What Does Best Execution Look Like?*

Thomas Ernst[†], Andrey Malenko[‡], Chester Spatt[§], and Jian Sun[¶]

January 8, 2025

Abstract

U.S. retail brokers have a "best execution" legal mandate in executing their clients' orders, but the specific way best execution operates is unknown. Using data on retail order routing from three large brokers, we examine how they interact with wholesalers and establish three results. First, brokers allocate order flow based upon past performance of wholesalers, though they do so in various ways as they create future incentives. Second, wholesalers recognize that future order allocation depends on current performance and therefore compete on price improvement. Finally, there are significant differences between stocks and brokers in the way competition occurs.

*For helpful comments and feedback we are thankful to Simcha Barkai, Robert Battalio, Kevin Crotty, Vincent Fardeau, Slava Fos, Shiyang Huang, Xing Huang, Mina Lee, Nadya Malenko, Albert Menkveld, Dermot Murphy, Christopher Schwarz, Andriy Shkilko, along with seminar and conference participants at Boston College, Carnegie Mellon University, European Finance Association, NES and Bank of Russia workshop on retail investors, the University of Maryland, University of Melbourne, the Microstructure Exchange, the National Bureau of Economic Research, Singapore Management University, the SFS Cavalcade, the University of Wisconsin, and numerous anonymous industry participants.

[†]University of Maryland, Robert H. Smith School of Business: ternst@umd.edu

[‡]Boston College, Carroll School of Management: malenkoa@bc.edu

[§]Carnegie Mellon University, Tepper School of Business: cspatt@andrew.cmu.edu

[¶]Singapore Management University, Lee Kong Chian School of Business: jiansun@smu.edu.sg

I. Introduction

Almost all retail brokerages in the U.S. offer commission-free stock trading. Market retail orders are rarely executed on stock exchanges, but instead get routed by the broker to one of several wholesalers, often for a fee that the wholesaler pays to the broker, a practice known as "payment for order flow." Academics and market observers recognized potential costs of routing retail orders away from exchanges, such as increased adverse selection at the exchange due to the cream-skimming effect (e.g., Easley, Kiefer, and O'Hara (1996) and Bessembinder and Kaufman (1997)), as well as potential benefits, such as increased competition for order execution (e.g., Battalio (1997) and Battalio and Holden (2001)).

The benefit side of the trade-off relies on competition among wholesalers operating efficiently, so that the rents that wholesalers make on order execution get passed to retail investors via lower trading commissions and price improvement. However, how brokers ensure competition among wholesalers is not well understood. The Securities and Exchange Commission (SEC) has announced several proposals to restructure or refine the industry, including more disclosure of execution quality, requirements for order-by-order auctions, reforms to trading fees and tick size, and a legal best-execution obligation for brokers, which would replace the existing FINRA rule for best execution.[1] Motivated in part by these proposals, we seek to develop an understanding of a fundamental question. How do retail brokers obtain best execution for their clients?

Although the goal of best execution–obtaining the best price possible–is straightforward, achieving it presents a formidable strategic challenge. This paper investigates the economics of achieving best execution, including the decisions brokers make and how those decisions shape both immediate outcomes and the market structure of retail internalization. Best execution is a classic principal-agent problem: brokers (the principals) must design a system that induces wholesalers (the agents) with private information about their willingness to trade to offer competitive prices. As an example, suppose that a broker uses two market centers and ranks them based on past performance. For future order flow, a broker should route more orders to the better market center, but exactly how much order flow to route in this manner is a difficult question. Sending 100% of the order flow to the better market center eliminates any incentive for further marginal improvement from the

[1] Lewis (2024) argues that the best-execution proposal of the SEC does not identify a market failure to justify the proposal.

better market center, as well as the ability to evaluate the weaker market center, while sending a proportionate 50% of the order flow does not reward market centers for price improvement, nor does it maximize the benefit of superior prices offered by that market center. Moreover, the frequency of evaluation, as well as whether to evaluate trades in different symbols separately or collectively, form important dimensions of the strategic choice.

In practice, retail brokers work with several wholesalers, who are market makers that specialize in executing retail order flow. These wholesalers offer better prices for retail order flow than open market centers (such as exchanges),[2] due to the order flow's lower adverse selection or correlation.[3] Retail brokers do not communicate with wholesalers on each individual trade, but instead evaluate wholesalers on overall past performance. This practice reflects that substantial liquidity, both from wholesalers and alternative trading systems, is non-displayed and even discretionary in nature. Rather than focusing on a trade-by-trade view, brokers entice competition for aggregate order flow. Brokers make strategic choices, informed by past history, to dynamically route more orders to wholesalers offering better past prices.

To gain insight into these strategic choices of brokers, we obtained proprietary data from three large retail brokers (henceforth Broker A, Broker B, and Broker C), which collectively comprise over 50% of the U.S. equity market's retail brokerage industry. This data offers us considerable insight into the specific practices of these brokers, including best-execution statistics on each of the wholesalers to whom they route, and provides three main results. First, brokers are very responsive to wholesaler performance, allocating more order flow to wholesalers who have offered better average prices. Second, wholesalers are responsive to brokers, changing the amount of price improvement provided as market conditions change or brokers' focuses change. Third, brokers focus on different histories and routing algorithms, which shapes the competitive landscape.

Brokers primarily measure past performance through effective-over-quoted (EFQ) spread, which captures the ratio of spreads paid by retail investors relative to the publicly available quoted spreads; lower EFQ means larger savings by retail investors. For every one percentage point increase in EFQ,

[2]Battalio and Jennings (2022) show that even brokers who do not take payment for order flow (PFOF) use wholesalers due to better prices offered by wholesalers.

[3]Easley et al. (1996) provide empirical evidence suggesting that retail order flow is less informed, Battalio and Holden (2001) model lower adverse selection among retail investors, while Baldauf, Mollner, and Yueshen (2024) model the implications of lower correlation of retail orders, an opinion also stated in the 2021 SEC staff report (SEC (2021)).

we find that wholesalers obtain between 1.0 and 1.2% less order flow allocation. If we consider the ordinal ranking of wholesalers, a worse ordinal ranking leads to 6 to 9% lower order flow allocation. We also show that these results are specific to the history considered by the broker; if we evaluate wholesaler performance based on longer or shorter windows, or for differently sized orders, the explanatory power of history, measured by the R^2 of the regression of order allocation based on past history, drops considerably. Wholesaler rankings can change based on the sample; when we randomly select subsets of days or subsets of trades, we show that wholesaler rankings in the sub-sample can differ from full sample rankings, even for fairly large sub-samples. Brokers route based on a specific history, and that specific history is essential to understand order allocation decisions made by the broker.

Having established that brokers route order flow based on past performance, we next show evidence suggesting that wholesalers respond to incentives provided by brokers and market conditions when deciding how much price improvement to offer. While wholesalers compete on aggregate order flow rather than individual orders, market conditions which allow greater price improvement apply to their competitors as well as to themselves. Using daily execution data, we show that on volatile market days, EFQ ratios decrease, consistent with wholesalers offering greater price improvement to retail investors. Effective spreads charged to retail investors are less volatile than publicly quoted spreads, indicating that when quoted spreads increase, the opportunity for wholesalers to provide price improvement to retail investors is greater and competitive pressures lead them to pass these improvements to retail investors.

During the sample we study, Broker B made a focus change in their routing decisions, facilitating an event study on how broker priorities shape wholesaler behavior. Broker B considers all aspects of past performance, but expanded their primary focus to include particularly small (odd-lot) orders. Wholesalers respond rapidly, with EFQs for odd-lot orders declining from an average above 80% to an average below 30%. We find a contemporaneous rise in EFQ of large orders, however, suggesting that improvement in the performance on small orders is not a "free lunch": any change in priorities for improvement along one dimension may bring disimprovement along other dimensions, as implied by the multitasking principal-agent theory (Holmstrom and Milgrom (1991)). The change in prices obtained when the broker routing behavior changes is critical to understanding the effect of proposed changes in the market structure. In the empirical analysis of the SEC's order competition proposal,

for example, the SEC measures substantial mid-quote liquidity that retail investors do not always access. Such attempts to access this liquidity, however, particularly in a way that gives rise to a winner's curse or information leakage, may lead to the liquidity changing in availability. What may seem like a tempting 'obvious' improvement to routing algorithms fails to take into account that the price improvement is offered conditional on the current routing algorithm. Any change to the algorithm would potentially change the conditional liquidity offered.

Wholesalers typically have detailed information on how their performance compares to that of their competitors. Two brokers in our sample evaluate and change routing allocations at a monthly frequency, typically at the beginning of the month. This creates an opportunity to observe how wholesalers behave when changes to their order allocation are imminent. We find that wholesalers who are far behind their closest competitor give worse prices. This result is much stronger at the beginning of the month compared to the end of the month, however, consistent with wholesalers awareness of the evaluation window and making strategic responses to offer price improvement closer to the end of the window, when evaluation and re-adjustment are closer in time.

Our last set of analysis focuses on the competitive landscape for order routing. A large market maker, henceforth wholesaler A5, entered the wholesaling business in late 2021, which allows us to conduct an event study around the entry of a competitor. We use two competition measures, $First\_To\_Second$, the arithmetic difference in EFQ offered by the first-best vs. second-best wholesaler and $First\_To\_Average$, the difference between the first-best and volume-weighted average wholesaler. Wholesaler A5 works with Broker A, who routes each symbol on an individual basis. We find that wholesaler A5 gains more market share in securities with a larger $First\_To\_Second$ difference, and a larger EFQ. Following the entry of Wholesaler A5, we find that $First\_To\_Second$ declines when measured among the incumbent wholesalers, but does not decline when including wholesaler A5, consistent with a displacement story, where wholesaler A5 offers superior improvement to gain market share, but displaced competitors respond by reducing the price improvement that they give. We find some evidence consistent with an increase in competition, however, as $First\_To\_Average$ and EFQ decrease in both the incumbent wholesalers and the wholesalers including wholesaler A5.

Decisions by brokers in how they divide their order flow change the competitive landscape for wholesalers. One broker in our sample routes all orders according to past performance across

stocks, while another broker routes each individual stock according to the history in that specific stock. We show that there is greater difference in wholesaler EFQ in trading large stocks than small stocks, consistent with greater economies of scale in large stocks. We note a similar set of differences between small and large orders, with greater differences in EFQ among wholesalers in orders in the largest stocks. Routing each symbol in individual size and symbol bins may create more competition in smaller symbols and orders, where wholesalers are more equal in ability, but less competition in large stocks or large orders, where wholesalers may have greater differences in their economies of scale.

We also analyze non-marketable limit orders. In contrast to market orders, limit orders must be displayed on the exchange. We show that wholesaler performance in limit order fill rates is far less well explained by wholesaler-specific prior performance, consistent with there being smaller differentiation among wholesalers in their relative ability to handle non-marketable limit orders. Nonetheless, there are considerable differences in the exchange rebates wholesalers would collect for posting these limit orders, as exchanges tie rebates to volume tiers that market makers conduct. In September of 2024, the SEC adopted Rule 6b-1, which would restrict such tiering as well as require additional data on tiering. Our results highlight a connection between this rule and wholesaler competition for retail orders, as lower fees or higher rebates allow wholesalers to offer more price improvement to retail orders, while differences in exchange volume tiers may make it difficult for smaller wholesalers to offer price improvement comparable to that offered larger wholesalers.

The SEC has adopted an update to Rule 605 with individual broker-wholesaler performance data; our proprietary data resembles this, and highlights the valuable insights such data offer as well as the importance of quoted spreads in making sense of a measure of effective-over-quoted spreads. Contemporaneously, the SEC had also proposed a best-execution rule. We show that for equities the existing FINRA Best-Execution rule has created a relatively competitive market system, but that "Betrand competition" on price improvement is not the right model of competition among wholesalers. Instead, we argue that the empirical patterns point to a model of dynamic imperfect competition in which privately informed agents (wholesalers) compete with prices of services they already obtained for allocation of future business, and the principal (the broker) designs the allocation rule to balance the provision of incentives to agents to give good prices today with rewarding the agents for offering good prices in the past (see Section VI for a more detailed

discussion). These interactions combine features of all-pay auctions (a wholesaler exercises the current order at any price improvement of its choice) and optimal dynamic contracting, where provision of incentives via continuation value is central. To put it differently, brokers develop policies to manage competition among strategic and privately informed wholesalers, and the price improvement offered by wholesalers is conditional on these policies and changes as these policies change.

II. Prior Literature and Contribution

Our paper relates to the literature that studies execution of trades by brokers on behalf of their clients. Macey and O'Hara (1997) discuss early academic literature and a number of legal and economic issues associated with the duty of best execution. Angel, Harris, and Spatt (2011) and Angel, Harris, and Spatt (2015) suggest that retail equity trading costs are very low and greatly improved after the adoption and implementation of Regulation NMS. Adams, Kasten, and Kelley (2024) analyze execution quality around the introduction of zero-commission trading, and find that current PFOF (viewed as a potential implicit commission) is substantially smaller than historical fixed commissions.

Our work complements closely related work by Dyhrberg, Shkilko, and Werner (2022) which uses public SEC Rule 605 and 606 data to evaluate broker routing to wholesalers. Their paper shows that the market is competitive, that price improvement is much larger than payment for order flow, and that the scale of price improvement wholesalers give to retail investors closely matches the level of price improvement institutions obtain by timing their marketable orders to trade when spreads are narrower. While their focus is on market-wide competition, our data allows us to focus on the *individual* broker-wholesaler relationships. We characterize the specific choices individual brokers make in obtaining best execution, and how these specific choices influence the prices obtained, as well as the broader competitive landscape. In a related vein, Battalio and Jennings (2023) have data from one or more large wholesalers, and show that wholesalers give substantial price improvement to retail trades. We similarly find substantial price improvement for retail investors, and complement their work with an analysis of how wholesalers respond to specific broker focuses.

Our work is also closely related to Huang, Jorion, Lee, and Schwarz (2023), who use self-generated trading data to evaluate broker-wholesaler competition. They find the entry of Jane Street improves order execution quality. Similarly, we find the entry of a large anonymous wholesaler improves execution quality and decreases market concentration overall, though we see some consolidation in market share among the incumbent wholesalers. Huang et al. (2023) also find that "a majority of our brokers seem to hardly change their routing for our trades based on past execution ... only one does so at a statistically significant level." In contrast, we find that all three brokers in our sample are highly responsive to wholesaler performance. While the anonymity of our data prevents us from drawing specific contrasts, we note that within our data, very large sample sizes of trades and somewhat detailed knowledge of what historical window a broker measures are needed to accurately calculate the true rankings of wholesaler performance. At the same time, we also see one of our brokers adjusts how they adjust order routing based on odd-lot performance; prior to this adjustment, our measures for this broker (restricted to only odd-lot trades) align with the general pattern of Huang et al. (2023).

Better-than-NBBO liquidity is by no means unique to wholesalers. Bartlett and O'Hara (2024) and Levy (2022) highlight the prevalence of hidden liquidity, better than the NBBO and available to all market participants on exchanges. Ernst, Sun, and Spatt (2024b) examine retail liquidity programs on exchanges, which offer better prices only to retail investors. Nor is NBBO liquidity all equal–some exchanges charge take fees, while others pay rebates.[4] The NBBO, however, provides a single yardstick for calculating EFQ against which all wholesalers are evaluated. While the NBBO itself may not be an ideal measure of expected trading costs, it is a perfectly standardized benchmark against which all market centers can be equally compared.[5]

Several papers analyze the performance of wholesalers and market segmentation. Easley et al. (1996), Battalio and Holden (2001), and Hu and Murphy (2022) theoretically model cream-skimming by wholesalers, while van Kervel and Yueshen (2023) model anti-competitive reference pricing. Battalio and Jennings (2022) empirically analyze the execution quality given one or more large wholesalers, showing that wholesalers consistently provide price improvement relative to the exchange. Comerton-Forde, Malinova, and Park (2018) find that introduction of a minimum price improvement

[4] Li, Ye, and Zheng (2021) discuss how fees and rebates can change the true quoted spread.

[5] Brokers send a small number of orders to each market center (e.g. exchanges) to test market quality. These orders do not appear on SEC 606 reports so long as they comprise less than 5% of a broker's total order flow.

rule by Canadian regulators reduced retail order segmentation, improved liquidity, but lowered price improvement that retail traders received. Our focus is on the competitive broker-wholesaler relationship, with brokers monitoring wholesaler performance and allocating order flow accordingly.

III. Broker Responsiveness

We begin our analysis by describing our data and best-execution strategies of brokers; we show that brokers are extremely responsive to wholesalers' performance, with brokers sending more order flow to wholesalers offering superior prices. We show that brokers are selective in the historical window they evaluate, and that evaluating differences in wholesaler performance requires considerable data.

A. Data and Best-Execution Strategies

We obtain proprietary data from three large retail brokers, henceforth, Broker A, Broker B, and Broker C. Collectively, these firms comprise over 50% of the U.S. equity market's retail brokerage industry. From each broker, we obtain proprietary data on their routing practices, including the best-execution statistics for each of the wholesalers to whom they route. While brokers target all specific features of best execution highlighted under FINRA Rule 5310, they do place emphasis on distinct features of wholesaler performance; we highlight some of these focus-points of their best-execution practices in Table I.

All U.S. brokers have a best-execution duty under FINRA Rule 5310, which requires brokers to "use reasonable diligence" to obtain a price "as favorable as possible under prevailing market conditions." For retail orders, brokers are able to obtain far better prices than the prevailing NBBO, as wholesalers are willing to offer superior prices to retail investors. This price improvement, i.e., prices better than the public exchange bid or ask, is non-displayed liquidity; consequently, brokers do not simply route to the best quote, but must define a routing practice which maximizes the extent of non-displayed price improvement their orders obtain. In practice, brokers do not negotiate on individual trades, but send future order flow to wholesalers based on past performance.[6] Brokers

[6]Wholesalers do not stream quotes in advance of trade. While the exchange NBBO is visible, wholesaler quotes are not. We note that single dealer platforms are allowed to stream private quotes under Reg NMS, but we are not aware of any U.S. wholesalers who do so for retail trades.

regularly evaluate the execution quality they receive, and adjust routing accordingly. Dyhrberg et al. (2022), using aggregate SEC 606 data, show that all major U.S. retail brokerages adjust order flow from month to month, consistent with performance of their best-execution duties. Execution quality can cover many aspects of trading, with FINRA Rule 5310 suggesting a holistic process in which brokers consider price improvement, execution likelihood, speed, and order fill size.

The brokers in our sample work with wholesalers who will always match or improve upon the public national best bid or offer (NBBO). Consequently, brokers route orders to wholesalers based on the past history of how much price improvement, relative to the NBBO, that wholesalers have offered. This quantity is typically measured as an effective-over-quoted spread. An effective spread measures the difference between trade price and mid-quote, while a quoted spread is half the difference between the bid and ask.[7] Brokers in our sample measure EFQ as a ratio of totals: $EFQ = \frac{\sum Eff}{\sum Quo}$, where $\sum Eff$ is the sum of effective spreads (distance from trade price to mid-quote price) while $\sum Quo$ is the sum of half-quoted spreads (half the distance from the best ask to the best bid price). These sums are measured over their evaluation period (e.g. daily, monthly, etc.). Lower EFQ ratios imply larger savings for retail investors.

The first dimension along which our brokers differ is the amount of history they focus upon when making routing decisions. When routing an order, brokers will consider prior EFQ charged by each wholesaler. All brokers holistically evaluate past performance, and consider the entire history of wholesaler performance, but place a primary focus on different periods. Broker A, for example, places a primary focus on 30 days of history, while Brokers B and C both place a primary focus on 90 days of history. Conversations with brokers highlighted a trade-off between focusing on shorter history, which provides a rapid reward to improvement, and a longer history, which provides a reward to persistence in performance.

When using past history to inform routing decisions, brokers differ not just in the length, but in what history they use to route orders. Broker C measures wholesaler performance in all securities, and routes all orders based on the past performance in all securities. In contrast, Broker

[7] As an example for a single trade, suppose the best bid is \$10.00 while the best ask is \$10.10, the mid-quote is \$10.05. If a retail investor uses a market order and buys at \$10.07, they received 3 cents of price improvement relative to the best ask price, and are charged an effective-over-quoted spread of $\frac{10:07 \quad 10:05}{1=2(10:10 \quad 10:00)} = \frac{2}{5} = 40\%$. In this example, the retail investor's EFQ of 40% means that they paid 40% of what they would have paid in spread if they had traded at the NBBO. Rather than for a single trade, brokers calculate the ratio over the sum of trades routed to the wholesaler over a time period. For example, a series of trades filled with effective spread of 3, 3, and 4 cents when half-quoted spreads are 3, 6, 7 cents would produce an EFQ ratio of $\frac{3+3+4}{3+6+7} = \frac{10}{16} = 62{:}5\%$*:*

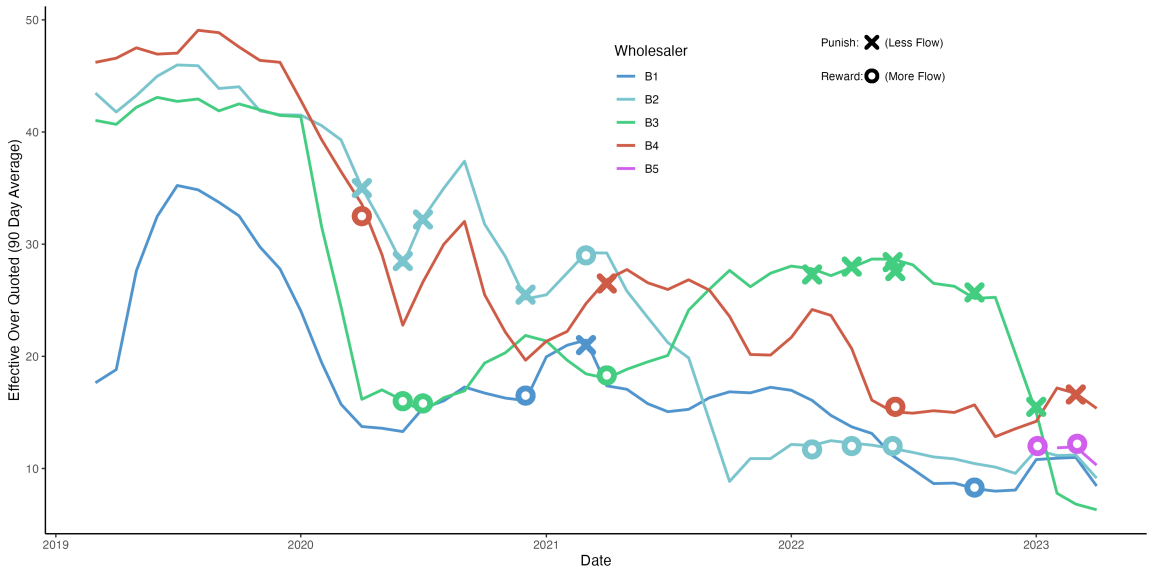
A evaluates wholesaler performance for past orders in an individual symbol, further divided into five size categories. Within each symbol, routing in each order-size category is determined based on past performance in that order-size category. Broker B takes an intermediate approach, evaluating history in four specific security bins divided into three size categories. When considering past history to determine a routing decision, there is a trade-off between the quantity and relevance of the information. The choice of routing history also impacts market structure, as the specific history a broker considers defines distinct categories in which wholesalers compete.

When brokers adjust future order flow, they face a further strategic choice on how to allocate flow across wholesalers. The wholesalers offering the best performance naturally obtain the most order flow, while wholesalers offering the worst performance obtain the least. The reward to the best wholesaler, however, must strike a balance between rewarding past performance while also maintaining the ability to increase the reward (order allocation) should the wholesaler offer even greater price improvement.

Figure 1 plots wholesaler performance over our sample period for the set of wholesalers who work with Brokers B and C. For Broker B, wholesalers who reduce their EFQ obtain more order flow, while wholesalers who increase their effective-over-quoted spread obtain less order flow. For Broker C, we use a proprietary score metric (of which effective-over-quoted spread is an important component), and note a similar pattern: wholesalers with improved scores obtain more order flow while wholesalers with impaired scores obtain less order flow.

Figure 1. Broker Execution and Routing. Brokers route based on past performance. We plot the 90-day-average of wholesaler effective-over-quoted spread. Brokers allocate more order flow to wholesalers charging lower EFQ spreads (adjustments depicted with circles) and allocate less flow to wholesalers charging higher EFQ spreads (adjustments depicted with crosses). For Broker B, we plot EFQ. For Broker C, we plot the negative of their proprietary score in Panel B, with a more negative score awarded to higher performance on a variety of characteristics, including EFQ.

Panel A: Broker B Data



Panel B: Broker C Data

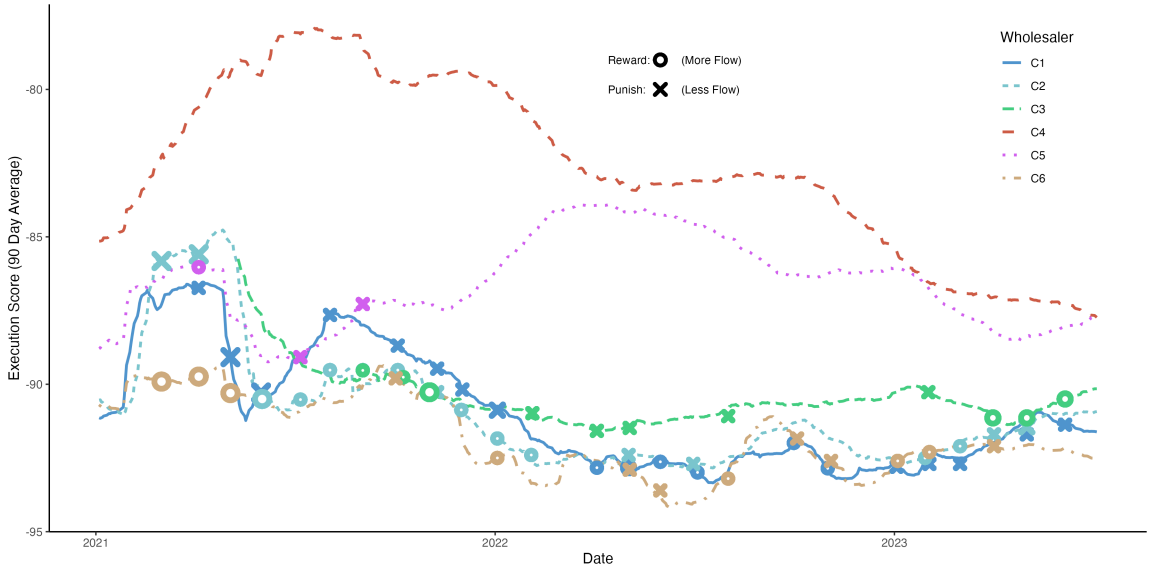


Table I: Best Execution Focus Points. All brokers use a holistic process and consider all aspects of wholesaler performance, but have slightly different primary focuses in measuring wholesaler performance. We summarize some of the basic differences here. Category refers to the bin that a broker uses, with all orders in the bin being routed in the same fashion. History refers to the past data which forms the focal point for future routing decisions.

	Broker A	Broker B	Broker C
Category:	Each Symbol 5 Size Categories	Four Security Bins 3 Size Categories	One Bin
History:	30 Days	90 Days	90 Days
Decisions:	Daily (Rolling)	Monthly	Monthly
Market Order Data:	Individual Stock Data: Nov-Dec 2021 March-Nov 2022	Aggregate: Jan 2019-May 2023 Individual Stock Data: Jan 2020-May 2023	Aggregate Data: Jan 2021-July 2023 Individual Stock Data: Jan-Feb 2023
Sample:	500 Stock Sample	All Stocks	All Stocks

B. Routing-Performance Relationship

Brokers route orders based on past performance, routing more orders to wholesalers who have offered better prices. Routing creates competition among wholesalers to offer the best prices: wholesalers offering superior prices are rewarded with more order flow in the future, while wholesalers offering inferior prices are punished with reduced order flow in the future. To evaluate the routing-performance relationship, we estimate the following regression:

REGRESSION 1: *For each wholesaler j, time period t and security bin k, we estimate:*

$$OrderShare_{jkt} = \alpha_0 + \alpha_1 Prior\_EFQ_{jkt} + X_{jk} + \epsilon_{jkt}$$

We estimate 1 separately using data from each broker.[8] We use two measurements of $Prior\_EFQ$. The first measure of $Prior\_EFQ$ is the prior period's effective-over-quoted ratio, which is the average effective spread paid divided by the average quoted spread, measured as a percentage. The second measure of $Prior\_EFQ$ is the wholesaler's ordinal rank (with the wholesaler having the best EFQ score in that period receiving a rank of one for that period, the wholesaler with the second-

[8]The security bin k depends on the broker. For broker C, there are no bins. For Broker B, there are four bins. For broker A, each stock and order size combination is a unique bin.

best EFQ score in that period receiving a score of two for that period, and so on). *OrderShare* measures the share of orders routed to a wholesaler, measured as a percentage of the broker's total order flow.

Our unit of observation depends on the broker, as there are differences in how brokers categorize past wholesaler performance. Brokers B and C typically make order share adjustments on a monthly basis, so we use monthly data. For Broker A, adjustments are made on a daily basis so we use daily data. Across all specifications, we fit a fixed effect for each time period and cluster standard errors by wholesaler. For Broker B, we also fit a fixed effect for each asset category. For Broker A, we also fit fixed effects both for each stock symbol and each size category.

Results of Regression 1 are presented in Table II. Across all three brokers, there is a strong relationship between performance and order share, with wholesalers with worse EFQ obtaining smaller order share allocations in future periods. For every 1% increase in EFQ (i.e., charging a higher price, as a percentage of the quoted spread), we estimate that wholesalers receive 1.2% less order share from Broker A and 0.96% less order share from Broker B. For each lower rank, wholesalers receive 8.9% less order share from Broker A and 5.6% less order share from Broker B. With the proprietary Execution Score from Broker C (of which EFQ is an important component), brokers similarly receive 3.0% (cardinal score measure) to 7.3% (ordinal score rank) less order flow from Broker C.

C. Historical Speci city and Self-Evaluation

When routing, all brokers must make a choice in determining how much history to consider, and any choice will involve a trade-off. Considering more history gives more information and encourages consistency in performance, while a shorter history allows more rapid adjustments to reflect more recent market conditions. We investigate the importance of window history in explaining broker routing history by re-estimating Regression 1 with Broker A's data on four different historical windows: 5 days, 10 days, 30 days, and 45 days. For each window, we examine the explanatory power of $PriorEFQ$ on $OrderShare$. Results are presented in Table III. The R^2 of Regression 1 peaks at 9.3% for the 30-day-window history, which matches the true length of history on which Broker A focuses. We conduct a similar exercise using performance across order size bins. We compare routing of orders in size category 3 (trades between 500 and 1,999 shares), and how it

Table II: Order Share and Execution Quality. We estimate Regression 1 which regresses the *OrderShare* allocated to each wholesaler on prior market performance. Order Share is the fraction of orders going to each wholesaler. *PriorEFQ* refers to effective-over-quoted spread of the wholesaler in the prior period. *PriorEFQRank* is the ordinal rank of each wholesaler in the prior period, with 1 being the best-ranked wholesaler, 2 the second-ranked wholesaler, and so on. *PriorScore* and *PriorScoreRank* refer to a proprietary score used by Broker C, of which EFQ is a key component. Observations are at the symbol-size-wholesaler-date level for Broker A with a fixed effect for each trade date, symbol, and order size and we cluster standard errors by wholesaler; observations are at the category-wholesaler-date level for Broker B with a fixed effect for each trade date and asset, and we cluster standard errors by wholesaler; and observations are at the wholesaler-date level for Broker C, with a fixed effect for each date and we cluster standard errors by wholesaler.

	Dependent variable: OrderShare					
	Broker A Data		Broker B Data		Broker C Data	
	(1)	(2)	(3)	(4)	(5)	(6)
Prior EFQ	1.230		0.958			
	(0.129)		(0.286)			
Prior EFQ Rank		8.882		5.553		
		(0.754)		(1.898)		
Prior Score					3.015	
					(0.646)	
Prior Score Rank						7.294
						(0.653)
Observations	129,526	129,526	786	786	170	170
R^2	0.316	0.339	0.248	0.253	0.613	0.766
Note:					p<0.1;	p<0.05; p<0.01

relates to prior EFQ performance of trades in Size 1 (< 100 shares), Size 3, or Size 5 (5,000 to 9,999). The R^2 is 12.5% for the predictability of Size 3 Order Shares based on Size 3 $PriorEFQ$, and less than 2.5% for Size 1 or Size 5 $PriorEFQ$, consistent with Broker A's practice of routing orders for each symbol of order size 3 based on past EFQ of Size 3 trades in that symbol. We obtain similar results with other cross-comparisons (such as order share for Size 2 trades regressed against Size 1, 3, 4, or 5 historical performance). Our results highlight the limited predictive power of small samples for understanding routing decisions: routing history from the last 5 days will have limited predictive power for understanding a broker's decision if that decision is made on data from the last 30 days; similarly, order performance for small trades will have limited predictive power for broker decisions based on order performance of large trades.

Motivated by the specificity of wholesaler routing in Table III, we next consider whether a retail investor could, using their own trades, understand how a broker routes orders to wholesalers. Brokers route to wholesalers using all information available, including how a wholesaler has performed on every trade in the historical window. A retail customer may observe their own trades,[9] but these trades would be only a fraction of all trades that the broker routed. Brokers make routing decisions on the population of orders in the order history, while an individual retail customer would observe at most a small sample. To gain insight into how well a sample of orders would reflect the population of orders, we conduct a sampling procedure, across orders and across days.

We first consider how many orders are necessary to understand wholesaler rankings on a single day. From Broker A, we have a sample of all trades in JP Morgan stock for odd-lot orders (between 1 and 100 shares) placed on March 13, 2023, a total of 1,315 trades. For each wholesaler, we plot the cumulative distribution of EFQ in Figure 2, Panel A, and calculate the true wholesaler rankings on this population of orders. From this total population of orders, we draw 5,000 samples of orders and calculate wholesaler rankings on each sub-sample. Figure 2, Panel B plots the percentage of sub-samples which have wholesaler rankings which match the population wholesaler rankings, as a function of sample size. With a sample of 50 trades, the third-vs-fourth-best wholesaler ranking is correct in around 60% of the 5,000 sample draws; this accuracy improves to 80% when the sample size is 500 trades. Even with relatively large sample sizes, some wholesaler differences are very

[9] SEC Rule 606(b)(1) allows a retail customer to request routing information for where each individual trade they place was routed.

Table III: History Windows. When brokers route orders, they focus on a specific amount of past history. We estimate Regression 1 with Broker A's data with a variety of windows. Our interest is in how the R^2 changes with historical window measured, so we fit a constant term but do not fit any fixed effects. In Panel A, we estimate using 4 different windows: 5 days, 10 days, 30 days, and 45 days; broker A uses a 30-day window. In Panel B, we estimate order share for Size 3 (between 500 and 1,999 shares) trades using prior EFQ on Size 1 (< 100 shares), Size 3, or Size 5 (5,000 - 9,999 shares); broker A routes Size 3 orders based on Size 3 performance.

Panel A: Different History Windows

	Dependent variable:			
	OrderShare			
	(1)	(2)	(3)	(4)
Prior 5 Days EFQ	0.469			
	(0.006)			
Prior 10 Days EFQ		0.635		
		(0.006)		
Prior 30 Days EFQ			0.835	
			(0.007)	
Prior 45 Days EFQ				0.469
				(0.006)
Observations	129,526	129,526	129,526	129,526
R^2	0.051	0.072	0.093	0.051

Panel B: Different Order Sizes

	Dependent variable:		
	OrderShare For Trades Size 3		
	(1)	(2)	(3)
Prior EFQ - Size 1	0.518		
	(0.031)		
Prior EFQ - Size 3		0.975	
		(0.024)	
Prior EFQ - Size 5			0.003
			(0.002)
Observations	11,420	11,420	11,420
R^2	0.024	0.125	0.0002
Note:		p<0.1; p<0.05;	p<0.01

difficult to detect. For example, with a sample size of over 250 orders (all drawn from a single stock, single size bin, and on a single day), the sample will have correct wholesaler rankings of the first-vs-second-best wholesaler less than 60% of the time. For small differences between wholesalers, even large numbers of observations may fail to accurately detect differences in underlying performance. This is true both for large wholesalers and small wholesalers, which have an additional observability problem: because orders are allocated based on past performance, wholesalers with the worst past performance would obtain a very small portion of total orders.[10] While the broker will observe wholesaler performance on all trades routed to a wholesaler, a retail customer will see a tiny individual fraction, particularly for a small wholesaler.

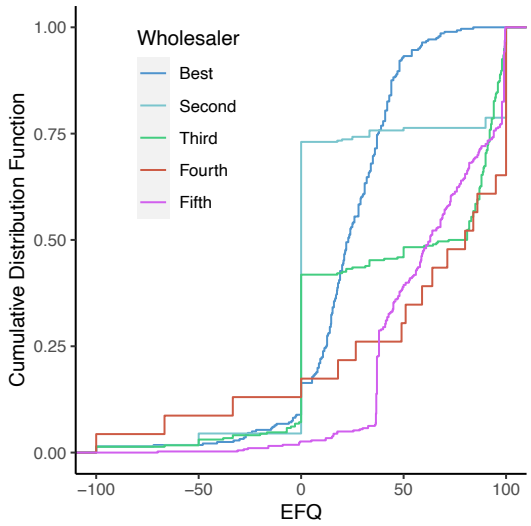
This same difficulty of estimating wholesaler performance persists across days as well as across orders. From Broker B performance data, we measure performance at the daily level for orders of between 1 and 1,999 shares in the "All Nasdaq" order category, and plot wholesaler performance at the daily level in Figure 2, Panel C. There is considerable variation in day-to-day performance. As before, we first calculate wholesaler performance on the entire 90-day history to obtain a true ranking of wholesalers. We then take samples, this time samples of individual trading days from the last 90 calendar days, and calculate wholesaler performance on each sub-sample. For each of 5,000 sub-samples, we calculate the proportion of wholesaler sample rankings which match the true population rankings, and plot these results in Figure 2, Panel D. Even when observing true performance on 20 out of the last 62 trading days, the first-vs-second wholesaler ranking is correct in only 80% of the 5,000 sub-samples, and the third-vs-fourth wholesaler ranking is correct in less than 60% of the 5,000 sub-samples.

These simulations offer a potential explanation for the difference between our results and those of Huang et al. (2023), who find that the brokers they use "hardly change their routing for our trades based on past execution only one does so at a statistically significant level." Brokers focus on best execution for all trades, and we show that even fairly large subsamples of trades may not capture the true mean of wholesaler performance. Alternative explanations include a potentially different set of brokers, or that brokers in their sample route order flow based on aggregate performance, where aggregate performance differs considerably from odd-lot performance.

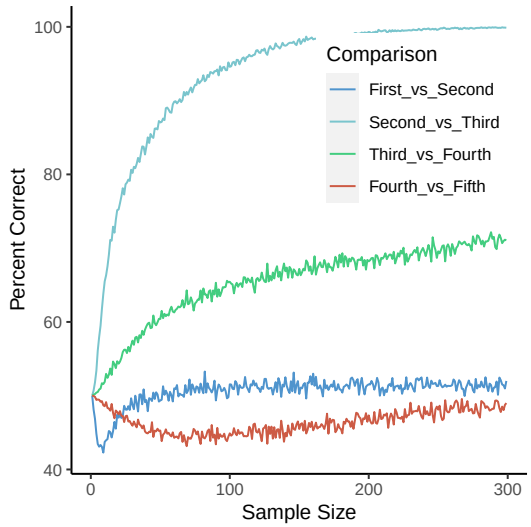
[10] When we calculate wholesaler performance in the sub-sample, if we are comparing two wholesalers and do not observe any trades by one or both of the wholesalers, we assign a 50% chance to correctly specifying the ranking of wholesalers.

Figure 2. Sub-Sample Estimation of Wholesaler Performance. Wholesaler performance is incredibly variable both across orders in a single day (Panel A) and across days of the month (Panel C). True wholesaler rankings are calculated on the population of orders. We then take sub-samples of data, and calculate wholesaler rankings in each sub-sample. In Panels B and D, we take sub-samples of varying sizes (x-axis), and for each possible sample size, we calculate the proportion (y-axis) of 5,000 sub-samples which have sample wholesaler rankings which match the true wholesaler ranking. Panel B presents samples of orders within a single day of trading, while Panel D presents samples of trading days over the prior 90 calendar days.

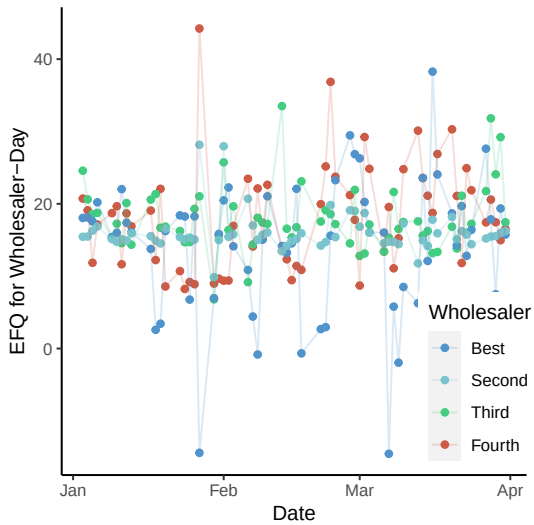
Panel A: Stock EFQ CDF (Broker A)



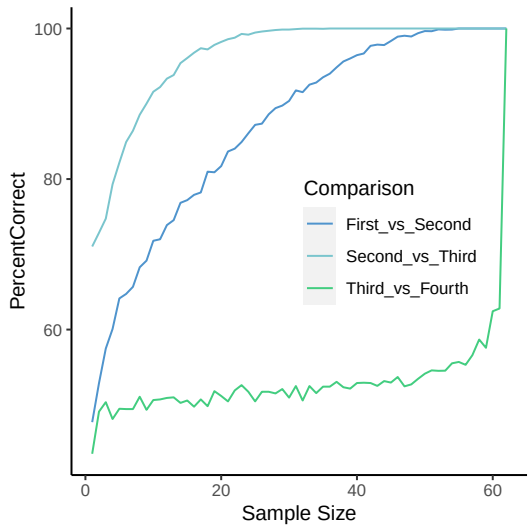
Panel B: Comparison Probability (Broker A)



Panel C: Daily EFQ (Broker B)



Panel D: Comparison Probability (Broker B)



IV. Wholesaler Responsiveness

Having established that brokers route to wholesalers based on past history, we next show evidence suggesting that wholesalers are responsive to incentives provided by the market and by the evaluation method used by brokers. Specifically, we show that improvement offered by wholesalers varies with market volatility, that wholesalers are responsive to broker focuses, and that wholesalers change performance near the end of evaluation windows.

A. Routing and Market Volatility

Brokers route orders based on past performance, meaning that each order a wholesaler receives will factor into future evaluation and subsequent routing, by the broker sending the order. While wholesalers also face competition from the public NBBO, as they must match or improve upon the NBBO for any orders they internalize, their primary source of competition is the degree to which they improve upon the NBBO in aggregate relative to competing wholesalers. This creates pressure for day-by-day competition, as any market conditions in which wholesalers can offer more improvement implies that their competitors could offer more improvement.

Retail trading flow has lower adverse selection than non-retail flow. Days with higher market volatility will have greater rewards to timing trades; as a result, wholesalers should offer more price improvement. To evaluate the relationship between wholesaler improvement and market conditions, we estimate the following regression:

REGRESSION 2: *For each wholesaler j, time period t and security bin k, we estimate:*

$$\begin{aligned} Spread_{jkt} = \quad & \alpha_0 + \alpha_1 Volatility_{jkt} + \alpha_2 Variance_{ijkt} + Depth_{jkt} \\ & + Return_{jkt} + X_{jk} + \epsilon_{ijkt} \end{aligned}$$

We estimate this regression separately with data from each broker. We estimate three specifications with different values for *Spread*: the effective-over-quoted spread, the effective spread paid by retail customers, or the public quoted spread. Results of Regression 2 are presented in Table IV (for Broker C) and Table V (for Broker A). Higher volatility is associated with larger effective spreads for retail investors as well as higher quoted spreads on exchange trades. The relationship between

volatility and the EFQ of retail trades is not significant. We note that brokers typically measure EFQ not as a weighted average of the individual EFQs, but as the total effective spread charged divided by the total quoted spread, which avoids any incentive by wholesalers to offer improvement only when spreads are narrow. Instead, wholesalers are always in competition with each other benchmarked against public quoted spreads. When quoted spreads increase, all wholesalers are evaluated against the wider quoted spreads, while when quoted spreads decrease, all wholesalers are evaluated against the narrower quoted spreads.

For wholesalers, competition over the entire window of history means that price improvement on volatile days has the same weight in evaluation as price improvement on calm days. For retail investors, the effect is a more consistent trading experience, with greater price improvement on more volatile days. Comparing Columns (2) and (3) of Table IV and V, one possible interpretation is that effective spreads paid by retail investors are considerably less volatility-sensitive than public quoted spreads. An alternative interpretation, which we do not have data to evaluate, is that the adverse selection of anonymous exchange trades increases with volatility more rapidly than the adverse selection of retail investor trades.

These results highlight two important aspects of EFQ. The first is that total spreads, rather than averaged spreads, avoid incentives by wholesalers to only offer price improvement when spreads are narrow. Brokers evaluate the total effective spread their customers paid benchmarked against the total quoted spread prevailing when their customers trade. Any action to narrow the spread by wholesalers is rewarded equally.[11] Second, the prevailing quoted spreads at the time retail traders place their orders are very important in evaluating and benchmarking performance. With the updates to Rule 605, the interpretation of the reports would be far more meaningful with benchmarking information on each broker's total quoted spreads, as this offers crucial context to

[11] In contrast, if EFQ is averaged across trades, rather than calculated as total effective divided by total quoted, market makers are incentivized to concentrate their improvement when quoted spreads are narrow, or to concentrate improvement in stocks with narrow quoted spreads. Suppose, for example, that a retail trader places two buy orders, each for 100 shares, and that the half-bid-ask spread is 2 cents at the time of the first order, and 4-cents at the time of the second order. If a market maker fills these orders with 2 cents improvement on the first order (EFQ of 0%), and none on the second (EFQ of 100%), the simple average is an EFQ of 50%. If instead, a wholesaler gives 0 cents improvement on the first order (EFQ of 100%) and 2 cents improvement on the second order (EFQ of 50%), the simple average is an EFQ of 75%. Under a simple weighted average, wholesalers have incentives to avoid giving price improvement when bid-ask spreads are wide. In both cases, however, the wholesaler gave 2 cents of improvement, and the total effective spread over total quoted spread is $\frac{4}{6} = 66\%$, meaning the total effective over quoted spread metric avoids this perverse incentive structure. Battalio and Jennings (2022) provide further discussion of EFQ statistics and the related issue of size improvement.

interpreting an EFQ measure.

Table IV: Market Volatility and Retail Outcomes - Broker C. This table presents results of Regression 2. Volatility is the daily standard deviation of log returns over 30-second returns, using the midquote to calculate returns. Column (1) uses EFQ, the effective spread divided by quoted spread. Column (2) uses the effective spread paid by retail customers, in basis points. Column (3) uses the public quoted spreads, in basis points.

	Dependent variable:		
	EFQ (%)	Effective Spread (BPS)	Public Quoted Spread (BPS)
	(1)	(2)	(3)
Volume-Weighted Average Price	0.016	0.097	0.145
	(0.029)	(0.031)	(0.029)
Total Volume (Millions $)	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Volatility	19.663	730.146	1,726.279
	(37.136)	(39.758)	(37.548)
Depth (Millions $)	0.029	0.483	1.599
	(1.797)	(1.924)	(1.814)
Observations	176,359	176,359	176,369
R^2	0.075	0.328	0.736
Note:		p<0.1;	p<0.05; p<0.01

Table V: Market Volatility and Retail Outcomes - Broker A. This table presents results of Regression 2. Volatility is the daily standard deviation of log returns over 30-second returns, using the midquote to calculate returns. Column (1) uses EFQ, the effective spread divided by quoted spread. Column (2) uses the effective spread paid by retail customers, in basis points. Column (3) uses the public quoted spreads, in basis points.

	Dependent variable:		
	EFQ (%)	Effective Spread (BPS)	Public Quoted Spread (BPS)
	(1)	(2)	(3)
Volume-Weighted Average Price	0.001	0.157	0.110
	(0.006)	(0.022)	(0.032)
Total Volume (Millions $)	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Volatility	34.316	3,491.487	6,818.475
	(20.981)	(79.565)	(115.214)
Depth (Millions $)	0.688	0.416	2.027
	(0.332)	(1.257)	(1.821)
Observations	59,701	59,701	59,701
R^2	0.124	0.615	0.551
Note:		p<0.1;	p<0.05; p<0.01

B. Broker Focus Change

While brokers evaluate wholesalers based on specific criteria, wholesalers in turn are aware and extremely responsive to these criteria. Any changes brokers make in how they allocate orders will impact the responses of wholesalers, including the average price improvement that they can provide. We examine a focus change by Broker B, whereby Broker B changed the weighting of different historical information, and evaluate the strategic responses of wholesalers to this focus change.

We gain critical insight into how a change to existing evaluation criteria influences wholesaler behavior, as well as a deeper view of the competitiveness of wholesalers. Contemporaneous work by Huang et al. (2023) argues that brokers could improve price improvement by changing their routing focus, but the authors "assume that such rerouting would not alter our trade execution nor the competitive dynamics of the wholesaler market." We offer a direct empirical test of that assumption: when a broker changes routing criteria, how do prices change? We reach the opposite conclusion: changes in routing practices change the prices obtained, and we note that there is a trade-off. While there are improvements along one dimension (execution quality of small orders), they are accompanied by dis-improvements along another dimension (execution quality of large orders), consistent with a rational response to the change in the incentive scheme of an agent in multitasking principal-agent models (Holmstrom and Milgrom (1991)).

Broker B implemented its focus change on January 1, 2020. While Broker B uses a holistic process which considers all aspects of order execution both before and after this date, beginning January 1, Broker B includes odd-lot orders (orders for less than 100 shares) in their primary focus.[12] Prior to the focus change, all orders executed within the NBBO, but odd-lot orders frequently executed at a price only slightly better than the public quoted spread (an EFQ close to 100%). Following the change, the EFQ spread charged on odd-lot orders rapidly diminishes, as illustrated in Figure 3, with the average EFQ on odd-lot orders declining to below 50% over the first six months, and below 30% over the next six months.

At the same time that EFQ declines dramatically for odd-lot orders, however, EFQ rises for large orders (orders for more 2,000 or more shares). Figure 3, Panel B illustrates the gradual rise

[12] Broker B monitors executions for all order sizes but placed particular emphasis on orders between 100 and 1,999 shares. Following the focus change, executions between 1 and 1,999 shares became a primary focus.

in EFQ for large orders, with most wholesalers charging an average effective spread approximately equal to the quoted spread on January 1, 2020, and charge an effective spread closer to 150% of the average quoted spread six months after the focus change.[13] To confirm the intuition of Figure 3, we estimate the following regression:

REGRESSION 3: *For each wholesaler j, time period t and security bin k, we estimate:*

$$EFQ_{jkt} = \alpha_0 + \alpha_1 Category_{jkt} + \alpha_2 FocusChange_t + \alpha_3 Category_{jkt} \quad FocusChange_t + X_{jk} + \epsilon_{jkt}$$

Results of Regression 3 are presented in Table VI. Following the introduction of the focus change, odd lots obtain an EFQ spread which is 61% better, while large orders obtain an EFQ spread which is 68% worse.[14] Relative to the category of orders between 1 and 1,999 shares, small orders (less than 100 shares) obtain an EFQ spread which is 44% better while large orders obtain an EFQ spread which is 85% worse.

Any change to market structure will change the behavior of participants. Ernst, Spatt, and Sun (2024a) highlight how order-by-order auctions expose market makers to a winner's curse, and consequently the price improvement they provide in auctions can sometimes be worse than that obtained in the current system of broker's routing. Within the SEC's proposal for order competition, for example, the SEC highlights the existence of widespread mid-quote liquidity, but cannot measure what the availability of such liquidity would be under a different market structure. Our results here provide a unique opportunity to directly observe how prices offered change under different broker routing priorities, and offer further empirical support for the competitiveness of

[13] Note that our data from Broker B has the EFQ measured against the NBBO and not the depth-weighted NBBO; thus any marketable order which "walks the book" will be measured as paying an EFQ larger than 100%. Nonetheless, these orders may still receive price improvement relative to the depth-weighted NBBO: for example, if the best bid is \$9.92 while the best ask consists of 100 shares offered at \$10.00 and 100 shares offered at \$10.10, a market order to buy 200 shares filled by a wholesaler at an average price of \$10.04 would have received price improvement of 1 cent. Compared to the half-bid-ask spread of 4 cents, however, the order would be measured as paying an effective-over-quoted spread of $\frac{10.04 - 9.96}{\frac{1}{2}(10.00 - 9.92)} = 200\%$ despite receiving price improvement. Separately, we obtain exception reports from Broker B, which track any time an order of any size trades at a price worse than the NBBO. Most months have zero such orders, with an occasional month having a low single digit of such orders; both before and after the focus change, we detect no change in the number of order exceptions, indicating that any orders with an EFQ greater than 100% are trading at worse prices than the top of the NBBO, but equal or superior prices to the appropriate depth-weighted NBBO which would be obtained by "walking the book."

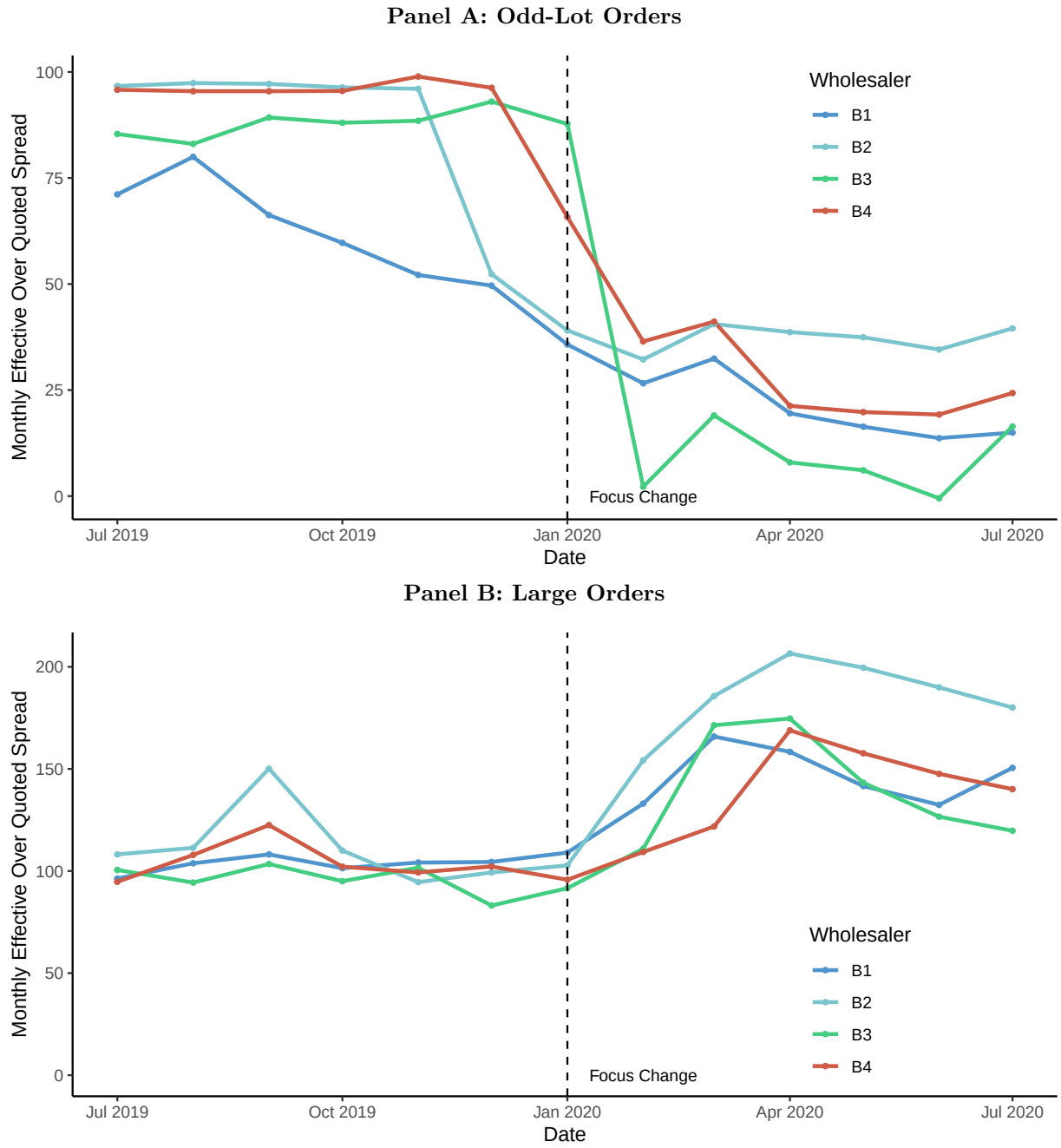
[14] Our regression has the structure of a difference-in-difference estimation. Our data from Broker B has performance in the focus category (1 to 1999 shares), the odd-lot category (1 to 99 shares), and the large-stock category (2000+ shares). Our omitted group is the focus category (1 to 1999 shares); data availability prevents us from estimating an alternative specification where orders for between 100 to 1999 shares are the omitted group.

broker's routing. Prioritizing one dimension of improvement leads to price improvement in that dimension, but that priority comes at the expense of a reduced priority along other dimensions, and potential price dis-improvement, which constrasts with the idea that brokers could achieve obvious improvements by changing their routing practices.

Table VI: Broker Focus-Change - Estimation of Regression 3. FocusChange is a variable which takes the value 1 after Broker B institutes a focus change on January 1, 2020. OddLot is a variable which takes the value 1 for orders of between 1 and 99 shares; LargeOrder is a variable which takes the value 1 for orders over 2,000 shares. We regress EFQ for odd-lot orders in Column (1) and for large orders in Column (3). In Columns (2) and (4), we include order performance on orders between 1 and 1,999 shares as an omitted group. We include fixed effects for each asset and wholesaler, and cluster standard errors by wholesaler.

	Dependent variable: EFQ			
	Odd Lot		Large Orders	
	(1)	(2)	(3)	(4)
OddLot		44.685		
		(7.731)		
LargeOrder				66.695
				(4.483)
FocusChange	61.821	17.085	68.992	16.847
	(8.034)	(1.641)	(11.205)	(1.591)
OddLot FocusChange		44.790		
		(6.642)		
LargeOrder FocusChange				85.549
				(10.749)
Observations	820	1,650	820	1,650
R^2	0.841	0.809	0.544	0.886
Note:				p<0.1; p<0.05; p<0.01

Figure 3. Focus Change by Broker B. Broker B holistically evaluates all aspects of order execution, but began placing more emphasis on odd-lot orders on January 1, 2020. We plot monthly EFQ for each wholesaler for the "all listed" category of stocks for two size categories: odd lots (less than 100 shares) in Panel A and large orders (over 2,000 shares) in Panel B. While odd lots always execute within the NBBO (i.e. customers paid smaller effective spreads than the publicly quoted spreads, paying an EFQ ratio less than 100%), EFQs narrow further once the focus change goes into effect. Conversely, the EFQ ratio widens for large orders. Note that for large orders, the order size is often larger than the available depth at the NBBO; an EFQ of 150% therefore represents a worse price than the NBBO but can still represent a superior price than if the order were to "walk the book" on a public exchange.



C. Strategic Wholesaler Behavior

On every order, wholesalers face a trade-off. Charging a higher bid-ask spread increases their revenue on an individual trade, but harms their average EFQ which may lead to reductions in their order flow allocation in future months. Brokers frequently share information with a wholesaler about its performance, including where it falls relative to its competitors, and each wholesaler also will have access to public SEC 605 and 606 reports. Consequently, wholesalers have abundant information about their position relative to competitors.

Figure 4 plots the difference between the first-ranked and second-ranked wholesalers, across months. We plot the difference in overall score, the proprietary measure of Broker C which includes EFQ as a component, between the first-ranked and second-ranked wholesaler in Panel A, and the difference between the second-ranked and third-ranked wholesaler in Panel B. The month-end for each period is plotted via a gray vertical line. Order allocation decisions are made in the beginning of each month, thus any improvement at the end of the month can immediately lead to a larger allocation share, while improvement in the middle of the month will not pay off with a larger allocation share until the beginning of the next month. We note that there are some rapid shifts in the 90-day-average execution scores around the month-ends. To formally analyze how wholesalers use information about their position to change their execution quality, particularly around the month-end, we estimate the following regression.

REGRESSION 4: *For each wholesaler j, trading date t and security bin k, we estimate:*

$$EFQ_{jkt} = \alpha_0 + \alpha_1 Distance\_Below\_Better_{jkt} + \alpha_2 Distance\_Above\_Worse_{jkt} + X_{jk} + \epsilon_{jkt}$$

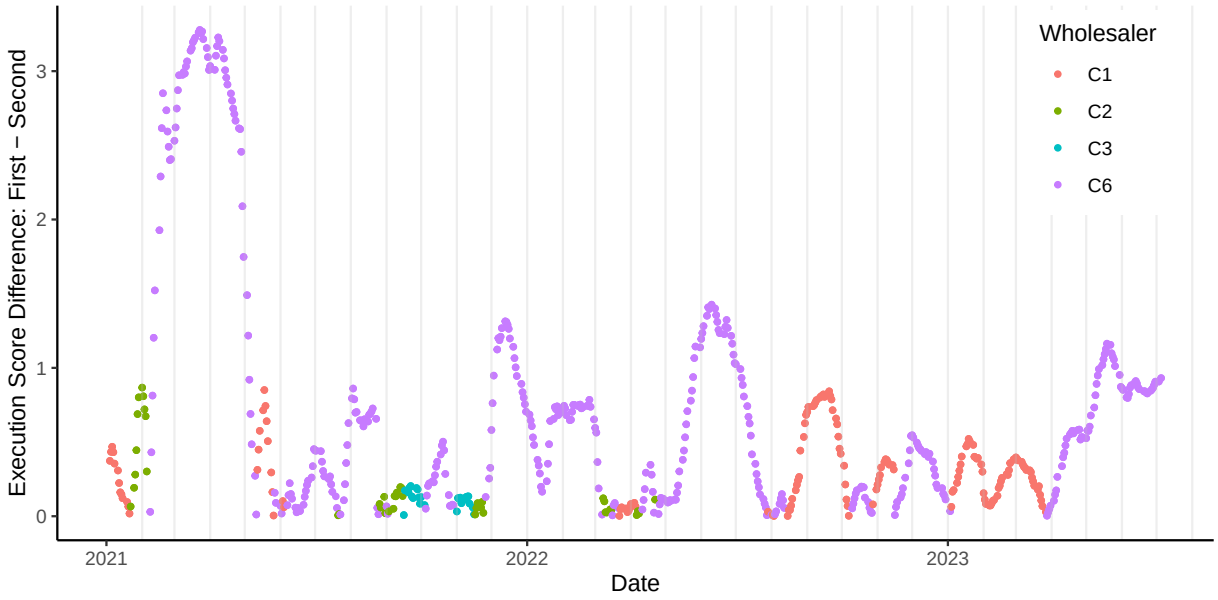
For each wholesaler j, we define $Distance\_Below\_Better_{jt}$ as the absolute value of the EFQ difference between wholesaler j and the closest better-ranked wholesaler (or zero in the case that wholesaler j is the best ranked) on trading date t. We define $Distance\_Above\_Worse_{jt}$ as the absolute value of the EFQ difference between wholesaler j and the closest worse-ranked wholesaler (or zero in the case that wholesaler j is the worst ranked) on trading date t. We also include fixed effects for each wholesaler rank and each trade date, and cluster standard errors by wholesaler. For Broker B, which has four separate routing tables based on security bins, we also include a fixed

effect for each security bin and wholesaler rank.

Results of Regression 4 are presented in Table VII. For both Brokers B and C, wholesalers offer worse EFQ when their nearest leading competitor has a larger 90-day-average lead. This relationship is much stronger in the beginning of the month (more than ten trading days before month-end) relative to the end of the month (last ten trading days), however, consistent with wholesalers having increased incentives to improve performance toward the end of the month. For Broker C, larger leads over competitors are also associated with worse price improvement, though the effect is again much stronger in the second half of the month compared to first half of the month. For Broker B, we find the opposite effect, with larger leads over worse competitors associated with better EFQ, though the effect disappears in the last ten days of the month. For days from evaluation, we find no significance when we estimate using Broker B data. With Broker C data, we see no significance over the full month, but when we restrict to the last ten days, we see wholesalers offering better EFQ the longer the time interval which passes from the previous evaluation, consistent with wholesalers offering price improvement right before decisions on order allocation are made.

Figure 4. Wholesaler Score Differences. Broker C uses a proprietary execution score to rank wholesalers, of which EFQ is an important component. In Panel A, we plot the execution score difference between the first-ranked and second-ranked wholesaler, while in Panel B we plot the execution score difference between the second-ranked and third-ranked wholesaler. We highlight each month-end with a gray vertical line, and note that order allocation decisions are made near the beginning of each month.

Panel A: Difference between First and Second-Best Wholesaler



Panel B: Difference between Second and Third-Best Wholesaler

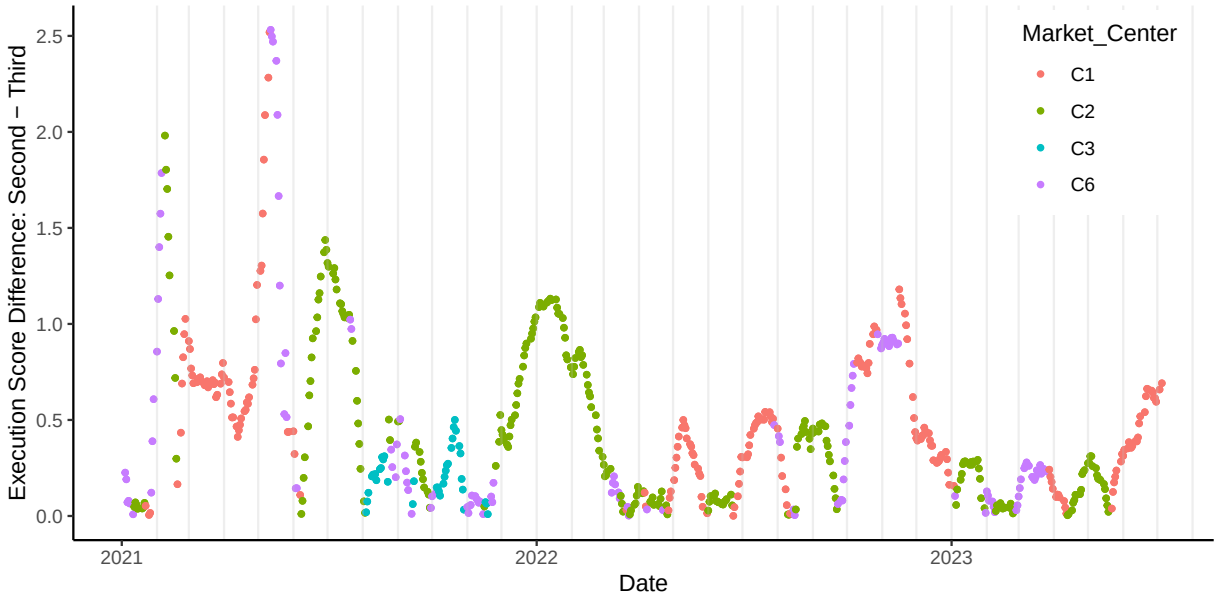


Table VII: Wholesaler Strategic Improvement. We present estimation of Regression 4. *Distance_Below_Better* refers to the absolute value of the EFQ difference between a wholesaler and the closest better-ranked wholesaler (and takes the value 0 if the wholesaler is the best-ranked wholesaler). For Broker C, we use the proprietary score rather than the EFQ. *Distance_Above_Worse* refers to the absolute value of the EFQ difference between a wholesaler and the closest worse-ranked wholesaler (and takes the value 0 if the wholesaler is the worst-ranked). *DaysFromEvaluation* refers to the number of days until the broker re-allocates order flow. We estimate using three groups: the full sample in column (1), the sub-sample of dates at least 10 trading days from the month end in column (2), and the sub-sample of dates at most 10 trading days from the month end in column (3). We fit a fixed effect for each wholesaler and, for Broker B's data, each category and wholesaler rank. All standard errors are clustered by wholesaler.

Panel A: Broker B Data

	Dependent variable: EFQ		
	Full	$>$ 10 Days	$<$ 10 Days
	(1)	(2)	(3)
Distance Below Better	0.472	0.578	0.368
	(0.083)	(0.079)	(0.081)
Distance Above Worse	0.103	0.092	0.116
	(0.040)	(0.026)	(0.055)
Days From Evaluation	0.005	0.044	0.008
	(0.005)	(0.041)	(0.047)
Observations	8,889	4,585	4,304
R^2	0.101	0.089	0.121

Panel B: Broker C Data

	Dependent variable: Proprietary Score		
	Full	$>$ 10 Days	$<$ 10 Days
	(1)	(2)	(3)
Distance Below Better	0.929	1.133	0.594
	(0.081)	(0.058)	(0.189)
Distance Above Worse	0.578	0.631	0.367
	(0.147)	(0.211)	(0.095)
Days From Evaluation	0.020	0.002	0.085
	(0.013)	(0.006)	(0.030)
Observations	2,518	1,698	736
R^2	0.124	0.153	0.101
Note:		p<0.1; p<0.05;	p<0.01

V. Competitive Landscape

Having examined how brokers respond to wholesaler performance, and how wholesalers respond to broker evaluation, we now examine the competitive landscape for order execution. We evaluate the entry of firm into the wholesaling business, differences between wholesalers in liquid and illquid securities, and competition for limit orders, which have strict display requirements.

A. Competition and Wholesaler Entry

Broker A began working with a new wholesaler, referred to as A5, on December 15, 2021. We examine the impact the entry of wholesaler A5 has on competition for Broker A's order flow. We take as exogenous the exact date, December 15, 2021, on which wholesaler A5 begins receiving order flow from Broker A, and compare market competition before and after this date. The extent to which wholesaler A5 enters the market is endogenous, particularly as Broker A routes each security and order size bin separately. We investigate the extent to which the existing competition impacts wholesaler A5's entry with the following regression:

REGRESSION 5: *For each asset j in order size category k in time period t:*

$$A5\_Post\_Share_{jk(t+1)} = \alpha_0 + \alpha_1 Competition_t + \epsilon_{ijkt}$$

We consider two measures of *competition*: $First\_To\_Second$, the difference in EFQ between the first- and second-ranked wholesalers, and $First\_To\_Average$, the difference in EFQ between the first wholesaler and the volume-weighted average EFQ. We calculate each competition measure in the period from November 15 to December 15, 2021. *A5_Post_Share* is the order share of wholesaler A5 in the period from March 3 to April 3, 2022.

Results of Regression 5 are presented in Table VIII. Wholesaler A5 enters categories with a larger gap between the first and second-ranked wholesalers; for every 1% increase in $First\_To\_Second$ (the arithmetic difference in EFQ between the first-ranked to second-ranked wholesaler), wholesaler A5 obtains an additional 0.07% of order share in the post-entry period.[15] Similarly, when the top wholesaler has a 1% larger order share, wholesaler A5 has a 0.15% larger order share in the post-

[15] In relative terms, a one standard deviation increase in *First_To_Second* is associated with a 5% increase in the order share of wholesaler A5 in the post-entry period.

entry period. We find no significant relationship between $First\_To\_Average$ and A5's order share in the post-entry share, however, suggesting the competitiveness of the average wholesaler is less important than that of the highest-ranked wholesaler.

We also consider how competition changes after wholesaler A5 enters with the following regression:

REGRESSION 6: *For each asset j in order size category k in time period t:*

$$Competition_{jkt} = \alpha_0 + \alpha_1 Post_t + \epsilon_{jkt}$$

$Competition$ is either the $First \quad To \quad Second$ or $First \quad To \quad Average$ competition measure, calculated excluding wholesaler A5 as our baseline specification. $Post$ is an indicator for which takes the value 1 for the periods after the entry of wholesaler A5. We estimate using data from November 15 to December 15, 2021 (prior to wholesaler A5's entry), and two time periods following wholesaler A5's entry: March 3 to April 3, 2022 and May 15 to June 15, 2022.

Results are presented in Table IX. Following the entry by wholesaler A5, the gap between the first and second-ranked wholesaler narrows by 1.4%, while the first-to-average gap narrows by 2%. EFQ decreases by 6%, consistent with an increase in competition, though we note that we cannot control for time trends due to the nature of the data.

We find some evidence consistent with displacement, whereby entry by wholesaler A5 aligns with exit by another wholesaler. From Panel A of Table IX, the HHI index increases for the firms excluding A5, and from Panel B of Table IX, there is no change to the gap between the $First \quad to \quad Second$ EFQ offered by wholesalers, while the top wholesaler market share increases by 9.6% in the post-entry period. Thus in categories where wholesaler A5 gains market share, they often gain considerable market share from competitors. Nonetheless, EFQ and the gap between the $First \quad To \quad Average$ spread decrease.

We also estimate Regression 6 with the level or order share (as opposed to a $Post$ indicator) obtained by wholesaler A5 in the March 3 to April 3, 2022 time period. While category-specific order share obtained by wholesaler A5 is endogenous (depending on the EFQ provided by A5 and by competing wholesalers), in categories where wholesaler A5 obtains a larger proportion of the order flow, there are larger decreases in the $First \quad To \quad Average$ difference in EFQ, the HHI, and EFQ itself.

Table VIII: Wholesaler Entry Decision. We estimate Regression 5, which measures the impact of a new wholesaler entering the market. Our outcome variable is $A5\_PostShare$ measures the share of wholesaler A5 obtained in each category in the March 3 to April 3, 2022 data. Categories are the unique symbol and order size bins for each stock. Within each category, we measure $First\ \ To\ \ Second$ (the difference between the effective-over-quoted spread of the first vs. second wholesaler), $First\ \ To\ \ Average$ (the difference between the EFQ of the first wholesaler vs. volume-weighted average), and $FirstFirmOrderShare$ (the share of orders obtained by the top wholesaler in that category). These variables are measured between November 15 to December 15, 2021 and March 3 to April 3, 2022, and we fit a fixed effect for each stock, order size bin, and each top wholesaler.

	Dependent variable: A5_PostShare					
	(1)	(2)	(3)	(4)	(5)	(6)
First-To-Second	0.073					0.001
	(0.033)					(0.057)
First-To-Avg		0.054				0.060
		(0.045)				(0.073)
First Firm Order Share			0.147			0.095
			(0.030)			(0.120)
HHI				17.102		9.990
				(2.809)		(13.297)
Effective-Over-Quoted Spread					0.053	0.148
					(0.036)	(0.062)
Observations	1,461	1,461	1,586	1,586	1,586	1,461
R^2	0.444	0.442	0.396	0.403	0.384	0.447
Note:			p<0.1; p<0.05; p<0.01			

Table IX: Wholesaler Entry and Competition. We estimate Regression 6, which measures the impact of a new wholesaler entering the market. We use data from November 15 to December 15, 2021 (before the wholesaler enters) and contrast it with data from March 3 to April 3, 2022. Post takes the value 1 for the periods after wholesaler A5 enters, while A5_Share is the order share of wholesaler A5.

Panel A: Excluding Wholesaler A5

	Dependent variable:				
	First-To-Second	First-To-Avg	First Firm Order Share	HHI	EFQ
	(1)	(2)	(3)	(4)	(5)
Post	1.367	2.021	0.159	0.112	6.020
	(0.589)	(0.405)	(0.599)	(0.013)	(0.499)
Observations	3,133	3,133	3,467	2,100	3,467
R^2	0.262	0.233	0.384	0.470	0.555

Panel B: Including Wholesaler A5

	Dependent variable:				
	First-To-Second	First-To-Avg	First Firm Order Share	HHI	EFQ
	(1)	(2)	(3)	(4)	(5)
Post	0.674	2.977	9.617	0.070	7.109
	(0.586)	(0.432)	(1.188)	(0.006)	(0.478)
Observations	3,157	3,157	2,106	3,467	3,467
R^2	0.293	0.205	0.441	0.421	0.560

Panel C: Wholesaler A Entry As Independent Variable

	Dependent variable:				
	First-To-Second	First-To-Avg	First Firm Order Share	HHI	EFQ
	(1)	(2)	(3)	(4)	(5)
A5_Share	0.027	0.068	149.499	0.004	0.198
	(0.020)	(0.015)	(302.094)	(0.0002)	(0.017)
Observations	3,157	3,157	2,106	3,467	3,467
R^2	0.294	0.197	0.418	0.458	0.548
Note:					p<0.1; p<0.05; p<0.01

B. Symbol History Bundling and Wholesaler Economics

When brokers route an order, they may consider historical wholesaler performance in past orders in the same symbol, or they may consider historical wholesaler performance in past orders in a set of related symbols, or historical wholesaler performance in all symbols. Each choice changes the competitive landscape for wholesalers.

Broker A routes each symbol separately, with each symbol further divided into separate size bins, which allows us to gain insight into the nature of wholesaler competition across symbols. For each stock-day, we calculate the average level of price improvement given by each wholesaler. We then plot the cumulative distribution function of price improvement in Figure 5. Panel A plots the distributions for the lowest volume stocks, and has a tightly clustered pattern with each wholesaler offering a very similar level of price improvement. Panel B plots the distributions for the highest volume stocks, with much more variation in outcome: wholesalers 3 and 5 give considerably better price improvement than wholesalers 1 and 2, for example. A similar pattern occurs between the smallest and largest orders, as illustrated in Figure 5, Panels C and D. For the smallest orders, differences in price improvement between wholesalers are small, while for the largest orders, differences in price improvement are sometimes large.[16] To gain further insight into the variation of outcomes across stocks, we estimate the following regression:

REGRESSION 7: *For each asset j and order size category k:*

$$Competition_{jk} = \alpha_0 + \alpha_1 Order\_Size\_k + \alpha_2 Trading\_Volume_j + Price_j + X + \epsilon_{jk}$$

$Competition$ is either the $First \quad To \quad Second$ or $First \quad To \quad AverageEFQ$, order share of the first firm, HHI (the Herfindahl–Hirschman Index), and the effective-over-quoted spread. $Order\_Size$ is an indicator for the order-size bin, $Trading\_Volume$ is the decile of trading volume, and $Price$ is the average share price.

Results of Regression 7 are presented in Table X. Larger order sizes are associated with EFQ, larger differences between the first-ranked and either second-ranked or average wholesaler, and

[16] We also note that for the largest orders, price improvement is measured relative to best bid or ask, and not the depth-weighted NBBO; consequently, an order which would "walk the book" would obtain negative price improvement relative to the best bid or ask, even if it executes at a price superior to that of the depth-weighted NBBO.

greater concentration in order share routing (as measured by either the first-firm order share or HHI). When stocks are sorted into deciles by trading volume, stocks in the deciles with more trading volume are associated with smaller differences between wholesalers and lower concentration in order routing.

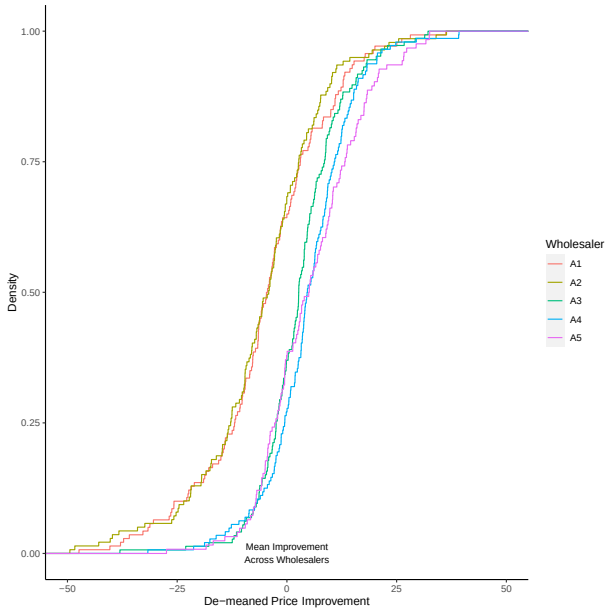
Small stocks have larger bid-ask spreads, but may also have larger inventory holding costs. Dyhrberg et al. (2022) highlight this trade-off between spreads and inventory holding costs, and separately postulate that fixed costs are high in the wholesaling industry, with economies of scale rewarding the largest wholesalers. These economies of scale, however, must also be measured against prevailing spreads, as larger spreads in some stocks may be partially offset by larger spreads. In the cross section of stocks in our sample, we observe greater wholesaler concentration in larger order sizes, and in lower trading volume deciles, suggesting economies of scale may lead to more concentration in these categories.

Table X: Competitive Variation across Stocks. Estimation of Regression 7. $First \quad To \quad Second$ and $First \quad To \quad Average$ are the EFQ difference between the first-ranked and either second-ranked or average wholesaler. $FirstFirmOrderShare$ is the order share of the first-ranked wholesaler. HHI is the Herfindahl–Hirschman index. $OrderSize$ is an indicator for the order-size bin, Trading Volume Decile is the decile of trading volume (with 1 as the lowest volume and 10 as the highest-volume), and Price is the average share price.

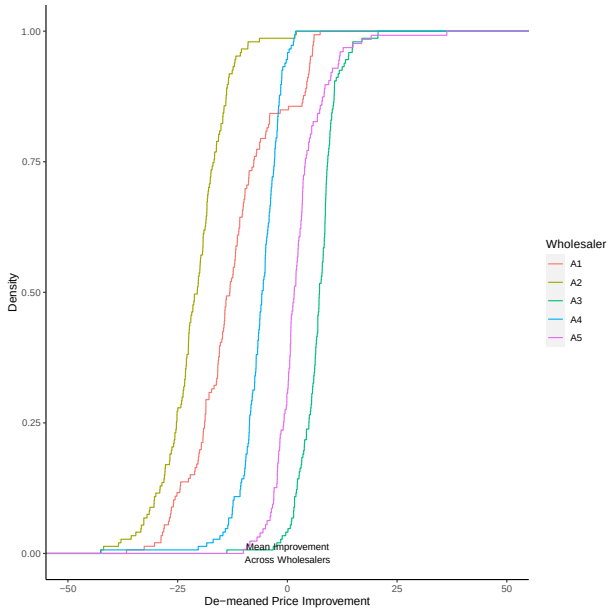
	Dependent variable:				
	First-To-Second	First-To-Avg	First Firm Order Share	HHI	EFQ
	(1)	(2)	(3)	(4)	(5)
Order Size Bin	5.149	1.296	6.274	0.064	9.559
	(0.760)	(0.176)	(0.606)	(0.007)	(0.702)
Trading Volume Decile	0.896	0.553	1.495	0.017	0.265
	(0.126)	(0.028)	(0.156)	(0.001)	(0.246)
Observations	1,840	1,840	1,887	2,130	2,130
R^2	0.131	0.029	0.189	0.174	0.321
Note:					p<0.1; p<0.05; p<0.01

Figure 5. Price Improvement by Stock Characteristics. We plot the distribution of de-meaned price improvement, with 0 corresponding to the mean price improvement across all wholesalers. We divide our data into ten deciles by trading volume. Panel A presents the distribution for the least active stocks, while Panel B presents the distribution for the most active stocks. We also divided data into order size bins. Panel C presents the distribution for the smallest category of order sizes (odd lots), while Panel D presents the distribution for the largest category of order sizes (greater than 5,000 shares). Note that the mean is value-weighted, and wholesalers with superior performance will obtain a larger portion of order flow. Note that A1, A2, ... A5 represent pseudonyms for each wholesaler; the pseudonym-wholesaler pairings for this data sample may or may not match the pseudonym-wholesaler pairings for other data samples from Broker A.

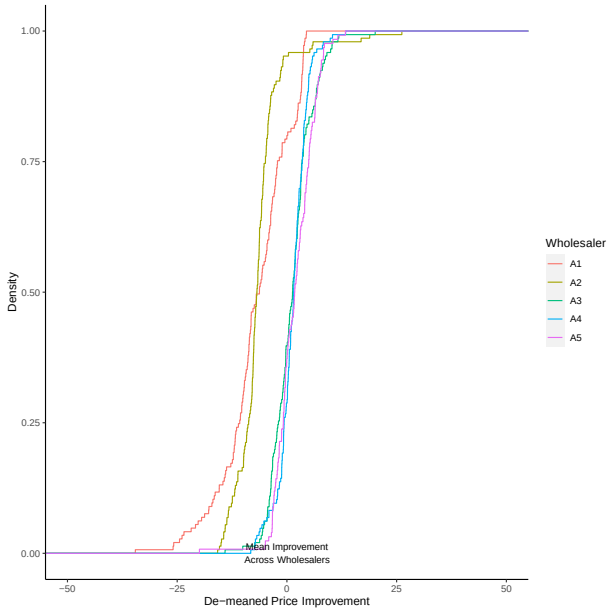
Panel A: Lowest Volume Stocks



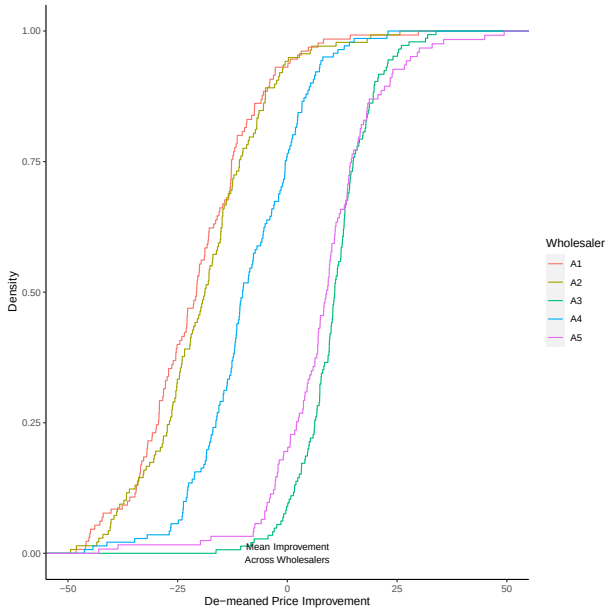
Panel B: Highest Volume Stocks



Panel C: Smallest Orders



Panel D: Largest Orders



C. Limit Orders and Fill Rates

While our primary focus is on the routing of market orders, we also examine the routing of limit orders. In addition to best-execution obligations of brokers, limit orders are also regulated by the Order Display Rule.[17] Any non-marketable limit orders which a wholesaler receives and does not immediately fill must be displayed on an exchange. Effectively, the wholesalers serve largely as a pass-through, receiving limit orders from brokers and posting them to exchanges. Spatt (2019) highlights an important reason for this pass-through behavior: exchange pricing is often volume-tiered. Limit orders are eligible for a market-making rebate, and the more a market participant trades, the larger the rebate they earn. Market makers who are active in the market may qualify for the best rebate tier, while a small or even medium broker may not qualify for the best rebate tier. We compare the persistence of performance between market and limit orders with the following regression:

REGRESSION 8: *For each wholesaler j, time period t and security bin k, we estimate:*

$$RelativePerformance_{jkt} = \alpha_0 + \alpha_1 Prior\_Relative\_Performance_{jkt} + X_{jk} + \epsilon_{jkt}$$

$RelativePerformance$ measures the arithmetic difference between wholesaler j and the average wholesaler performance, while $Prior\_RelativePerformance$ is the same value at a lag of one month. For market orders, performance is measured via effective-over-quoted spreads, as a percentage, while for limit orders, performance is measured as the average monthly fill rate, as a percentage. Controls include a fixed effect for each wholesaler and we cluster standard errors by wholesaler. Results of Regression 8 are presented in Table XI. Of primary interest is the share of explained variation of performance, which we measure with R^2. For limit orders, our simple regression explains just 1% of future performance variation, while for market orders, our simple regression explains 40% of future performance variation.

Our results are consistent with wholesaler variation in limit order performance being minimal, which is natural given the order display rule applies equally to all wholesalers. Whatever orders they do not immediately fill are simply displayed on exchange, with no wholesaler having a

[17] There are three rules covering the display of limit orders. SEC Rule 242.604, SEC Rule 11AC1-4 and FINRA Rule 6460: Display of Customer Limit Orders.

Table XI: Limit Order Persistence. We estimate Regression 8, which measures the persistence of wholesaler performance, using data from Broker B. For limit orders, performance is defined as the arithmetic difference between the fill rate for wholesaler i compared to the average fill rate across all wholesalers. For market orders, performance is defined as the arithmetic difference between the EFQ for wholesaler i compared to the average EFQ across all wholesalers. We include wholesaler fixed effects and cluster standard errors by wholesaler.

	Dependent variable: Performance	
	Limit Orders	Market Orders
	(1)	(2)
Prior Relative Performance	0.001	0.694
	(0.006)	(0.023)
Observations	618	800
R^2	0.013	0.402
Note:	p<0.1;	p<0.05; p<0.01

better or worse display technology. We note, however, that to the extent that there is variation in exchange volume-based pricing between wholesalers, different wholesalers will obtain different financial rewards from posting limit orders to the exchange. Even if the performance, as measured by fill rates obtained by customers, is the same, wholesaler profits will not be.

Our results also highlight an important facet of exchange volume tiering. In September 2024, the SEC adopted Rule 6b-1, which introduces restrictions on exchange volume tiering as well as requiring additional disclosure of volume tiers. Although these results have an obvious immediate impact on exchange customers, wholesalers are a key demographic of exchange customers, and reductions in the exchange price differentials that wholesalers face have the potential to change the nature of competition among wholesalers for retail orders. The largest wholesalers may shoulder a larger portion of the costs of exchange trading, potentially reducing their ability to offer price improvement to retail traders, while smaller wholesalers may see an increased ability to compete for retail order flow if they obtain the same fee and rebate levels.

VI. Interpretation of Results

Our empirical results are clearly inconsistent with two benchmark models, the Bertrand competition model among wholesalers and the model of cooperation between brokers and wholesalers. The Bertrand competition model is rejected because brokers do not route all order flow to the wholesaler

with the strongest recent performance. The model of cooperation is rejected because broker order routing is sensitive to historical performance, and because wholesalers respond to changes in the evaluation focus of the broker. Rejection of these two models leads to the question: What is the theoretical model of the broker-wholesaler relationship that is consistent with the empirical patterns that we document?

We argue that these empirical patterns point to a model of dynamic imperfect competition among wholesalers with private information about their execution costs at each point in time, in which the broker operates as a mechanism designer who designs the order allocation rule to provide incentives to wholesalers to give significant price improvement today and to reward them for price improvement in the past. To make this case, we explore the predictions of relevant theories. There are two related classes of models. Models of the first class are models of auctions, particularly models of all-pay auctions, which are auctions in which every player pays her bid even if the bid loses (e.g., Krishna and Morgan (1997)),[18] The execution price at which a wholesaler executes order flow today can be viewed as a bid in an all-pay auction, and the allocation of future order flow can be viewed as a prize that goes to winning bidders. At the same time, as in auctions of shares, each wholesaler gets a fractional share in future order flow based on past execution quality of this wholesaler and competing wholesalers. The second related literature addresses dynamic contracting with multiple agents. In particular, several papers have studied dynamic contracting problems in which a buyer allocates future business among multiple suppliers based on past performance in various contractual settings, such as complete contracting under hidden effort (Li, Zhang, and Fine (2013)), relational contracting under hidden effort and liquidity constraints of suppliers (Board (2011)), and relational contracting under hidden effort and limited commitment by the principal (Andrews and Barron (2016)). These papers feature the conclusion that good past performance gets rewarded by future allocation of business. Indeed, the interaction between brokers and wholesalers is conceptually similar: a unit of business (order flow) is allocated every period based on historical performance, and the allocation rule is a choice variable of the broker.

Our first finding establishes that brokers allocate the order flow based on the historical performance

[18] An all-pay auction is a common approach to model lobbying and innovation races. Closely related are models of a war of attrition (Krishna and Morgan (1997); Bulow and Klemperer (1999)) and rank-order tournaments (Lazear and Rosen (1981)) and models of auctions of shares, which are auctions in which bidders receive fractional shares of the item for sale (Wilson (1979); Bernheim and Whinston (1986)).

of the wholesalers. This prediction is uniformly shared by both auction theory and dynamic contracting literatures. In both types of models, the allocation of future business is a tool that provides incentives to 'behave well', such as bidding aggressively or engaging in costly hidden effort. Interestingly, while this prediction is common among theories, they differ in how wholesalers should discriminate among bidders. Classic auction theory often predicts that the optimal auction mechanism should be biased in the direction of weaker bidders (e.g. Bulow and Roberts (1989)). Intuitively, such a bias promotes stronger competition by inducing the ex ante stronger bidder to bid more aggressively. In contrast, dynamic contracting models usually predict that the allocation rule should be biased in the direction of an incumbent, that is, the agent that has received more business in the past (Board (2011); Andrews and Barron (2016)).

Our second result establishes that wholesalers realize how they are evaluated by brokers and behave accordingly. This prediction is also uniformly shared by both auction theory and dynamic contracting literatures: In both cases, a foundational assumption is that the agents know the allocation rule (the rules of the auction or the contract) and optimize their behavior. Interestingly, our result that a shift in the broker focus by Broker B in the direction of odd-lot orders leads to wholesalers' competing more aggressively on spreads in odd-lot orders at the expense of less aggressive competition in large orders is consistent with a central prediction of the multi-tasking theory (Holmstrom and Milgrom (1991)). Indeed, its key prediction is that agents will direct their attention into tasks whose performance is measured and rewarded by the principal at the cost of reallocation of attention away from the other tasks. This is exactly what we saw empirically when Broker B had a change in its focus, effectively making incentives to perform on odd-lot orders more high-powered.

Lastly, it is interesting that some brokers bundle different stocks together, while others do not. Theory models suggest two reasons for bundling. First, bundling may be optimal if performance on one stock is informative about a wholesaler's execution quality on the other stock. This idea is related to Holmstrom's informativeness principle in contracting, which states that an information metric should be included in the contract as long as it provides additional marginal information about the agent's effort (Holmstrom (1979)). Second, bundling may be optimal even if wholesalers' execution qualities are independent across stocks in order to induce more aggressive competition among wholesalers. One insight from the literature on optimal bundling in auctions is that bundling

is optimal if there are few competing bidders and suboptimal if there are many competing bidders (Palfrey (1983)). It is therefore interesting that we see variation in bundling at the broker level, even though brokers interact with a similar set of wholesalers. One possible explanation is that different brokers face different investor clienteles, resulting in different correlation among order flow of different stocks and order sizes, which can plausibly result in different optimal bundling decisions.

Lastly, it is worth noting that none of the existing papers, to our knowledge, capture the setting of a broker repeatedly interacting with wholesalers perfectly. Unlike Board (2011), Li et al. (2013), and Andrews and Barron (2016), the key agency friction is hidden information of wholesalers about their holding and execution costs, rather than hidden action. In addition, while our application has features of all-pay auctions and auctions of shares, none of the existing auction theory papers capture it very closely, because interactions are repeated and the allocation rule can depend on bids in multiple past periods. Thus, it remains an open theoretical question what the optimal allocation rule looks like in a setup that closely resembles repeated interactions among a broker and wholesalers. We pursue this question in an ongoing project.

VII. Conclusion

U.S. retail brokers need to repeatedly decide on how to divide order flow among wholesalers. This decision is non-trivial and has both information and incentive components. From the information point of view, past performance provides information to a broker about quality of different wholesalers. From an incentive point of view, allocation of future order flow provides incentives for wholesalers to offer price improvements today. There is some conflict between these two roles: while the information role of order flow allocation favors routing almost all order flow to the wholesaler the broker perceives to be the best, such allocation is likely to be detrimental for incentives, since the best wholesaler may have weak incentives to offer price improvement today if it knows that the beliefs of the broker do not change much. In this paper, we provide empirical analysis of this question using proprietary data from three large retail brokers. We find that brokers frequently adjust flow based on wholesaler performance, that wholesalers are responsive to broker focuses, and that best-execution algorithms have the potential to alter the competitive landscape between liquid-vs-illiquid stocks and large-vs-small orders.

The proprietary data we analyze is similar to the updated Rule 605 that the SEC has adopted, and our work highlights the rich information contained in this data. As the SEC considers changing exchange volume tiering, our work highlights how this also can impact wholesaler competition for retail orders. Our results also indicate that under the current system brokers allocate flow based on past wholesaler performance, which runs contrary to the narrative behind the SEC proposals for order-by-order auctions and a separate best-execution rule which would overlap with the existing FINRA rule.

Our results suggest that Bertrand competition on price improvement is not the right benchmark to model competition among wholesalers. Instead, they suggest a model in which brokers manage competition strategically by designing order flow reallocation rules to balance rewarding past price improvement with creating future incentives to improve, and that wholesalers offer price improvements optimizing subject to such rules. In ongoing work, we are developing a dynamic principal-agent model to study this interaction theoretically. Our goal is to capture the key empirical patterns and use the model to shed light on the pros and cons of various reallocation rules.

REFERENCES

Adams, Samuel W, Connor Kasten, and Eric K Kelley, 2024, How Free is Free? Retail Trading Costs with Zero Commissions, *Journal of Banking & Finance* 165, 107226.

Andrews, Isaiah, and Daniel Barron, 2016, The Allocation of Future Business: Dynamic Relational Contracts with Multiple Agents, *American Economic Review* 106, 2742–2759.

Angel, James J, Lawrence E Harris, and Chester S Spatt, 2011, Equity trading in the 21st century, *The Quarterly Journal of Finance* 1, 1–53.

Angel, James J, Lawrence E Harris, and Chester S Spatt, 2015, Equity trading in the 21st century: An update, *The Quarterly Journal of Finance* 5, 1–39.

Baldauf, Markus, Joshua Mollner, and Bart Zhou Yueshen, 2024, Siphoned Apart: A Portfolio Perspective on Order Flow Segmentation, *Journal of Financial Economics* 154.

Bartlett, Robert P, and Maureen O'Hara, 2024, Navigating the Murky World of Hidden Liquidity.

Battalio, Robert, and Craig W Holden, 2001, A Simple Model of Payment for Order Flow, Internalization, and Total Trading Cost, *Journal of Financial Markets* 4, 33–71.

Battalio, Robert, and Robert Jennings, 2022, Why do Brokers Who do Not Charge Payment for Order Flow Route Marketable Orders to Wholesalers? .

Battalio, Robert H., 1997, Third Market Broker-Dealers: Cost Competitors or Cream Skimmers?, *Journal of Finance* 52, 341–352.

Battalio, Robert H, and Robert H Jennings, 2023, Wholesaler Execution Quality, *Available at SSRN 4304124* .

Bernheim, B. Douglas, and Michael D. Whinston, 1986, Menu Auctions, Resource Allocation, and Economic Influence, *Quarterly Journal of Economics* 101, 1–32.

Bessembinder, Hendrik, and Herbert M. Kaufman, 1997, A cross-exchange comparison of execution costs and information flow for NYSE-listed stocks, *Journal of Financial Economics* 46, 293–319.

Board, Simon, 2011, Relational Contracts and the Value of Loyalty, *American Economic Review* 101, 3349–3367.

Bulow, Jeremy, and Paul Klemperer, 1999, The Generalized War of Attrition, *American Economic Review* 89, 175–189.

Bulow, Jeremy, and John Roberts, 1989, The Simple Economics of Optimal Auctions, *Journal of Political Economy* 97, 1060–1090.

Comerton-Forde, Carole, Katya Malinova, and Andreas Park, 2018, Regulating Dark Trading: Order Flow Segmentation and Market Quality, *Journal of Financial Economics* 130, 347–366.

Dyhrberg, Anne Haubo, Andriy Shkilko, and Ingrid M Werner, 2022, The Retail Execution Quality Landscape, *Fisher College of Business Working Paper* 014.

Easley, David, Nicholas M Kiefer, and Maureen O'Hara, 1996, Cream-skimming or Profit-Sharing? The Curious Role of Purchased Order Flow, *Journal of Finance* 51, 811–833.

Ernst, Thomas, Chester S Spatt, and Jian Sun, 2024a, Would Order-By-Order Auctions Be Competitive?, *Journal of Finance, Forthcoming* .

Ernst, Thomas, Jian Sun, and Chester S Spatt, 2024b, Why Did Retail Liquidity Programs Fail?, *Available at SSRN* .

Holmstrom, Bengt, 1979, Moral Hazard and Observability, *Bell Journal of Economics* 10, 74–91.

Holmstrom, Bengt, and Paul Milgrom, 1991, Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design, *Journal of Law, Economics, & Organization* 7, 24–52.

Hu, Edwin, and Dermot Murphy, 2022, Competition for Retail Order Flow and Market Quality, *Available at SSRN 4070056* .

Huang, Xing, Philippe Jorion, Mina Lee, and Christopher Schwarz, 2023, Who Is Minding the Store? Order Routing and Competition in Retail Trade Execution, *Working Paper. Available at SSRN: 4609895* .

Krishna, Vijay, and John Morgan, 1997, An Analysis of the War of Attrition and the All-Pay Auction, *Journal of Economic Theory* 72, 343–362.

Lazear, Edward P., and Sherwin Rosen, 1981, Rank-Order Tournaments as Optimum Labor Contracts, *Journal of Political Economy* 89, 841–864.

Levy, Bradford, 2022, Price Improvement and Payment for Order Flow: Evidence From a Randomized Controlled Trial, *SSRN Electronic Journal, Available at: https://ssrn. com/abstract* 4189658.

Lewis, Craig, 2024, "Rethinking the Economic Analysis in the SEC's Best Execution Proposal", Pennsylvania Wall, Washington, D.C. [Accessed: 2024 08 25].

Li, Hongmin, Hao Zhang, and Charles H. Fine, 2013, Dynamic Business Share Allocation in a Supply Chain with Competing Suppliers, *Operations Research* 62, 280–297.

Li, Sida, Mao Ye, and Miles Zheng, 2021, Refusing the Best Price?, Technical report, National Bureau of Economic Research.

Macey, Jonathan R, and Maureen O'Hara, 1997, The Law and Economics of Best Execution, *Journal of Financial Intermediation* 6, 188–223.

Palfrey, Thomas R., 1983, Bundling Decisions by a Multiproduct Monopolist with Incomplete Information, *Econometrica* 51, 463–483.

SEC, 2021, Staff report on equity and options market structure conditions in early 2021.

Spatt, Chester S, 2019, Is Equity Market Exchange Structure Anti-Competitive?

van Kervel, Vincent, and Bart Zhou Yueshen, 2023, Anticompetitive Price Referencing, *Available at SSRN: 4545730* .

Wilson, Robert, 1979, Auctions of Shares, *Quarterly Journal of Economics* 93, 675–689.