worse off.

Second, we consider the SEC's proposal to require retail orders to be sent to auctions for order-by-order competition. The proposal assumes that non-professional liquidity providers such as hedge funds, mutual funds, and pension funds would demonstrate significant interest in engag-

[8] Proposals to shift retail flow to exchanges are often heard in current U.S. market structure discussions. The SEC has received numerous comment letters advocating for a significant reduction in off-exchange retail trading. See, for instance, the letter promulgated by the We The Investors group (https://bit.ly/434ceeE).

ing with retail flow and offer superior price improvement compared to wholesalers. However, our analysis of institutional trading data indicates that this assumption may only hold true for large stocks and may not apply to many stocks currently traded by retail investors. Additionally, order-by-order auctions would eliminate bundling and could reduce the incentives for intermediaries to engage with retail traders in small stocks. Therefore, we caution that many retail investors would likely experience lower execution quality if the proposal were to be implemented. Ernst, Spatt, and Sun (2023) come to a similar conclusion based on a theoretical model.

Related literature. We aim to contribute to the expanding literature that examines possible positive and negative outcomes of the current market structure for handling retail flow. On the positive side, Adams, Kasten, and Kelley (2021), Kothari, So, and Johnson (2021), and Battalio and Jennings (2023a) suggest that wholesalers deliver low trading costs for retail investors. Jain, Mishra, O'Donoghue, and Zhao (2022) further argue that wholesale revenues may boost the ability of wholesalers like Citadel and Virtu to compete on exchanges where they act as market makers. Baldauf, Mollner, and Yueshen (2023) show theoretically that market makers may use retail flows to offset inventory imbalances, improving exchange execution quality. On the negative side, Eaton, Green, Roseman, and Wu (2022) report that wholesalers may be negatively impacted by retail herding, causing them to perform worse as market makers. In addition, in theoretical settings Hu and Murphy (2023) and van Kervel and Yueshen (2023) posit that if firms like Citadel wield market power they may maintain wide exchange spreads to enhance their wholesale profits.

Three concurrent studies come to overall execution quality conclusions similar to ours, although they do not analyze competition among wholesalers, interactions between retail brokerages and wholesalers, or recent market reform proposals. These studies also have several empirical limitations that we are able to bypass. First, 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. Barber, Huang, Jorion, Odean, and Schwarz (2023) and Battalio, Jennings, Saglam, and Wu (2023) show that the algorithm misses a number

of retail trades and also often identifies institutional trades as retail. Second, Kothari, So, and Johnson (2021) use proprietary data from the Robinhood brokerage, while Battalio and Jennings (2023a) use proprietary data from one or more anonymous wholesaler(s) in May 2022. While these two studies deliver valuable insights, a complementary comprehensive analysis across multiple months, wholesalers, and brokerages may help shed light on the external validity of their inferences. This is especially true since Eaton, Green, Roseman, and Wu (2022) and Schwarz, Barber, Huang, Jorion, and Odean (2023) show that Robinhood trader behavior and execution quality tend to differ from those observed for other retail brokerages, while our study shows that there exist substantial differences among wholesalers.

We complement these studies by analyzing a multi-year comprehensive public dataset that academic researchers have not examined under the current retail market structure.[9] According to industry participants, this dataset allows for the clearest view into retail execution quality that is possible without proprietary data. It also enables us to speak to the external validity of contemporary findings in the literature and conduct an analysis of competitive forces by observing interactions between wholesalers and retail brokerages over an extended period of time marked by changes in competitive dynamics. Since our paper was first made public, two new working papers that examine broker monitoring and wholesaler incentives by Huang, Jorion, Lee, and Schwarz (2023) and Ernst, Malenko, Spatt, and Sun (2023) have recently been circulated. These studies complement our study and shed further light on our findings. We discuss these studies in detail in Sections 3.5 and 3.6.

Compared to the literature on equities, the literature that examines retail trading in options is in relative consensus. Ernst and Spatt (2022) suggest that options markets provide less price improvement compared to equity markets, and that retail brokerages are incentivized to nudge their customers into options trading, which is more profitable for the brokerages yet detrimental

[9] A flurry of papers was written after the SEC mandated in 2001 that market centers publicly disclose execution quality metrics, known as Dash 5 reports (e.g., Bessembinder (2003), Lipson (2004), Boehmer (2005), and Boehmer, Jennings, and Wei (2007)).

to customer investment returns. Along similar lines, Bryzgalova, Pavlova, and Sikorskaya (2023) argue that options market makers behave non-competitively and disproportionately benefit from the growth in retail trading. Finally, Hendershott, Khan, and Riordan (2023) 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 overall liquidity costs.

2. Data and Sample

We obtain monthly order execution quality, routing, and fee data for all equities traded in the U.S. from publicly available Rule 605 and Rule 606 reports filed by trading venues. Our main results derive from Rule 605 execution quality data, which cover a four-year period from January 2019 through December 2022 and are detailed in Internet Appendix A.1. Rule 606 routing and fee data are not available in 2019, so for the few analyses that require these data, we use a shorter three-year period from January 2020 through December 2022.

Rule 605 data include a wide range of securities. We restrict our main sample to ordinary and class A, B, and C shares that merge with monthly stock data from the Center for Research in Securities Prices (CRSP) for a total of 8,165 symbols and refer to them as *stocks*. We then divide the stocks into four sub-samples. The first sub-sample includes S&P 500 stocks, and the remaining three sub-samples are formed from size-based terciles of non-S&P 500 stocks.

The data cover close to 70 execution venues, including all stock exchanges, all major wholesalers, and many dark pools and crossing networks. We focus on the first two venue categories, that is, 14 stock exchanges and 8 largest wholesalers. According to industry consensus, Rule 605 reports submitted by wholesalers contain only retail orders, covering virtually all such orders, and are therefore the most comprehensive information source about retail execution quality. Note that the 605 reports cover execution quality for all orders wholesalers receive, not only the orders they internalize. Given that the majority of retail flow is directed to wholesalers, exchange 605

reports primarily contain execution quality for institutional orders.

We focus exclusively on liquidity-demanding orders because wholesalers are required to forward liquidity-providing orders to exchanges. Rule 605 applies to orders of size 100 to 9,999 shares that are executed during regular trading hours and contain no special instructions.[10] A comparison with the data from CRSP shows that our sample accounts for about 45% of U.S. trading volume. The unaccounted 55% corresponds primarily to the activities of institutions that use orders with special instructions, institutions that trade in the dark, as well as volume executed in opening and closing auctions and outside of regular trading hours. Because the data do not cover all institutional activity, we exercise caution and avoid making broad conclusions applicable to institutions. Instead, we focus on retail execution quality and use the institutional portion of the data only for benchmarking.

Rule 605 data exclude odd lots, but data from an industry initiative titled Financial Information Forum (FIF) include odd-lots and show that retail odd-lot market quality is similar to that reported for orders of other sizes.[11] Even though odd lots are popular among retail investors, statistics reported by Battalio and Jennings (2023a) suggest that they comprise only about 6% of all retail share volume.[12]

Panel A of Table 1 shows that exchanges execute 59.5% of volume in our sample, while

[10]For details, see "Final Rule: Disclosure of Order Execution and Routing Practices," 17 CFR Part 240 (https://bit.ly/3zyrpB1). As an example of special instruction, a trader may request that the receiving venue does not forward the order to other venues if it is unable to provide the best price, an order known as *Do Not Ship*. Li, Ye, and Zheng (2023) show that orders that contain special instructions are typically submitted by sophisticated market participants such as high-frequency traders.

[11]See, for instance, "Q1-2019 FIF Supplemental Retail Execution Quality Statistics" (https://bit.ly/3m2RC33). For a study that illustrates the experience of retail traders using odd lots, see Schwarz, Barber, Huang, Jorion, and Odean (2023).

[12]Panel A of Table 2 in Battalio and Jennings (2023a) reports the value of price improvement per share and the total value of price improvement for (i) round lots and (ii) a combined sample of round and odd lots. From these statistics, we back out the per-share price improvement for odd lots ($0.0139) and then divide the dollar volume in odd lots ($7.99M) by the per-share price improvement to obtain the total share volume in odd lots (575M). Meanwhile, the per-share price improvement for round lots is reported as $0.0085 and dividing round lot dollar volume ($77.86M) by this number gives total share volume in round lots (9,160M). Hence, odd lot volume is 5.91% (= 575/(9,160+575)) of total retail volume. The share of odd lots remains virtually the same when we account for the values in Panels B and C of the authors' Table 2.

wholesalers capture the remaining 40.5%. Panel B contains statistics for individual exchanges and wholesalers. Among exchanges, Nasdaq and the NYSE play the leading roles. Among wholesalers, Citadel and Virtu stand out as the largest, capturing 71% of retail flow, while other wholesalers are considerably smaller.

[Table 1]

Market structure studies typically rely on a set of execution quality metrics that includes quoted, effective, and realized spreads, as well as price impacts. The *quoted spread* is the difference between the national best offer (the offer quote that is the lowest across all exchanges) and the national best bid (the bid quote that is the highest across exchanges). It represents trading costs advertised by liquidity providers. Liquidity demanders do not always incur these costs exactly as advertised. They may time liquidity, consuming it when relatively cheap, and their orders may be price improved as is often done by wholesalers, or interact with better-priced non-displayed orders on exchanges such as odd lots and hidden orders (Bartlett, McCrary, and O'Hara (2023)). To assess trading costs actually incurred by liquidity demanders, the data contain the *effective spread* computed as twice the difference between the traded price and the midquote (the average of the best offer and the best bid) for buyer-initiated trades and the difference between the midquote and the traded price for seller-initiated trades.

Effective spreads are typically further divided into two components. The first component, the *price impact*, captures adverse selection (toxicity) generated by a trade and is computed as 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 and trading fees), and (ii) market maker profits. Due to the composite

nature of this metric, its interpretation is nuanced, and we carefully take this into account.

Rule 605 requires that price impacts are estimated over five-minute horizons. This timeframe may seem rather long to some readers, as they may believe that horizons relevant to modern market makers are considerably shorter, measured in seconds or even sub-seconds. In Internet Appendix A.2, we use intraday Trade and Quote data to demonstrate that the five-minute price impacts serve as a suitable proxy for adverse selection costs. The intraday data reveal that prices react to trades exceptionally quickly before levelling off. Meanwhile, new orders that could offset a liquidity provider's position and mitigate some of the price impact rarely arrive before the levelling off. As a result, liquidity providers typically incur a significant portion, and often the entirety, of price impacts. Consequently, even though we observe price impacts after they have leveled off, they tend to represent the true adverse selection cost borne by liquidity providers.

When working with the metrics, we remove outliers by trimming all variables at the 0.1 and 99.9 percentiles. Reporting of the quoted spreads is not required by Rule 605, and we derive them as discussed in Internet Appendix A.1. We scale all execution quality metrics by the CRSP closing stock price (results without scaling are in Internet Appendix A.4) and calculate stock-level statistics using share-volume weights. When further aggregating across stocks, we use a simple average, which enables us to assess execution quality in the average stock and allows for a view of the entire equity landscape. As we demonstrate shortly, retail traders are quite active across the landscape, particularly in smaller stocks.[13]

[13] In its recent analysis of retail execution quality, the SEC uses dollar-volume weights. This approach skews the execution quality metrics towards high-price high-volume stocks. See Release 34-96495 "Order Competition Rule" from December 14, 2022 (https://bit.ly/3v1Z96V). In Internet Appendix A.5, we show that our results are similar to the Commission's when we use the same weighting technique.

3. Empirical Results

3.1 Execution Quality

Table 2 reports summary execution quality metrics. Wholesalers price-improve a substantial portion, 66.10%, of order flow they receive, whereas exchanges only price-improve 9.00%. This said, exchanges fare better in their ability to match the NBBO, executing 98.35% of shares at the NBBO prices or better versus 93.12% by the wholesalers.[14] Institutional traders typically split large orders to avoid walking the book, 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]

Wholesaler executions tend to occur when the NBBO is relatively wide, 69.60 bps, vs. the exchange equivalent of 52.94 bps, a 31% difference. This difference cannot be attributed to wholesaler choices, as arrangements with retail brokerages do not allow wholesalers to choose which retail orders to execute and when. Instead, the disparity must be driven by trader decisions and is perhaps expected given that many institutional liquidity-taking algorithms time their activities to periods of narrow spreads, while retail traders are less likely to engage in such timing.

Even though retail executions occur when quoted spreads are relatively wide, the differential is reduced significantly once we switch to effective spreads, which account for price improvement. The wholesaler effective spreads are much closer to those reported by exchanges, at 53.16 bps and 51.31 bps, respectively, highlighting that price improvement is an important feature of execution quality. To account for this feature, researchers often use the ratio of effective to quoted spreads. 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. Wholesalers therefore appear

[14] As should perhaps be expected, wholesalers excel at NBBO matching and improvement for smaller orders. For instance, 99% of orders below $5,000 (approximately the 25th order size percentile) obtain NBBO or better execution, while this figure is only 86% of orders exceeding $64,000 (approximately the 75th percentile).

to play a valuable role within the existing market structure. They provide substantial, rather than *de minimis*, price improvement that may not be available from the exchanges.

Some market structure commentators argue that while price improvement provided by wholesalers is small, PFOF payments made by them to retail brokerages are large. Combining data from Rule 605 and 606 reports suggests that the opposite is true. Figure 1 shows that for every dollar of spread revenue, a typical wholesaler returns 24 cents to retail customers in price improvement while paying only 1 cent to the routing brokers. The remaining 75 cents are retained to cover the wholesaler's costs and potentially earn profits.

[Figure 1]

Next, we turn to the wholesaler costs and focus on the two components of the effective spread: price impact and realized spread. Table 2 confirms our earlier assertion that wholesalers obtain order flow that generates considerably less adverse selection, characterized by price impacts of 36.80 bps, compared to exchanges where price impacts are 52.73 bps. These figures align with the widely held view that retail traders tend to be less informed than institutional traders.

The table further shows that wholesalers collect larger realized spreads than exchanges, 16.36 vs. -1.42 bps. At first glance, this large difference may seem suggestive of excess profits earned by wholesalers. However, it is important not to over-interpret these figures. Liquidity on exchanges is only partly provided by professional market makers. For instance, Nasdaq attributes only 16% of liquidity provision to pure market making strategies.[15] The remaining liquidity-providing orders are submitted by non-market makers, whose primary objective is to manage investment positions rather than profit from liquidity provision. Consequently, while professional market makers on exchanges may charge realized spreads similar to those charged by wholesalers, the overall exchange realized spreads will remain low. Recognizing this fact, in subsequent analyses we refrain from making direct comparisons between wholesaler and exchange realized spreads and use

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

exchange statistics only for benchmarking.[16]

3.2 Cross-Sectional Differences

Because we are working with a large cross-section, we separately examine four sub-samples: the S&P 500 stocks and three size-based groups of non-S&P 500 stocks labeled terciles 1 through 3. During the sample period, there are 514 stocks in the S&P 500 sub-sample (the 14 stocks account for turnover within the index). Terciles 1, 2, and 3 include 2,550, 2,550, and 2,551 stocks. Table 3 shows noticeable differences across the sub-samples. Wholesalers represent 32.11% of share volume in S&P 500 stocks, and their share increases monotonically as firm size declines reaching a high of 63.94% in tercile 3 stocks. Thus, retail flow plays an out-sized role in less liquid stocks, a point we will return to shortly.

[Table 3]

Prior market structure literature has linked execution quality to several market characteristics. Among these are the stock price, trading volume, and volatility. A higher price is typically related to lower execution costs because of the fixed tick size. Greater volatility is typically associated with greater information flows and may negatively affect execution quality through the adverse selection channel. With volatility controlled for, a greater volume is typically associated with uninformed balanced order flow that benefits liquidity. In Table 4, we examine how execution quality differs between orders routed to wholesalers and exchanges while controlling for the above-mentioned characteristics:

$$DepVar_{ijt} = \alpha_i + \gamma_t + \beta_1 WHOL_j + \beta_2 price_{it} + \beta_3 volatility_{it} + \beta_4 volume_{it} + \varepsilon_{ijt}, \quad (1)$$

[16]The realized spreads reported in this section are not adjusted for PFOF and exchange liquidity rebates. Ingredients for this adjustment would come from Rule 606 data that are not available in 2019, the first year of our four-year sample period. Rule 606 data from 2020-2022 show that one would need to subtract 0.65 bps from the realized spreads charged by wholesalers to account for PFOF and add 0.70 bps to the realized spreads charged by exchange liquidity providers to account for rebates. When we redo our analyses using the three-year sample and adjusting for PFOF and rebates, our conclusions are unchanged.

where $DepVar_{it}$ is one of the following execution quality variables for stock i intermediary j in month t: the ratio of effective to quoted spread, quoted spread, effective spread, price impact, and realized spread, $WHOL$ 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. We use CRSP values for volatility and volume instead of Rule 605 values to capture market-wide activity. The models control for stock and month fixed effects and use double-clustered standard errors.

We estimate equation (1) for the full sample in Panel A of Table 4. We are primarily interested in the coefficient on the $WHOL$ dummy but note that the control coefficients are significant and of the expected signs. The univariate findings we discussed earlier continue to hold and the magnitudes change only slightly.

[Table 4]

We next augment the regression by interacting the $WHOL$ dummy with dummies for terciles 1, 2, or 3 in Panel B.[17] The coefficient on the $WHOL$ dummy captures the difference between the outcome variables for orders in S&P 500 stocks routed to wholesalers compared to exchanges. The interaction terms, e.g., $WHOL \times T3$, test whether the outcomes for orders routed to wholesalers are significantly different for tercile 3 stocks relative to S&P 500 stocks. To obtain the total difference in outcome variables between wholesalers and exchanges for T3 stocks, we add the coefficient on the $WHOL$ dummy to the coefficient on the $WHOL \times T3$ dummy.

In all sub-samples, the data confirm that wholesalers provide greater price improvement compared to exchanges. In S&P 500 stocks, the difference between exchange and wholesaler effective-to-quoted spread ratios is 0.44, a 44 percentage points larger price improvement. Wholesaler price improvement declines as we move from large to small firms. Still, even for tercile 3,

[17] The univariate version of this analysis is reported in Internet Appendix A.3.

price improvement is 19.5 (= -0.438 + 0.243) percentage points larger for wholesalers.

Finally, we confirm for all four sub-samples that toxicity of wholesaler-bound flow is lower than that of the exchange-bound flow, and that wholesalers earn larger realized spreads. For instance, price impacts facing wholesalers in S&P 500 (tercile 3) stocks are 2.73 (40.16) bps lower than those facing liquidity providers on exchanges, whereas realized spreads earned by wholesalers are 1.91 (43.81) bps greater than those earned by exchange liquidity providers. We note that although the realized spreads obtained by wholesalers appear quite large, particularly for tercile 3, they may be representative of substantial inventory and fixed costs incurred in these relatively infrequently traded stocks. We return to inventory costs in Section 3.7.

3.3 Differences Across Wholesalers

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 two*, which includes Citadel and Virtu, and the *others*. Recall that the top two are considerably larger than their peers and execute 71% of marketable order flow that is routed to wholesalers.

In Table 5, we use the following panel regressions to ask if execution quality systematically differs for the top two compared to the other wholesalers:

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

where $DepVar_{ijt}$ is one of the following execution quality metrics for stock i wholesaler j in month t: the ratio of effective to quoted spread, quoted, effective, and realized spread, and price impact, $top2$ is a dummy variable that has a value of 1 for orders executed by Citadel and Virtu and 0 for orders executed by the other wholesalers, and the control variables are defined as previ-

ously. The regressions control for stock and month fixed effects and use double-clustered standard errors. Note that we only use wholesaler data for these regressions.

[Table 5]

Panel A shows that the top two wholesalers provide two percentage points lower price improvement than their competitors. This statistic may appear concerning, especially considering that these two likely have the most market power. We however note that the top two face more toxic order flow (the difference is 2.68 bps) and charge lower realized spreads (the difference is -1.70 bps). The results are similar when we examine the differences across sub-samples in Panel B. The top two offer a 7 percentage point lower price improvement for S&P 500 stocks on average, but this does not suffice to compensate for the fact that they face significantly greater adverse selection. While the differences in price improvements shrink as we go from tercile 1 to 3, the differences in toxicity and realized spreads are magnified. Consider tercile 3 stocks, where price improvement is 0.3 percentage points smaller for the top two than for the others, yet toxicity they face is 6.9 bps greater, and realized spreads earned by them are 5.0 bps lower.

Considering that toxicity of order flow varies across wholesalers, Table 5 suggests that using price improvement as a comparison metric might not be the most appropriate for our analyses. If wholesaler 1 offers a slightly smaller price improvement than wholesaler 2 but receives considerably more toxic flow, comparing price improvements would be unfair. In contrast, the realized spread is a metric that accounts for both price improvement and toxicity. Therefore, in subsequent discussions, we primarily focus on realized spreads as a toxicity-adjusted liquidity cost metric.

We also note that individual brokerages assess wholesaler performance using metrics that implicitly control for order flow toxicity, as each brokerage's order flow carries its own unique toxicity levels. For example, if brokerage 1 generates order flow with higher toxicity than brokerage 2, it is likely to receive smaller price improvement from all wholesalers. Hence, when making comparisons across wholesalers, the brokerage implicitly evaluates toxicity-adjusted per-

formance metrics. This rationale becomes particularly crucial in subsequent analyses where we ask if a wholesaler can increase its market share based on prior performance.

3.4 Economies of Scale

We next ask if the differences in realized spreads between the top two and other wholesalers can be explained by the scale of a wholesaler's operation. The idea is that a wholesaler handling more retail flow can more efficiently internalize orders, leading to lower inventory costs as well as lower per-share fixed costs. To understand the inventory cost argument, it is important to mention how retail brokerages distribute orders among wholesalers. Most brokerages use a wholesaler wheel, which rotates to wholesaler 2 after a preset quantity has been routed to wholesaler 1. In a hypothetical scenario with two wholesalers assigned 60% and 40% of the flow, the brokerage would send 6 orders to the first wholesaler, 4 orders to the second wholesaler, and then revert to the first wholesaler for the next 6 orders. By design, the wheel system aims to achieve a random allocation of orders. Industry insiders suggest that the variation in toxicity among wholesalers results from the mix of brokerages and the types of brokerage accounts they serve, rather than the brokerages' flow allocation decisions.

Returning to inventory costs, let us assume that for every ten orders a retail brokerage receives, it sends four orders, or 40%, to Citadel. Given the typical balance of retail flow, with buys arriving as frequently as sells, the orders will tend to reconcile against each other, resulting in either zero or only a small inventory imbalance. Even in the case of a leftover imbalance, Citadel benefits from the shortest waiting time (only six orders) before the wheel rotates to it again. In contrast, a wholesaler with a 10% market share faces a waiting time of nine orders before having an opportunity to undo the imbalance created by the first order it receives. The ability to balance the flow quickly and the relatively short wait between rotations enjoyed by large wholesalers therefore seems essential for sustaining low inventory costs, fostering economies of scale.

To approximate the size of a wholesaler's operation, we use the retail volume processed by a wholesaler in a specific stock during a particular month. To ensure comparability across securities, we normalize the aforementioned retail volume by the total CRSP volume in that stock for that month. If economies of scale enable the top two wholesalers to generate liquidity at a lower cost, controlling for operation size should reduce or entirely eliminate the observed differential between them and the other wholesalers. To verify this, we run the following regression:

$$realized\ spread_{ijt} = \alpha_i + \gamma_t + \beta_1 top2_j + \beta_2 price_{it} + \beta_3 volatility_{it} + \beta_4 op.size_{ijt} + \varepsilon_{ijt}, \qquad (3)$$

where $realized\ spread_{ijt}$ is realized spread as defined previously for stock i wholesaler j in month t, $top2$, $price$, and $volatility$ are as previously defined, and a new control variable $op.size$ is the natural log of retail volume captured by a wholesaler in stock i scaled by the natural log of total CRSP trading volume in this stock. As a result, the $op.size$ variable absorbs the volume control used in the earlier models. As previously, 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.

Table 6 confirms our expectations, showing a significant association between a wholesaler's operation size and lower realized spreads across all sub-samples, except for S&P 500 stocks. Notably, when $op.size$ is included as a control, the coefficient on $top2$ becomes insignificant in all sub-samples. This finding aligns with the idea that the top two wholesalers face reduced inventory and fixed costs due to the scale of their operations, consequently offering lower liquidity costs.

[Table 6]

3.5 Broker Routing and Wholesaler Performance

Rule 606 data enable us to reconstruct order flow patterns from brokerages to wholesalers, but only for the brokerages that consistently accept payments for order flow. Internet Appendix A.6 explains how we calculate routed volumes. Figure 2 shows that the four largest retail brokerages – TD Ameritrade, Robinhood, E*Trade, and Schwab – route to at least six of our eight wholesalers during the sample period. Not surprisingly, the majority of order flow goes to Citadel and Virtu. Inspection of the time series suggests that there is some variation in routing patterns over time, but all four retail brokers route to at least three wholesalers at all times. It is worth noting that TD Ameritrade, the largest brokerage by flow, is considerably larger than the largest wholesaler, which casts some doubt on the notion that wholesalers are the sole wielders of market power.

[Figure 2]

How do retail brokers decide which wholesaler to route to and how much to allocate to a specific wholesaler? Industry participants suggest that retail brokerages regularly evaluate wholesaler performance.[18] Such evaluations typically occur on a monthly basis.[19] We suggest that in a competitive market for retail order flow, brokerages should adjust their routing to favor wholesalers with superior past performance. In other words, a wholesaler's market share should increase if the liquidity cost it charges to retail customers is lower than that of competing wholesalers.

Industry participants also emphasize that retail brokerages aim to achieve the highest execution quality for every individual customer order. As we showed in Table 3, retail customers are noticeably active in small stocks, where liquidity is naturally limited. To meet their best execution obligations and objectives, brokerages mandate that wholesalers provide low liquidity costs across the entire range of securities, rather than focusing on specific securities that wholesalers

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

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

may prefer to intermediate.

To investigate whether brokerages actively monitor and reward high-performing wholesalers, and whether they require that wholesalers deliver superior execution quality across all stocks, we use the econometric framework for order routing proposed by Boehmer, Jennings, and Wei (2007). The framework was developed for data similar to ours and uses a combination of geometric and arithmetic means to allow predicted wholesaler market shares to lie between zero and one and to allow the sum of market shares across wholesalers to equal one. Specifically, we estimate the following regression:

$$\begin{aligned} \textit{market share}_{ijt} = \alpha_i + \theta_j + \gamma_t + \beta_1 \textit{stock realiz. spr.}_{ijt-1} + \beta_2 \textit{portfolio realiz. spr.}_{jt-1} \\ + \beta_3 \textit{price}_{it} + \beta_4 \textit{volatility}_{it} + \beta_5 \textit{volume}_{it} + \varepsilon_{ijt}, \end{aligned} \tag{4}$$

where $\textit{market share}_{ijt}$ is the share of retail volume in stock i executed by wholesaler j in month t expressed as the deviation from the geometric mean across all wholesalers, $\textit{stock realiz. spr.}_{ijt-1}$ is the average realized spread charged by wholesaler j in stock i in month *t-1* expressed as the deviation from the arithmetic mean across all other wholesalers, $\textit{portfolio realiz. spr.}_{jt-1}$ is the average realized spread charged by wholesaler j in all stocks routed to it in month *t-1* expressed as a deviation from the arithmetic mean across all other wholesalers, and *price*, *volatility*, and *volume* are as previously defined. The realized spread variables are scaled so the economic significance corresponds to basis points. We run these regressions for the full sample and then for each sub-sample using stock, wholesaler, and month fixed effects and clustered standard errors.

Table 7 shows that if a wholesaler charges a relatively low realized spread across all stocks, retail brokerages respond by granting the wholesaler a larger market share in the following month. Conversely, wholesalers charging relatively high spreads face a reduction in their allocations. This result holds for the full sample and for all sub-samples. A one basis point reduction in a wholesaler's realized spread relative to the average across wholesalers is associated with a 3.1%

greater market share for the full sample and between 2.8 and 3.2% greater market shares for the sub-samples.

[Table 7]

A relatively low realized spread charged by a wholesaler in one particular stock is also associated with a significantly greater future market share. This result is evident across all stocks except the S&P 500 constituents. However, the economic magnitude of this effect is negligible. When considered together, the findings in Table 7 are consistent with wholesalers competing by offering lower liquidity costs across all stocks rather than on a security-by-security basis, and with wholesalers with superior performance receiving more order flow.

Two new working papers also examine broker monitoring and wholesaler incentives. Huang, Jorion, Lee, and Schwarz (2023) use a sample of self-generated small trades and find no evidence that retail brokers reallocate order flow among wholesalers based on execution quality in these trades. Ernst, Malenko, Spatt, and Sun (2023) use a proprietary dataset containing all retail transactions from three large brokerages that together handle close to 50% of all retail flow. They find results similar to ours in that brokerages monitor wholesalers and reallocate future flows based on past performance. Taken together, these new studies complement each other and our findings. It appears that brokerages do not base their reallocation decisions on execution quality in specific trade size groups (small odd-lots in the case of Huang, Jorion, Lee, and Schwarz (2023)), but rather value execution quality across all trade sizes.

3.6 A Competitive Shock

The dynamics of competition in the retail segment change during our sample period due to the entry of a new wholesaler, Jane Street. If wholesalers wielded market power prior to this entry, we expect that competitive forces would intensify, placing greater pressure on wholesalers to provide lower trading costs post-entry, and realized spreads would decrease.

Panel A of Figure 3 illustrates Jane Street's entry and market share growth over time. The firm enters into the wholesale business in the middle of 2019, but throughout 2020 it still has a very small market share. Its market share begins to increase more rapidly in the late summer of 2021, reaching a substantial level by October 2021.[20] By the end of 2021, all brokerages in our sample route to Jane Street, and by the end of our sample period (the end of 2022), it has a market share in the 12-14% range. All incumbent wholesalers, large and small, experience a market share loss of 11% or greater to Jane Street.

[Figure 3]

In the same panel, we also plot the ratio of Jane Street's realized spreads to the incumbent wholesaler's realized spreads. A month after entering, Jane Street starts charging realized spreads similar to those of its competitors, but from July 2021 it competes considerably more aggressively, charging realized spreads that are about 50% lower than the incumbents. The firm's market share stabilizes in 2022, yet Figure 3 suggests that it continues offering relatively low realized spreads. Recall that our spread statistics are volume-weighted up to the cross-section and then equal-weighted across stocks, allowing for a clearer picture of the entire securities landscape. An examination of the cross-section in a later table reveals that Jane Street offers low spreads only in tercile 3 stocks. In the remaining stocks its realized spreads are similar to those of the incumbents. Because most retail volume occurs in large stocks, Jane Street's revenues are likely comparable to those of its competitors. It appears that Jane Street is contractually committed to providing highly competitive executions in small stocks for an extended time period.

To understand if the entry of Jane Street leads to lower average realized spreads charged by wholesalers, we run a difference-in-differences regression of wholesalers against exchanges with the pre-period being April-June 2021, when Jane Street has a small market share, and the

[20] Rule 606 data suggest that upon entry, Jane Street first contracts with smaller brokerages such as APEX, Tradestation, and Webull. In 2021, having established itself as a reliable wholesaler, it wins contracts with large brokerages such as E*TRADE, Robinhood, Schwab, and TD Ameritrade, significantly expanding its market share.

post-period being the last three months of 2021, by which time Jane Street has become a sizeable wholesaler. Note that we use exchange spreads solely to control for potential market-wide confounding events, and not for making direct comparisons between two platform types. Table 8 reports the results from the following regression:

$$\begin{aligned} \textit{realized spread}_{ijt} = \alpha_i + \gamma_t + \beta_1 WHOL_j + \beta_2 WHOL \times POST_{jt} + \beta_3 price_{it} + \beta_4 volume_{it} \\ + \beta_5 volatility_{it} + \varepsilon_{ijt}, \end{aligned} \tag{5}$$

where $\textit{realized spread}_{ijt}$ is the realized spread in stock i for intermediary j (wholesaler or exchange) in month t, $WHOL$ is a dummy variable that has a value of 1 for orders executed by wholesalers and 0 for orders executed by exchanges, $POST$ is a dummy variable that has a value of 1 after Jane Street market share capture and 0 otherwise, and $price$, $volatility$, and $volume$ are as previously defined. 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 8]

Panel A of Table 8 contains no indication of wholesaler realized spreads decreasing following Jane Street's entry in any sub-sample. For S&P 500 stocks, the entry does not lead to changes in realized spreads, while for tercile 1-3 stocks, realized spreads actually increase. Panels B and C report the results for the incumbents and Jane Street separately. Recall that Jane Street was present but at a much lower market share in the pre-period. The incumbents do not change their realized spreads for S&P 500 stocks but increase spreads for tercile 1-3 stocks after Jane Street's entry. In contrast, as we mention earlier, Jane Street keeps the S&P 500 and tercile 1 and 2 spreads unchanged while reducing the spreads for tercile 3 stocks.

To understand the result for the incumbents, let us consider what happens when Jane Street enters. Order flow is now divided among a larger number of wholesalers resulting in less flow for

each incumbent. For instance, Citadel and Virtu each lose more than 11% of their flow to Jane Street, while smaller wholesalers generally lose even more. With less flow, the incumbents likely face increased inventory costs, especially in less liquid securities in terciles 1-3. Meanwhile, the S&P 500 securities are highly liquid, making it easier for the incumbents to manage inventories and reducing the need to increase liquidity costs.

During our sample period, one wholesaler also exits. Panel B of Figure 3 shows how Wolverine, a wholesaler that presently operates only in options, leaves the equity business. Rule 606 data suggest that in 2019, three retail brokers – Ally, Robinhood, and Tastyworks – are routing to Wolverine. In the meantime, Wolverine charges realized spreads that are 42.5% greater than those charged by the incumbents. Perhaps not surprisingly, Ally drops Wolverine by April 2020, followed by Tastyworks in the following month. Robinhood gradually reduces its routing over the following year, stopping entirely by April 2021, causing Wolverine to exit the equity wholesale business. Importantly, as Wolverine's market share declines, its already high realized spreads increase dramatically, consistent with the loss of economies of scale. Rule 605 data show that the increase in realized spreads cannot be explained by a change in the toxicity of flow received by Wolverine, leading us to conclude that the importance of economies of scale in the wholesale business is difficult to overstate.

Classic economic theory offers various predictions regarding the competitive effects of entry. Contestable market theories, for instance, argue that a mere threat of entry can restrain market power. In contrast, entry barrier models assign a smaller role to threats, arguing that only actual entry has an impact on competition (see Bresnahan and Reiss (1991) for a review). In our setting, even actual entry does not appear to positively affect the competitive outcome, suggesting that the absence of market power abuse may have been inherent from the beginning.

In this context, it is also useful to consider the possibility of implicit collusion among wholesalers. Colliard, Foucault, and Lovo (2022) conduct experiments in which algorithms learn to make the market. Under certain conditions, without human encouragement, the algorithms settle

on non-competitive prices, essentially overcharging for liquidity. Could wholesalers be engaging in similar behavior? While we cannot rule out this possibility entirely, two factors lead us to believe that it is relatively unlikely. First, Colliard, Foucault, and Lovo (2022) observe collusive behavior primarily when only two algorithms are present. When they increase the number of competing algorithms beyond two, overcharging is reduced and eventually eliminated bringing the market to a Bertrand-Nash competitive state. This result echoes Brogaard and Garriott (2019), who report that the competitiveness of liquidity provision on a new exchange increases as the number of liquidity providers goes beyond two and reaches an equilibrium level after four competitors enter. Second, the monitoring by retail brokerages is likely to reduce the ability of wholesalers to charge non-competitive prices.

In a new working paper, Huang, Jorion, Lee, and Schwarz (2023) examine execution quality in self-generated small trades after Robinhood begins routing these trades to Jane Street. In a non-difference-in-differences setting, they find evidence of execution quality improvements. Their first trades appear to reach Jane Street in late February 2022, whereas Rule 606 data show that Robinhood starts routing to Jane Street in December 2021. We view the results of these two event studies as complementary. While Huang, Jorion, Lee, and Schwarz (2023) examine the effects of Jane Street entry into competition for small trades submitted through Robinhood, our event study mainly focuses on Jane Street's capture of a sizeable market share resulting from brokerage partnerships that occur several months earlier, in the second half of 2021, and examines execution quality across the full spectrum of order sizes.

3.7 Inventory and Fixed Costs

Our earlier discussions of the cross-section hinge on the concept of inventory and fixed costs to explain why realized spreads in smaller stocks are higher than those in larger stocks. Although measuring these costs is notoriously challenging, we propose using processed retail volume as a

proxy for a wholesaler's ability to manage them. We posit that, when controlling for volatility, a stock-month with lower volume may be associated with greater inventory costs, as outstanding inventory positions take longer to offload. Returning to our previous example with Citadel waiting for six orders for the wholesaler wheel to cycle back to it, the wheel will turn quicker when trading volume is high and slower when it is low, directly impacting Citadel's inventory holding costs. Such a stock-month would also be associated with greater fixed costs per share. To delve deeper into this matter, we run the following regression:

$$\begin{aligned} \textit{realized spread}_{it} = \alpha_i + \gamma_t + \beta_1 T1_i + \beta_2 T2_i + \beta_3 T3_i + \beta_4 \textit{price}_{it} + \beta_5 \textit{volatility}_{it} \\ + \beta_6 \textit{retail volume}_{it} + \varepsilon_{it}, \end{aligned} \tag{6}$$

where $\textit{realized spread}_{it}$ is the realized spread in stock i in month t, $T1$, $T2$, and $T3$ are dummies indicating whether a stock is in tercile 1, 2, or 3, with the intercept capturing S&P 500 stocks, *price* and *volatility* are as previously defined, and *retail volume* is the natural log of total retail volume across all wholesalers as reported in Rule 605 data. We also estimate a specification that uses *total volume*, defined as the natural log of CRSP trading volume, replacing retail volume. This specification allows for the possibility that the vertically integrated wholesalers use non-retail flow to manage retail inventory positions. Since the tercile dummies are unique for each security, the regression controls only for month fixed effects, yet uses double-clustered standard errors. Note that we only use wholesaler data for these regressions.

Column [1] of Table 9 reports the base specification without controls, confirming that tercile 1, 2, and 3 stocks have significantly greater realized spreads than S&P 500 stocks. Introducing controls for price and volatility in column [2], eliminates the difference between S&P 500 stocks and tercile 1 stocks, but realized spreads for the remaining size terciles remain significantly higher. In column [3], we include retail volume as a proxy for wholesalers' ability to manage inventory and cover fixed costs. Now, realized spreads for tercile 1 and tercile 2 stocks

are significantly lower than for S&P 500 stocks, while spreads for tercile 3 stocks remain higher, albeit at a considerably lower magnitude than in the base specification. Finally, column [4] considers the possibility that wholesalers may use non-retail volume to manage inventory. In this specification, all coefficients on the size terciles are negative and significant.

[Table 9]

Taken together, these results suggest not only that inventory and fixed cost considerations are important components of trading costs, but also that, *ceteris paribus* these considerations, wholesalers charge less in less liquid stocks than S&P 500 stocks. The latter result dovetails with our earlier finding that retail brokers demand superior execution quality for the entire portfolio of securities being routed to wholesalers. Additionally, it highlights a new observation – that the portfolio approach to wholesaler evaluation leads to a cross-subsidy of small-stock liquidity at the expense of the largest stocks. Whether such a cross-subsidy is optimal from the social welfare standpoint is largely beyond the scope of this study; however, in the next section we relate it to the outstanding congressional mandate to improve liquidity in small stocks.

4. Market Structure Proposals

The results presented thus far provide no evidence that wholesaler intermediation harms retail investors. Could these investors be even better off with a different market structure? In this section, we discuss two possible alternative setups. First, we use our data to infer what would happen if retail orders were routed to exchanges. Needless to say, this is a counterfactual, and we must make assumptions that the reader may or may not agree with. Second, we discuss some pitfalls with the current SEC market structure proposal that aims to create an auction process for retail orders. Our data do not enable us to delve into the proposal in detail. Still, the cross-sectional evidence reported earlier, along with the institutional trading interest we will discuss shortly, implies

that the Commission may need to exercise caution when considering this proposal.

4.1 Moving Retail Flow to Exchanges

The potential move of retail flow to exchanges is a recurring topic in U.S. market structure discussions. In response to the SEC's request for comments on the proposed reform of retail order flow handling, retail investors submitted over 2,600 comment letters advocating for the relocation of their flow to exchanges.[21] To understand the potential impact of such a relocation, we examine its two likely effects. On the positive side, and based on the statistics reported in Table 2, retail flow would be charged lower realized spreads. On the negative side, when pooled with institutional flow, retail investors would bear the cost of the resulting mix's higher toxicity, which would exceed that of pure retail flow. Using the data at our disposal and a set of assumptions grounded in our understanding of the market, in what follows we deduce that the combined impact of these two effects would likely be detrimental for retail investors.

To explain our reasoning, we first focus on the amount of toxicity that retail investors would pay for if pooled with institutional flow. Given that exchange flow is anonymous, liquidity providers cannot differentiate between low- and high-toxicity customers. Therefore, all demanders of exchange liquidity would incur the same toxicity cost, which would be the weighted average of retail and institutional toxicities:

$$t_{pool} = w_{ret} \times t_{ret} + w_{ins} \times t_{ins}, \tag{7}$$

where t_{pool} is toxicity resulting from pooling retail and institutional flows, w_{ret} (w_{ins}) is the share of retail (institutional) volume in exchange volume, and t_{ret} (t_{ins}) is retail (institutional) toxicity.

Of the four variables required to calculate t_{pool}, three are relatively straightforward to obtain.

[21] "Main Street investors pressure SEC, confront Wall Street on stock plan," by J. McCrank, Reuters, March 10, 2023 (https://bit.ly/41n25sK) and "Main Street investors plan to keep pressure on Gensler as Citadel Securities fights auction reform," by C. Matthews, MarketWatch, April 10, 2023 (https://on.mktw.net/3zVGtIg).

First, Rule 605 data provide us with t_{ret}. Second, the same data contain retail volume, easily converted into w_{ret} by dividing it by the estimate of pooled exchange volume. To estimate the pooled exchange volume, we use 0.7 $\times$ CRSP volume, with the excluded 30% accounting for dark institutional volume that executes off-exchange (20%) and volume executing during opening and closing auctions and outside regular exchange hours (10%).[22] Third, w_{ins} is simply $1 - w_{ret}$.

The fourth variable, t_{ins}, necessitates further clarification. We note that t_{ins} is not simply the institutional price impact from Rule 605 data, because these data are limited to simple unconditional orders. Meanwhile, orders that are ineligible for Rule 605 reporting are more sophisticated conditional orders, likely associated with greater toxicity.[23] Consequently, we anticipate t_{ins} to be greater than the toxicity derived from Rule 605 data.

Keeping this in mind, t_{ins} can be deduced from the Rule 605 quoted spreads encountered by retail investors. In this derivation, we assume that retail traders do not time liquidity, and the quoted spreads they encounter represent the typical on-exchange spread. This quoted spread is, in turn, the sum of the exchange realized spread, price impact, and price improvement. Since we have data on exchange realized spreads and price improvement, obtaining the institutional price impact, t_{ins}, involves subtracting these two variables from the quoted spreads. Thus, $t_{ins} = qs_{ret} - rs_{ex} - primpr_{ex}$, where qs_{ret} is the quoted spread faced by retail traders, rs_{ex} is the exchange realized spread, and $primpr_{ex}$ is the exchange price improvement.

Combining the four components of equation 7 produces the expected toxicity cost t_{pool} that retail investors would face if pooled with institutional investors. To finalize the calculation of the resulting total liquidity costs, we assume that the current exchange realized spread will remain

[22] According to industry estimates, approximately 60% of U.S. volume executes on exchanges, while 40% executes off-exchange. Approximately one half of off-exchange volume is retail volume, and the other half is dark institutional volume (see "An Update on Retail Market Share in U.S. Equities," by Rosenblatt Securities (https://bit.ly/4a2MhAS)). In addition, closing auction volume is around 7.5% of total volume, while opening auction and after-hours volumes approximate to another 2.5% (see Bogousslavsky and Muravyev (2023) and "Auction Volume Report: Americas," by Tethys Technology, Inc. (https://bit.ly/47Pv9MO)).

[23] Li, Ye, and Zheng (2023) report that high-frequency traders (HFTs) are the primary users of conditional orders. Meanwhile, prior research indicates that HFTs are typically among the most informed market participants (e.g., Brogaard, Hendershott, and Riordan (2014), Brogaard, Hendershott, and Riordan (2019)).

unchanged. We will explore the relaxation of this assumption shortly. Consequently, the total liquidity cost incurred by retail investors upon pooling will be the sum of the pooled toxicity cost and the current exchange realized spreads. We compute this cost starting at the stock-month level and then aggregate to the final average using share volume weights. This aggregation approach differs somewhat from that used in the earlier sections, focusing on understanding the outcome for an average share traded rather than the average stock. We believe this shift in focus offers better insights into the wealth implications of the move for retail investors.

Our calculations indicate that retail investors would experience a 30.0% increase in effective spreads, resulting in an additional monthly trading cost of $157 million if their orders were routed to exchanges. On the flip side, institutional effective spreads would decrease by 12.3%, as their toxicity costs in the pooled setting would be lower than in the current status quo on exchanges. These results point to a challenging regulatory choice. On one hand, moving retail investors to exchanges may harm them. On the other hand, it would benefit institutional investors, many of whom oversee the savings of numerous individual investors. Ultimately, the decision of which group to prioritize depends on the regulator's objective function. In this study, our aim is to inform regulatory decision-making rather than to prescribe a definitive solution.

One of the assumptions that goes into these calculations is that realized spreads would not change if retail volume moved to the exchanges. The literature provides mixed guidance on the validity of this assumption. On the one hand, Comerton-Forde, Malinova, and Park (2018) demonstrate that when some retail flow shifts from an off-exchange facility to an exchange in Canada, quoted and effective spreads on the exchange do not change (the authors do not report realized spreads). On the other hand, Bessembinder, Carrion, Tuttle, and Venkataraman (2016) find that the arrival of uninformed volume is accompanied by additional liquidity coming off the sidelines and improving market quality. In our setting, even if realized spreads were to decline, this change would need to be rather large (a reduction of about 210%) to outweigh the additional costs retail traders face because of having to pay for higher toxicity.

Finally, our calculations omit one additional possibility, as institutions may strategically respond to being pooled with retail flow in the sense of Kyle (1985) by increasing information-based trading. The resulting increase in adverse selection would lead to an even greater increase in retail liquidity costs.

4.2 Order-by-Order Competition

In December 2022, the SEC proposed the Order Competition Rule, asserting that retail traders would benefit from order-by-order competition. The rule suggests that retail orders be directed to open-participation auctions, enabling institutions to interact with retail flow.[24] The SEC contends that this approach could result in greater price improvement for retail flow compared to the current offerings by wholesalers.

How feasible is this proposal, and what impact would it have on the cross-section of stocks? The answer hinges on whether there is institutional interest in trading the stocks favored by retail investors. In Table 3, we illustrate that retail investors are presently responsible for the majority of volume in less liquid stocks. This finding is however based on Rule 605 data and does not account for institutional trades originating from complex orders. To explore this issue further, we estimate institutional trading volume using changes in quarterly holdings from 13F reports. To account for intra-quarter trading, we adjust the inferred volume by a factor of 1.17, as suggested by Chakrabarty, Moulton, and Trzcinka (2017). As institutional short selling is not reflected in 13F reports, we add changes in short interest to 13F-based estimates. Short interest changes are obtained from bi-weekly Compustat estimates aggregated to quarterly values to match 13F data. Subsequently, we calculate the ratio of retail volume to the institutional volume proxy.

Column [1] ([2]) of Table 10 shows that the mean (median) retail volume represents 85% (20%) of our proxy for institutional volume in S&P 500 stocks, so for these stocks there is sig-

[24] For other analyses of the proposal, see Battalio and Jennings (2023b), Ernst, Spatt, and Sun (2023), and van Kervel and Yueshen (2023).

nificant institutional trading interest. As we move to the less liquid stocks, retail volume swamps institutional volume. Already for tercile 1 stocks, institutional activity is insufficient on average as the ratio of retail to institutional volume exceeds one. For tercile 2, mean retail volume is more than double the institutional volume, and for tercile 3, mean retail volume is almost ten times larger than institutional volume. The ratios are highly skewed, suggesting that retail investors focus on particular stocks and that institutions do not always favor these stocks. When we switch our attention to the median values for a more conservative view, we observe that institutional interest is substantially below retail interest only in tercile 3 stocks.

[Table 10]

In addition to institutional volume potentially being an inadequate liquidity source in small stocks, results in Section 3.7 suggest that wholesalers cross-subsidize liquidity in such stocks, thanks to brokers using a portfolio approach to evaluate wholesaler performance. The proposed auctions would bring an end to the portfolio system, as wholesalers would lose exclusive access to retail flow. This shift would eliminate the cross-subsidy, leading to increased liquidity costs in small stocks. Considering Congress's concern about the lack of liquidity and the resulting challenges in raising capital for emerging growth firms, as emphasized in its 2012 Jumpstart Our Business Startups (JOBS) Act, the SEC's decision on whether to proceed with the auctions appears challenging.

5. Conclusion

The U.S. retail trading volume, which constitutes nearly 20% of total volume, is primarily executed off-exchange by intermediaries known as wholesalers. This practice has sparked a debate, mainly due to the apparent concentration of market power in the wholesale environment. Critics argue that wholesalers hold excessive influence and offer limited benefits, prompting the

SEC to contemplate introducing new rules to encourage additional competition for retail executions. Conversely, wholesalers contend that retail brokers choose to execute through them in the best interest of their clients.

Our data tend to support the latter claim, suggesting that it is the retail brokerages who have the power in this ecosystem. The brokerages are large, with the largest brokerage surpassing the largest wholesaler. They closely monitor wholesaler performance, rewarding the best performers with more order flow and reducing allocations to the underperformers. Furthermore, the wholesale environment is characterized by economies of scale. The largest wholesalers appear to generate liquidity at a lower cost, and brokerage oversight ensures that the cost savings transfer to retail customers. The wholesale market is also contestable, as evidenced by a new wholesaler entering and gaining a substantial market share during our sample period. Upon the arrival of this wholesaler, retail customer trading costs do not decrease, contradicting the idea that incumbents were exploiting market power before the entry.

We also discuss two alternatives to the status quo. One such alternative is to route retail flow to exchanges mixing it with institutional flow. While doing so may reduce realized spreads faced by retail investors, it would force them to pay for higher adverse selection costs, ultimately harming their overall welfare. The other alternative is the SEC's outstanding proposal for order-by-order auctions. Our results suggest two possible issues with such auctions. First, institutional investors are unlikely to participate in auctions for thousands of small stocks, as institutional interest in such stocks is lower than retail interest. Second, the status quo involves bundling, where retail brokerages compel wholesalers to price improve all stocks traded by retail investors. This practice causes wholesalers to undercharge for liquidity in small stocks. Eliminating bundling through auctions is likely to result in an increase in retail trading costs for small stocks.

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. 5

Baldauf, M., J. Mollner, and B. Z. Yueshen, 2023, "Siphoned Apart: A Portfolio Perspective on Order Flow Segmentation," *Working paper*, University of British Columbia. 5

Barber, B. M., X. Huang, P. Jorion, T. Odean, and C. Schwarz, 2023, "A (Sub) penny For Your Thoughts: Tracking Retail Investor Activity in TAQ," *Journal of Finance*, forthcoming. 5

Bartlett, R. P., J. McCrary, and M. O'Hara, 2023, "The Market Inside the Market: Odd-Lot Quotes," *Review of Financial Studies*, forthcoming. 9

Battalio, R., R. Jennings, M. Saglam, and J. Wu, 2023, "Identifying Market Maker Trades as 'Retail' from TAQ: No Shortage of False Negatives and False Positives," *Working paper*, University of Notre Dame. 5

Battalio, R. H., and R. H. Jennings, 2023a, "Absolute and Relative Wholesaler Execution Quality in May 2022," *Working paper*, University of Notre Dame. 5, 6, 8

——— , 2023b, "On the Potential Cost of Mandating Qualified Auctions for Marketable Retail Orders," *Working paper*, University of Notre Dame. 31

Bessembinder, H., 2003, "Selection Biases and Cross-market Trading Cost Comparisons," *Working paper*, University of Utah. 6

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. 30

Boehmer, E., 2005, "Dimensions of Execution Quality: Recent Evidence for US Equity Markets," *Journal of Financial Economics*, 78(3), 553–582. 6

Boehmer, E., R. Jennings, and L. Wei, 2007, "Public disclosure and private decisions: Equity market execution quality and order routing," *Review of Financial Studies*, 20(2), 315–358. 6, 20

Boehmer, E., C. M. Jones, X. Zhang, and X. Zhang, 2021, "Tracking Retail Investor Activity," *Journal of Finance*, 76(5), 2249–2305. 5

Bogousslavsky, V., and D. Muravyev, 2023, "Who Trades at the Close? Implications for Price Discovery and Liquidity," *Journal of Financial Markets*, p. 100852. 29

Bresnahan, T. F., and P. C. Reiss, 1991, "Entry and Competition in Concentrated Markets," *Journal of Political Economy*, 99(5), 977–1009. 24

Brogaard, J., and C. Garriott, 2019, "High-Frequency Trading Competition," *Journal of Financial and Quantitative Analysis*, 54(4), 1469–1497. 25

Brogaard, J., T. Hendershott, and R. Riordan, 2014, "High-Frequency Trading and Price Discovery," *Review of Financial Studies*, 27(8), 2267–2306. 29

——— , 2019, "Price Discovery Without Trading: Evidence from Limit Orders," *Journal of Finance*, 74(4), 1621–1658. 29

Bryzgalova, S., A. Pavlova, and T. Sikorskaya, 2023, "Retail Trading in Options and the Rise of the Big Three Wholesalers," *Journal of Finance*, forthcoming. 7

Chakrabarty, B., P. C. Moulton, and C. Trzcinka, 2017, "The Performance of Short-Term Institutional Trades," *Journal of Financial and Quantitative Analysis*, 52(4), 1403–1428. 31, 47

Colliard, J.-E., T. Foucault, and S. Lovo, 2022, "Algorithmic Pricing and Liquidity in Securities Markets," *Working paper*, HEC Paris. 24, 25

Comerton-Forde, C., K. Malinova, and A. Park, 2018, "Regulating Dark Trading: Order Flow Segmentation and Market Quality," *Journal of Financial Economics*, 130(2), 347 – 366. 30

Eaton, G. W., T. C. Green, B. S. Roseman, and Y. Wu, 2022, "Retail Trader Sophistication and Stock Market Quality: Evidence from Brokerage Outages," *Journal of Financial Economics*, 146(2), 502–528. 5, 6

Ernst, T., A. Malenko, C. Spatt, and J. Sun, 2023, "What Does Best Execution Look Like?," *Working paper*, University of Maryland (available as part of the authors' letter to the SEC at: https://tinyurl.com/bdhrm5k8). 6, 21

Ernst, T., and C. S. Spatt, 2022, "Payment for Order Flow and Asset Choice," *Working paper*, University of Maryland. 6

Ernst, T., C. S. Spatt, and J. Sun, 2023, "Would Order-by-Order Auctions Be Competitive?," *Working paper*, University of Maryland. 5, 31

Hendershott, T., S. Khan, and R. Riordan, 2023, "Option Auctions," *Working paper*, University of California at Berkeley. 7

Hu, E., and D. Murphy, 2023, "Competition for Retail Order Flow and Market Quality," *Working paper*, New York University. 5

Huang, X., P. Jorion, J. Lee, and C. Schwarz, 2023, "Who Is Minding the Store? Order Routing and Competition in Retail Trade Execution," *Working paper*, Washington University in St. Louis. 6, 21, 25

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. 5

Kothari, S., E. So, and T. Johnson, 2021, "Commission Savings and Execution Quality for Retail Trades," *Working paper*, Massachusetts Institute of Technology. 5, 6

Kyle, A., 1985, "Continuous Auctions and Insider Trading," *Econometrica*, 53, 1315–1336. 31

Li, S., M. Ye, and M. Zheng, 2023, "Refusing the Best Price?," *Journal of Financial Economics*, 147(2), 317–337. 8, 29

Lipson, M., 2004, "Competition Among Market Centers," *Working paper*, University of Virginia. 6

Schwarz, C., B. M. Barber, X. Huang, P. Jorion, and T. Odean, 2023, "The'Actual Retail Price'of Equity Trades," *Working paper*, University of California, Irvine. 6, 8

van Kervel, V., and B. Z. Yueshen, 2023, "Anticompetitive Price Referencing," *Working paper*, Pontificia Universidad Católica de Chile. 5, 31

Table 1
Market Shares

The table contains the list of 22 trading venues that execute Rule 605 liquidity-demanding orders during the sample period (2019-2022). 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	2,148	59.50
	WHOL	1,462	40.50
Panel B: by venue			
NASDAQ	EXCH	623.18	17.26
Citadel	WHOL	578.75	16.03
Virtu	WHOL	437.94	12.13
NYSE	EXCH	396.95	11.00
NYSE ARCA	EXCH	279.78	7.75
EDGX	EXCH	247.28	6.85
BATS	EXCH	213.54	5.92
G1	WHOL	190.39	5.27
BYXX	EXCH	83.78	2.32
Jane Street	WHOL	83.16	2.30
EDGA	EXCH	75.57	2.09
Two Sigma	WHOL	74.07	2.05
IEX	EXCH	70.44	1.95
UBS	WHOL	58.59	1.62
NYSE NAT	EXCH	45.51	1.26
NSDQ BOS	EXCH	32.81	0.91
MEMX	EXCH	30.42	0.84
Merrill Lynch	WHOL	26.93	0.75
NSDQ PHIL	EXCH	26.38	0.73
NYSE AMER	EXCH	19.63	0.54
Morgan Stanley	WHOL	12.56	0.35
NYSE CHI	EXCH	2.23	0.06
Total		3,609.89	100.00

Table 2
Execution Quality

The table contains execution quality statistics for Rule 605 liquidity-demanding orders. We compute the statistics separately for orders executed by wholesalers (WHOL) and exchanges (EXCH). We report the average stock price in a sample stock during the sample period, followed by the percentage of shares that are price improved or executed at or better than the corresponding NBBO. We report the quoted and effective spreads in basis points and compute the ratio of the effective to the quoted spread. Finally, we compute the components of the effective spread: price impact and realized spread, also in basis points. All variables are share-volume-weighted up to the cross-section and then equal-weighted across stocks. Asterisks *** and * in column [3] indicate statistical significance of differences between columns [1] and [2] at the 1% and 10% levels.

	WHOL	EXCH	diff. [1]-[2]
	[1]	[2]	[3]
price, $	30.04	30.59	
improved, %	66.10	9.00	***
at or better, %	93.12	98.35	***
quoted spread, bps	69.60	52.94	***
effective spread, bps	53.16	51.31	*
effective / quoted	0.76	0.97	***
price impact, bps	36.80	52.73	***
realized spread, bps	16.36	-1.42	***

Table 3
Market Shares: Sub-samples

The table reports market shares in Rule 605 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.

	S&P 500	Tercile 1	Tercile 2	Tercile 3
	[1]	[2]	[3]	[4]
WHOL	32.11	33.47	51.03	63.94
EXCH	67.89	66.53	48.97	36.06
No. Stocks	514	2,550	2,550	2,551

Table 4
Execution Quality: Regression

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

$$DepVar_{it} = \alpha_i + \gamma_t + \beta_1 WHOL_j + \beta_2 price_{it} + \beta_3 volatility_{it} + \beta_4 volume_{it} + \varepsilon_{ijt},$$

where $DepVar_{it}$ is one of the following execution quality variables for stock i intermediary j in month t: the ratio of effective to quoted spread, quoted spread, effective spread, price impact, and realized spread, $WHOL$ 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. We use CRSP values for volatility and volume instead of Rule 605 values to capture market-wide activity. Panel B augments the original 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	-0.278***	15.543***	1.807***	-16.901***	18.706***
	(0.01)	(0.42)	(0.40)	(0.71)	(0.97)
price	-0.018***	-23.319***	-22.804***	-17.921***	-4.879***
	(0.00)	(1.22)	(1.14)	(0.89)	(0.58)
volatility	0.000***	0.200***	0.180***	0.163***	0.017*
	(0.00)	(0.02)	(0.02)	(0.02)	(0.01)
volume	-0.002***	-29.634***	-26.626***	-16.280***	-10.352***
	(0.00)	(1.18)	(1.18)	(0.96)	(0.61)
intercept	1.036***	440.852***	404.174***	276.092***	128.142***
	(0.01)	(15.01)	(14.85)	(12.16)	(7.56)
Adj. R^2	0.658	0.756	0.739	0.538	0.199
Panel B: Specification with Interaction Terms					
WHOL	-0.438***	2.068***	-0.819***	-2.731***	1.911***
	(0.01)	(0.15)	(0.07)	(0.18)	(0.18)
WHOL×*T1*	0.120***	5.831***	1.432***	-4.038***	5.468***
	(0.01)	(0.22)	(0.15)	(0.27)	(0.32)
WHOL×*T2*	0.194***	15.628***	3.546***	-12.662***	16.207***
	(0.01)	(0.49)	(0.45)	(0.59)	(0.80)
WHOL×*T3*	0.243***	28.326***	4.479***	-37.425***	41.897***
	(0.01)	(0.90)	(1.04)	(1.59)	(2.42)
price	-0.018***	-23.319***	-22.804***	-17.922***	-4.879***
	(0.00)	(1.22)	(1.14)	(0.89)	(0.58)
volatility	0.000***	0.200***	0.180***	0.163***	0.017*
	(0.00)	(0.02)	(0.02)	(0.02)	(0.01)
volume	-0.003***	-29.639***	-26.626***	-16.275***	-10.358***
	(0.00)	(1.18)	(1.18)	(0.96)	(0.61)
intercept	1.037***	440.901***	404.182***	276.032***	128.211***
	(0.01)	(15.01)	(14.85)	(12.16)	(7.57)
Adj. R^2	0.694	0.759	0.739	0.548	0.223

Table 5
Execution Quality Across Wholesalers

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

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

where $DepVar_{ijt}$ is one of the following market quality variables for stock i wholesaler j in month t: the ratio of effective to quoted spreads, quoted spread, effective spread, price impact, and realized spread as defined previously, $top2$ is a dummy variable that has a value of 1 for orders executed by Citadel and Virtu and 0 for orders executed by other wholesalers, *price* is the natural log of the stock price, *volatility* is the difference between the high and low prices scaled by the high price, and *volume* is the natural log of trading volume. We use CRSP values for volatility and volume instead of Rule 605 values to capture market-wide activity. Panel B augments the specification with interaction terms between the $top2$ dummy and indicator variables for the size-based terciles of non-S&P 500 stocks: $T1$, $T2$, and $T3$. The models are estimated with stock and month fixed effects, and the standard errors are double-clustered across stocks and months. Note that we only use wholesaler data for these regressions. 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	0.020***	-1.297***	0.980	2.676***	-1.697**
	(0.01)	(0.13)	(0.61)	(0.29)	(0.68)
price	-0.032***	-23.263***	-22.348***	-10.550***	-11.796***
	(0.00)	(1.34)	(1.16)	(0.69)	(0.77)
volatility	-0.007***	-33.808***	-28.130***	-11.818***	16.314***
	(0.00)	(1.26)	(1.29)	(0.75)	(0.80)
volume	0.000***	0.257***	0.213***	0.146***	0.067***
	(0.00)	(0.02)	(0.02)	(0.01)	(0.01)
intercept	0.827***	503.585***	420.780***	187.698***	233.099***
	(0.02)	(16.33)	(16.46)	(8.90)	(10.32)
Adj. R^2	0.313	0.760	0.702	0.388	0.260
Panel B: Specification with Interaction Terms					
top2	0.070***	-0.011	0.466***	0.675***	-0.210*
	(0.01)	(0.02)	(0.06)	(0.08)	(0.11)
top2×*T1*	-0.041***	-0.598***	0.001	0.301**	-0.299
	(0.01)	(0.05)	(0.16)	(0.12)	(0.18)
top2×*T2*	-0.063***	-1.456***	0.582	1.375***	-0.794
	(0.01)	(0.12)	(0.59)	(0.31)	(0.65)
top2×*T3*	-0.067***	-2.691***	1.461	6.214***	-4.754***
	(0.01)	(0.39)	(1.62)	(0.73)	(1.76)
price	-0.032***	-23.263***	-22.348***	-10.550***	-11.796***
	(0.00)	(1.34)	(1.16)	(0.69)	(0.77)
volatility	-0.007***	-33.808***	-28.130***	-11.817***	-16.315***
	(0.00)	(1.26)	(1.29)	(0.75)	(0.80)
volume	0.000***	0.257***	0.213***	0.146***	0.067***
	(0.00)	(0.02)	(0.02)	(0.01)	(0.01)
intercept	0.827***	503.592***	420.775***	187.686***	233.106***
	(0.02)	(16.33)	(16.46)	(8.90)	(10.32)
Adj. R^2	0.316	0.760	0.702	0.388	0.260

Table 6
Wholesaler Economies of Scale

To assess the effect of economies of scale on liquidity costs charged by wholesalers, we estimate the following regression:

$$realized\ spread_{ijt} = \alpha_i + \gamma_t + \beta_1 top2_j + \beta_2 price_{it} + \beta_3 volatility_{it} + \beta_4 op.\ size_{ijt} + \varepsilon_{ijt},$$

where $realized\ spread_{ijt}$ is realized spread as defined previously for stock i wholesaler j in month t, $top2$, *price*, and $volatility$ variables are as previously defined, and a new control variable $op.size$ is the natural log of retail volume captured by a wholesaler in stock i scaled by the natural log of total CRSP trading volume in this stock. As a result, the $op.size$ variable absorbs the volume control used in the earlier models. We run these regressions for the full sample and then separately for each sub-sample, use stock and month fixed effects, and cluster standard errors by stock and month. Note that we only use wholesaler data for these regressions. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	Full sample	S&P 500	Tercile 1	Tercile 2	Tercile 3
	[1]	[2]	[3]	[4]	[5]
top2	1.089	-0.014	0.645	0.882	1.370
	(0.66)	(0.24)	(0.39)	(0.65)	(1.87)
price	-10.809***	0.072	-1.791***	-8.521***	-17.815***
	(0.77)	(0.18)	(0.39)	(0.90)	(1.36)
volatility	-0.038***	-0.008	-0.046	0.010	-0.077***
	(0.01)	(0.02)	(0.03)	(0.01)	(0.02)
op. size	-3.687***	-0.327	-1.547***	-2.492***	-7.628***
	(0.34)	(0.23)	(0.27)	(0.31)	(1.01)
intercept	53.351***	1.155	12.867***	38.826***	92.380***
	(2.11)	(0.78)	(1.36)	(2.16)	(2.64)
Adj. R^2	0.218	0.080	0.127	0.175	0.155

Table 7
Wholesaler Order Flow Determinants

To infer if the share of retail order flow received by a wholesaler depends on its prior performance, we estimate the following regression:

$$\begin{aligned} \textit{market share}_{ijt} = \alpha_i + \theta_j + \gamma_t + \beta_1 \textit{stock realiz. spr.}_{ijt-1} + \beta_2 \textit{portfolio realiz. spr.}_{jt-1} \\ + \beta_3 \textit{price}_{it} + \beta_4 \textit{volatility}_{it} + \beta_5 \textit{volume}_{it} + \varepsilon_{ijt}, \end{aligned}$$

where $\textit{market share}_{ijt}$ is the share of retail volume in stock i executed by wholesaler j in month t expressed as the deviation from the geometric mean across all wholesalers, $\textit{stock realiz. spr.}_{ijt-1}$ is the average realized spread charged in stock i by wholesaler j in month $t-1$ expressed as the deviation from the arithmetic mean across all other wholesalers, $\textit{portfolio realiz. spr.}_{jt-1}$ is the average realized spread charged by wholesaler j in all stocks routed to it in month $t-1$ expressed as a deviation from the arithmetic mean across all other wholesalers, and *price*, *volatility*, and *volume* are as previously defined. The realized spread variables are scaled, so the economic significance corresponds to basis points. We run these regressions for the full sample and then separately for each sub-sample using stock, wholesaler, and month fixed effects and clustered standard errors. Note that we only use wholesaler data for these regressions. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	Full sample	S&P 500	Tercile 1	Tercile 2	Tercile 3
	[1]	[2]	[3]	[4]	[5]
stock realiz. spr.	-0.000***	-0.000	-0.000*	-0.000***	-0.000*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
portfolio realiz. spr.	-0.031***	-0.028***	-0.032***	-0.032***	-0.029***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
price	-0.005	0.188*	0.017	-0.033***	-0.023**
	(0.01)	(0.10)	(0.01)	(0.01)	(0.01)
volatility	-0.039**	-0.283	-0.097	-0.030**	-0.012
	(0.02)	(0.19)	(0.06)	(0.01)	(0.01)
volume	0.048***	0.181*	0.047***	0.034***	0.042***
	(0.01)	(0.10)	(0.02)	(0.01)	(0.00)
intercept	-0.081	-2.810	-0.177	0.137***	0.097**
	(0.10)	(1.75)	(0.22)	(0.05)	(0.04)
Adj. R^2	0.685	0.716	0.722	0.668	0.674

Table 8
Jane Street Entry

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

$$realized\ spread_{ijt} = \alpha_i + \gamma_t + \beta_1 WHOL_j + \beta_2 WHOL \times POST_{jt} + \beta_3 price_{it} + \beta_4 volatility_{it} + \beta_5 volume_{it} + \varepsilon_{ijt},$$

where $realized\ spread_{ijt}$ is the realized spread in stock i for intermediary j (wholesaler or exchange) in month t, $WHOL$ is a dummy variable that has a value of 1 for orders executed by wholesalers and 0 for orders executed by exchanges, $POST$ is a dummy variable that has a value of 1 after Jane Street market share capture and 0 otherwise, and $price$, $volatility$, and $volume$ variables are as previously defined. We run the regressions separately for each subsample. In Panel A, the model is estimated for all wholesalers, while in Panels B and C, it is estimated separately for the incumbents and Jane Street. The models are estimated with stock and month fixed effects, which is why the standalone $POST$ variable is omitted. The standard errors are double-clustered across stocks and months. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	S&P 500	Tercile 1	Tercile 2	Tercile 3
	[1]	[2]	[3]	[4]
Panel A: All Wholesalers				
WHOL	1.168***	5.614***	11.525***	20.349***
	(0.17)	(0.29)	(0.54)	(0.97)
WHOL×POST	0.335	1.746*	5.562**	14.382***
	(0.18)	(0.74)	(1.72)	(3.39)
price	-0.396	-0.128	1.178	-2.349
	(0.47)	(0.54)	(0.92)	(2.15)
volatility	0.013	0.014	0.029	-0.022
	(0.01)	(0.02)	(0.02)	(0.03)
volume	-0.470	-2.341***	-6.183***	-2.732**
	(0.25)	(0.50)	(1.31)	(0.80)
intercept	7.351	26.525**	61.986***	35.422***
	(5.11)	(6.92)	(14.46)	(8.03)
Adj. R^2	0.219	0.304	0.245	0.218
Panel B: Incumbents				
WHOL	1.171***	5.689***	11.574***	20.106***
	(0.16)	(0.30)	(0.57)	(1.02)
WHOL×POST	0.327	1.932**	6.525**	16.863***
	(0.17)	(0.75)	(1.82)	(3.79)
Panel C: Jane Street				
WHOL	1.183**	4.563***	8.752***	26.885***
	(0.34)	(0.16)	(0.40)	(2.15)
WHOL×POST	0.421	0.617	0.206	-12.680***
	(0.35)	(0.59)	(0.96)	(2.18)

Table 9
Inventory Costs

The table reports coefficient estimates from the following regression:

$$\begin{aligned} \textit{realized spread}_{it} = \alpha_i + \gamma_t + \beta_1 T1_i + \beta_2 T2_i + \beta_3 T3_i + \beta_4 \textit{price}_{it} + \beta_5 \textit{volatility}_{it} \\ + \beta_6 \textit{retail volume}_{it} + \varepsilon_{it}, \end{aligned}$$

where $\textit{realized spread}_{it}$ is the realized spread in stock i in month t, $T1$, $T2$, and $T3$ are dummies indicating whether a stock is in tercile 1, 2, or 3, with the intercept capturing S&P 500 stocks, *price* and $\textit{volatility}$ are as previously defined, and *retail volume* is the natural log of total retail volume across all wholesalers as reported in Rule 605 data. We also estimate a specification that uses *total volume*, defined as the natural log of CRSP trading volume, replacing retail volume. The latter specification allows for the possibility that wholesalers use non-retail flow to manage retail inventory positions. Since the tercile dummies are unique for each security, the regressions control only for month fixed effects, yet use double-clustered standard errors. Note that we only use wholesaler data for these regressions. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	[1]	[2]	[3]	[4]
T1	3.453***	0.988	-14.312***	-20.244***
	(0.19)	(0.75)	(1.41)	(1.63)
T2	14.126***	9.116***	-19.654***	-35.816***
	(0.50)	(1.32)	(2.26)	(2.99)
T3	45.991***	38.502***	4.603**	-20.363***
	(2.64)	(1.90)	(2.06)	(2.53)
price		-2.017***	-8.940***	-9.391***
		(0.61)	(0.80)	(0.79)
volatility		0.059***	0.056**	0.116***
		(0.02)	(0.03)	(0.02)
retail volume			-7.200***	
			(0.35)	
total volume				-10.597***
				(0.51)
intercept	0.973	9.446***	148.730***	182.411***
	(0.68)	(2.38)	(8.10)	(9.75)
Adj. R^2	0.124	0.127	0.183	0.208

Table 10
Institutional Trading Interest

The table reports descriptive statistics for stock-quarter ratios of retail liquidity-demanding volume to institutional volume. We estimate institutional volume in the sample stocks based on changes in reported quarterly holdings from 13F reports and add to that changes in short interest computed from bi-weekly Compustat estimates aggregated to quarterly values to match 13F data. To account for intra-quarter trading, we gross-up institutional volume inferred from 13F reports by a factor of 1.17 based on Chakrabarty, Moulton, and Trzcinka (2017).

	mean	*median*	*st. dev.*	*p25*	*p75*
	[1]	[2]	[3]	[4]	[5]
S&P 500	0.850	0.199	3.081	0.125	0.406
Tercile 1	1.601	0.165	5.860	0.081	0.523
Tercile 2	2.639	0.196	7.930	0.041	1.092
Tercile 3	9.798	1.568	15.594	0.150	10.785

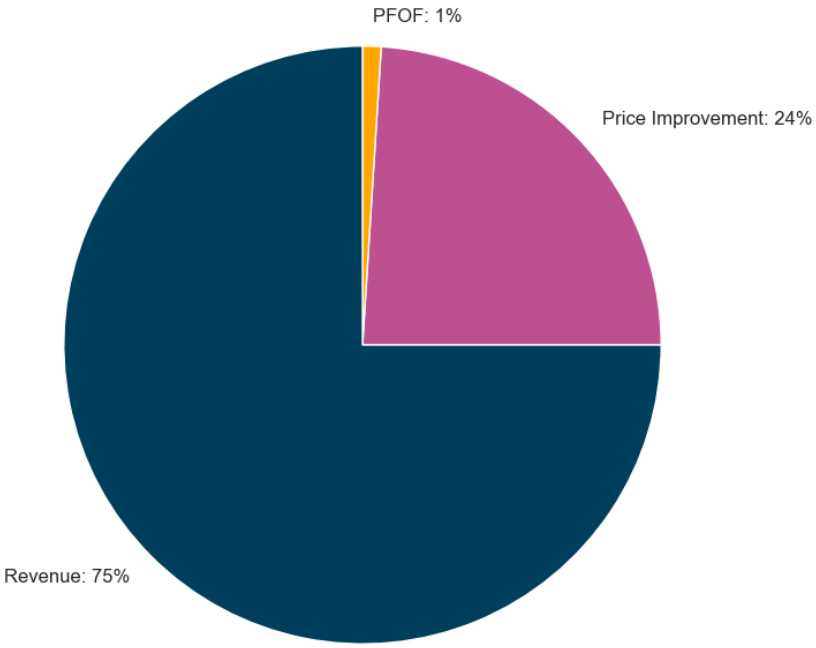

Figure 1. Allocation of Spread Revenue
The figure depicts the allocation of revenue obtained from quoted spreads between price improvement provided to retail investors, payments for order flow from wholesalers to retail brokers, and the portion of revenue retained by wholesalers. The retained revenue is used by wholesalers to cover the adverse selection, inventory, and fixed costs, with the balance being their profit. The sample includes all market and marketable orders in the sample stocks during January 2020-December 2022 period, during which we have access to both Rule 605 and 606 data.

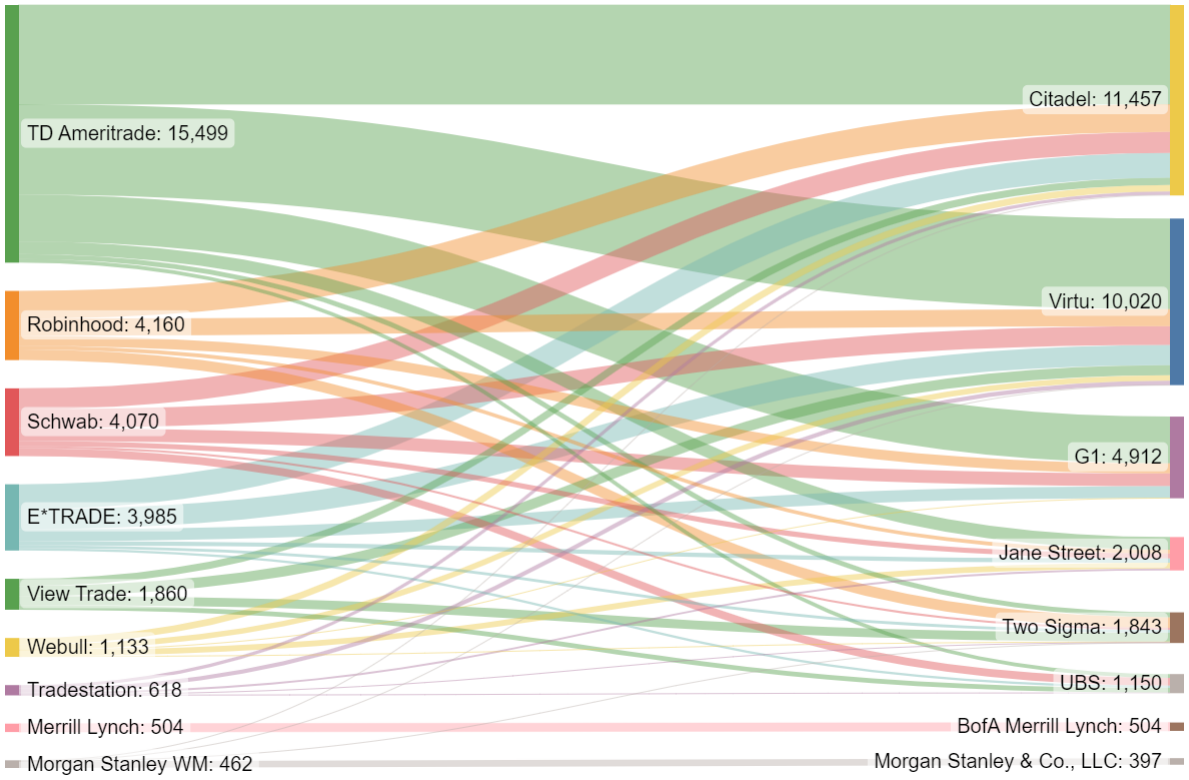

Figure 2. Retail Broker Routing

The figure reports order routing patterns (in millions of shares per month) by select major retail brokers to wholesalers and exchanges. The sample covers all sample stocks, and the sample period is January 2020 through December 2022. To obtain routed volumes, we use two variables available from Rule 606 data: the total PFOF dollar amounts received by retail brokerages and the PFOF amounts in cents per one hundred shares. Dividing the former by the latter allows us to estimate the share amounts sent by the brokerages to the wholesalers. For brokerages such as Fidelity and Vanguard that do not accept PFOF, we are unable to compute the share amounts, so these brokerages are not included in the figure. Interactive Brokers accepts PFOF for some orders submitted by IBKR Lite customers, but we are unable to separate this flow from their IBKR PRO flow in Rule 606 data and therefore also exclude this broker from the figure. We note that because the figure does not include flows from all brokerages, and because Rule 606 data are unavailable for 2019, the total volumes and wholesaler market shares in this figure do not perfectly align with those in the main sample.

Figure 3. Wholesaler Entry and Exit

This figure reports the market share (shaded area, left axis) and the realized spread ratio (line, right axis) for wholesalers Jane Street in Panel A and Wolverine in Panel B. Market share is the percentage of total retail volume executed by wholesalers, and the realized spread ratio is the ratio of Jane Street's and Wolverine's realized spreads divided by the volume-weighted average realized spread for the other wholesalers. The sample covers all stocks during the 2019-2022 period, and the data source is Rule 605 reports.

Internet Appendix to "The Retail Execution Quality Landscape"

A.1 Data Details

Rule 605 data are reported monthly, by close to 70 market centers, and we use data from January 2019 to December 2022. We obtain the majority of our data from a firm that specializes in trading cost analyses, and complement these data by downloading original Rule 605 reports directly from the market centers.

Market centers occasionally use their own symbol coding schemes, and many include securities that are not National Market System stocks (e.g., warrants and convertible bonds) in their Rule 605 reports, even though these securities are not covered by the rule. As a result, the data contain over 16,000 unique symbols even after we recode the symbols to a more uniform notation. We focus on ordinary stocks and Class A, B, and C shares, and are able to match 94.8% of the symbol-months in our Rule 605 data for these stocks to CRSP. The final sample consists of 8,165 ordinary stocks.

To study cross-sectional differences we divide our sample into four sub-samples. The first sub-sample is S&P 500 stocks which we define based on all stocks indicated as being part of the index at any time between January 2019 and December 2022. This results in 514 unique symbols. We divide the remaining 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 stocks, and tercile 3 includes 2,551 stocks.

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

- *Executed shares (EXshs)* are the cumulative number of shares executed at the receiving

market center.

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

After data cleaning to correct for inconsistent coding of missing vs. 0 in share volume fields across market centers, we calculate the below variables.

$$shares\ executed \equiv SHS = EXshs + AWshs \tag{8}$$

$$quoted\ spread \equiv \$QS = \$ES + 2 \cdot \frac{1}{SHS} \cdot (\$PI \cdot PIshs + 0 \cdot AQshs - \$OQ \cdot OQshs) \tag{9}$$

$$price\ impact = \$ES - \$RS \tag{10}$$

$$effective\ /\ quoted = \frac{\$ES}{\$QS} \cdot 100 \tag{11}$$

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

$$at\ or\ better = \frac{AQshs+PIshs}{SHS} \cdot 100 \tag{12}$$

$$price\ improved = \frac{PIshs}{SHS} \cdot 100 \tag{13}$$

We merge Rule 605 data with CRSP monthly data to obtain information on closing monthly price (*prc*) and shares outstanding (*shrout*) so we can calculate firm size ($prc \times shrout$), and *askhi* and *bidlo* so we can calculate monthly price range ((*askhi-bidlo)* $\div$ *askhi*). We trim the following variables at 0.1 and 99.9% separately for market and marketable limit orders: quoted spread, effective spread, realized spread, price impact, and CRSP closing price. Finally, we calculate the quoted, effective, realized spreads and price impact in basis points relative to the monthly price from CRSP. We discuss the details of Rule 606 data in Section A.6 of this Appendix.

A.2 Price Impact Horizons

Rule 605 requires that price impacts are estimated over five-minute horizons. In the meantime, modern technology allows market makers to operate at sub-second horizons, raising doubts about the relevance of Rule 605 statistics. To better understand this issue, we analyze one month of intraday Trade and Quote (TAQ) data from April 2022, estimating price impacts at horizons of 1, 5, 15, 30, 60, and 300 seconds (5 minutes). For this analysis, we randomly select 50 stocks from each of the four sub-samples (the S&P 500 and terciles 1 through 3), totaling 200 stocks. Figure A1 shows that prices react to trades rather quickly. In S&P 500 stocks, nearly 80% of the total 5-minute price impact occurs within the first second after a trade. Price adjustments are slower in smaller stocks, with tercile 3 seeing only 40% of the price reaction in the first second.

[Figure A1]

To avoid the adverse selection cost, a market maker needs to close a position before price

impact accumulates. Let us consider a scenario where the exchange market maker opens a short position to provide liquidity to a buyer in an S&P 500 stock, and the buyer's trade results in a price impact of one cent. If the market maker offsets the position in one second, she will bear a price impact cost of 0.8 cents. To exit, she must possess the ability to quickly cover the short position, ideally without paying for liquidity.

For such an offset to take place, another exchange customer must be willing to sell and pay for liquidity. How often do such sellers arrive? When the flow is balanced, TAQ data indicate that in S&P 500 stocks a seller typically arrives once every 1.27 seconds. In terciles 1, 2, and 3, they arrive once every 6.59, 21.68, and 86.60 seconds, respectively. Thus, even if the market maker is a monopolist, she will bear between 70% and 80% of the price impact of the original trade. However, a typical market maker is not a monopolist. As discussed in the main text, professional market makers on exchanges compete not only with each other, but also with a large number of institutions trading via limit orders. Consequently, the market maker does not engage with every incoming order, resulting in her incurring (nearly) the entirety of the price impact. Considering this, the five-minute horizons used in Rule 605 data are likely appropriate for our analyses.

A.3 Execution Quality Statistics for Sub-samples

In Table A1, we ask if market capitalization affects execution quality using univariate analysis. We begin with the S&P 500 sub-sample. Wholesalers price-improve 76% of marketable orders and provide price improvement corresponding to 47% of the quoted spread. By comparison, only 12% of marketable orders receive price improvement on exchanges, and price improvement is a modest 5%. The adverse selection that accrues on exchanges is 103% (= 6.34/3.13-1) greater than that accruing to wholesalers. Wholesalers earn larger realized spreads than liquidity providers on exchanges, 1.23 vs. -1.18 bps.

[Table A1]

When it comes to terciles 1 through 3, the pattern discussed for the S&P 500 stocks is generally preserved. First, wholesalers price improve a substantially larger portion of marketable orders than exchanges for each sub-sample (e.g., 64% vs. 9% in tercile 2). Note also that the fraction of price improved orders falls as we move from larger to smaller size firms, both for wholesalers and exchanges. Second, the magnitude of price improvement continues to be significantly larger for marketable orders routed to wholesalers for all terciles (e.g., 26% vs. 4% of the quoted spread for tercile 2). This metric is generally declining as we move from larger to smaller size firms for orders routed to wholesalers but is relatively constant for orders routed to exchanges.

Order flow toxicity is substantially greater on exchanges, with exchange price impacts 54%, 39%, and 43% greater than those at wholesalers for terciles 1, 2, and 3, respectively. Finally, the exchange realized spreads are even more negative for tercile 1 and 2 stocks than for S&P 500 stocks but turn positive for tercile 3 stocks. By contrast, wholesalers earn positive realized spreads that increase as we move from larger to smaller size firms.

A.4 Execution Quality Statistics in Cents

In the main analysis, we report spread statics in basis points. For completeness, here we report the same statistics in cents. Table A2 contains results for the full sample, and Table A3 for the sub-samples.

[Tables A2 and A3]

A.5 Dollar-Volume-Weighted Execution Quality Statistics

In the main analysis, we share-volume-weight execution quality statistics up to the cross-section and then use a simple average across stocks to obtain the final statistics representing the securities landscape. By contrast, in the analysis accompanying the proposed Order Competition Rule, the SEC uses dollar-volume-weighted statistics. To assure the reader that our sample is not

very different from the one analyzed by the SEC, we report the dollar-volume-weighted statistics in Table A4. For completeness, Table A5 reports dollar-volume-weighted statistics for the four sub-samples. Note that we report full spreads, whereas the SEC reports half-spreads.

[Tables A4 and A5]

A.6 Retail Broker Routed Volume

To calculate routed volumes for Figure 2 in the main analysis, we use two variables from Rule 606 data reported by major retail brokerages: the total dollar amounts of PFOF received by retail brokerages and the PFOF amounts in cents per one hundred shares. Dividing the former by the latter, we estimate the number of shares sent by each brokerage to each wholesaler. Brokerages such as Fidelity and Vanguard, which do not charge PFOF, do not have sufficient data for us to reconstruct the routed volume. Interactive Brokers receives PFOF for their IBKR Lite flow, but not for the IBKR PRO flow, and we are unable to see the entire picture of its routing. These three brokerages are therefore excluded from Figure 2.

Rule 606 mandates that each retail brokerage submits monthly summary routing and fee data for two stock groups: the S&P 500 and non-S&P 500. As a result, these data do not have the cross-sectional richness of Rule 605 data, which are reported by stock. Still, they offer valuable insights into broker routing. An alternative to measuring routed volumes, a metric often used in the industry to compare the size of brokerages is Daily Average Revenue Trades (DART). We note that with the advent of commission-free trading for much of the retail equity brokerage business, these numbers are no longer comparable across brokers. Some brokers report DART only for commission trades and therefore effectively do not report the retail volume. Others combine equity with options and futures and often report DART based on their global activity (e.g., Interactive Brokers). Yet others combine retail and institutional trades when reporting DART (e.g., Fidelity).

Table A1
Execution Quality: Sub-samples

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 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 than 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] as well as [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		
price, $	151.79	152.47		52.20	54.00	
improved, %	76.11	12.00	***	69.76	11.15	***
at or better, %	94.71	97.82	***	92.86	98.34	***
quoted spread, bps	8.16	5.45	***	26.60	17.31	***
effective spread, bps	4.36	5.16	***	17.65	16.42	**
effective / quoted	0.53	0.95	***	0.66	0.95	***
price impact, bps	3.13	6.34	***	13.15	20.20	***
realized spread, bps	1.23	-1.18	***	4.50	-3.78	***
	Tercile 2			Tercile 3		
price, $	14.43	14.61		7.03	7.09	
improved, %	63.56	8.52	***	63.16	6.84	***
at or better, %	93.20	98.65	***	93.02	98.27	***
quoted spread, bps	60.77	42.43	***	132.03	107.34	***
effective spread, bps	44.77	40.56	**	105.42	104.98	
effective / quoted	0.74	0.96	***	0.80	0.98	***
price impact, bps	31.79	44.22	***	71.24	101.89	***
realized spread, bps	12.98	-3.66	***	34.18	3.09	***

Table A2
Execution Quality (Cents)

The table contains execution quality statistics in cents. We compute the statistics separately for orders executed by wholesalers (WHOL) and exchanges (EXCH). We report 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 than the corresponding NBBO. We report the quoted and effective spreads, 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, also in cents. Statistics are share-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]
price, $.	30.04	30.59	
improved, %	66.10	9.00	***
at or better, %	93.12	98.35	***
quoted spread, cents	10.52	7.28	***
effective spread, cents	7.10	6.88	
effective / quoted	0.67	0.95	***
price impact, cents	4.69	6.99	***
realized spread, cents	2.41	-0.11	***

Table A3
Execution Quality (Cents): Sub-samples

The table contains execution quality statistics measured in cents. 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 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 than the corresponding NBBO. Further, we report the quoted and effective spreads in cents, 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 share-volume-weighted. Asterisks *** in columns [3] and [6] indicate statistical significance of differences between columns [1] and [2] as well as [4] and [5] at the 1% level.

	WHOL	EXCH	diff.	WHOL	EXCH	diff.
	[1]	[2]	[3]	[4]	[5]	[6]
	S&P 500			Tercile 1		
price, $	151.79	152.47		52.20	54.00	
improved, %	76.11	12.00	***	69.76	11.15	***
at or better, %	94.71	97.82	***	92.86	98.34	***
quoted spread, cent	15.27	9.97	***	12.30	7.90	***
effective spread, cent	8.25	9.26		7.67	7.35	
effective / quoted	0.54	0.93	***	0.63	0.93	***
price impact, cent	4.51	9.98	***	5.32	7.96	***
realized spread, cent	3.74	-0.72	***	2.4	-0.61	***
	Tercile 2			Tercile 3		
price, $	14.43	14.61		7.03	7.09	
improved, %	63.56	8.52	***	63.16	6.84	***
at or better, %	93.20	98.65	***	93.02	98.27	***
quoted spread, cent	9.58	6.60	***	8.97	6.96	***
effective spread, cent	6.73	6.21		6.75	6.72	
effective / quoted	0.70	0.94	***	0.75	0.97	***
price impact, cent	4.49	6.25	***	4.36	6.25	***
realized spread, cent	2.24	-0.04	***	2.39	0.47	***

Table A4
Execution Quality (Dollar-volume-weighted)

We compute the statistics separately for orders executed by wholesalers (WHOL) and exchanges (EXCH). We report 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 than the corresponding NBBO. We report the quoted and effective spreads, 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, also in basis points. Statistics are dollar-volume-weighted and expressed in basis points.

	WHOL	EXCH
	[1]	[2]
price, $.	171.22	134.14
improved, %	81.32	10.12
at or better, %	95.05	97.57
quoted spread, bps	10.04	6.86
effective spread, bps	6.09	6.60
effective / quoted	0.61	0.96
price impact, bps	4.68	8.45
realized spread, bps	1.41	-1.85

Table A5
Execution Quality (Dollar-volume-weighted): Sub-samples

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 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 than 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 dollar-volume-weighted.

	WHOL	EXCH	WHOL	EXCH
	[1]	[2]	[3]	[4]
	S&P 500		Tercile 1	
price, $	214.83	171.99	124.40	83.50
improved, %	83.17	10.06	80.12	10.54
at or better, %	95.41	97.48	94.73	97.66
quoted spread, bps	3.92	3.50	13.09	9.60
effective spread, bps	1.87	3.36	7.63	9.19
effective / quoted	0.48	0.96	0.58	0.96
price impact, bps	1.42	4.19	5.99	12.19
realized spread, bps	0.46	-0.83	1.64	-2.99
	Tercile 2		Tercile 3	
price, $	12.4	14.60	5.87	6.46
improved, %	70.98	7.66	67.49	5.77
at or better, %	93.65	98.44	93.01	97.89
quoted spread, bps	40.68	28.55	80.93	66.50
effective spread, bps	28.48	26.86	61.89	65.70
effective / quoted	0.70	0.96	0.76	0.99
price impact, bps	21.7	37.29	46.22	70.49
realized spread, bps	6.78	-8.74	15.68	-4.79

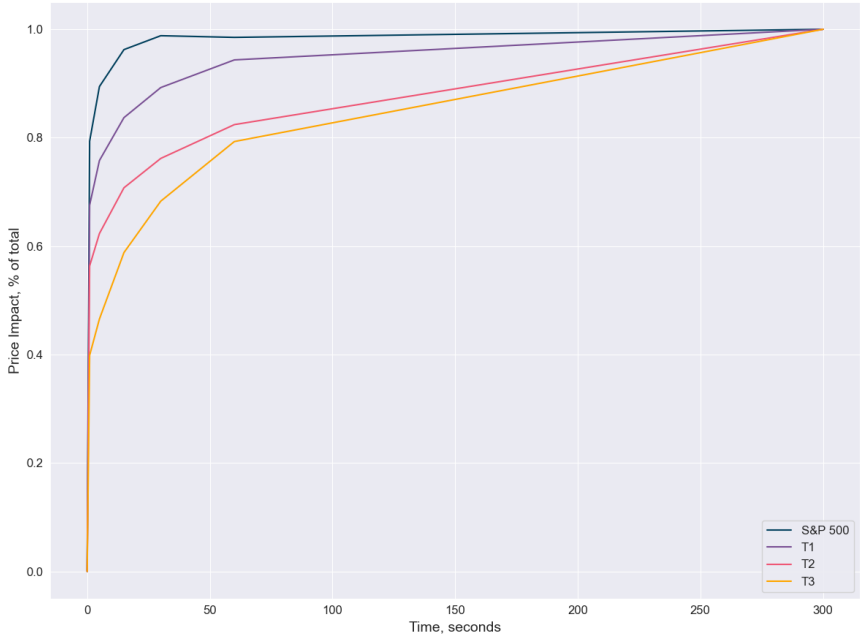

Figure A1. Price Impact Horizons

This figure reports trade price impacts at time horizons of 1, 5, 15, 30, 60, and 300 seconds (5 minutes). The data are from TAQ in April 2022 and cover 50 stocks in each of the four sub-samples, that is S&P 500 and terciles 1 through 3 (T1-T3), for a total of 200 stocks. To facilitate comparison across sub-samples, we assume that the 5-minute horizon captures the entirety of price impact.