The Retail Execution Quality Landscape*

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

September 29, 2023

Abstract. Using a comprehensive multi-year U.S. dataset, we show that off-exchange (wholesaler) executions tend to provide significant trading cost savings, benefiting retail investors. Although the wholesale industry is concentrated, three findings suggest that wholesalers do not abuse market power. Firstly, brokers closely monitor wholesalers and reward those who offer lower liquidity costs with more order flow. Secondly, the largest wholesalers offer the lowest costs, due to economies of scale. Finally, the entry of a new large wholesaler does not lead to a reduction of liquidity costs. We discuss the advantages and drawbacks of the two proposed alternatives to the current system: (i) to pool retail and institutional flows on exchanges and (ii) to send retail flow to order-by-order auctions.

Key words: Retail Trading, Wholesalers, Execution Quality

JEL: G20; G24; G28

*We thank James Brugler, Sabrina Buti, Doug Clark, Carole Comerton-Forde, Laurence Daures, Tom Ernst, Marinela Finta, Corey Garriott, Carole Gresse, Peter Haynes, David Hecht, Bob Jennings, Travis Johnson, S.P. Kothari, Phil Mackintosh, Josh Mollner, Dmitriy Muravyev, Peter Reiss, Vincent Skiera, Chen Yao, Marius Zoican, and conference/seminar participants at Brock University, Case Western Reserve University, Chicago Quantitative Alliance, China International Conference in Finance, European Finance 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 Toronto, and Université Paris Dauphine-PSL for valuable comments. The most recent version may be downloaded from: https://bit.ly/3ATa2v5.

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

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

§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 goes to the retail broker as payment for order flow (PFOF), and yet another portion is passed on to the retail trader as price improvement. While wholesalers internalize liquidity demanding orders, they send liquidity providing orders to exchanges since regulation requires such orders to be displayed.

How retail orders are handled is currently actively debated. Some observers argue that wholesalers wield (and abuse) market power and provide limited benefits to retail investors.[2] They suggest that price improvement offered by wholesalers tends to be *de minimis*, or the smallest amount possible.[3] 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 provide the largest amount of price improvement. Wholesalers instead argue that it is retail brokerage firms that have market power and that they route orders to wholesalers because it is in the best interest of their retail clients. Wholesalers claim that they offer significant price improvement and that retail investors are well-served by the current system.

Reconciling these conflicting views is an empirical task, and we do so using four years of SEC Rule 605 reports from 2019 to 2022.[4] Each order-handling venue must file such reports

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

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

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

[4] A recent analysis by the SEC shows that these reports are highly consistent with the audit trail data avail-

on a monthly basis to maintain a public record of execution quality. Our analyses reveal that wholesalers and exchanges offer unique execution benefits. Wholesalers, in particular, provide substantial price improvement, executing liquidity-demanding orders at prices better than those quoted on exchanges. Wholesaler price improvement is far from *de minimis*, accounting for 24% of the quoted spread in the full sample, which translates to 12% (1.7 cents) for both the retail buyer and the seller.[5] What's even more striking is that a retail order in an average S&P 500 stock receives 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 clearly offer a significant advantage.[6]

Another important execution quality dimension is the cost of supplying liquidity. Insofar as professional market makers, including wholesalers, supply liquidity, two considerations come into play. First, market makers incur three costs: the adverse selection cost, the cost of holding inventory, and the technology cost. Second, they aim to make profits. Researchers typically measure the adverse selection cost by estimating trade *price impacts*, the difference between the midquote at the time of a trade and the midquote at a future time. Adverse selection cost arises when midquotes increase after a buyer-initiated trade and decrease after a seller-initiated trade. Wholesalers are professional market makers facing notably lower adverse selection cost than liquidity providers on exchanges. In our sample, the order flow received by them generates 30% less price impact than the exchange-bound flow.

Adverse selection costs aside, the inventory and technology costs, as well as market making profits, are captured by a conventional metric, the *realized spread*, which is the difference

able to the agency. See Release 34-96495 "Order Competition Rule" from December 14, 2022 (https://bit.ly/3v1Z96V).

[5] It is commonly assumed that quoted spreads are close to 1 cent, especially for large firms. In reality, the average stock has a quoted spread of 9.1 cents, and the average S&P 500 stock has a quoted spread of 11.66 cents, largely due to the presence of high-priced stocks.

[6] 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. However, in our data, price improvement by a wholesaler 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.

between the effective spread (the market maker's liquidity provision revenue) and the price impact. The data show that exchanges offer realized spreads that are substantially lower than those offered by wholesalers. How does the difference in realized spreads arise? On exchanges, limit orders submitted by market making algorithms compete with limit orders submitted by non-market making algorithms. The latter operate multiple strategies: from limit order legs of latency arbitrage (Aquilina, Budish, and O'Neill (2021)) to managing institutional investment positions (O'Hara (2015)). Covering market making costs and earning liquidity provision profits is not as important to non-market making algorithms as to their market making counterparts, if at all. Therefore, exchange realized spreads should be smaller than those observed in a pure market maker setting. According to industry estimates, pure market making algorithms represent about 16% of all liquidity provision on modern exchanges.[7] With such a split, exchange liquidity should be cheaper than wholesaler liquidity.

Our data reveal that the wholesale industry exhibits a relatively high concentration level, with Citadel and Virtu capturing over 70% of retail flow. In this light, it is not surprising that some market observers express concerns about the potential market power abuses by wholesalers. Although we cannot directly observe whether such abuses occur, the data offer several indicators that lead us to think otherwise. First, the largest retail brokerage is larger than the largest wholesaler. This suggests that market power could potentially reside with retail brokerages. To examine this possibility, we ask if retail brokerages actively manage their relationships with wholesalers by favoring those that offer lower liquidity costs. The data indicate that they indeed do so. Wholesalers that provide lower costs today are rewarded with additional order flow in the future. Interestingly, brokerages appear to evaluate wholesalers not on a stock-by-stock basis but rather on a bundled basis. In other words, if Citadel offers the cheapest liquidity in AAPL, it will not necessarily receive more future AAPL flow. Instead, Citadel must outperform its competitors across the entire range of stocks to attract more order flow in AAPL or any other stock.

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

This finding highlights an intriguing aspect of the retail ecosystem, where the brokerages compel wholesalers to compete in stocks that have relatively low trading frequency and high inventory costs. Typically, small low-volume stocks are less attractive to intermediaries due to their lower profitability, and market regulators and exchanges often seek ways to improve liquidity in such stocks (Foley, Liu, Malinova, Park, and Shkilko (2023)). When we account for inventory costs, our analysis suggests that wholesalers tend to charge relatively low liquidity costs in small stocks compared to large stocks, pointing to cross-subsidization facilitated by bundling.

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

Lastly, the dynamics of wholesaler competition undergo a transformation during our sample period with the entry of a new player, Jane Street. Within a few months, Jane Street gains a significant market share, capturing over 12% of retail flow. If wholesalers were exploiting their market power and enjoying economic rents before this entry, we would expect competitive pressures to intensify, leading to lower liquidity costs. The data, however, do not support this conjecture; we find no evidence of a decrease in liquidity costs. In fact, in low-volume stocks, the costs increase, likely due to the incumbents' loss of economies of scale.

In summary, despite the concerns expressed by critics, wholesalers do not seem to control the marketplace for retail order flow. Instead, retail brokers seem to exert control over execution quality by routing orders to wholesalers that require lower compensation for providing liquid-

ity. The marketplace is also contestable, as evidenced by a new entrant successfully capturing a sizable market share from the incumbents in a surprisingly short time.

These observations make it worth contemplating why the market has not naturally evolved toward an equilibrium with a single 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 a perpetual credible threat, discouraging any potential rogue behavior.

Our final analyses discuss alternatives to the current U.S. market structure for retail orders. First, we contemplate what would happen if retail order flow were to be routed to exchanges instead of the current practice of wholesalers receiving the vast majority of retail order flow from brokers. Second, we discuss potential concerns with the current SEC proposal to introduce a system with order-by-order competition for retail flow.

The richness of our data enables us to consider what would happen if the retail flow were to shift to exchanges. Such proposals are often heard in current market structure discussions.[8] Moving to exchanges would benefit retail investors by reducing realized spreads, but retail investors would face higher adverse selection costs from being pooled with institutional flow. When we analyze the overall impact of such a move on retail traders, we find that they would generally be worse off. Retail investor losses would be accompanied by gains to institutional traders, whose costs would decline due to lower on-exchange toxicity. Consequently, the transfer of retail flow to exchanges would essentially subsidize institutional flow at the expense of retail flow.

Our analysis and conclusions are generally conservative, as we focus exclusively on execution costs. However, we note that the current retail order flow handling system is inextricably linked to commission-free trading for retail investors. The PFOF payments that brokers receive

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

from wholesalers subsidize brokerage businesses, allowing some brokerages to operate without charging commissions.[9] If the current system were to be dismantled, commissions may again be necessary, increasing the overall cost of retail market participation. If the return of commissions leads to a decline in retail volumes, moving retail flow to exchanges will be less effective in diluting the current toxicity levels making the move even less attractive.[10]

Finally, we revisit the SEC's proposal to overhaul retail trading practices, which involves directing retail flow to auctions for order-by-order competition. The proposal assumes that non-professional liquidity providers such as hedge funds, mutual funds, and pension funds would demonstrate significant interest in engaging with retail flow and offer superior price improvement compared to wholesalers. However, our analysis of institutional trading data indicates that this assumption may only hold true for large stocks and may not apply to many stocks currently traded by retail investors. Additionally, our empirical analysis suggests that eliminating bundling could potentially 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 SEC 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 a growing literature that examines retail execution quality and the effects of wholesaler internalization on lit market quality. Notably, this literature has not yet reached a consensus. On the one hand, Adams, Kasten, and Kelley (2021), Kothari, So, and Johnson (2021), and Battalio and Jennings (2023) argue that wholesalers deliver retail trading costs that are lower than those offered by exchanges. Also, Jain, Mishra, O'Donoghue, and Zhao (2022) suggest that internalization revenues may boost the ability of

[9] Robinhood has obtained an average of 71% of its revenue from PFOF in 2020-2022 (Robinhood Markets, Inc. 10-K for the fiscal year ended December 31, 2022 (`https://bit.ly/3Ok8MYW`)).

[10] A separate issue is whether retail investors trade excessively and whether curbing their trading would be welfare-improving. While this matter falls outside the scope of our study, it is worth noting that Barber, Huang, Odean, and Schwarz (2022) and Welch (2022) show that even Robinhood investors, often perceived as less sophisticated, do not typically underperform the standard benchmarks.

market makers such as Citadel and Virtu to compete on exchanges, thereby improving overall liquidity.[11] Similarly, Baldauf, Mollner, and Yueshen (2023) show theoretically that market makers may use retail flows to offset inventory imbalances.

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

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

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

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

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

We complement these studies by carefully analyzing a multi-year comprehensive public dataset that academic researchers have only cursorily examined. According to industry participants, this dataset allows for the cleanest identification of retail order flow that is possible without proprietary data. This dataset enables us to speak about the external validity of the results and allows for an analysis of competitive forces by observing interactions between multiple wholesalers and retail brokerages.

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

2. Data and Sample

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

SEC Rule 605 applies to market and limit orders that are executed during regular trading hours and contain no special handling instructions. We refer to them as basic orders.[12] As an example of special instruction, a trader may ask that a limit order is not forwarded to venues other than the original receiving venue, an order known as *Do Not Ship*. Li, Ye, and Zheng (2020) show that orders that contain special instructions are typically submitted by sophisticated traders such as high-frequency traders. Therefore, Rule 605 data cover (i) virtually all retail orders and (ii) some institutional orders submitted largely by buy-side entities. In the following analyses, we use the execution quality of this institutional flow solely for comparison with the execution quality of retail flow. We refrain from drawing conclusions about overall institutional execution quality because we do not examine institutional transactions on ATSs and institutional flow that executes via non-basic orders.

We focus exclusively on liquidity-demanding orders because wholesalers are required to forward liquidity-providing retail orders to exchanges. Furthermore, motivated by our discussions with industry participants, we group market orders and marketable limit orders together as marketable orders. Rule 605 data include a wide range of securities (over 16,400 unique symbols). We restrict our sample to non-ETF ordinary and Class A, B, and C shares for a total of 8,493 symbols and refer to them as *stocks*.[13] We also create four sub-samples consisting of S&P 500

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

[13]Summary statistics for 3,577 ETFs are provided in the Appendix.

stocks and size-based terciles (T1-T3) of non-S&P 500 stocks (see the Appendix for details).

The data cover 70 execution venues, including all stock exchanges, all major wholesalers, many dark pools, crossing networks, etc. We focus on the first two venue categories, that is, fourteen stock exchanges and the eight largest wholesalers. Table 1 reports trading volumes and market shares of all exchanges and wholesalers. Panel A shows that exchanges execute the majority, 59.5%, of Rule 605 orders, with wholesalers capturing the remaining 40.5%. Our conversations with industry participants indicate that the flow routed to wholesalers consists predominantly of retail orders, while the flow routed to exchanges is mainly institutional orders. In later tests, we provide empirical support to this view.

[Table 1]

Panel B of Table 1 contains statistics for the individual exchanges and wholesalers. Among exchanges, the leading roles are played by Nasdaq and the NYSE/NYSE Arca respectively executing 17.26% and 18.75% (=11.00+7.75) of order flow. Among wholesalers, Citadel and Virtu stand out as the largest, capturing respectively 16.03% and 12.13% of order flow. Other wholesalers are considerably smaller, with the third largest, G1, processing 5.27%, and the next three, Jane Street, Two Sigma, and UBS, processing 2.30%, 2.05% and 1.62%, respectively.[14] Overall, the dataset contains information on execution quality for orders representing more than 3.6 trillion executed shares, which amounts to about 45% of trading volume reported by CRSP during the sample period.

Market structure studies typically rely on a set of execution quality metrics that consists of quoted, effective, and realized spreads, as well as price impacts. The *quoted spread* is the difference between the national best offer (the offer quote that is the lowest across all lit markets) and the national best bid (the bid quote that is the highest across lit markets). It represents trading costs advertised by liquidity providers. Liquidity demanders do not always incur these costs

[14] Jane Street enters the wholesale business several months after our sample period begins and captures over 12% of retail flow by the time the sample period ends. We discuss Jane Street in detail in a subsequent section.

exactly as advertised. Their orders may be price improved as is often done by wholesalers, or interact with better-priced non-displayed orders on exchanges such as odd lots and hidden orders (Bartlett, McCrary, and O'Hara (2022)). To assess trading costs actually incurred by liquidity demanders, Rule 605 data contain the *effective spread* computed as twice the signed difference between the traded price and the midquote (the average of the best offer and the best bid) at the time of the trade. Trade signs are observed by the filers and therefore do not need to be inferred using an algorithm such as Lee and Ready (1991).

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

Rule 605 requires that price impacts and realized spreads are estimated at the 5-minute horizon. This horizon is an eternity in the modern marketplace, and we approach the interpretation of these statistics with due caution. This said, an analysis of the Trade and Quote data in the Appendix shows that adverse selection costs are incurred quickly and that the frequency of arrival of offsetting trades is not sufficiently high to enable them to exit positions prior to incurring most of the adverse selection cost. As a result, price impacts estimated at 5-minute horizons are rather effective at capturing adverse selection costs.

Rule 605 data exclude odd lots, so our analysis is restricted to orders of 100 shares or more.[15] When working with the metrics, we remove outliers by trimming all variables at the 0.1 and 99.9 percentiles. Reporting of the quoted spread is not required by Rule 605, and we derive it from the other metrics as discussed in the Appendix. We scale all metrics by the CRSP closing stock price and use share volume-weighted averages. We weigh by share volume to ensure that our execution quality metrics reflect the experience in the average stock. This is particularly important as retail volume represents over two-thirds of trading activity for smaller stocks (see Table 3).[16]

3. Execution Quality

3.1 Wholesalers vs. Exchanges

Table 2 reports summary execution quality metrics for our sample. During the sample period, wholesalers execute 146.36 million shares in an average sample stock, whereas exchanges execute 207.65 million shares. Rule 605 requires that venues report how their executions compare to the NBBO. Wholesalers price-improve a substantial portion, 65.71%, of order flow they receive, whereas exchanges only price-improve 9.49%. However, exchanges fare better than wholesalers with respect to their ability to match the existing NBBO, executing 98.34% of shares at the NBBO prices or better versus 92.98% by the wholesalers.[17] Institutional traders typically split larger orders to avoid walking the book (slippage), and since their orders are predominantly routed to

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

[16] The SEC instead uses dollar-volume-weighted averages in their recent analysis of retail execution quality. This approach skews the execution quality metrics towards high-price, large capitalization stocks. See Release 34-96495 "Order Competition Rule" from December 14, 2022 (https://bit.ly/3v1Z96V).

[17] 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 only 86% of orders exceeding $64,000 (approximately the 75th percentile) attain NBBO or better execution.

exchanges, this may account for the difference in the proportion of flow that matches the NBBO.

[Table 2]

Notably, wholesalers tend to execute when the NBBOs are relatively wide, 69.6 bps vs. the exchange equivalent of 52.94 bps, a 31% difference. This difference cannot be attributed to wholesaler choices because commercial agreements with retail brokerages do not allow wholesalers to choose what orders to execute and when. Rather, wholesalers are required to execute all orders routed to them. As such, the difference in quoted spreads must be driven by trader decisions and is perhaps expected given the clienteles served by wholesalers and exchanges. Many institutional liquidity-taking algorithms time their activity to periods of narrow quoted spreads. When spreads are wide, they either switch from liquidity demand to liquidity supply or reduce trading altogether. Retail traders are much less likely to engage in such timing. Since the metrics in Table 2 are volume-weighted, it is not surprising that liquidity-demanding exchange trades (institutional flow) tend to occur when spreads are relatively narrow.

Even though retail trade executions occur when quoted spreads are relatively wide, the differential is reduced significantly once we account for the substantial price improvement wholesalers provide to retail flow. Consequently, effective spreads reported by wholesalers are much closer to those reported by their exchange counterparts, at 49.06 bps and 46.98 bps, respectively.[18] With this in mind, we suggest that an execution quality metric appropriate for our setting should account for both quoted and effective spreads. We adopt a ratio of effective to quoted spreads as such a metric. In Table 2, this ratio is 0.76 for wholesalers, suggesting that orders executed by them pay 76% of the prevailing quoted spread and 0.97 for exchanges. Wholesalers, therefore, appear to play a valuable role within the existing market structure. They provide substantial,

[18] In a report titled "The good, the bad, and the ugly of payment for order flow," (https://bit.ly/3EAVeTk) BestEx Research criticizes assertions that wholesalers offer substantial price improvement off of NBBO. It highlights the limitation of NBBOs as benchmarks, which do not account for hidden orders and odd lots. We circumvent this criticism using effective spreads that account for both price improvement from wholesalers and price improvement achieved through exchange odd lots and hidden orders.

rather than *de minimis*, price improvement that may not be available from the exchanges when retail trades are executed. In fact, if we simply add up the dollars of price improvement provided by wholesalers across stocks and months in our sample, it amounts to $400 million on average per month.

Market structure commentators occasionally argue that while price improvement provided by wholesalers is small, PFOF payments made by wholesalers 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. We turn to the wholesaler costs next.

[Figure 1]

As we discuss above, market structure studies typically distinguish between two components of the effective spread. One such component is price impact which captures the adverse selection cost associated with a trade. The other is the realized spread that reflects three important market making considerations: inventory costs, fixed costs, and profits. Table 2 confirms our earlier assertion that wholesalers obtain order flow that is considerably less toxic (price impact of 36.80 bps) than that routed to exchanges (price impact of 52.73 bps). These figures are consistent with the statements by the industry participants that retail order flow is predominantly routed to wholesalers, while exchanges end up receiving mainly institutional flow.

Given similar effective spreads and lower price impacts, wholesalers earn substantially larger realized spreads compared to those earned by exchanges, 16.36 vs. -1.42 bps. At first glance, this large difference may appear suggestive of excess profits earned by wholesalers; however, it is important not to over-interpret these figures. Liquidity on exchanges is only partially provided by professional market makers. For instance, Nasdaq attributes only 16% of liquidity provision to

pure market making strategies. The remaining liquidity-providing orders are submitted by non-market markers, whose main goal is to manage positions rather than earn spread revenue. The realized spreads that non-market makers earn are, therefore, not reflective of market making costs and profits. Since non-market makers' share of exchange liquidity provision is significant, caution should be used when comparing exchange realized spreads to wholesaler realized spreads. Put differently, the 16.36 bps realized spread earned by wholesalers may represent either a substantial profit or a combination of inventory and fixed costs that allows only for a zero profit or anything in-between. We examine this issue in more detail later in the manuscript.

3.2 Cross-Sectional Differences

Because we have a large cross-section, including many illiquid securities, we separately examine four sub-samples, the S&P 500 and size-based terciles of non-S&P 500 stocks labeled Tercile 1, Tercile 2, and Tercile 3. During the sample period, there are 514 stocks in the S&P 500 sub-sample[19] and the Tercile 1, Tercile 2, and Tercile 3 sub-samples include 2,550, 2,550, and 2,551 stocks respectively. In this section, we investigate if wholesaler involvement and execution quality differ between the four sub-samples.

Table 3 shows noticeable differences across sub-samples. Wholesalers represent 32.11% of share volume for S&P 500 stocks, but their share increases monotonically in size, reaching a high of 63.94% for Tercile 3 stocks. In other words, retail flow plays an out-sized role for less liquid stocks, a point we will return to below.

[Table 3]

Prior market structure literature has linked execution quality to several market characteristics. Among these are price, trading volume, and volatility. A higher price is typically related to

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

lower execution costs because of the fixed tick size in the U.S. Greater volatility is typically associated with greater fundamental information flows and may negatively affect execution quality through the adverse selection channel. With volatility controlled for, a greater volume is typically associated with lower adverse selection as it is thought to represent uninformed flow. In Table 4, we examine how execution quality differs between orders routed to wholesalers and exchanges while controlling for these characteristics in the following regression model:

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

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

We estimate equation (1) for the full sample in Panel A of Table 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 hold. Wholesaler executions tend to occur when quoted spreads are relatively wide. For example, the quoted spreads that prevail during wholesaler executions are 15.54 bps wider than those that prevail during exchange executions. For comparison, the univariate results in Table 2 suggest that this difference is 17.78 bps. Price improvements offered by wholesalers are 27.8% larger, but the effective spreads facing retail investors are still slightly larger, by 1.81 bps. These results confirm our earlier assertion that due to differences in quoted spreads that prevail at the time of wholesaler and exchange

executions, effective spreads are not the optimal execution quality comparison metric. With this in mind, we omit effective spreads from subsequent discussions. Finally, price impacts are 16.90 bps lower and realized spreads are 18.71 bps higher for orders routed to wholesalers.

[Table 4]

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

In all sub-samples, the data confirm that wholesalers provide greater price improvement compared to exchanges. For S&P 500 stocks, Panel B shows that the difference between exchange and wholesaler effective-to-quoted spread ratios is 0.44, a 44 percentage points larger price improvement relative to the quoted spread. As noted earlier, wholesaler price improvements decline as we move from large to smaller size firms. Still, even for tercile 3, we estimate that price improvements are 20%

Finally, we confirm for all four sub-samples that toxicity of wholesaler-bound flow is lower than that of the exchange-bound flow and that wholesalers earn larger realized spreads. For instance, column [4] in Panel B shows that price impacts for wholesalers in S&P 500 stocks are 2.73 bps lower than for their exchange counterparts, whereas the realized spreads earned by wholesalers are 1.91 bps greater than those earned by exchange liquidity providers. The corresponding numbers for tercile 3 stocks are a 40.16 bps (= -2.731 - 37.425) lower price impact and a 43.81 bps (= 1.911 + 41.897) greater realized spread. We note that although the realized spreads ob-

tained 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 therefore refrain from linking these figures to excessive profits earned by wholesalers and return to inventory costs shortly.

So far, we have shown that retail order execution quality varies across the sub-samples of stocks. Yet the data allow for an even more detailed examination. Rule 605 reports are filed by individual venues, and therefore we are able to examine execution quality across wholesalers. To keep this analysis manageable, we divide wholesalers into two groups, the *top 2*, which includes Citadel and Virtu, and the *others*. Recall that Table 1 shows that these two firms are considerably larger than their peers and execute 70% of marketable order flow that is routed to wholesalers.

In Table 5, we use panel regressions to ask if execution quality is systematically different for the $top2$ compared to the $other$ wholesalers overall (Panel A) and for the sub-samples (Panel B). The regressions are of the following form:

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

where $DepVar_{ijt}$ is one of the following execution quality variables for stock i wholesaler group j (top2 vs. the rest) in month t: the ratio of effective to quoted spread, price impact, and realized spread as defined previously; $top2$ is a dummy variable that has a value of 1 for orders executed by Citadel and Virtu, and 0 for orders executed by other wholesalers; *price* is the natural log of the stock price; $volatility$ is the difference between the high and low prices scaled by the high price, and *volume* is the natural log of CRSP trading volume. The regression controls for stock and month fixed effects and uses double-clustered standard errors. Note that we only use wholesaler data for these regressions.

The results show that price improvement is roughly the same for the two groups in the overall

sample. However, $top2$ wholesalers face more toxic order flow (the difference is 2.68 bps) and earn lower realized spreads (a difference of -1.70 bps). We explore differences across sub-samples in Panel B, where we augment regression (3) by adding interaction variables between $top2$ and tercile dummies. The coefficient on $top2$ shows that Citadel and Virtu 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 the price improvements are 0.3 percentage points (= 0.070 - 0.067) greater for $top2$ than for other wholesalers. Toxicity facing $top2$ in tercile 3 stocks is 6.9 bps (= 0.675 + 6.214) greater, and realized spreads are 5.0 bps (= -0.210 - 4.754) lower than for other wholesalers trading the same stocks.

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

We run the following regression to explore whether wholesaler economies of scale can explain the $top2$ wholesaler's ability to charge lower realized spreads:

$$realized\ spread_{ijt} = \alpha + \beta_1 top2_j + \beta_2 price_{it} + \beta_3 volatility_{it} + \beta_4 share_{ijt} + \varepsilon_{ijt}, \tag{3}$$

where $realizedspread_{ijt}$ is realized spread as defined previously for stock i wholesaler group j (top2 vs. the rest) in month t; $top2$ 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 *share* is the natural log of retail volume captured by wholesalers in group j in stock i divided by overall trading volume in stock i. 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]

Table 6 confirms our expectations that a wholesaler's operation size is associated with significantly lower realized spreads for all sub-samples except for S&P 500 stocks. Importantly, including this variable as a control renders the coefficient on $top2$ insignificant for all sub-samples. This evidence is consistent with the top two wholesalers being better able to lower inventory and fixed costs due to the size of their operations and as a result being able to offer lower realized spreads.

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

3.3 Broker Routing and Wholesaler Performance

Retail brokerages are a key source of order flow for wholesalers (Figure 2). The four largest retail brokers that we see in the data – TD Ameritrade, Robinhood, E*Trade, and Schwab – route to at least six of our eight wholesalers during the sample period. The Appendix explains how we calculate routed volumes based on Rule 606 data. Not surprisingly, given the market shares reported earlier, the majority of order flow goes to Citadel and Virtu. Other brokers use very different routing tables. Inspection of the time series suggests that there is some variation 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 on which wholesaler to route to, and how much to route to a particular wholesaler? Industry participants suggest that retail brokerages regularly evaluate wholesaler performance.[20] Such evaluations typically occur on a monthly basis.[21] We propose that if the market for retail order flow is competitive, brokerages should adjust their routing to favor wholesalers with better past performance.

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

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

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

Specifically, we estimate the following regression:

$$\begin{aligned} mkt.\ share_{ijt} = \alpha + \beta_1 stock\text{-}specific\ realiz\ spr_{ijt-1} + \beta_2 portfolio\ realiz\ spr_{jt-1} \\ + \beta_3 price_{it} + \beta_4 volatility_{it} + \beta_5 volume_{it} + \varepsilon_{ijt}, \end{aligned} \tag{4}$$

where $mkt.\ share_{ijt}$ is the market share of volume in stock i executed by wholesaler j in month t expressed as the deviation from the geometric mean across all wholesalers; $stock-specific\ realiz\ spr_{ijt-1}$ is the average realized spread earned in stock i by wholesaler j in month $t-1$ expressed as the deviation from the arithmetic mean across all other wholesalers; $portfolio\ realiz\ spr_{jt-1}$ is the average realized spread earned by wholesaler j in all stocks routed to it in month $t-1$ expressed as a deviation from the arithmetic mean across all other wholesalers; *price* is the natural log of the stock price; *volatility* is the difference between the high and low prices scaled by the high price, and *volume* is the natural log of CRSP trading volume. The realized spread variables are scaled, so the economic significance corresponds to basis points. We run these regressions for the full sample and then separately for each sub-sample using stock, wholesaler, and month fixed effects and clustering standard errors by stock and month.

Table 7 shows that if wholesaler j offers a relatively low realized spread across all stocks, retail brokerages will reward the wholesaler with a greater market share next month. This result holds both for the full sample and for all sub-samples. The full-sample coefficient indicates that a one basis point reduction in a wholesaler's realized spread relative to the mean is associated with a 3.1% greater market share for the full sample and between 2.8 and 3.2% greater market share for the sub-samples.

[Table 7]

In addition, a relatively low realized spread for wholesaler j in stock i is associated with a significantly greater market share next month for the full sample and for stocks in terciles 2

and 3. However, the economic magnitude of this effect is relatively small. Taken together, the evidence in Table 7 is consistent with wholesalers competing by offering lower liquidity costs across all stocks as opposed to on a security-by-security basis, and retail brokerages playing pivotal roles in ensuring the best-performing wholesalers are awarded more order flow. As such, even though wholesalers have the right of first refusal when handling retail flow (meaning they are not obligated to internalize every transaction), the regular execution quality reviews compel them to strive for overall excellence in serving retail customers.

3.4 A Competitive Shock

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

Panel A of Figure 3 illustrates Jane Street's entry and market share growth over time. Jane Street enters into the wholesale business in 2019, but throughout 2020 the firm still has a very small market share. Its market share increases gradually in the late summer of 2021, reaching a substantial level by October 2021.[22] 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), Jane Street 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

[22]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.

wholesaler's realized spreads. A month after entering the business, Jane Street starts offering realized spreads similar to those of its competitors, but from July 2021, it competes considerably more aggressively, offering realized spreads that are 40-50% lower than the incumbents. The figure illustrates that a sizable increase in market share ensues.[23]

To evaluate whether the entry of Jane Street results in lower average realized spreads offered by wholesalers relative to exchanges, we run a difference-in-differences regression of wholesalers against exchanges with the pre-period being April-June 2021, when Jane Street still has a small market share, and the post-period being the last three months of 2021, during which Jane Street has already established itself as a sizeable wholesaler.

Table 8 reports the results from running the following regression:

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

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

[23]The data show that Jane Street's market share stabilizes in 2022, yet the firm continues offering relatively low realized spreads. An examination of the cross-section reveals that Jane Street offers low spreads exclusively in tercile 2 and 3 stocks. In S&P 500 and tercile 1 stocks, it charges realized spreads comparable to 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.

[Table 8]

Based on the β_2 coefficients in Table 8 Panel A, we do not find that realized spreads for wholesalers relative to exchanges decline following Jane Street's entry for any sub-sample. For the index and tercile 1 stocks, the entry did not cause realized spreads to change, consistent with the possibility that they were already at a competitive level. For tercile 2 and 3 stocks, realized spreads actually increase.

Panels B and C of Table 8 report regression results for the incumbent wholesalers and Jane Street separately. Recall that Jane Street was present but at a much lower market share in our pre-period. The results show that the incumbents did not change their realized spreads for the largest stocks but significantly increased the spreads for tercile 1-3 stocks after Jane Street's entry. By contrast, Jane Street, which already offered lower costs in terciles 1 and 2 compared to the other wholesalers, significantly reduced its costs for tercile 3 stocks.

To understand this result, consider first what happens when Jane Street enters. Order flow is now divided among a larger number of wholesalers resulting in less flow for each of the incumbents. For instance, Citadel and Virtu lose more than 11% of their flow to Jane Street, while smaller wholesalers generally lose even more. With less flow, the incumbents' inventory costs likely increase. This may be particularly true for less liquid securities such as those in terciles 1-3. By contrast, the S&P 500 securities are highly liquid, and this likely makes it easier for the incumbents to manage inventory risk, reducing the need to increase liquidity costs. We return to the topic of inventory risk in the next section. Finally, note that Jane Street appears to be competing particularly fiercely in tercile 3 stocks, where it lowers its realized spreads significantly, cutting them by almost fifty percent.

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, Tastyworks, and Robinhood – are routing

to Wolverine. In the meantime, Wolverine charges realized spreads that are 42.5% greater than those of the incumbents. Perhaps not surprisingly, Ally drops Wolverine by April 2020 (likely due to a contract expiry), 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 declined, 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.

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

In this context, it is interesting 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, the algorithms settle on non-competitive prices, i.e., they overcharge for liquidity. Is it possible that wholesalers do the same? While we cannot dismiss this possibility outright, two factors make us think that it is relatively unlikely. Firstly, Colliard, Foucault, and Lovo (2022) observe collusive behavior primarily when only two algorithms are present. Increasing the number of competing algorithms beyond two reduces and eventually eliminates overcharging, 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. Secondly, the monitoring by retail brokerages is likely to reduce the ability of wholesalers to charge non-competitive prices.

3.5 Inventory Costs

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

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

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

where $\textit{realized spread}_{jt}$ is the realized spread in stock i in month t; $T1$, $T2$, and $T3$ are dummies indicating whether a stock is in size-based tercile 1, 2, or 3 of non-S&P 500 stocks; *price* is the natural log of the stock price; $\textit{volatility}$ is the difference between the high and low prices scaled by the high price; and *retail volume* is volume as reported in 605 data. We also run a specification with *total volume*, defined as the natural log of CRSP trading volume, replacing retail volume. Because the tercile dummies are unique for each security, the regressions control only for month fixed effects, and we use two-way clustered standard errors.

[Table 9]

Column [1] of Table 9 reports the tercile results without the controls and, as expected, shows that tercile 1, 2, and 3 stocks have significantly greater realized spreads than S&P 500 stocks.

When we control for price and volatility in column [2], there is no longer a significant difference between S&P 500 stocks and tercile 1 stocks, but realized spreads for the remaining size terciles are still significantly higher. Column [3] includes retail volume, which is our first proxy for inventory 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 significantly higher. Finally, column [4] accounts for the possibility that wholesaler inventory management may not rely solely on retail volumes and replaces retail volume with total volume. This is our second proxy for inventory costs, and including it makes all the coefficients on the size terciles turn negative and significant. Note also that the coefficients on retail and CRSP volume are highly significant and negative, as predicted. After controlling inventory costs, we conclude that wholesalers earn significantly lower realized spreads for less liquid stocks than they do for the S&P 500 stocks. This result once again points to a cross-subsidy from large stocks to small stocks that arises due to the portfolio approach used by retail brokers to evaluate wholesaler performance.

van Kervel and Yueshen (2023) propose a model in which wholesalers quote wider spreads on exchanges, aiming to collect greater profits from internalized flow whose executions reference exchange prices. They argue that such anticompetitive behavior will be most evident in stocks with a lot of internalization, specifically small ones. While we cannot directly test their predictions, our results suggest that broker monitoring of retail execution quality tends to counteract potential anticompetitive behavior.

4. Market Structure Proposals

Our data show that wholesaler intermediation delivers good execution quality for retail investors, but could it be even better 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 have to make assump-

tions that the reader may or may not agree on to conjecture the outcome of such a change. 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 allow us to study the proposal in detail, but the cross-sectional evidence reported above, coupled with the assessed institutional interest that we report below, suggests that caution is in order.

4.1 Moving Retail to Exchanges

In Table 2 above we identified two important differences between wholesaler- and exchange-intermediated executions. On the one hand, wholesalers provide sizeable price improvement, while exchange price improvement is noticeably smaller. On the other hand, exchange liquidity providers earn considerably lower realized spreads compared to wholesalers. With these differences in mind, what would happen if retail order flow were to be moved to the exchanges? To grasp the potential impact of this move, it is helpful to consider its positive and negative effects. On the positive side, the retail flow will pay considerably less in realized spreads. However, there is also a negative effect to consider. When combined with institutional flow, retail investors will bear the cost of the resulting mix's higher toxicity, which exceeds that of pure retail flow. The sum of these two effects is negative for retail investors, meaning they would be worse off if their order flow were routed to exchanges. We elaborate below.[24]

Consider three groups of investors that trade U.S. equities: retail traders, institutions using held orders, and institutions using not-held orders. We have detailed data at the stock-month level for the first two groups, but Rule 605 does not report market quality for the third group. However, we can back out the data we need from the data we have by making a few assumptions.

First, we assume that not-held institutional volume at the stock-month level is CRSP volume

[24] A report by BestEx Research that we referenced earlier also examines the potential implications of shifting retail flow to exchanges. BestEx, however, limits its analysis to showing that exchange spreads are likely to decrease as a result of pooling low-toxicity retail flow with institutional flow. Consequently, it does not examine the entirety of the consequences of such a move on retail investors.

minus the sum of retail volume and held institutional volume. Second, we assume that retail investors trade smoothly throughout the day, so that the quoted spread we observe when they trade is the average exchange quoted spread, whereas we know from our analysis that institutions using held orders time their trades to when quoted spreads are narrow. Third, the average exchange quoted spread is equal to the volume-weighted average of the quoted spread when held institutional orders execute and the quoted spread when not-held institutional orders execute (which we do not observe). This means that the quoted spread when not-held institutional orders execute has to be larger than the quoted spread when retail orders execute. It can easily be backed out from the data. Finally, we assume that the current (low) liquidity costs (realized spread) and (minimal) price improvement facing held institutional orders also apply for not-held institutional orders. This set of assumptions enables us to compute the share-volume-weighted-average toxicity (price impact) for institutional order flow on exchanges at the stock-month level.

Now consider what would happen if retail orders were routed to exchanges. Assuming liquidity provision is competitive, we expect exchange quoted spreads to adjust so that they cover the toxicity of the combined retail and institutional order flows (held and not-held). Furthermore, we assume that retail orders will enjoy the current (low) realized spreads faced by institutional traders.

This set of assumptions enables us to compute the change in total trading costs facing retail traders when their orders are pooled with institutional orders, and compare it to the current total trading costs facing retail traders when their orders are routed to wholesalers. In this calculation, we define total trading costs as share volume times the dollar effective spreads summed across stocks and months. We can also compute the percent change in the share-weighted average dollar effective spreads. Changes in total institutional trading costs and percent changes in dollar effective spreads are computed analogously.

Our calculations show that retail investors in our sample of equities would lose $221 million per month if their orders were to be routed to exchanges. They also face an increase in effective

spreads of 38.5%. As mentioned above, the explanation is that although retail investors benefit from the lower realized spreads in the pooled setting, this is more than compensated for by the higher toxicity they have to pay for when their order flow is pooled with that of institutions. By contrast, institutional investors gain a total of $518 million per month because the toxicity they have to pay for in the pooled setting is lower than it is currently, and their effective spreads decline by 9.3%.

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 the 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 institutional volume is typically accompanied by additional liquidity coming off the sidelines and improving market quality. It is important to note that even if realized spreads were to decline when retail flow moves to exchanges, this decline would need to be exceptionally large to outweigh the additional costs retail traders face because of having to pay for higher toxicity.

Our calculations omit two additional possibilities. First, consider what would happen if institutional order flow patterns do not change in response to the arrival of retail traders. We assume that retail investors when mixed with institutional flow trade in proportion to held and not-held institutional volumes. This means that the reduction of toxicity is larger when institutions using not-held orders trade. From the perspective of exchange liquidity providers, this means that retail orders offer additional benefits because they help lower average adverse selection costs exactly when these costs would otherwise be high. With competitive liquidity provision on exchanges, quoted spreads may fall further. To the extent that this effect is significant, our calculation may underestimate the benefits to retail from moving to exchanges.

Second, our calculations omit the possibility that institutions may strategically respond to the arrival of retail flow. Retail trades disproportionately reduce quoted spreads for institutions using not-held orders (high adverse selection), implying that these institutions may respond by trading more raising adverse selection costs (Kyle (1985)), resulting in an increase in pooled effective spreads. If quoted spreads facing institutions using held orders also fall when low toxicity retail flow is mixed in, they may also trade more pushing pooled effective spreads further upwards. If both types of institutional traders increase their activity in response to lower quoted spreads, our base calculations may underestimate the cost to retail from moving to exchanges.

4.2 Order-by-Order Competition

In December 2022, the SEC proposed rules that would significantly change the equity markets. To analyze these comprehensive rules is beyond the scope of the current study, but we believe our results may shed some light on SEC's conjecture that retail traders would be better off if there were order-by-order competition for retail orders (labeled segmented orders in the rule) as envisioned in the proposed Order Competition Rule. In a nutshell, the rule proposes a requirement that segmented orders be forwarded, either by the retail broker directly or by the wholesaler receiving retail order flow, to auctions run by exchanges and/or certain ATSs where institutions can interact with the order flow.[25] The SEC believes that since retail order flow has lower toxicity (as we document above), it should get a larger price improvement than what is currently offered by wholesalers and that institutions would be willing to trade with retail at the NBBO midquote.

How realistic is this proposal, and how would it impact the cross-section of stocks currently managed by wholesalers? The answer depends on whether there is institutional interest in trading the stocks favored by retail investors. On this matter, the Commission itself is not entirely

[25] See Ernst, Spatt, and Sun (2023) and van Kervel and Yueshen (2023) for theoretical analyses of the auction proposal.

convinced. For instance, Commissioner Hester Peirce has expressed concern that "institutional investors may not expend much effort to participate regularly in auctions."[26] We document in Table 3 that wholesalers currently execute the bulk of retail share volume in less liquid stocks (Tercile 2 and Tercile 3). Is there sufficient institutional interest to do without the intermediation offered by wholesalers for these stocks?

To answer this question, we estimate institutional trading in the sample stocks based on changes in reported quarterly holdings from 13F reports and add to that changes in short interest (which are available bi-monthly). To account for intra-quarter trading, we gross up institutional volume inferred from 13F reports by a factor of 1.17 based on Chakrabarty, Moulton, and Trzcinka (2017). This gives us a proxy for institutional trading volume in a particular stock. We then calculate the ratio of retail trading as reflected in Rule 605 data divided by our proxy for institutional volume. Table 10 reports the across stock means, medians, and quartiles for each sub-sample, that is S&P 500, Tercile 1, Tercile 2, and Tercile 3 stocks.

[Table 10]

Column [1] ([2]) shows that average (median) retail order flow represents 85% (20%) of institutional volume for S&P 500 stocks, so for index stocks, there is significant institutional interest. Importantly, as we move to less liquid stocks, it becomes clear that retail order flow swamps institutional trading interest. Already for Tercile 1 stocks, institutional trading interest starts becoming insufficient on average as the ratio of retail to institutional interest exceeds one. For Tercile 2 stocks, retail interest is more than double the institutional interest, and for Tercile 3 stocks, the average retail order flow is almost ten times larger than institutional interest. The ratios are highly skewed, suggesting that retail interest tends to be focused on particular stocks and that institutions do not favor those stocks. If we switch our attention to the median values for a more conservative view, we observe that institutional interest is substantially below retail

[26] "Statement on Ordering Competition" by Hester M. Peirce, December 14, 2022 (https://bit.ly/3qSe3OR).

interest only for Tercile 3.

We conclude that institutional trading interest may be low for some of the cross-section of securities traded by retail investors. At best, the effect of the proposed auctions for these securities would be to delay executions. However, the auctions could actually have even more detrimental consequences for retail investors in less liquid stocks. Our earlier results suggest that realized spreads may be insufficient to cover inventory costs for less liquid stocks in the current environment and that wholesalers may cross-subsidize small stocks with their large-stock revenues. If that is indeed the case, and wholesalers end up losing a significant fraction of order flow in liquid stocks through the proposed auctions, they may be unable to offer price improvements at the level we observe today for less liquid stocks. In other words, we could see execution quality deteriorate for some of the universe of securities retail investors currently trade.[27]

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 over its impact on retail investors, 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 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, showing that it is the retail brokerages who have the power in this ecosystem. The brokerages are large, with the largest brokerage surpassing the

[27]The U.S. Congress expressed concern about the lack of liquidity and the resulting difficulties in raising capital for emerging growth firms in its 2012 Jumpstart Our Business Startups (JOBS) Act. Our results suggest that the abandonment of the status quo bundling approach may lead to a deterioration in small firm liquidity, which is inconsistent with the spirit of the Act.

largest wholesaler in size. They continuously monitor the performance of wholesalers, rewarding the best performers with increased order flow, while reducing allocations to the underperformers.

Furthermore, the wholesaler environment is characterized by economies of scale. The largest wholesalers are able to provide liquidity at a reduced cost, and the brokerage oversight ensures that these cost savings benefit 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. Notably, upon the entry of this new wholesaler, retail customer trading costs in large stocks remain unchanged, while trading costs in small stocks increase. This result is inconsistent with the notion that incumbents were extracting excessive rents prior to entry.

We also discuss the two alternatives to the status quo currently under consideration. 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 expose them to significantly 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. Firstly, institutional investors are unlikely to participate in auctions for thousand small stocks, as institutional trading interest in such stocks is notably lower than retail trading interest. Secondly, 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. 6, 7

Aquilina, M., E. B. Budish, and P. O'Neill, 2021, "Quantifying the high-frequency trading "arms race": A simple new methodology and estimates," *Quarterly Journal of Economics,* forthcoming. 3

Baldauf, M., J. Mollner, and B. Z. Yueshen, 2023, "Siphoned Apart: A Portfolio Perspective on Order Flow Segmentation," *Available at SSRN 4173362*. 7

Barber, B. M., X. Huang, P. Jorion, T. Odean, and C. Schwarz, 2022, "A (Sub) penny For Your Thoughts: Tracking Retail Investor Activity in TAQ," *Available at SSRN 4202874*. 7

Barber, B. M., X. Huang, T. Odean, and C. Schwarz, 2022, "Attention-induced trading and returns: Evidence from Robinhood users," *The Journal of Finance*, 77(6), 3141–3190. 6

Bartlett, R. P., J. McCrary, and M. O'Hara, 2022, "The Market Inside the Market: Odd-Lot Quotes," *Available at SSRN 4027099*. 11

Battalio, R., and R. Jennings, 2023, "Why do Brokers who do not Charge Payment for Order Flow Route Marketable Orders to Wholesalers?," *Available at SSRN 4304124*. 6, 8

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. 31

Boehmer, E., R. Jennings, and L. Wei, 2007, "Public disclosure and private decisions: Equity

market execution quality and order routing," *The Review of Financial Studies*, 20(2), 315–358. 21

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

Brogaard, J., and C. Garriott, 2019, "High-frequency trading competition," *Journal of Financial and Quantitative Analysis*, 54(4), 1469–1497. 26

Brogaard, J., B. Hagströmer, L. Nordén, and R. Riordan, 2015, "Trading fast and slow: Colocation and liquidity," *Review of Financial Studies*, 28(12), 3407–3443. 11

Bryzgalova, S., A. Pavlova, and T. Sikorskaya, 2022, "Retail Trading in Options and the Rise of the Big Three Wholesalers," *Working paper*, London Business School. 8

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. 33, 49

Colliard, J.-E., T. Foucault, and S. Lovo, 2022, "Algorithmic Pricing and Liquidity in Securities Markets," *HEC Paris Research Paper*. 26

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. 31

Eaton, G. W., T. C. Green, B. Roseman, and Y. Wu, 2022, "Retail Trader Sophistication and Stock Market Quality: Evidence from Brokerage Outages," *Journal of Financial Economics*, forthcoming. 7, 8

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

Ernst, T., C. S. Spatt, and J. Sun, 2023, "Would Order-by-Order Auctions Be Competitive?," *Available at SSRN*. 6, 32

Foley, S., A. Liu, K. Malinova, A. Park, and A. Shkilko, 2023, "Cross-subsidizing liquidity," *Working Paper*, Macquarie University. 4

Hendershott, T., C. M. Jones, and A. J. Menkveld, 2011, "Does algorithmic trading improve liquidity?," *Journal of Finance*, 66, 1–33. 11

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

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

Jain, P. K., S. Mishra, S. O'Donoghue, and L. Zhao, 2022, "Trading Volume Shares and Market Quality: Pre-and Post-Zero Commissions," *Working paper*, University of Memphis. 6, 58

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

Kyle, A. S., 1985, "Continuous Auctions and Insider Trading," *Econometrica*, 53, 1315–1336. 32

Lee, C. M., and M. J. Ready, 1991, "Inferring trade direction from intraday data," *Journal of Finance*, 46, 733–746. 11

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

O'Hara, M., 2015, "High frequency market microstructure," *Journal of Financial Economics*, 116(2), 257–270. 3

Schwarz, C., B. M. Barber, X. Huang, P. Jorion, and T. Odean, 2022, "The'Actual Retail Price'of Equity Trades," *Available at SSRN 4189239*. 8

van Kervel, V., and B. Z. Yueshen, 2023, "Anticompetitive Price Referencing," *Available at SSRN 4545730*. 28, 32

Welch, I., 2022, "The wisdom of the Robinhood crowd," *The Journal of Finance*, 77(3), 1489–1527. 6

Table 1
Market Shares

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

	venue type	shares executed, bil.	mkt. share, %
Panel A: by venue type			
	EXCH	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 held liquidity-demanding orders. We compute the statistics separately for orders executed by wholesalers (WHOL) and exchanges (EXCH). We report the average number of shares executed and the average stock price in a sample stock during the sample period, followed by the percentage share of shares that are price improved or executed at or better the corresponding NBBO. 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, also in basis points. All variables 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]
# shares, mil.	172.80	252.86	***
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 held liquidity-demanding orders for wholesalers and exchanges, with the sample divided into S&P 500 and size-based terciles of non-S&P 500 stocks labeled Tercile 1, Tercile 2, and Tercile 3.

	S&P 500	Tercile 1	Tercile 2	Tercile 3
	[1]	[2]	[3]	[4]
WHOL	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 of the table reports coefficient estimates from market quality regressions of the following form:

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

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

	eff. spr. / quot. spr.	quoted spr.	effective spr.	price imp.	realized spr.
	[1]	[2]	[3]	[4]	[5]
Panel A: Base Specification					
WHOL	-0.278***	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: Regressions

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

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

where $DepVar_{ijt}$ is one of the following market quality variables for stock i wholesaler group j in month t: the ratio of effective to quoted spreads, quoted spread, effective spread, price impact, and realized spread as defined previously; $top2$ is a dummy variable that has a value of 1 for orders executed by Citadel and Virtu and 0 for orders executed by other wholesalers; *price* is the natural log of the stock price; *volatility* is the difference between the high and low prices scaled by the high price, and *volume* is the natural log of CRSP trading volume. In the last specification, we add the natural log of retail volume executed by a wholesaler to control for the wholesaler's ability to manage inventory internally. Panel B augments the specification by including interaction terms between the $top2$ dummy and indicator variables for the size-based terciles of non-S&P 500 stocks; Tercile 1 ($T1$), Tercile 2 ($T2$), and Tercile 3 ($T3$). The models are estimated with stock and month fixed effects, and the standard errors are double-clustered across stocks and months. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	eff. spr. / quot. spr.	quoted spr.	effective spr.	price imp.	realized spr.
	[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

We estimate the following regression:

$$realized\ spread_{ijt} = \alpha + \beta_1 top2 + \beta_2 price_{it} + \beta_3 volatility_{it} + \beta_4 share_{ijt} + \varepsilon_{ijt},$$

where $realizedspread_{ijt}$ is the realized spread in basis points for stock i wholesaler group j in month t; $top2$ 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 $share_{ijt}$ is the natural log of retail volume in stock i in month t captured by wholesalers in group j divided by stock i total volume in month t. The realized spread variables are scaled, so the economic significance corresponds to basis points. We run these regressions for the full sample and then separately for each sub-sample, use stock and month fixed effects, and cluster standard errors by stock and month. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	Full sample	S&P 500	Tercile 1	Tercile 2	Tercile 3
	[1]	[2]	[3]	[4]	[5]
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)
share	-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: Regression

We estimate the following regression:

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

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

	Full sample	S&P 500	Tercile 1	Tercile 2	Tercile 3
	[1]	[2]	[3]	[4]	[5]
stock-specific realiz spr$_{ij}$	-0.000***	-0.000	-0.000*	-0.000***	-0.000*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
portfolio realiz spr$_j$	-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 reports coefficient estimates from the following regression:

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

where $realized\ spread_{it}$ is the realized spread in stock i in month t; $WHOL$ is a dummy variable that has a value of 1 for orders executed by wholesalers and 0 for orders executed by exchanges; *POST* is a dummy variable that has a value of 1 after the Jane Street entry (Panel A), *price* is the natural log of the stock price; *volatility* is the difference between the high and low prices scaled by the high price, and *volume* is the natural log of CRSP trading volume. In Panel A, the model is estimated for all wholesalers, while in Panels B and C, it is estimated separately for the incumbents and Jane Street. The models are estimated with stock and month fixed effects, which is why the standalone *POST* variable is omitted. The standard errors are double-clustered across stocks and months. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	S&P 500	Tercile 1	Tercile 2	Tercile 3
	[1]	[2]	[3]	[4]
Panel A: All Wholesalers				
WHOL	1.168***	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:

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

where $\textit{realized spread}_{jt}$ is the realized spread in stock i in month t; $T1$, $T2$, and $T3$ are dummies indicating whether a stock is in size-based tercile 1, tercile 2, or tercile 3 of non-S&P 500 stocks; *price* is the natural log of the stock price; *volatility* is the difference between the high and low prices scaled by the high price; and *retail volume* is the natural log of retail volume. In column [4], we replace *retail* with *total volume*, defined as the natural log of CRSP trading volume. The regressions control for month fixed effects, and we use two-way clustered standard errors. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	[1]	[2]	[3]	[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 Interest

The table reports descriptive statistics for stock-quarter ratios of Rule 605 liquidity-demanding volume (retail 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 (which are available bi-monthly). 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

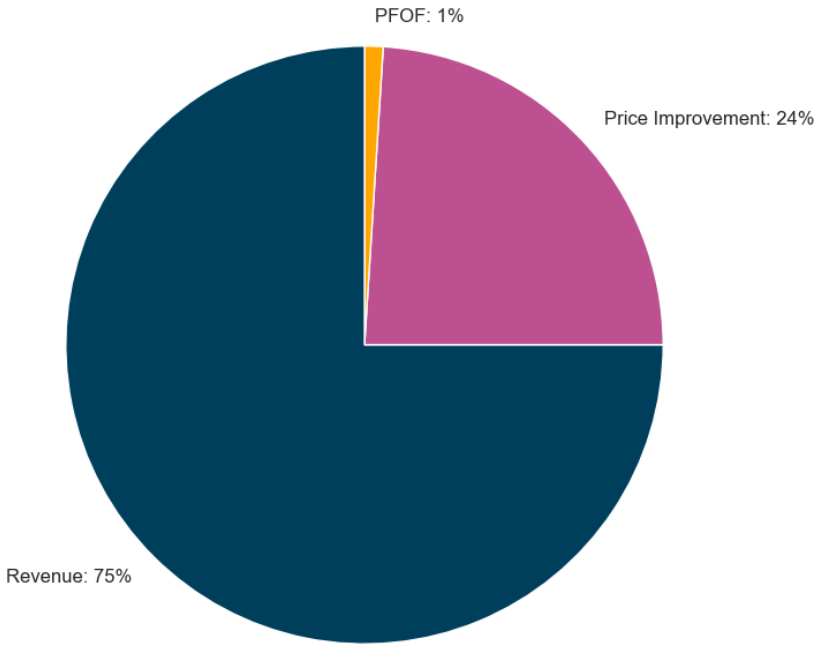

Figure 1. Allocation of Spread Revenue
The figure depicts the allocation of revenue obtained from quoted spreads between price improvement provided to retail investors, payments for order flow from wholesalers to retail brokers, and the portion of revenue retained by wholesalers. The retained revenue 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.

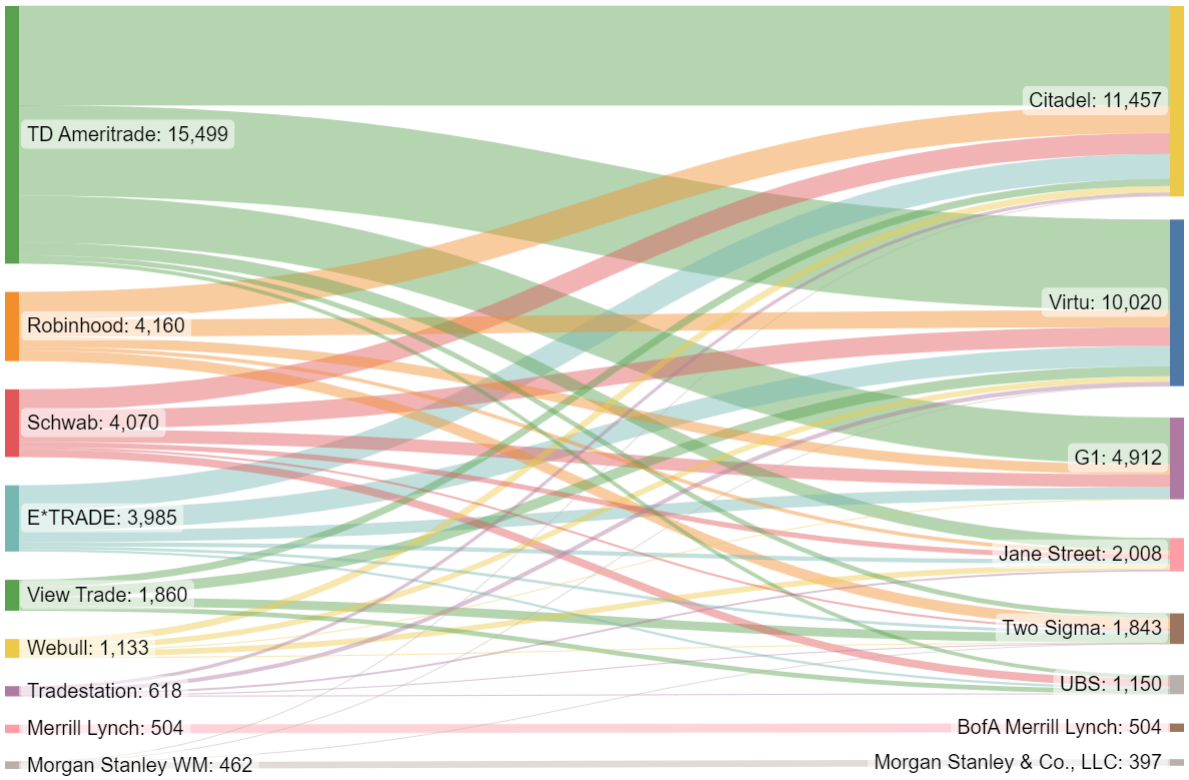

Figure 2. Retail Broker Routing

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 this 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 will not perfectly align with those in the main sample.

Figure 3. Wholesaler Entry and Exit

This figure reports the market share (left axis) and the realized spread ratio (right axis) for Jane Street in Panel A and Wolverine in Panel B. Market share is the percentage of all wholesaler volume, and the realized spread ratio is the ratio of Jane Street and Wolverine 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.

Appendix

A.1 Data details

We obtain our data from a service provider that specializes in compliance and trade analytics. The Rule 605 data we have access to cover 70 market centers for January 2016 - December 2022. While our service providers' Rule 605 data coverage is extensive it is not complete. To patch the missing data, we download Rule 605 data directly in order to add NYSE National (XCIS) and missing months.[28]

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

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

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

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

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

- *Executed shares (EXshs)* are the cumulative number of shares executed at the receiving market center.
- *Away executed shares (AWshs)* are the cumulative number of shares executed at another venue.
- *Average realized spread ($RS* is the share-weighted average spread in dollars using a five minute horizon.[29]
- *Average effective spread ($ES)* is the share weighted average in dollars.
- *Price improved shares (PIshs)* is the cumulative number of shares executed with a price improvement.
- *Price improved average amount ($PI)* is the per share share-weighted average dollar amount that prices were improved.
- *At the quote shares (AQshs)* is the cumulative number of shares executed at the quote.
- *Outside the quote shares (OQshs)* is the cumulative number of shares executed outside the quote.
- *Outside the quote average amount ($OQ)* is the per share share-weighted average dollar amount that prices were outside the quote.

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

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

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

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

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

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

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

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

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

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

A.2 Execution Quality Statistics for Sub-samples

In Table A1, we ask if market capitalization affects execution quality. 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 substantially 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% for tercile 2). Note also that the fraction of price improved orders falls as we move from larger to smaller size firms, both for wholesalers and exchanges. Second, the magnitude of price improvement continues to be significantly larger for marketable orders routed to wholesalers for all terciles (e.g., 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 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.3 Execution Quality Statistics in Cents

We report basis point spreads in the main analyses. For completeness, we here report the share-volume-weighted execution quality statistics measured in cents. Table A2 reports the statistics for the overall sample, while A3 reports share-volume-weighted execution quality statistics for sub-samples.

[Table A2]

[Table A3]

A.4 Dollar-Volume-Weighted Execution Quality Statistics

We use share-volume-weighted execution quality statistics in our paper. By contrast, 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 execution quality statistics Table A4 below. For completeness, Table A5 reports dollar-volume-weighted execution quality statistics for the four sub-samples. Note that we report full spreads, whereas the SEC reports half-spreads.

[Table A4]

[Table A5]

A.5 Retail Broker Routed Volume

To calculate routed volumes, we use two variables from Rule 606 data reported by major retail brokers: the total dollar amounts of PFOF received by retail brokerages and the PFOF amounts in cents per one hundred shares. By 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 flows. Interactive Brokers accept PFOF for their IBKR Lite flow, but we are unable to separate this flow from their IBKR PRO flow so we also drop this broker.

Each retail brokerage reports the 606 data on a monthly basis and categorizes the sample stocks into the S&P 500 and non-S&P 500. Consequently, these data lack the cross-sectional richness of Rule 605 data, yet they still provide valuable insights into broker routing.

Another metric often used in the industry to compare the size of brokerages is Daily Average Revenue Trades (DART). With the advent of commission free trading for much of the retail equity brokerage business (Jain, Mishra, O'Donoghue, and Zhao (2022)), 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 (Interactive Brokers). Yet others combine retail and institutional trades when reporting DART (e.g., Fidelity).

A.6 ETF Tables

For completeness, we repeat the analyses in the main paper also for the sample of XXXX ETFs. Table A6 illustrates that Citadel and Virtu are the largest wholesalers also for ETFs with a combined market share of 64.9%.

[Table A6]

Execution quality for ETFs is reported for the overall sample in Table A7. The differences between wholesalers and exchanges are similar to what we observe for non-ETFs: quoted spreads are wider when wholesalers execute retail orders, but they offer significantly larger price improvements (35% compared to 4% of the spread for exchanges) so effective spreads are significantly lower. Yet, since the price impact for held orders routed to wholesalers is so much lower, the

realized spread is significantly higher for wholesalers (14.72 bps) than for exchanges (5.12 bps) also for ETFs. We confirm these univariate results in regressions in Table A8.

[Table A7]

[Table A8]

We examine whether market quality in ETFs delivered by the $top2$ wholesalers (Citadel and Virtu) are significantly different from the market quality of other wholesalers in Table A9. The results show that $top2$ provide significantly less price improvement, have more toxic order flow, and charge higher realized spreads than their competitors.

[Table A9]

Just as for non-ETFs, broker routing is sensitive to past market quality for ETFs in or sample. Specifically, Table A10 shows that wholesalers that offered lower realized ETF spreads on average in the previous month receives more order flow the following month.

[Table A10]

Finally, we estimate the effects of moving ETF order flow from wholesalers to exchanges in Table A11. The base case based on the realized spread of 5.12 bps from Table A7 shows that retail would experience a 31.5% increase in spreads and a loss of $6.30 bn during our sample period if their orders were pooled with the more toxic order flow on exchanges. As for non-ETFs, the winners would be exchange liquidity providers who would gain $16.62 bn.

[Table A11]

Table A1
Execution Quality: Sub-samples

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

	WHOL	EXCH	diff.	WHOL	EXCH	diff.
	[1]	[2]	[3]	[4]	[5]	[6]
	S&P 500			Tercile 1		
# shares, mil.	515.96	1,106.88	***	195.45	381.42	***
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		
# shares, mil.	105.31	100.74		132.91	74.79	***
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 for held liquidity-demanding orders. We compute the statistics separately for orders executed by wholesalers (WHOL) and exchanges (EXCH). We report the average number of shares executed and the average stock price in a sample stock during the sample period, followed by the percentage share of shares that are price improved or executed at or better the corresponding NBBO. 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]
# shares, mil.	172.80	252.86	***
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 (Cent): Sub-samples

The table contains execution quality statistics measured in cents for held liquidity-demanding orders. The sample is divided into S&P 500 and size terciles T1, T2, and T3 of non-S&P 500 stocks. We report the average number of shares executed and the average stock price in a sample stock during the sample period, followed by the percentage share of orders that are price improved or executed at or better the corresponding NBBO. Further, we report the quoted and effective spreads in 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] and [4] and [5] at the 1% (5%) level.

	WHOL	EXCH	diff.	WHOL	EXCH	diff.
	[1]	[2]	[3]	[4]	[5]	[6]
	S&P 500			Tercile 1		
# shares, mil.	515.96	1,106.88	***	195.45	381.42	***
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		
# shares, mil.	105.31	100.74		132.91	74.79	***
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

The table contains execution quality statistics for held liquidity-demanding orders. We compute the statistics separately for orders executed by wholesalers (WHOL) and exchanges (EXCH). We report the average number of shares executed and the average stock price in a sample stock during the sample period, followed by the percentage share of shares that are price improved or executed at or better the corresponding NBBO. 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]
dollar volume, $B	49,389	83,122
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 table contains execution quality statistics for held liquidity-demanding orders. The sample is divided into S&P 500 and size terciles T1, T2, and T3 of non-S&P 500 stocks. We report the average number of shares executed and the average stock price in a sample stock during the sample period, followed by the percentage share of orders that are price improved or executed at or better the corresponding NBBO. Further, we report the quoted and effective spreads in basis points, and to better understand the magnitude of price improvement, we compute the ratio of the effective to the quoted spread. Finally, we compute the components of the effective spread: price impact and realized spread. All variables are dollar-volume-weighted. .

	WHOL	EXCH	WHOL	EXCH
	[1]	[2]	[3]	[4]
	S&P 500		Tercile 1	
dollar volume, $B	29,437	50,041	17,156	30,391
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	
dollar volume, $B	1,741	2,059	1,053	631
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

Table A6
Market Shares for ETFs

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

	venue type	shares executed, bil.	mkt. share, %
Panel A: by venue type			
	EXCH	552.14	65.65
	WHOL	288.85	34.35
Panel B: by venue			
NYSE ARCA	EXCH	140.25	16.68
NASDAQ	EXCH	137.34	16.33
Citadel	WHOL	108.16	12.86
Virtu	WHOL	79.50	9.45
BATS	EXCH	66.99	7.97
EDGX	EXCH	49.95	5.94
G1	WHOL	43.65	5.19
NYSE	EXCH	31.16	3.70
BYXX	EXCH	29.54	3.51
EDGA	EXCH	25.28	3.01
Jane Street	WHOL	19.92	2.37
NSDQ PHIL	EXCH	19.06	2.27
NYSE NAT	EXCH	16.81	2.00
UBS	WHOL	14.98	1.78
Two Sigma	WHOL	13.26	1.58
NSDQ BOS	EXCH	12.11	1.44
MEMX	EXCH	9.79	1.16
IEX	EXCH	8.29	0.99
NYSE AMER	EXCH	5.07	0.60
Merrill Lynch	WHOL	4.84	0.58
Morgan Stanley	WHOL	4.68	0.56
NYSE CHI	EXCH	0.51	0.06
Total		841.00	100.00

Table A7
Execution Quality for ETFs

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

	WHOL	EXCH	diff. [1]-[2]
	[1]	[2]	[3]
# shares, mil.	80.99	154.46	***
price, $.	41.06	40.95	
improved, %	75.73	10.34	***
at or better, %	95.23	98.82	***
quoted spread, bps	28.47	24.82	***
effective spread, bps	18.61	23.80	***
effective / quoted	0.65	0.96	***
price impact, bps	3.89	18.68	***
realized spread, bps	14.72	5.12	***

Table A8
ETF Execution Quality: Regression

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

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

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

	eff. spr. / quot. spr.	quoted spr.	effective spr.	price imp.	realized spr.
	[1]	[2]	[3]	[4]	[5]
WHOL	-0.414***	3.602***	-4.075***	-10.075***	6.000***
	(0.01)	(0.28)	(0.15)	(0.62)	(0.57)
price	-0.015***	-6.722***	-5.740***	-4.372***	-1.368**
	(0.00)	(0.76)	(0.63)	(0.67)	(0.61)
volatility	-0.000	-0.003	-0.003	0.003	-0.006**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
volume	-0.006***	-2.203***	-1.705***	-0.825***	-0.880***
	(0.00)	(0.18)	(0.15)	(0.13)	(0.10)
intercept	1.063***	62.684***	53.391***	35.858***	17.529***
	(0.01)	(3.49)	(3.01)	(3.56)	(2.79)
Adj. R^2	0.582	0.668	0.616	0.302	0.224

Table A9
ETF Execution Quality Across Wholesalers: Regressions

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

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

where $DepVar_{it}$ is one of the following market quality variables for ETF i in month t: the ratio of effective to quoted spreads, quoted spread, effective spread, price impact, and realized spread as defined previously; $top2$ is a dummy variable that has a value of 1 for orders executed by Citadel and Virtu and 0 for orders executed by other wholesalers; *price* is the natural log of the ETF price; *volatility* is the difference between the high and low prices scaled by the high price; and *volume* is the natural log of trading volume. The models are estimated with ETF and month fixed effects, and the standard errors are double-clustered across ETFs and months. Asterisks ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels.

	eff. spr. / quot. spr.	quoted spr.	effective spr.	price imp.	realized spr.
	[1]	[2]	[3]	[4]	[5]
top2	0.066***	-0.074	1.687***	0.581***	1.106***
	(0.01)	(0.05)	(0.15)	(0.12)	(0.21)
price	-0.032***	-7.025***	-5.070***	-1.417***	-3.653***
	(0.00)	(0.84)	(0.62)	(0.37)	(0.50)
volatility	-0.000	0.002	0.000	-0.003	0.004
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
volume	-0.006***	-2.607***	-1.522***	0.165**	-1.686***
	(0.00)	(0.20)	(0.13)	(0.07)	(0.12)
intercept	0.656***	71.209***	43.819***	5.318***	38.498***
	(0.02)	(3.98)	(2.86)	(1.61)	(2.45)
Adj. R^2	0.167	0.699	0.558	0.176	0.372

Table A10
Wholesaler ETF Order Flow Determinants: Regression

We estimate the following regression:

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

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

	$\textit{mkt. share}_{j,t}$
$\textit{abn. realized spread}_{ij}$	0.000
	(0.00)
$\textit{abn. realized spread}_{j}$	-0.039***
	(0.01)
price	0.023*
	(0.01)
volatility	-0.048***
	(0.01)
volume	0.043***
	(0.01)
intercept	-0.107
	(0.13)
Adj. R^2	0.594

Table A11
Moving ETF Retail Flow to Exchanges

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

realiz. spr.,	Panel A: Δ eff. spread, %		Panel B: gains, in $ bil.			
bps.	RET LDs	EXCH LDs	RET LDs	EXCH LDs	EXCH LPs	WHOL LPs
5.62	34.51	-8.30	-6.91	13.32	2.74	-15.83
5.52	33.91	-8.72	-6.79	13.98	1.96	-15.83
5.42	33.30	-9.13	-6.66	14.64	1.18	-15.83
5.32	32.70	-9.54	-6.54	15.30	0.40	-15.83
5.22	32.10	-9.95	-6.42	15.96	-0.39	-15.83
5.12	**31.49**	**-10.36**	**-6.30**	**16.62**	**-1.17**	**-15.83**
5.02	30.89	-10.78	-6.18	17.29	-1.95	-15.83
4.92	30.28	-11.19	-6.06	17.95	-2.73	-15.83
4.82	29.68	-11.60	-5.94	18.61	-3.51	-15.83
4.72	29.08	-12.01	-5.82	19.27	-4.29	-15.83
4.62	28.47	-12.42	-5.70	19.93	-5.07	-15.83