What Does Best Execution Look Like?*

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

June 6, 2025

Abstract

U.S. retail brokers must provide "best execution" for client orders, though how this works in practice remains unclear. We model the interaction between brokers and wholesalers executing orders as imperfect competition where privately-informed wholesalers compete for future order flow through current price improvement, with brokers designing allocation rules to enhance competition. Using data from three large retail brokers, we document three findings. First, brokers allocate more orders to wholesalers with better past performance. Second, wholesalers recognize this and respond strategically to broker routing criteria and competitive pressures. Finally, there are differences among stocks and brokers in how competition is managed.

*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, the Securities and Exchange Commission Economics Conference, 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

Conflict-of-interest disclosure statement

This paper utilizes data from three retail brokers. This data was provided on the condition of maintaining the anonymity of each broker, that the data only be used to study best execution, and that we would analyze separately each broker's best-execution data. Thomas Ernst and Chester Spatt were direct parties to these restrictions. They received no compensation and retained editorial control of the paper but provided the brokers the right to read the paper to ensure that confidentiality was maintained.

Thomas Ernst:
I have no further conflicts of interest to disclose.

Andrey Malenko:
I have no further conflicts of interest to disclose.

Chester Spatt:
I have no further conflicts of interest to disclose.

Jian Sun:
I have no further conflicts of interest to disclose.

I. Introduction

The vast majority of marketable retail order flow is routed to wholesalers—market makers who specialize in executing retail orders—and not to public exchanges. In routing orders, U.S. retail brokers have a duty of best-execution. This duty is to their customers, to whom they must obtain execution that is "as favorable as possible under prevailing market conditions."[1] The rise of zero-commission trading and the concomitant importance of payment-for-order flow (PFOF), where brokers are paid by wholesalers to route orders away from exchanges, has raised concerns among many market participants about the operation of best-execution. Attendant to these concerns, the SEC has proposed several new rules for equity trading, some recently adopted, motivated in no small part by these concerns about the handling of retail orders.[2]

Our paper contributes basic insights into the implementation of best-execution. We utilize proprietary data, as well as background conversations, from three major U.S. retail brokers to examine what execution looks like in practice: which data brokers measure and emphasize in evaluating wholesalers, how wholesalers respond to competition, and how the choices of brokers can shape the competitive environment. While the possible benefits of wholesalers and the practice of PFOF have been studied,[3] the specifics of how brokers practically implement best execution and competition among wholesalers are comparatively unexplored. To further bridge this gap in understanding, our paper also provides a simple stylized theoretical framework for how a broker operating in the best interest of a client would optimally organize competition among wholesalers, modeling the broker-wholesaler relationship as a mechanism design problem.

In practice, retail brokers work with several wholesalers. These wholesalers offer better prices for retail order flow than open market centers (such as exchanges),[4] due to the

[1]The legal duty of best-execution comes from FINRA Rule 5310.

[2]The SEC has adopted amendments to Rule 605, which requires brokers and market centers to disclose certain information on order execution performance. The SEC has also proposed, but not adopted, an order competition rule, which would mandate individual auctions for marketable retail orders, and a best-execution rule similar to that of the existing FINRA best-execution rule.

[3]Easley, Kiefer, and O'Hara (1996) first documents the practice of PFOF, along with Battalio (1997), while Battalio and Holden (2001) consider welfare implications of the practice. Empirically, Dyhrberg, Shkilko, and Werner (2022) and Battalio and Jennings (2023) examine the performance of wholesalers.

[4]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.

order flow's lower adverse selection or correlation.[5] 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.

To capture the essence of this interaction, we propose the following theoretical framework, which develops the idea that best execution is an example of a mechanism design problem in which the broker (the principal) must design a mechanism to allocate order flow among a fixed set of privately-informed wholesalers (the agents). In our model, wholesalers have private information about their future ability to execute trade and the broker designs the mechanism to minimize the expected price that its customer pays. We solve for the optimal mechanism in a simple two-period dynamic mechanism design problem and show that it can be implemented via an order allocation rule that allocates period-1 order flow contingent on price improvement that wholesalers offered in period 0 The optimal order allocation rule reflects the fundamental trade-off between efficiency, which calls for allocating much of the order flow to the strongest wholesaler, and limiting information rents of wholesalers, which implies allocating some order flow to weaker wholesalers to provide incentives to strong wholesalers to offer good prices in period 0.

We then use this theoretical model to derive a number of implications. Through comparative statics, we argue that in more volatile markets wholesalers offer worse prices, but since the exchange spread also goes up, the relative price improvement is indeterminate. We also examine how equilibrium price improvement responds to changes in the broker's implementation of the optimal allocation rule. Since only total price improvement matters, encouraging better execution for one subgroup will reduce it for others. The model also predicts that market entry by new wholesalers increases competition and reduces spreads, with the impact varying based on the new entrant's efficiency. Lastly, we argue that bundling of different orders (i.e., using the same order allocation for two different stocks or orders of different sizes) may or may not be optimal because it reflects the trade-off between greater allocative efficiency, which calls for unbundling, and limiting information rents of wholesalers, which

[5]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)).

calls for bundling.

Having described the theoretical framework, we next examine how brokers decide on order allocation in practice. 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 conclusions. First, brokers are 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 shape 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, 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 understanding 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, effective spreads charged to

retail investors increase, but so do public quoted spreads, and we cannot reject the null that EFQ ratios do not change.

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. While Broker B considers all aspects of past performance in routing decisions, at the start of the sample it places particular emphasis on performance for orders between 100 and 1,999 shares. Following the focus change, Broker B expanded this focus category to include odd-lot orders (orders between 1 and 99 shares). Wholesalers were informed of the change and respond rapidly, with EFQs for odd-lot orders declining from an average above 80% to an average below 30%. However, we find a contemporaneous increase in EFQ of large orders, suggesting that improving the performance of small orders is not a "free lunch": Any change in priorities for improvement along one dimension may bring disimprovement along other dimensions. This is consistent with the optimal allocation rule in our theoretical model, where only total price improvement matters, not its distribution across order types. 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. However, such attempts to access this liquidity, 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 in 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 then offer 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 individually. We find that the 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 they give. However, we find some evidence consistent with an increase in competition, such as $First-To-Average$ and EFQ, decrease in both incumbent wholesalers and wholesalers including wholesaler A5.

Broker decisions 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. Differences between the first-ranked and either the second-ranked or average wholesaler are larger for larger orders or symbols with less trading volume. Routing each symbol in individual size and symbol bins may create more allocative efficiency, particularly for large orders or low-volume symbols, but at the potential cost of larger information rents to wholesalers relative to routing with a higher degree of bundling.

We also analyze non-marketable limit orders. In contrast to marketable orders, limit orders must be displayed on the exchange. We show that wholesaler performance in limit order fill rates is not 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. However, there are differences in the exchange rebates that wholesalers would collect for posting these limit orders, as exchanges tie rebates to the volume tiers that market makers conduct. In September 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 requiring individual brokers to release performance data on their aggregate order flow; our proprietary data are more detailed than this future proposal, as they provide a breakdown of performance an individual broker obtains from each individual wholesaler. This improved data presents valuable insights on the broker-wholesaler relationship and the importance of quoted spreads in making sense of an effective-over-quoted spread measure. Contemporaneously, the SEC had also proposed a best-execution rule. We show that for equity, the existing FINRA Best-Execution rule has created a relatively competitive market system, but that "Betrand competition" on price improvement is not a good benchmark model for 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 allocative efficiency with limiting information rents of wholesalers. Put 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 the 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 et al. (2022), which uses public data from SEC Rule 605 and 606 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 allow 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 similar 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. Our paper and their paper study similar research questions but use very different data sets. Our paper uses detailed data from three large brokers covering orders of different sizes, and exactly what history (period, order sizes, stocks) each broker uses to evaluate the performance of wholesalers. The advantage of Huang et al. (2023) over our data is that they have data from six brokers, not three. The disadvantage is that their data is based only on small orders, and they do not observe how each broker evaluates performance of wholesalers. Huang et al. (2023) show that on average brokers route more order flow to wholesalers with stronger prior price improvement, but there is heterogeneity across brokers based on whether a broker routes each stock individually or in aggregate. In our paper, routing by all three major brokers is very sensitive to past performance of wholesalers. 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. We also find that the behavior of wholesalers changes when the broker has a focus change in its evaluation scheme, implying that any potential order reallocation would imply a strategic response from wholesalers in what price improvement they are offering. Overall, our evidence points to a mechanism design model, in which the broker designs an order allocation rule to manage competition by strategic and privately informed wholesalers, as a benchmark to understand the broker-wholesalers relationship.

Better-than-NBBO liquidity is by no means unique to wholesalers. Bartlett and O'Hara

(2024) and Levy (2022) highlight the prevalence of odd lot and 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. NBBO liquidity is not all equal–some exchanges charge take fees, while others pay rebates.[6] The NBBO, however, provides a single yardstick for calculating EFQ against which all wholesalers are evaluated. Although the NBBO itself may not be an ideal measure of expected trading costs, it is a standardized benchmark against which all market centers can be equally compared.[7]

Several articles 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 the introduction of a minimum price improvement rule by Canadian regulators reduced the segmentation of retail orders, improved liquidity, but lowered the 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. Theoretical Framework

In this section, we develop a benchmark model in which the broker operating in the best interests of its retail clients designs an optimal mechanism to promote competition among a number of privately informed wholesalers. We show that the optimal mechanism can be implemented via an order allocation rule, which is contingent on the past price improvement offered by wholesalers, and perform comparative statics of the order allocation rule in this mechanism in some fundamental parameters of the model.

Formally, consider the model with the following setup. There are two periods $t = 0, 1$. All players are risk neutral, and there is no discounting.

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

[7] 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 as long as they comprise less than 5% of a broker's total order flow.

Players and payoffs. There is a broker who interacts with $N \geq 2$ wholesalers. At the beginning of each period $t \in \{0, 1\}$, the broker receives a unit mass of buy orders[8] that she allocates between the wholesalers and the market. If wholesaler i executes an order mass $x_{i,t} \in [0, 1]$ at price $s_{i,t}$, his period-t payoff is

$$x_{i,t} \left[s_{i,t} - (\zeta_{i,t} + c x_{i,t}) \right],$$

where $\zeta_{i,t} + c x_{i,t}$ represents the average inventory cost per order for a wholesaler. We assume that there is a publicly observable price $\bar{s}$, and any wholesaler i can only choose $s_{i,t} \leq \bar{s}$. This corresponds to practice: price $\bar{s}$ corresponds to the National Best Bid and Offer (NBBO), and restriction $s_{i,t} \leq \bar{s}$ captures the Order Protection Rule. If the broker sends the order to the market, it is assumed that it is executed at the price $\bar{s}$. Then the wholesaler i's total payoff is

$$\mathcal{U}_i = \sum_{t=0}^{1} x_{i,t} \left[s_{i,t} - (\zeta_{i,t} + c x_{i,t}) \right].$$

The broker aims to maximize their customers' (retail investors') payoff, which is equivalent to minimizing the spread paid by the customers. The broker's total payoff can thus be written as:

$$V = \sum_{t=0}^{1} \left\{ - \sum_{i=1}^{N} x_{i,t} s_{i,t} - \left(1 - \sum_{i=1}^{N} x_{i,t} \right) \bar{s} \right\}. \tag{1}$$

Information environment and strategies. At the beginning of the game, both the period-0 order allocation $\{x_{i,0}\}_{i=1}^{N}$ and the cost parameters $\{\zeta_{i,0}\}_{i=1}^{N}$ are given exogenously and are public. However, only wholesalers observe their own period-1 cost parameter $\zeta_{i,1}$, which is drawn from mutually independent distributions $U\left[Z_i - \frac{b}{2}, Z_i + \frac{b}{2}\right]$.[9] For simplicity, we define $\epsilon_i = \zeta_{i,1} - Z_i \in \left[-\frac{b}{2}, \frac{b}{2}\right]$, and let $\vec{\epsilon} = (\epsilon_1, ...\epsilon_N)'$. We refer to ϵ_i as the type of wholesaler i. Outsiders only know the distributional information. Without loss of generality, we assume $Z_1 \geq Z_2 \geq ... \geq Z_N$, and let $\bar{Z} = \frac{1}{N} \sum_{i=1}^{N} Z_i$. Intuitively, Z_i captures the average

[8] The direction of the order does not matter in our setting.

[9] It is, of course, plausible that a wholesaler may not know its inventory cost next period with certainty. We can equivalently interpret $\zeta_{i,1}$ as wholesaler i's best guess (as of period 0) of its period-1 inventory cost. Since wholesalers are risk neutral and must execute the order flow they are allocated, the analysis is unchanged.

execution cost of wholesaler i, ϵ_i captures the shock to wholesaler i's execution cost (relative to its average), and $\bar{Z}$ captures the average execution cost across all N wholesalers.

A. Mechanism Design Problem

To solve the broker's problem, we assume that the broker can commit to any mechanism at the beginning of period 0. By the revelation principle, it is without loss of generality to restrict attention to direct truth-telling mechanisms. A direct mechanism consists of the broker committing to a period-1 order allocation rule:

$$\vec{x} : \times_{i=1}^{N} \left[-\frac{b}{2}, \frac{b}{2}\right] \rightarrow \left\{ (x_{1,1}, x_{2,1}, ...x_{N,1}) \,|x_{i,1} \geq 0 \quad \forall i, \sum_{i=1}^{N} x_{i,1} \leq 1 \right\},$$

and period-1 transfers from each wholesaler i to the broker, denoted as $t_{i,1}$:

$$\vec{t} : \times_{i=1}^{N} \left[-\frac{b}{2}, \frac{b}{2}\right] \rightarrow \left\{ (t_{1,1}, t_{2,1}, ...t_{N,1}) \in \mathbb{R}^N \right\},$$

as functions of the vector of types reported $\hat{\epsilon}$ by wholesalers. The optimal direct mechanism chooses $\left(\vec{x}(\vec{\epsilon}), \vec{t}(\vec{\epsilon})\right)$ to maximize the broker's total expected payoff (1) subject to providing incentives to each wholesaler to send a truthful message about its type $\hat{\epsilon}_i = \epsilon_i$. The next proposition shows that the order allocation rule in the optimal direct mechanism solves a simple optimization problem:

Proposition 1. *The broker's optimal period-*1 *allocation rule,* $\vec{x}^*(\vec{\epsilon})$*, solves the following problem:*

$$\max_{\vec{x}} \sum_{i=1}^{N} \left[x_i \left(\bar{s} - \frac{b}{2} - Z_i - 2\epsilon_i - cx_i \right) \right] - 2\bar{s} \tag{2}$$

subject to constraints:

$$x_i \geq 0 \, \forall i$$
$$\sum_{i=1}^{N} x_i \leq 1.$$

The derivation of this proposition follows the standard mechanism design argument, and

the underlying intuition is also similar. Since wholesalers are privately informed, they must earn some information rents, as otherwise they would not have incentives to report their types truthfully. Expression (2) captures the payoff of the broker net of information rents provided to wholesalers. It has an intuitive structure. In period 1, since the game ends, all wholesalers execute orders at the highest possible price $\bar{s}$, so the price paid for the unit of order flow in period 1 is $\bar{s}$. In period 0, the orders are executed at a discount relative to $\bar{s}$, as wholesalers compete for future order flow. The expression $\frac{b}{2} + Z_i + 2\epsilon_i$ is the "virtual cost" of type ϵ_i of the wholesaler i (i.e., the cost that takes into account the information rents that must be paid) and the term in the brackets of (2) captures the discount of $\bar{s}$ that the broker receives in period 0 if she allocates order x_i in period $t = 1$ to the wholesaler i of type ϵ_i. The optimal mechanism then chooses the order allocation rule to maximize the expected size of the discount.

For simplicity, we introduce the following assumption on the cost c.

Assumption 1. $c \in \left(\frac{N}{2}\left(Z_1 - \bar{Z}\right) + (N-1)\,b, \frac{N}{2}\left(\bar{s} - \bar{Z} - \frac{3}{2}b\right)\right)$.

Assumption 1 ensures that under the optimal mechanism, the optimal order allocation always satisfies $x_i > 0$ for all i and $\sum_{i=1}^{N} x_i = 1$. This appears to be the most empirically relevant case, as in our sample all wholesalers usually receive some positive allocation, and it is rare for the broker to pass order flow to the market. The following proposition solves the optimal mechanism design problem under this assumption.

Proposition 2. *Under Assumption 1, the broker's optimal allocation rule in Proposition 1,* $\vec{x}^*\left(\vec{\epsilon}\right)$*, is*

$$x_i^* = A_i + \frac{1}{c}\left[\frac{1}{N}\sum_{j=1}^{N}\epsilon_j - \epsilon_i\right], \quad \forall i,$$

where $A_i = \frac{\bar{Z} - Z_i}{2c} + \frac{1}{N}$.

Wholesaler i's expected profit is

$$\mathcal{U}_i = x_{i,0}\left[\bar{s} - \left(\zeta_{i,0} + c x_{i,0}\right)\right] + A_i\left(\frac{b}{2} - \epsilon_i\right) - \frac{N-1}{2Nc}\left(\frac{b^2}{4} - \epsilon_i^2\right).$$

The optimal transfers are shown in the Appendix.

B. Implementation

We next show that there is a realistic indirect mechanism that is equivalent to the optimal mechanism from Proposition 2. In this mechanism, the broker assigns period-0 order flow exogenously and determines the period-1 allocation based on the spreads (or price improvement) offered by wholesalers in period 0. With a slight abuse of notation, we use $\vec{x}(\vec{s})$ to denote the broker's period-1 allocation rule as a function of the vector of spreads $\vec{s}$. The next proposition shows that the optimal allocation rule implements exactly the same order allocation and transfers as the optimal mechanism from Proposition 2:

Proposition 3. *Let $\overrightarrow{s} = (s_1, s_2, ...s_N)$ denote the vector of spreads offered by wholesalers in period 0. There exists an order allocation rule $\vec{x}^*(\overrightarrow{s})$ that implements the optimal mechanism characterized by Proposition 2.*

This optimal order allocation is easy to implement and empirically resembles a broker that adjusts the order flow in response to past performance of wholesalers. The following lemma confirms that it is indeed optimal for the broker to adjust the order allocation in response to past performance.

Lemma 1. *The optimal order allocation rule in Proposition 3 satisfies $\frac{dx_i^*}{ds_i} < 0$ for all i.*

Our first empirical prediction follows from Lemma 1. Although intuitive, this result serves as a natural starting point for the subsequent analysis.

Prediction 1. *The broker allocates more order flow to wholesalers who offer lower spreads.*

C. Properties of the Optimal Order Allocation Rule

Next, we examine the properties of the optimal order allocation rule $\vec{x}^*(\overrightarrow{s})$ with respect to the key parameters of the model. In particular, we consider several scenarios that match our empirical setting. In this section, for simplicity, we assume that the period-0 (exogenous) order allocation is uniform: each wholesaler i receives $x_{i,0} = \frac{1}{N}$.

Effects of Volatile Market

First, we examine how market conditions, specifically volatility, affect equilibrium. In practice, higher volatility can increase execution costs (due to greater risk or adverse selection),

weaken economies of scale for wholesalers, and worsen information asymmetry between the broker and wholesalers. In our model, these effects correspond to increases in the average cost Z_i, the scale cost parameter c, and the information asymmetry parameter b. Higher volatility also tends to widen the exchange spread for the same reasons.

To capture these effects, we introduce a volatility measure σ, and let the model parameters depend on it: average cost is now $Z_i(\sigma)$, scale cost $c(\sigma)$, information asymmetry $b(\sigma)$, and exchange spread $\bar{s}(\sigma)$, where

$$Z_i'(\sigma) = Z_j'(\sigma) > 0 \quad \forall i, j; \quad c'(\sigma) > 0; \quad b'(\sigma) > 0; \quad \bar{s}'(\sigma) > 0.$$

We do not impose specific functional forms and keep these mappings general.

Let $s_i^*(\epsilon_i)$ denote the equilibrium spread under the optimal allocation policy. The average expected spread across all wholesalers is:

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[s_i^*],$$

where the expectation is taken over the vector $\vec{\epsilon}$. We establish the following results.

Lemma 2. *When market volatility σ increases:*

1. *The average spread offered by wholesalers, $\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[s_i^*]$, increases.*
2. *The change in the ratio $\frac{\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[s_i^*]}{\bar{s}}$ is indeterminate.*

Note that the ratio $\frac{\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[s_i^*]}{\bar{s}}$ corresponds to the EFQ spread, thus we obtain the following empirical predictions from Lemma 2.

Prediction 2. *When market volatility increases, the average spread offered by wholesalers increases, but the effect on the EFQ spread is ambiguous.*

Effects of Focus Change

In our data, broker B changes its allocation policy over time. Broker B always considers all aspects of execution quality, but defined a primary focus on orders between 100 and 1,999 shares: performance here had particular weight in decisions for routing order flow. During

our data sample, Broker B expanded the primary focus to include all orders between 1 and 1,999 shares. This change may affect execution quality for both the newly included orders and those that remain in the non-primary focus. To examine this question, we consider the following setting.

Suppose that there are two types of orders: those labeled as primary focus and those labeled as non-primary focus. These labels are purely nominal and do not affect any model primitives. We adopt a simple interpretation of the role of primary focus: once an order is labeled as such, the broker requires the price improvement provided by wholesalers for these orders to exceed a fixed, exogenous threshold. In this sense, wholesalers face a more stringent requirement on price improvement for primary focus orders than simply meeting the NBBO.

We analyze two cases, both assuming that the broker seeks to maximize total price improvement across all orders. In Case 1, only orders between 100 and 1,999 shares are labeled as primary focus, while orders with fewer than 100 shares or more than 2,000 shares are labeled as non-primary focus. In Case 2, the primary focus expands to include all orders between 1 and 1,999 shares, and only orders exceeding 2,000 shares are labeled as non-primary focus.

In our model, only the total price improvement in period 0 affects the period-1 order allocation. Since the order type does not enter any model primitives, the total price improvement offered by wholesalers remains the same in both cases. However, the distribution of price improvement among order types differs. When moving from Case 1 to Case 2, orders between 1 and 99 shares become part of the primary focus. As a result, wholesalers must offer greater price improvement for these orders to meet the broker's requirement. Given that the total price improvement remains unchanged under the optimal allocation, the price improvement for orders exceeding 2,000 shares must decline. Based on this observation, we derive the following empirical prediction:

Prediction 3. *Following the change in focus, price improvement increases for orders between 1 and 99 shares, but decreases for orders more than 2,000 shares.*

Effects of New Wholesaler Entry

In our data, a new wholesaler entered the market on December 15, 2021, potentially altering the competitive landscape of order routing. This setting provides a unique opportunity to examine the effect of new entry, which is central to understanding the broader impact of competition. In our baseline model, the number of wholesalers, N, captures the degree of market competitiveness, and we analyze how changes in this parameter affect the equilibrium.

In the optimal order allocation, the spread provided by wholesaler i in period 0 is given by

$$\bar{s} - \frac{\mathbb{E}[t_i(\vec{\epsilon})]}{1/N}.$$

Next, we present results that characterize the impact of the new entry.

Lemma 3. *Suppose the new wholesaler has an average cost of Z_{N+1}. Then*

1. *the average spread across all wholesalers decreases following its entry, and is an increasing function of Z_{N+1};*
2. *the new wholesaler's expected market share is a decreasing function of Z_{N+1}.*

The above results lead us to the following empirical predictions.

Prediction 4. *After the entry of the new wholesaler, the average spread decreases. In addition, the average spread is lower when the new wholesaler captures a larger market share.*

Effects of Bundling

It is interesting that the brokers in our sample differ in the extent to which they bundle different orders together when choosing their order routing strategies. Motivated by this difference, we consider an extension of the baseline model for multiple different stocks in which we ask about the optimal choice of order bundling.

Formally, we consider a pool with K stocks indexed by $j = 0, 1, 2, ..., K-1$. We compare two different order routing strategies: (1) the broker designs the optimal order allocation for each individual stock, or (2) the broker bundles all stocks together and designs the optimal order allocation for the entire pool. We refer the reader for the details of the model and the solution to the appendix. We show that either bundling or unbundling can be optimal

because of forces that operate in the opposite directions. The trade-off is similar to that in Palfrey (1983). On the one hand, greater allocational efficiency favors unbundling: since execution costs are not perfectly correlated between stocks, a wholesaler in the best position to execute stock j may differ from a wholesaler in the best position to execute stock i, and the bundling of stocks i and j negates this advantage. On the other hand, bundling of stocks reduces information rent that wholesalers collect, and thus promotes more aggressive competition among wholesalers for order flow. Depending on the parameters of the model, one force may dominate the other. Then our model rationalizes why different brokers may optimally choose different bundling decisions.

Wholesalers' different bundling choices also provide a unique setting to test our predictions about the optimal mechanism. Based on the mechanism derived earlier, we characterize the sensitivity of the incentive scheme, measured by $\frac{dx_i^*}{ds_i}$. This expression captures the marginal increase in orders that a wholesaler expects to receive for offering one additional unit of price improvement.

Lemma 4. *Consider the optimal order allocation $\vec{x}^*(\overrightarrow{s})$. The expected sensitivity of the period-1 order flow to the price at which wholesaler i chooses to execute the trade in period 0, $\mathbb{E}[dx_i^*/ds_i]$, is increasing in b.*

Notice that $dx_i^*/ds_i < 0$: lower price at which the wholesaler executes the order in period 0 means better terms for the retail customer, which, in turn, are rewarded with more order flow in period 1. To see the intuition, notice that there are two reasons for making period-1 order flow sensitive to period-0 prices: efficiency and incentives. First, it is efficient for wholesalers with lower execution costs to receive more order flow. Since period-0 low prices are indicative of low execution costs, they imply that it is efficient to give more order flow to such wholesalers. Second, making order flow sensitive to past prices is a way to reward wholesalers for offering good prices in the past. Greater private information of wholesalers (higher b) increases the importance of the incentive channel, which explains this result.

In our data, Broker A makes allocation decisions at the individual stock level, while Broker B groups all stocks into four security bins and allocates orders at the group level. Intuitively, bundling reduces information asymmetry, because in this case only the average cost across grouped stocks matters, and the optimal reward scheme becomes less sensitive in this case. This observation leads to the following empirical prediction.

Prediction 5. *All else being equal,* $\left|\frac{dx_i^*}{ds_i}\right|$ *is lower for a broker who bundles stocks than for a broker who unbundles.*

IV. Broker Responsiveness

We next turn to empirical analysis. We begin it by describing our data and best-execution strategies of brokers; we show that brokers are 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 about their routing practices, including the best-execution statistics for each of the wholesalers to whom they route. This proprietary data was provided for the purposes of studying how each broker individually obtains best-execution for their clients; as a condition of receiving the data, we agreed to analyze each broker's data in isolation.

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 much 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 nondisplayed 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.[10] Brokers 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 and Offer (NBBO). Consequently, brokers route orders to wholesalers based on the past history of how much price improvement, relative to the NBBO, 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.[11] 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 a shorter history, which provides a rapid reward

[10] 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.

[11] 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 $\frac{10.07-10.05}{1/2(10.10-10.00)} = \frac{2}{5} = 40\%$. 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 $\frac{3+3+4}{3+6+7} = \frac{10}{16} = 62.5\%$.

to improvement, and a longer history, which provides a reward to persistence in performance.

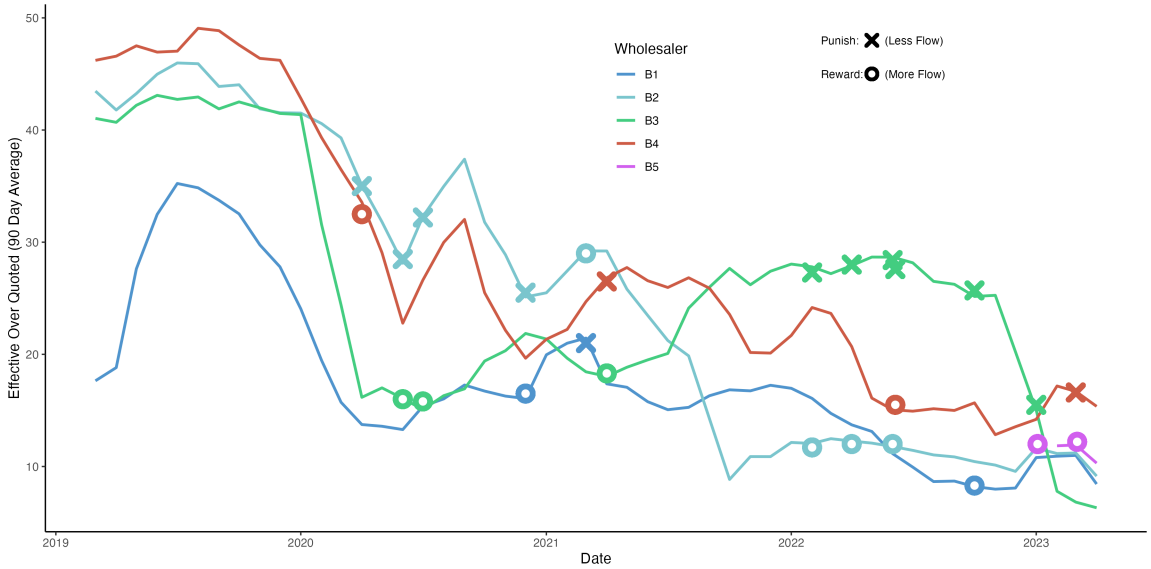
When using past history to inform routing decisions, brokers differ not just in the length but in what history they use to route orders. Broker C measures wholesaler performance in all securities, and routes all orders based on the past performance in all securities. In contrast, Broker A evaluates wholesaler 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 the Nasdaq 100 group of stocks. For Broker C, we plot the negative of their proprietary score (calculated based on all stocks) 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

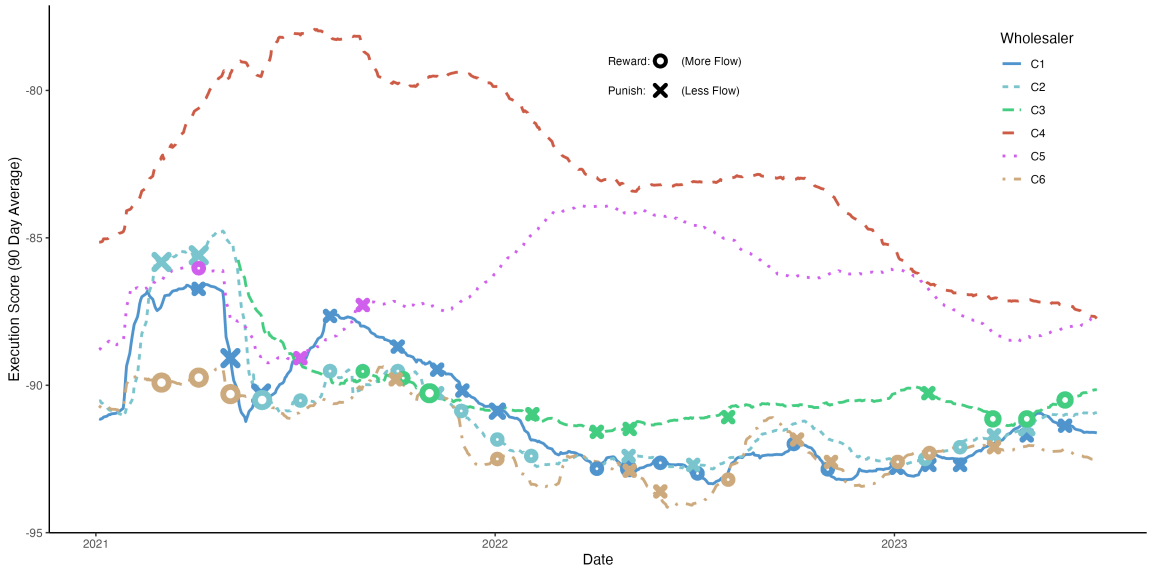


Table I: Best Execution Focus Points. All brokers use a holistic process and consider all aspects of wholesaler performance, but have slightly different focuses in measuring wholesaler performance. We summarize some of the basic differences here. Category refers to the bin that a broker uses, with all 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:*

$$OrderShare_{jkt} = \quad \alpha_0 + \alpha_1 Prior_EFQ_{jkt} + X_{jk} + \epsilon_{jkt}$$

We estimate Regression 1 separately using data from each broker.[12] We use two measurements of $Prior_EFQ$. The first measure of $Prior_EFQ$ is the prior period's effective-over-quoted

[12]The security bin of the fixed effect vector X_{jk} depends on the broker. For broker C, there is one aggregate bin (and thus no fixed effects). For Broker B, there is a single fixed effect for which of the four bins a security belongs to. For broker A, each stock and order size combination is a unique bin.

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-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 X_{kt} 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. To capture the basic fact that better-performing wholesalers obtain more order flow, we do not fit fixed effects for each wholesaler for this regression, though results are similar with wholesaler fixed effects.[13] Our main specification is without wholesaler fixed effects, rather than with, because theory unambiguously predicts the routing-performance relationship without wholesaler fixed effects, while the prediction of the theory with wholesaler fixed effects depends on the persistence of private information of wholesalers.[14]

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. This result is consistent with the optimal order allocation rule characterized in Proposition 3, and is a natural result of the broker's use of strategic order allocation to incentivize wholesalers to offer price improvement.

In detail, 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, which suggests that broker A is using a more sensitive incentive scheme in order routing. Our theoretical results in Lemma 4 suggest that

[13]With a wholesaler fixed effect, signs and magnitudes are extremely similar for Broker A and Broker B. For Broker C, sign and magnitude are similar for the EFQ rank regression, for the prior score regression, magnitude is similar but the standard errors are large so the result is not statistically significant.

[14]Consider two extreme cases. With fully persistent private information, there would be no response after including wholesaler fixed effects, while with i.i.d. private information there would be a response.

this pattern may be driven by the brokers' different approaches to order allocation: Broker A allocates orders at the individual stock level, while Broker B bundles stocks into four security bins and makes allocation decisions at the bin level. As discussed in our model, both bundling choices can be optimal. However, bundling stocks together reduces the private information held by wholesalers. In response, Broker B optimally adopts a less sensitive incentive scheme, whereas Broker A adopts a more sensitive one.

Table II also shows that 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 Specificity 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 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

Table II: Order Share and Execution Quality. We estimate Regression 1 which regresses the *OrderShare* allocated to each wholesaler on prior market performance. *OrderShare* 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

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,[15] 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 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.[16] While the broker will

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

[16]When we calculate wholesaler performance in the sub-sample, if we are comparing two wholesalers and

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	26		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

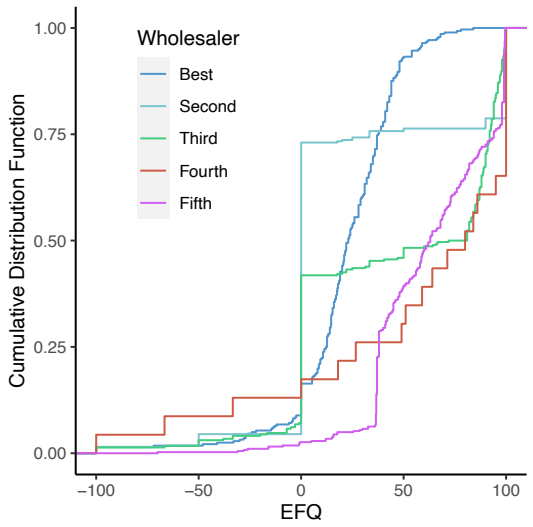
observe wholesaler performance on all trades routed to a wholesaler, a retail customer will see a tiny individual fraction, particularly for a small wholesaler.

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.

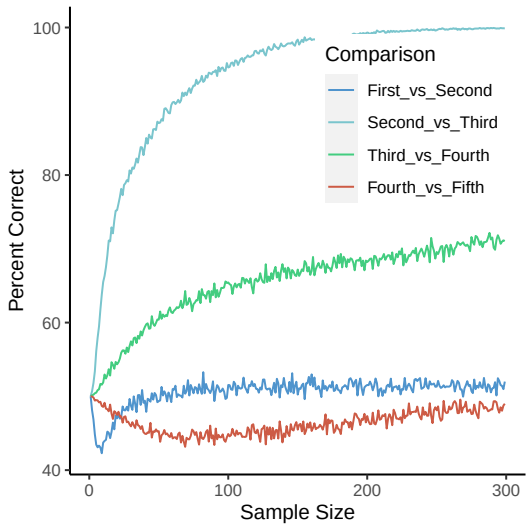
do not observe any trades by one or both of the wholesalers, we assign a 50% chance to correctly specifying the ranking of wholesalers.

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.

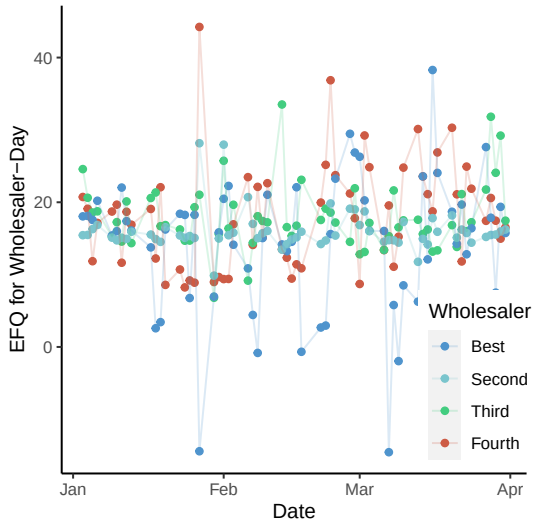
Panel A: Stock EFQ CDF (Broker A)



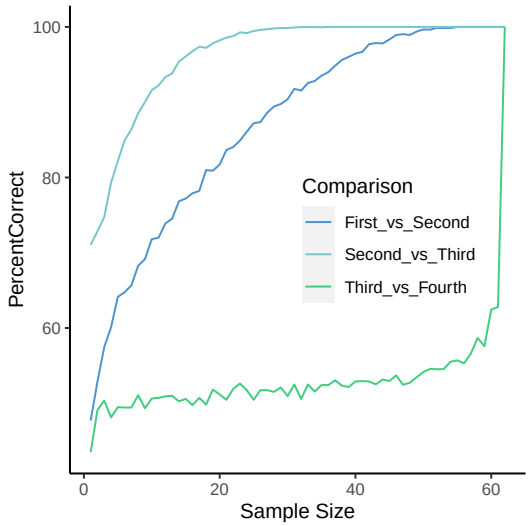
Panel B: Comparison Probability (Broker A)



Panel C: Daily EFQ (Broker B)



Panel D: Comparison Probability (Broker B)



V. Wholesaler Responsiveness

Having established that brokers route to wholesalers based on past history, we next show evidence suggesting that wholesalers respond to incentives provided by the market and by the evaluation method used by brokers. Specifically, we show that the 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 could 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 have potentially larger information asymmetries and thus higher spreads. To evaluate the relationship between wholesaler spreads and market conditions, we estimate the following regression:

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

$$Spread_{jkt} = \quad \alpha_0 + \alpha_1 Volatility_{jkt} + X_{jkt} + \epsilon_{ijkt}$$

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. $Volatility_{jkt}$ is the daily standard deviation of log returns over 30 second intervals, with returns measured using the midquote. X_{jkt} is a vector of controls including total trading volume, volume-weighted average price, and average quoted depth, along with fixed effects for each stock and trade date. 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 and higher quoted spreads on exchange trades. The relationship between volatility and the EFQ of retail trades is not significant. These results are consistent with our theoretical predictions in Lemma 2.

We note that brokers typically measure the 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 spreads on volatile days has the same weight in evaluation as spreads on calm days.

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.[17] 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 interpreting an EFQ measure.

[17]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. However, in both cases, the wholesaler gave 2 cents of improvement, and the total effective spread over the total quoted spread is $\frac{4}{6} = 66\%$, which means that 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.

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 intervals, 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. We include fixed effects for each symbol and trade date.

	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 intervals, 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. We include fixed effects for each symbol, trade date, and order size category.

	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 responsive to these criteria. Any changes brokers make in how they allocate orders may 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.

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.[18] 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 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.[19] To

[18]Broker B monitors executions for all order sizes but placed particular emphasis on orders between 100 and 1,999 shares (the primary focus). Following the focus change on January 1, 2020, the primary focus was expanded to include orders odd-lot orders, so that the primary focus was all orders between 1 and 1,999 shares. The evaluation process is holistic and all data is considered; there is no explicit formula used by Broker B, but they place particular attention to the focus category in making routing decisions.

In our theoretical framework, we model this simply: if an order type is part of the primary focus, the wholesaler must provide price improvement above a positive, exogenous threshold - not just meet the NBBO. Our empirical predictions are based on this interpretation.

[19]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 $\frac{10.04-9.96}{1/2(10.00-9.92)} = 200\%$ 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

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:*

$$EFQ_{jkt} = \alpha_0 + \alpha_1 Size_Category_{jkt} + \alpha_2 FocusChange_t + \alpha_3 Size_Category_{jkt} * FocusChange_t + X_{jk} +$$

$Size_Category$ is one of two indicator variables: either an indicator for odd lot orders (less than 100 shares) or an indicator for large orders (greater than or equal to 2,000 shares). $Focus_Change$ is an indicator which takes the value 1 after Broker B institutes a focus change on January 1, 2020. X is a vector of controls which includes fixed effects for each asset and wholesaler.

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.[20] 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.

The key conclusion of this analysis is that any substantial change in the routing practices of brokers is likely to trigger strategic response by wholesalers. In contemporaneous work Huang et al. (2023) provides a counterfactual analysis in which all their orders are rerouted to the wholesaler who had the best execution during the prior month. The authors find that such hypothetical rerouting would significantly lower spreads. To perform this analysis, the authors assume that "such rerouting would not alter our trade execution nor the competitive dynamics of the wholesaler market". Our analysis in this section offers a direct empirical test of that assumption: when a broker changes routing criteria, how do wholesalers respond? The results reject that assumption: changes in routing practices change wholesalers' behavior. In the focus change we analyze, there is a trade-off: While there are improvements along one

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."

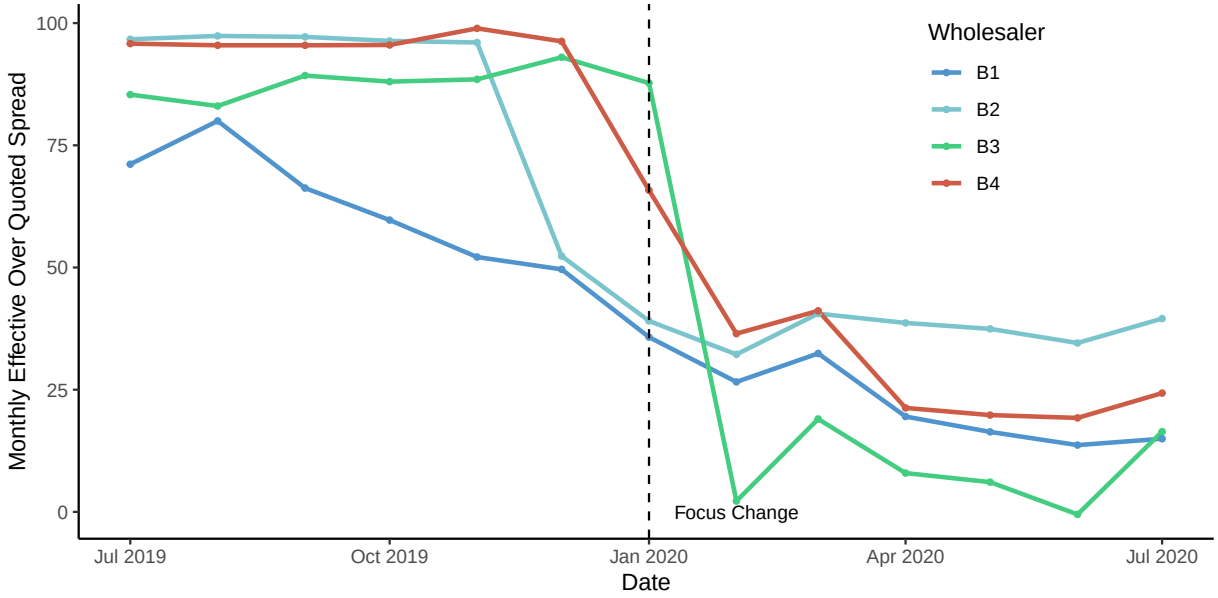
[20]Our regression has the structure of a difference-in-difference estimation. Our data from Broker B has performance in the focus category (1 to 1,999 shares), the odd-lot category (1 to 99 shares), and the large-stock category (2000+ shares). Our omitted group is the focus category (1 to 1,999 shares); data availability prevents us from estimating an alternative specification where orders for between 100 to 1,999 shares are the omitted group.

dimension (execution quality of small orders), they are accompanied by dis-improvements along another dimension (execution quality of large orders). The overall implication is that the analysis of costs and benefits of hypothetical changes in routing practices would generally require a structural model to account for the strategic response from wholesalers to the changes.

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 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. This is also consistent with both the intuitions and the predictions of our model.

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.

Panel A: Odd-Lot Orders



Panel B: Large Orders

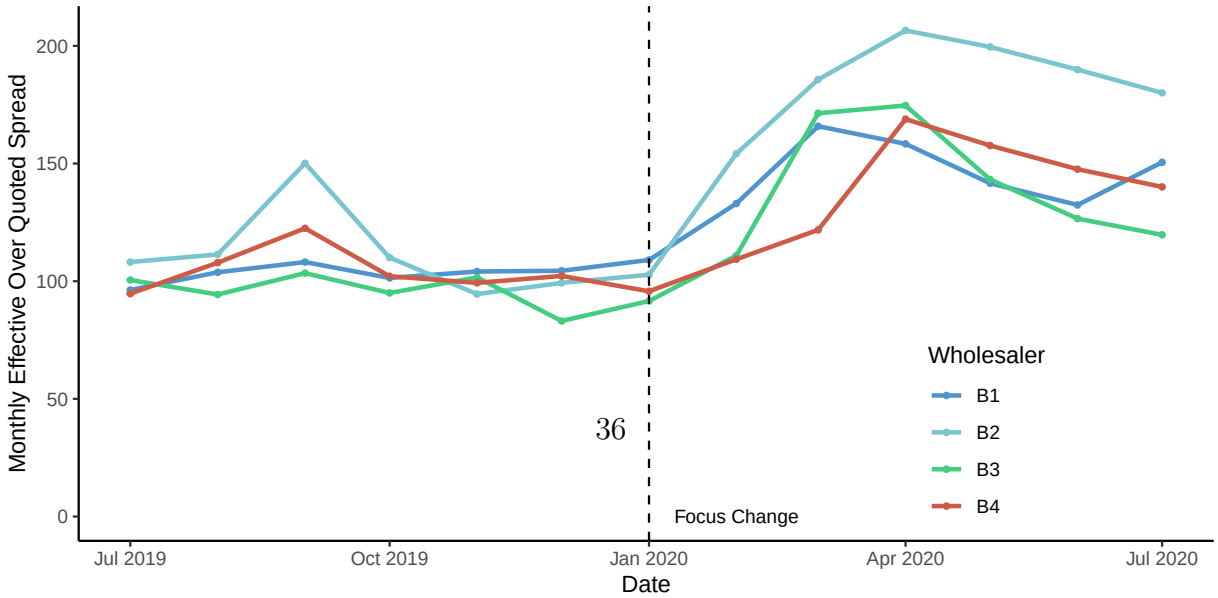


Table VI: Broker Focus-Change - Estimation of Regression 3. FocusChange is a variable which takes the value 1 after Broker B institutes a focus change on January 1, 2020. OddLot is a variable which takes the value 1 for orders of between 1 and 99 shares; LargeOrder is a variable which takes the value 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

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:*

$$EFQ_{jkt} = \alpha_0 + \alpha_1 Distance_Below_Better_{jkt} + \alpha_2 Distance_Above_Worse_{jkt} + X_{jk} + \epsilon_{jkt}$$

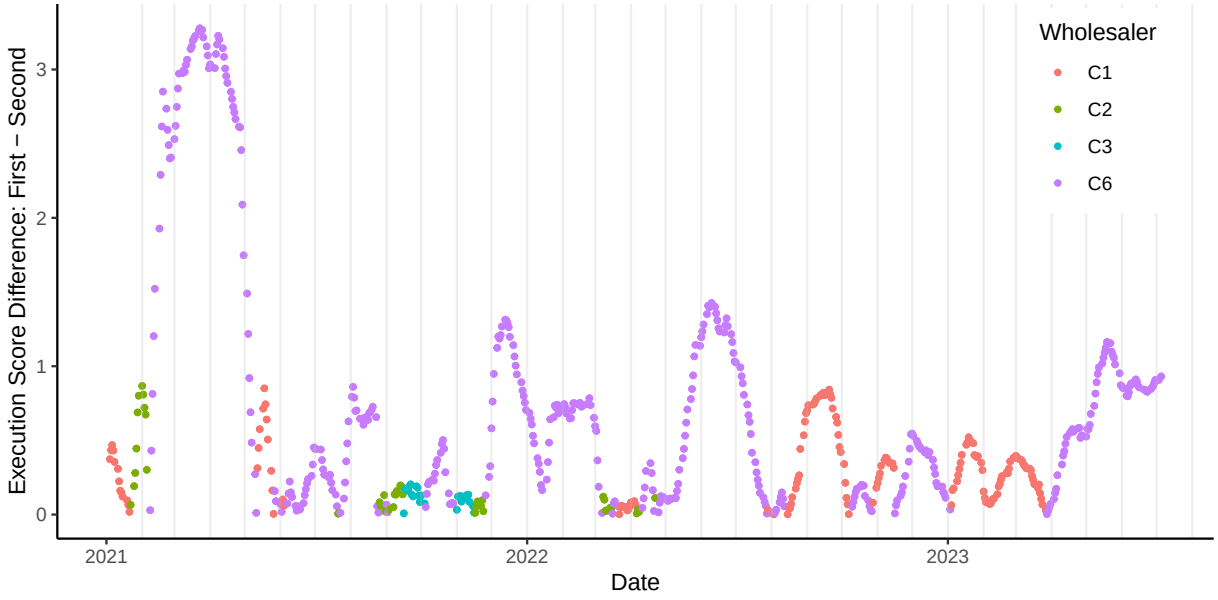
For each wholesaler j, we define $Distance_Below_Better_{jt}$ as the absolute value of the EFQ difference between wholesaler j and the closest better-ranked wholesaler (or zero in the case that wholesaler j is the best ranked) on trading date t. We define $Distance_Above_Worse_{jt}$ as the absolute value 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, 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. Wholesaler color represents the wholesaler in first place (Panel A) or second place (Panel B).

Panel A: Difference between First and Second-Best Wholesaler



Panel B: Difference between Second and Third-Best Wholesaler

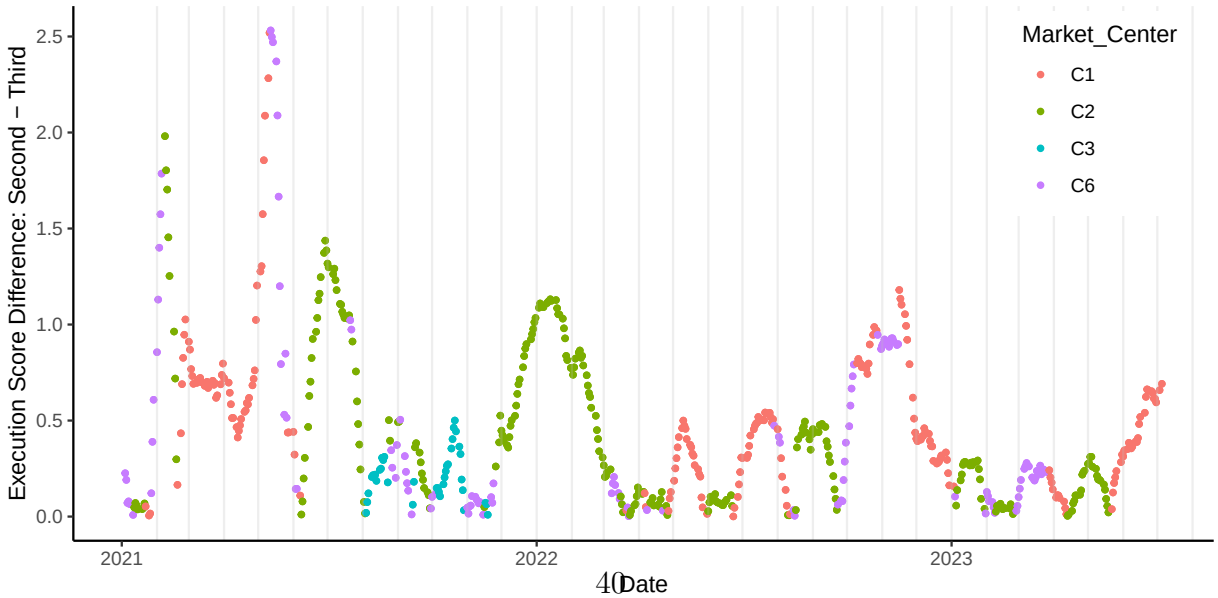


Table VII: Wholesaler Strategic Improvement. We present estimation of Regression 4. *Distance_Below_Better* refers to the absolute value of the EFQ difference between a wholesaler and the closest better-ranked wholesaler (and takes the value 0 if the wholesaler is the best-ranked wholesaler). For Broker 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

VI. 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 illiquid 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:*

$$A5_Post_Share_{jk(t+1)} = \alpha_0 + \alpha_1 Competition_t + X_{jk} + \epsilon_{ijkt}$$

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. We include a vector of fixed effects, X_{jk} which includes a fixed effect for each top wholesaler, each symbol, and each order size category. *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.[21] Similarly, when the top wholesaler has a 1% larger order share, wholesaler A5 has a 0.15% larger order share in the post-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:*

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

Competition is either the $First - To - Second$ or $First - To - Average$ 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. We include a vector of fixed effects, X_{jk} which includes a fixed effect for each lead wholesaler, each symbol, and each order size category.

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 $First - to - Second$ 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 $First - To - Average$ spread decrease.

We also estimate Regression 6 with the level or order share (as opposed to a *Post*

[21] 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.

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 $First - To - Average$ difference in EFQ, the HHI, and EFQ itself. The theoretical predictions in Lemma 3 are in line with this result. In particular, when the new entrant A5 is especially efficient in certain categories, it tends to capture a higher market share and also leads to a greater reduction in average spreads.

Table VIII: Wholesaler Entry Decision. We estimate Regression 5, which measures the impact of a new wholesaler entering the market. Our outcome variable $A5_PostShare$ 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, and cluster standard errors by stock symbol.

	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. We fit a fixed effect for each stock symbol, order size bin, and top wholesaler, and cluster standard errors by stock symbol.

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	3,15746	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. Bundling reduces information rents of wholesalers, but also reduces allocational efficiency. To gain insight into the nature of variation in wholesaler competition across bins, we estimate the following regression:

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

$$Competition_{jk} = \alpha_0 + \alpha_1 Order_Size_k + \alpha_2 Trading_Volume_Decile_j + Price_j + X_{jk} + \epsilon_{jk}$$

$Competition$ is either the $First-To-Second$ or $First-To-Average$, 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, $Trading_Volume_Decile$ is the decile of trading volume, and $Price$ is the average share price. X_{jk} is a fixed effect for the identity of the top wholesaler in asset j and order size category k.

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

liquid stocks, consistent with the costs found by Dyhrberg et al. (2022).

Table X: Competitive Variation across Stocks. Estimation of Regression 7. $First - To - Second$ and $First - To - Avg$ 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, $TradingVolumeDecile$ 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. We include a fixed effect for each top wholesaler and cluster standard errors by top wholesaler.

	Dependent variable:				
	First-To-Second	First-To-Avg	First Firm Order Share	HHI	EFQ
	(1)	(2)	(3)	(4)	(5)
Order Size Bin	5.346***	1.410***	6.358***	0.065***	9.453***
	(0.672)	(0.128)	(0.590)	(0.007)	(0.632)
Trading Volume Decile	−0.867***	−0.536***	−1.480***	−0.017***	−0.283
	(0.115)	(0.030)	(0.160)	(0.001)	(0.235)
Price	0.037*	0.021***	0.016**	0.0002*	−0.022
	(0.012)	(0.003)	(0.004)	(0.0001)	(0.011)
Observations	1,840	1,840	1,887	2,130	2,130
R^2	0.140	0.035	0.190	0.175	0.323
Note:					*p<0.1; **p<0.05; ***p<0.01

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.[22] 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

[22]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.

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:*

$$RelativePerformance_{jkt} = \quad \alpha_0 + \alpha_1 Prior_Relative_Performance_{jkt} + X_{jk} + \epsilon_{jkt}$$

$RelativePerformance$ 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. 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 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

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	

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.

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.

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.

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* .

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* .

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* .

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.

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.

Milgrom, Paul, and Ilya Segal, 2002, Envelope theorems for arbitrary choice sets, *Econometrica* 70, 583–601.

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* .

Appendix A. Extension in Section III: Order Bundling

In this extension, we consider a pool with K stocks indexed by $j = 0, 1, 2, ...K-1$. We compare two different order routing strategies: (1) the broker designs the optimal order allocation for each individual stock, or (2) the broker bundles all stocks together and designs the optimal order allocation for the entire pool. Our goal is to shed light on the optimal choice of order bundling within this simple model. Building on the baseline model, we further impose the following assumption on wholesalers' inventory costs.

When wholesaler i needs to execute $x_{i,j}$ units of stock j, let $x_i = \sum_{j=0}^{K-1} x_{i,j}$, then the total cost for wholesaler i is

$$x_i \left[\sum_{j=0}^{K-1} \zeta_{i,j} + c x_i \right],$$

where $\epsilon_{i,j} = \zeta_{i,j} - Z_{i,j}$ has the following structure:

$$\epsilon_{i,j} = (\theta_i + \lambda_{i,j}) \frac{b}{2K} - \frac{b}{2},$$

where $\theta_i \sim U[0,1]$, and $\lambda_{i,j}$ has the following distribution:

$$\begin{aligned}
&\text{Prob}\left(\lambda_{i,0} = 0 \& \lambda_{i,1} = 2 \& ... \lambda_{i,K-1} = 2K-2\right) = \frac{1}{2K} \\
&\text{Prob}\left(\lambda_{i,0} = 1 \& \lambda_{i,1} = 3 \& ... \lambda_{i,K-1} = 2K-1\right) = \frac{1}{2K} \\
&\qquad \vdots \\
&\text{Prob}\left(\lambda_{i,0} = 2K-1 \& \lambda_{i,1} = 1 \& ... \lambda_{i,K-1} = 2K-3\right) = \frac{1}{2K}.
\end{aligned}$$

Table XII and Table XIII provide two examples.

Prob	0.25	0.25	0.25	0.25
stock 0 $\lambda_{i,0}$	0	1	2	3
stock 1 $\lambda_{i,1}$	2	3	0	1
$\frac{1}{2}(\lambda_{i,0} + \lambda_{i,1})$	1	2	1	2

Table XII Example with 2 stocks

Prob	1/6	1/6	1/6	1/6	1/6	1/6
stock 0 $\lambda_{i,0}$	0	1	2	3	4	5
stock 1 $\lambda_{i,1}$	4	5	0	1	2	3
stock 2 $\lambda_{i,2}$	2	3	4	5	0	1
$\frac{1}{3}(\lambda_{i,0}+\lambda_{i,1}+\lambda_{i,2})$	2	3	2	3	2	3

Table XIII Example with 3 stocks

We can show that $\mathbb{E}[\theta_i + \lambda_{i,j}] = K$, so $\mathbb{E}[\epsilon_{i,j}] = 0$ for any i, j. We assume that $\left\{\theta_i, \{\lambda_{i,j}\}_{j=0}^{K-1}\right\} \perp \left\{\theta_r, \{\lambda_{r,j}\}_{j=0}^{K-1}\right\}$ for any $r \neq i$, and $\theta_i \perp \{\lambda_{i,j}\}_{j=0}^{K-1}$ for any i, j.

Lemma 5. *For any wholesaler i and stock j, $\epsilon_{i,j} \sim U\left[-\frac{b}{2}, \frac{b}{2}\right]$.*

This means that the optimal order allocation for any stock j remains unchanged under the new inventory cost specifications. Let $\hat{\epsilon}_i = \frac{1}{K}\sum_{j=0}^{K-1}\epsilon_{i,j}$, then

$$\hat{\epsilon}_i = \left[\theta_i + \frac{1}{K}\sum_{j=0}^{K-1}\lambda_{i,j}\right]\frac{b}{2K} - \frac{b}{2}.$$

Lemma 6. $\hat{\epsilon}_i \sim U\left[-\frac{b}{2K}, \frac{b}{2K}\right]$.

This implies that when bundling all stocks together, the distribution of the wholesaler's private information for the entire pool still follows a uniform distribution (with tighter bounds).

Lemma 7. *In the broker's order allocation problem, unbundling dominates bundling all stocks together if and only if*

$$\frac{b}{2}\left(1-\frac{1}{K}\right)\left[(N-1)\frac{b}{6}\left(1+\frac{1}{K}\right)-c\right] \geq \frac{N}{4}\left\{\left[\frac{1}{K}\sum_{j=1}^{K}\bar{Z}_k^2 - \bar{Z}^2\right] - \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{K}\sum_{j=1}^{K}Z_{i,k}^2 - Z_i^2\right)\right\},$$

where $\bar{Z}_k = \frac{1}{N}\sum_{i=1}^{N}Z_{i,k}$, $Z_i = \frac{1}{K}\sum_{k=0}^{K-1}Z_{i,k}$, and $\bar{Z} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{K}\sum_{k=0}^{K-1}Z_{i,k}\right)$. This constraint is more easily satisfied when b is larger or c is smaller.

Appendix B. Proofs

Proof of Proposition 1

First, let's rewrite players' payoffs. For wholesalers,

$$\begin{aligned}\mathcal{U}_i &= \sum_{t=0}^{1} x_{i,t}\left[s_{i,t} - (\zeta_{i,t} + cx_{i,t})\right] \\ &= x_{i,0}\left[\bar{s} - (\zeta_{i,0} + cx_{i,0})\right] - x_{i,0}\left(\bar{s} - s_{i,0}\right) + x_{i,1}\left[\bar{s} - (\zeta_{i,1} + cx_{i,1})\right] - x_{i,1}\left(\bar{s} - s_{i,1}\right).\end{aligned}$$

Since period 1 is the last stage of the game, wholesalers must choose the maximum spread $\bar{s}$ in period 1. Since $x_{i,0}\left[\bar{s} - (\zeta_{i,0} + cx_{i,0})\right]$ is exogenously given, we just need to consider the following equivalent utility function for each wholesaler i:

$$U_i = -x_{i,0}\left(\bar{s} - s_{i,0}\right) + x_{i,1}\left[\bar{s} - (\zeta_{i,1} + cx_{i,1})\right].$$

For the broker,

$$\begin{aligned}V &= \sum_{t=0}^{1}\left\{-\sum_{i=1}^{N} x_{i,t}s_{i,t} - \left(1 - \sum_{i=1}^{N} x_{i,t}\right)\bar{s}\right\} \\ &= \sum_{i=1}^{N} x_{i,0}\left(\bar{s} - s_{i,0}\right) - 2\bar{s}.\end{aligned}$$

Then $x_{i,0}\left(\bar{s} - s_{i,0}\right)$ is just the transfer (price improvement) from wholesaler i to the broker in period 0. Consider the optimal mechanism design problem by the broker. By the revelation principle, the optimal mechanism cannot do better than the optimal direct revelation mechanism. Specifically, suppose wholesaler i sends report $\hat{\epsilon}_{i,1}$. The mechanism prescribes allocation $\vec{x}_1 = \vec{x}\left(\hat{\epsilon}_{1,1}, \hat{\epsilon}_{2,1}, ..., \hat{\epsilon}_{N,1}\right)$ and transfers (the total price improvement, from the wholesalers to the broker) $\vec{t}_1 = \vec{t}(\hat{\epsilon}_{1,1}, \hat{\epsilon}_{2,1}, ..., \hat{\epsilon}_{N,1})$. Given a direct mechanism $\left(\vec{x}, \vec{t}\right)$, define

$\bar{x}_i\left(\hat{\epsilon}_{i,1}\right) = \int x_i\left(\hat{\epsilon}_{1,1}, \hat{\epsilon}_{2,1}, ..., \hat{\epsilon}_{N,1}\right) dF_{-i}\left(\hat{\epsilon}_{-i,1}\right)$ and $\bar{t}_i\left(\hat{\epsilon}_{i,1}\right) = \int t_i\left(\hat{\epsilon}_{1,1}, \hat{\epsilon}_{2,1}, ..., \hat{\epsilon}_{N,1}\right) dF_{-i}\left(\hat{\epsilon}_{-i,1}\right)$ to be the expected allocation and transfer from the wholesaler i to the broker after reporting $\hat{\epsilon}_{i,1}$, where F_{-i} is the joint distribution of $(\epsilon_{1,1}, \epsilon_{2,1}, ..\epsilon_{i-1,1}, \epsilon_{i+1,1}.., \epsilon_{N,1})$. For simplicity, we omit all time subscript 1 in the rest of the proof. Then, a wholesaler's payoff with type ϵ_i

and report $\hat{\epsilon}_i$ is:

$$U_i\left(\hat{\epsilon}_i, \epsilon_i\right) = \int_{\epsilon_{-i}} x_i\left(\hat{\epsilon}_i, \epsilon_{-i}\right)\left(\bar{s} - Z_i - \epsilon_i - cx_i\right) dF\left(\epsilon_{-i}\right) - \bar{t}_i\left(\hat{\epsilon}_i\right).$$

Incentive compatibility implies that truthful reporting of ϵ_i must be optimal:

$$U_i\left(\epsilon_i\right) = \max_{\hat{\epsilon}_i}\left\{\int_{\epsilon_{-i}} x_i\left(\hat{\epsilon}_i, \epsilon_{-i}\right)\left(\bar{s} - Z_i - \epsilon_i - cx_i\right) dF\left(\epsilon_{-i}\right) - \bar{t}_i\left(\hat{\epsilon}_i\right)\right\}.$$

By the generalized envelope theorem (Milgrom and Segal (2002)),

$$U_i'\left(\epsilon_i\right) = -\int_{\epsilon_{-i}} x_i\left(\epsilon_i, \epsilon_{-i}\right) dF\left(\epsilon_{-i}\right) = -\bar{x}_i\left(\epsilon_i\right). \tag{B1}$$

Given (B1), we can write down each wholesaler i's expected payoff in the integral form:

$$U_i\left(\epsilon_i\right) = U_i\left(\frac{b}{2}\right) + \int_{\epsilon_i}^{\frac{b}{2}} \bar{x}_i\left(\tau\right) d\tau. \tag{B2}$$

This expression ensures that truthtelling is a local extremum. To guarantee global optimality, we must also require weak monotonicity of $\bar{x}_i\left(\epsilon_i\right)$.

Given (B2), we can express transfers via wholesalers' utilities:

$$\begin{aligned}\bar{t}_i\left(\epsilon_i\right) &= \int_{\epsilon_{-i}} x_i\left(\hat{\epsilon}_i, \epsilon_{-i}\right)\left(\bar{s} - Z_i - \epsilon_i - cx_i\right) dF\left(\epsilon_{-i}\right) - U_i\left(\epsilon_i\right) \\ &= \bar{x}_i\left(\epsilon_i\right)\left(\bar{s} - Z_i - \epsilon_i\right) - c\mathbb{E}\left[x_i\left(\epsilon_i, \epsilon_{-i}\right)^2 | \epsilon_i\right] - U_i\left(\frac{b}{2}\right) - \int_{\epsilon_i}^{\frac{b}{2}} \bar{x}_i\left(\tau\right) d\tau.\end{aligned}$$

Substituting into the broker's objective function, we can write her program as

$$\max_{\vec{x}(\vec{\epsilon})} \sum_{i=1}^{N} \mathbb{E}\left[\bar{t}_i\left(\epsilon_i\right)\right] - 2\bar{s}.$$

It will be convenient to rewrite the expectation using integration by parts:

$$\begin{aligned}\int_{-\frac{b}{2}}^{\frac{b}{2}}\int_{\epsilon_i}^{\frac{b}{2}}\bar{x}_i(\tau)\,d\tau f(\epsilon_i)\,d\epsilon_i &= \int_{-\frac{b}{2}}^{\frac{b}{2}}\int_{-\frac{b}{2}}^{\tau}\bar{x}_i(\tau)\,f(\epsilon_i)\,d\epsilon_i d\tau \\ &= \int_{-\frac{b}{2}}^{\frac{b}{2}}\bar{x}_i(\tau)\,F(\tau)\,d\tau.\end{aligned}$$

Hence, we have the following problem:

$$\sum_{i=1}^{N}\mathbb{E}\left[\bar{x}_i(\epsilon_i)\left(\bar{s}-\left(Z_i+\epsilon_i+\frac{F(\epsilon_i)}{f(\epsilon_i)}\right)\right)-c\mathbb{E}\left[x_i(\vec{\epsilon})^2\,|\epsilon_i\right]\right]-2\bar{s}.$$

Here, $Z_i+\epsilon_i+\frac{F(\epsilon_i)}{f(\epsilon_i)}$ is the virtual cost of wholesaler i.

Next, solve the broker's problem for each realization of the vector of types $\vec{\epsilon}$. For a given realization, the broker's problem is

$$\max_{\vec{x}}\sum_{i=1}^{N}\left(x_i\left(\bar{s}-\left(Z_i+\epsilon_i+\frac{F(\epsilon_i)}{f(\epsilon_i)}\right)\right)-cx_i^2\right)-2\bar{s}$$

subject to constraints:

$$\begin{aligned}x_i &\geq 0\,\forall i \\ \sum_{i=1}^{N}x_i &\leq 1.\end{aligned}$$

Since $\epsilon_i\sim U\left[-\frac{b}{2},\frac{b}{2}\right]$, we know $Z_i+\epsilon_i+\frac{F(\epsilon_i)}{f(\epsilon_i)}=Z_i+2\epsilon_i+\frac{b}{2}$. Then the objective is equivalent to

$$\max_{\vec{x}}\sum_{i=1}^{N}\left(x_i\left(\bar{s}-\left(Z_i+2\epsilon_i+\frac{b}{2}\right)\right)-cx_i^2\right)-2\bar{s}.$$

It is clear that $\bar{x}_i(\epsilon_i)$ is weakly decreasing in ϵ_i.

Internet Appendix

Abstract. This Internet Appendix provides supplementary material for the main paper, including technical proofs and additional theoretical results that support the analysis presented in the main text.

Proofs of Main Results

Proof of Proposition 2

Let's first conjecture (and verify later) that the optimal allocation $\vec{x}^*$ satisfies $x_i^* > 0$ and $\sum_i x_i^* = 1$ for any $\vec{\epsilon}$. Consider the problem in Proposition 1, let λ_i, $i \in \{1, ..., N\}$ be the Lagrange multipliers on the first constraints and μ be the Lagrange multiplier on the second constraint. Then, the Lagrangian is

$$\mathcal{L} = -\sum_{i=1}^{N} \left(x_i \left(\bar{s} - \frac{b}{2} - Z_i - 2\epsilon_i \right) - c x_i^2 \right) + 2\bar{s} - \sum_{i=1}^{N} \lambda_i x_i - \mu \left(1 - \sum_{i=1}^{N} x_i \right).$$

Differentiating in x_i:

$$-\left(\bar{s} - \frac{b}{2} - Z_i - 2\epsilon_i \right) + 2cx_i - \lambda_i + \mu = 0,$$

and we get

$$x_i = \frac{\bar{s} - \mu + \lambda_i - \left(\frac{b}{2} + Z_i + 2\epsilon_i \right)}{2c}.$$

Since we conjecture that $x_i^* > 0$ and $\sum_i x_i^* = 1$ for any $\vec{\epsilon}$, we have $\lambda_i = 0$ for any i and $\mu > 0$. By $\sum_j x_j = 1$, we get

$$\frac{2c}{N} = \bar{s} - \mu - \left(\frac{b}{2} + \bar{Z} + 2\bar{\epsilon}\right) \Longleftrightarrow \mu = \bar{s} - \left(\frac{b}{2} + \bar{Z} + 2\bar{\epsilon}\right) - \frac{2c}{N}.$$

where $\bar{\epsilon} = \frac{1}{N}\sum_{j=1}^{N} \epsilon_j$. Then

$$\begin{aligned} x_i^* &= \frac{\bar{s} - \left(\bar{s} - \left(\frac{b}{2} + \bar{Z} + 2\bar{\epsilon}\right) - \frac{2c}{N}\right) + \lambda_i - \left(\frac{b}{2} + Z_i + 2\epsilon_i\right)}{2c} \\ &= \frac{\bar{Z} - Z_i}{2c} + \frac{1}{N} + \frac{\bar{\epsilon} - \epsilon_i}{c}. \end{aligned} \tag{IA01}$$

Now let us verify that $x_i^* > 0$ and $\sum_{i=1}^{N} x_i^* = 1$ for all $\vec{\epsilon}$. Note that

$$\begin{aligned} x_i^* > 0 \quad \forall i, \vec{\epsilon} &\Longleftrightarrow \frac{\bar{Z} - Z_i}{2c} + \frac{1}{N} + \frac{\bar{\epsilon} - \epsilon_i}{c} > 0 \quad \forall i, \vec{\epsilon} \\ &\Longleftrightarrow \frac{2c}{N} > \left(Z_i - \bar{Z}\right) + 2\left(\epsilon_i - \bar{\epsilon}\right) \quad \forall i, \vec{\epsilon} \\ &\Longleftrightarrow \frac{2c}{N} > \left(Z_1 - \bar{Z}\right) + 2\left(\frac{b}{2} - \frac{1}{N}\frac{b}{2} + \frac{N-1}{N}\frac{b}{2}\right) \\ &\Longleftrightarrow c > \frac{N}{2}\left(Z_1 - \bar{Z}\right) + (N-1)\, b. \end{aligned}$$

Besides, we also require $\sum_{i=1}^{N} x_i^* = 1$ for all $\vec{\epsilon}$, this is equivalent to

$$\mu = \bar{s} - \left(\frac{b}{2} + \bar{Z} + 2\bar{\epsilon}\right) - \frac{2c}{N} > 0$$

for all $\vec{\epsilon}$, which is equivalent to $c < \frac{N}{2}\left(\bar{s} - \bar{Z} - \frac{3}{2}b\right)$. Based on Assumption 1, we know that the optimal x^* solved by (IA01) indeed satisfies $x_i^* > 0$ and $\sum_i x_i^* = 1$ for any $\vec{\epsilon}$.

Wholesaler i's expected payoff is

$$\mathcal{U}_i = x_{i,0}\left[\bar{s} - \left(\zeta_{i,0} + c x_{i,0}\right)\right] + U_i\left(\epsilon_i\right),$$

where

$$\begin{aligned} U_i(\epsilon_i) &= U_i\left(\frac{b}{2}\right) + \int_{\epsilon_i}^{\frac{b}{2}} \bar{x}_i(\tau)\, d\tau \\ &= U_i\left(\frac{b}{2}\right) + \int_{\epsilon_i}^{\frac{b}{2}} \left(A_i - \frac{N-1}{N}\frac{\tau}{c}\right) d\tau \\ &= A_i\left(\frac{b}{2} - \epsilon_i\right) - \frac{N-1}{2Nc}\left(\frac{b^2}{4} - \epsilon_i^2\right). \end{aligned}$$

The wholesaler i's price improvement (the transfer) is

$$\begin{aligned} \bar{t}_i(\epsilon_i) &= \bar{x}_i(\epsilon_i)(\bar{s} - Z_i - \epsilon_i) - c\mathbb{E}\left[x_i(\epsilon_i, \epsilon_{-i})^2 \,|\epsilon_i\right] - U_i(\epsilon_i) \\ &= \bar{x}_i(\epsilon_i)(\bar{s} - Z_i - \epsilon_i) - c\mathbb{E}\left[x_i(\epsilon_i, \epsilon_{-i})^2 \,|\epsilon_i\right] - A_i\left(\frac{b}{2} - \epsilon_i\right) + \frac{N-1}{2Nc}\left(\frac{b^2}{4} - \epsilon_i^2\right). \end{aligned}$$

Proof of Proposition 3

We've solve the optimal allocation rule $x^*(\vec{\epsilon})$ and transfer $\bar{t}_i(\epsilon_i)$ for all i in Proposition 2. Let $h_i = \bar{t}_i^{-1}$ be the inverse function. Then the optimal allocation rule can be implemented by

$$x^*(\vec{s}) = x^*\left(h_1\left((\bar{s} - s_1)x_{1,0}\right), h_2\left((\bar{s} - s_2)x_{2,0}\right), ...h_N\left((\bar{s} - s_N)x_{N,0}\right)\right).$$

Proof of Lemma 1

The order allocation rule is

$$x_i^* = A_i + \frac{1}{c}\left[\frac{1}{N}\sum_{j=1}^{N}\epsilon_j - \epsilon_i\right], \quad \forall i,$$

and

$$\bar{x}_i(\epsilon_i) = A_i - \frac{1}{c}\frac{N-1}{N}\epsilon_i$$

the price improvement is

$$\begin{aligned}
\bar{t}_i(\epsilon_i) &= \bar{x}_i(\epsilon_i)(\bar{s}-Z_i-\epsilon_i) - c\mathbb{E}\left[x_i(\epsilon_i,\epsilon_{-i})^2|\epsilon_i\right] - A_i\left(\frac{b}{2}-\epsilon_i\right) + \frac{N-1}{2Nc}\left(\frac{b^2}{4}-\epsilon_i^2\right) \\
&= \left[A_i - \frac{1}{c}\frac{N-1}{N}\epsilon_i\right](\bar{s}-Z_i-\epsilon_i) - c\mathbb{E}\left[\left(A_i+\frac{1}{c}\left[\frac{1}{N}\sum_{j=1}^{N}\epsilon_j-\epsilon_i\right]\right)^2|\epsilon_i\right] - A_i\left(\frac{b}{2}-\epsilon_i\right) + \frac{N-1}{2Nc}\left(\frac{b^2}{4}-\epsilon_i^2\right) \\
&= \left[A_i - \frac{1}{c}\frac{N-1}{N}\epsilon_i\right](\bar{s}-Z_i-\epsilon_i) - c\mathbb{E}\left[\left(A_i^2 - 2A_i\frac{1}{c}\frac{N-1}{N}\epsilon_i + \frac{(N-1)^2\epsilon_i^2+\sum_{j\neq i}\epsilon_j^2}{N^2c^2}\right)|\epsilon_i\right] - A_i\left(\frac{b}{2}-\epsilon_i\right) + \frac{N-1}{2Nc}\left(\frac{b^2}{4}-\epsilon_i^2\right) \\
&= \left\{A_i(\bar{s}-Z_i) - cA_i^2 - \frac{(N-1)b^2}{12N^2c} - A_i\frac{b}{2} + \frac{(N-1)b^2}{8Nc}\right\} + \left\{2A_i - \frac{1}{c}(\bar{s}-Z_i)\right\}\frac{N-1}{N}\epsilon_i - \left\{\frac{N^2-3N+2}{2N^2c}\right\}\epsilon_i^2.
\end{aligned}$$

Taking derivative on both sides, we have

$$\begin{aligned}
\frac{d\bar{t}_i}{d\epsilon_i} &= \left\{2A_i - \frac{1}{c}(\bar{s}-Z_i)\right\}\frac{N-1}{N} - \left\{\frac{N^2-3N+2}{N^2c}\right\}\epsilon_i \\
&= \frac{N-1}{Nc}\left[2A_ic - (\bar{s}-Z_i) - \left\{\frac{N-2}{N}\right\}\epsilon_i\right].
\end{aligned}$$

Since

$$x_i^* = A_i + \frac{1}{c}\left[\frac{1}{N}\sum_{j=1}^{N}\epsilon_j(\bar{t}_j) - \epsilon_i(\bar{t}_i)\right], \quad \forall i,$$

we have

$$dx_i^* = \frac{1}{c}\left[\frac{1}{N}\sum_{j\neq i}d\epsilon_j(\bar{t}_j) - \frac{N-1}{N}d\epsilon_i(\bar{t}_i)\right].$$

Then

$$\begin{aligned}
\frac{dx_i^*}{d\bar{t}_i} &= -\frac{N-1}{Nc}\frac{d\epsilon_i(\bar{t}_i)}{d\bar{t}_i} \\
&= -\frac{N-1}{Nc}\frac{1}{\left[2A_i - \frac{1}{c}(\bar{s}-Z_i)\right]\frac{N-1}{N} - \left(\frac{N^2-3N+2}{N^2c}\right)\epsilon_i} \\
&= \frac{1}{\left(1-\frac{2}{N}\right)\epsilon_i + \left(\bar{s}-\bar{Z}\right) - \frac{2c}{N}}.
\end{aligned}$$

It is clear that $\frac{dx_i^*}{d\bar{t}_i}$ is positive, then we conclude that $\frac{dx_i^*}{ds_i}$ is negative.

Proof of Lemma 2

Note that

$$\mathbb{E}\left[s_i^*\right] = \bar{s} - N\mathbb{E}\left[\bar{t}_i\left(\epsilon_i\right)\right].$$

In the baseline model,

$$\begin{aligned}
\mathbb{E}\left[\bar{t}_i\left(\epsilon_i\right)\right] &= \left\{A_i\left(\bar{s} - Z_i\right) - cA_i^2 - \frac{\left(N-1\right)b^2}{12N^2c} - A_i\frac{b}{2} + \frac{\left(N-1\right)b^2}{8Nc}\right\} - \left\{\frac{N^2-3N+2}{2N^2c}\right\}\frac{b^2}{12} \\
&= \left\{\left(\bar{s} - Z_i - \frac{b}{2}\right) - cA_i\right\}A_i + \frac{1}{12}\frac{N-1}{N}\frac{b^2}{c} \\
&= \left\{\bar{s} - \frac{Z_i + \bar{Z} + b}{2} - \frac{c}{N}\right\}\left(\frac{\bar{Z} - Z_i}{2c} + \frac{1}{N}\right) + \frac{1}{12}\frac{N-1}{N}\frac{b^2}{c}.
\end{aligned}$$

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\bar{t}_i\left(\epsilon_i\right)\right] &= \frac{1}{N}\sum_{i=1}^{N}\left[\left(\bar{s} - \frac{Z_i + \bar{Z} + b}{2}\right)\frac{\bar{Z} - Z_i}{2c} - \frac{c}{N^2} + \frac{1}{N}\left(\bar{s} - \frac{Z_i + \bar{Z} + b}{2}\right) - \frac{\bar{Z} - Z_i}{2N}\right] + \frac{1}{12}\frac{N-1}{N}\frac{b^2}{c} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\left(-\frac{Z_i + \bar{Z}}{2}\right)\frac{\bar{Z} - Z_i}{2c} + \frac{1}{N}\left(-\frac{Z_i + \bar{Z}}{2}\right)\right] + \frac{1}{12}\frac{N-1}{N}\frac{b^2}{c} - \frac{c}{N^2} + \frac{1}{N}\left(\bar{s} - \frac{b}{2}\right).
\end{aligned}$$

The average expected spread is proportional to

$$\begin{aligned}
\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[s_i^*\right] &= \frac{\bar{s}}{N} - \left[\frac{1}{12}\frac{N-1}{N}\frac{b^2}{c} - \frac{c}{N^2} + \frac{1}{N}\left(\bar{s} - \frac{b}{2}\right)\right] + \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{Z_i + \bar{Z}}{2}\right)\frac{\bar{Z} - Z_i}{2c} + \frac{1}{N}\left(\frac{Z_i + \bar{Z}}{2}\right)\right] \\
&= -\left[\frac{1}{12}\frac{N-1}{N}\frac{b^2}{c} - \frac{c}{N^2} - \frac{1}{N}\frac{b}{2}\right] + \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{Z_i + \bar{Z}}{2}\right)\frac{\bar{Z} - Z_i}{2c} + \frac{1}{N}\left(\frac{Z_i + \bar{Z}}{2}\right)\right].
\end{aligned}$$

Then

$$\begin{aligned}
&\frac{d\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[s_i^*\right]}{d\sigma} \\
=&\frac{1}{N}\frac{dZ_i}{d\sigma} + \left(-\frac{1}{6}\frac{N-1}{N}\frac{b}{c} + \frac{1}{2N}\right)\frac{db}{d\sigma} + \left(\frac{1}{c^2}\left[\frac{1}{4}\left(\frac{1}{N}\sum_{i=1}^{N}Z_i^2 - \bar{Z}^2\right) + \frac{1}{12}\frac{N-1}{N}b^2\right] + \frac{1}{N^2}\right)\frac{dc}{d\sigma}.
\end{aligned}$$

Our assumption $c > \frac{N}{2}\left(Z_1 - \bar{Z}\right) + (N-1)\,b$ implies that $-\frac{1}{6}\frac{N-1}{N}\frac{b}{c} + \frac{1}{2N} > 0$, and

$$\frac{1}{c^2}\left[\frac{1}{4}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i^2 - \bar{Z}^2\right) + \frac{1}{12}\frac{N-1}{N}b^2\right] + \frac{1}{N^2}$$

must be positive by the Cauchy–Schwarz inequality. Then we must have

$$\frac{d\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[s_i^*\right]}{d\sigma} > 0.$$

However, the effect on EFQ, $\frac{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[s_i^*\right]}{\bar{s}}$, is

$$\frac{d\left[\frac{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[s_i^*\right]}{\bar{s}}\right]}{d\sigma} = \frac{\bar{s}\left[\frac{d\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[s_i^*\right]}{d\sigma}\right] - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[s_i^*\right]\frac{d\bar{s}}{d\sigma}}{\bar{s}^2},$$

which can be positive or negative.

Proof of Lemma 3

Following the baseline model's solution method, the broker's objective function after the entry of a new wholesaler can be written as:

$$\sum_{i=1}^{N+1}\mathbb{E}[\bar{t}_i(\epsilon_i)] - 2\bar{s} = \sum_{i=1}^{N+1}\left(x_i\left(\bar{s} - \left(Z_i + 2\epsilon_i + \frac{b}{2}\right)\right) - cx_i^2\right) - 2\bar{s},$$

subject to the constraints:

$$x_i \geq 0 \quad \forall i, \quad \sum_{i=1}^{N+1} x_i \leq 1.$$

Let the broker's optimal payoff be denoted by $V^*(Z_{N+1})$, written explicitly as a function of the new wholesaler's cost parameter Z_{N+1}. Consider two possible values: $Z_{N+1}^1 < Z_{N+1}^2$. Let the corresponding optimal allocation rules be $\vec{x}^{*1}$ and $\vec{x}^{*2}$, respectively.

Define $V(Z_{N+1}; \vec{x})$ as the broker's payoff when the cost of the new wholesaler is Z_{N+1},

and the allocation rule is $\vec{x}$. Then:

$$V^*(Z^1_{N+1}) = V(Z^1_{N+1}; \vec{x}^{*1}), \quad V^*(Z^2_{N+1}) = V(Z^2_{N+1}; \vec{x}^{*2}).$$

By optimality of $\vec{x}^{*1}$ under Z^1_{N+1}, we have:

$$V^*(Z^1_{N+1}) \geq V(Z^1_{N+1}; \vec{x}^{*2}).$$

Moreover, since $V(Z_{N+1}; \vec{x})$ is decreasing in Z_{N+1} (when all $x_i > 0$), we also have:

$$V(Z^1_{N+1}; \vec{x}^{*2}) > V(Z^2_{N+1}; \vec{x}^{*2}) = V^*(Z^2_{N+1}).$$

Combining these two inequalities gives:

$$V^*(Z^1_{N+1}) > V^*(Z^2_{N+1}).$$

Now consider the allocation before the new wholesaler enters, denoted by $\vec{x}^{*0}$. By the uniqueness and optimality of solutions, we also have:

$$V^*(Z^1_{N+1}) > V(Z^1_{N+1}; \vec{x}^{*0}),$$

where $V(Z^1_{N+1}; \vec{x}^{*0})$ represents the broker's payoff before the entry.

Since the broker's payoff is the negative of the average spread, a higher broker payoff implies a lower average spread. Therefore, we conclude that:

- The average spread is increasing in Z_{N+1},
- The entry of a new wholesaler reduces the average spread.

Finally, based on the baseline model, the expected market share of the new wholesaler is:

$$\mathbb{E}[x^*_{N+1}] = \frac{\frac{1}{N+1}\sum_{i=1}^{N+1} Z_i - Z_{N+1}}{2c} + \frac{1}{N+1},$$

which is clearly decreasing in Z_{N+1}.

Proof of Lemma 4

Based on the proof in Lemma 1, we have

$$\mathbb{E}\left[\frac{dx_i^*}{d\bar{t}_i}\right] = \int_{-\frac{b}{2}}^{\frac{b}{2}} \frac{1}{b} \frac{1}{\left(1-\frac{2}{N}\right)\epsilon_i + \left(\bar{s}-\bar{Z}\right) - \frac{2c}{N}} d\epsilon_i$$

Since $\frac{1}{\left(1-\frac{2}{N}\right)\epsilon_i+\left(\bar{s}-\bar{Z}\right)-\frac{2c}{N}}$ is a positive (because we assume $c < \frac{N}{2}\left(\bar{s}-\bar{Z}-\frac{3}{2}b\right)$), decreasing, convex function of ϵ_i, $\mathbb{E}\left[\frac{dx_i^*}{d\bar{t}_i}\right]$ must be increasing in b. This can be proved by Jensen's inequality.

Proof of Lemma 7

Let $\bar{Z}_k = \frac{1}{N}\sum_{i=1}^{N} Z_{i,k}$, $\bar{\epsilon}_k = \frac{1}{N}\sum_{i=1}^{N} \epsilon_{i,k}$, $\bar{Z} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{K}\sum_{k=0}^{K-1} Z_{i,k}\right)$, $Z_i = \frac{1}{K}\sum_{k=0}^{K-1} Z_{i,k}$. Based on our results in the baseline model, the optimal order allocation rule for any specific stock k is

$$x_{i,k} = A_{i,k} + \frac{1}{c}\left(\frac{1}{N}\sum_{j=1}^{N}\epsilon_{j,k} - \epsilon_{i,k}\right),$$

and

$$\bar{x}_{i,k}\left(\epsilon_{i,k}\right) = \mathbb{E}_{\epsilon_{-i,k}}\left[x_{i,k}\left(\epsilon_{1,k}, \epsilon_{2,k}, ...\epsilon_{N,k}\right)\right] = A_{i,k} - \frac{1}{c}\frac{N-1}{N}\epsilon_{i,k},$$

where $A_{i,k} = \frac{\bar{Z}_k - Z_{i,k}}{2c} + \frac{1}{N}$.

The broker's payoff for any stock k under the optimal allocation is

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{N}\left[x_{i,k}\left(\bar{s}-\frac{b}{2}-Z_{i,k}-2\epsilon_{i,k}\right)-cx_{i,k}^2\right]\right]-2\bar{s}\\
=&-\mathbb{E}\left[\sum_{i=1}^{N}\left(x_{i,k}\left(Z_{i,k}+2\epsilon_{i,k}\right)+cx_{i,k}^2\right)\right]-\frac{b}{2}-\bar{s}\\
=&-\mathbb{E}\left[\sum_{i=1}^{N}\left(A_{i,k}-\frac{1}{c}\frac{N-1}{N}\epsilon_{i,k}\right)\left(Z_{i,k}+2\epsilon_{i,k}\right)\right]-c\mathbb{E}\left[\sum_{i=1}^{N}\left(A_{i,k}+\frac{1}{c}\left(\frac{1}{N}\sum_{j=1}^{N}\epsilon_{j,k}-\epsilon_{i,k}\right)\right)^2\right]-\frac{b}{2}-\bar{s}\\
=&-\left[\sum_{i=1}^{N}A_{i,k}Z_{i,k}\right]+2\frac{1}{c}\frac{N-1}{N}\mathbb{E}\left[\sum_{i=1}^{N}\epsilon_{i,k}\epsilon_{i,k}\right]-c\sum_{i=1}^{N}\mathbb{E}\left[A_{i,k}^2+\frac{1}{c^2}\frac{\sum_{j\neq i}\epsilon_{j,k}^2+\left(N-1\right)^2\epsilon_{i,k}^2}{N^2}\right]-\frac{b}{2}-\bar{s}\\
=&-c\sum_{i=1}^{N}\left[\left(\frac{\bar{Z}_k-Z_{i,k}}{2c}+\frac{1}{N}\right)\frac{Z_{i,k}}{c}+\left(\frac{\bar{Z}_k-Z_{i,k}}{2c}+\frac{1}{N}\right)^2\right]+2\frac{N-1}{c}\mathbb{E}\left[\epsilon_{i,k}^2\right]-\frac{N-1}{c}\mathbb{E}\left[\epsilon_{i,k}^2\right]-\frac{b}{2}-\bar{s}\\
=&-c\sum_{i=1}^{N}\left[\left(\frac{\bar{Z}_k-Z_{i,k}}{2c}+\frac{1}{N}\right)\left(\frac{\bar{Z}_k+Z_{i,k}}{2c}+\frac{1}{N}\right)\right]+\frac{N-1}{c}\mathbb{E}\left[\epsilon_{i,k}^2\right]-\frac{b}{2}-\bar{s}\\
=&-c\left[N\left(\frac{\bar{Z}_k}{2c}+\frac{1}{N}\right)^2-\sum_{i=1}^{N}\left(\frac{Z_{i,k}}{2c}\right)^2\right]+\frac{N-1}{c}\mathbb{E}\left[\epsilon_{i,k}^2\right]-\frac{b}{2}-\bar{s}.
\end{aligned}
$$

Then the total payoff for these K stocks will be

$$
\begin{aligned}
&K\left\{-c\frac{1}{K}\sum_{j=1}^{K}\left[N\left(\frac{\bar{Z}_k}{2c}+\frac{1}{N}\right)^2-\sum_{i=1}^{N}\left(\frac{Z_{i,k}}{2c}\right)^2\right]+\frac{N-1}{c}\mathbb{E}\left[\epsilon_{i,k}^2\right]-\left(\frac{b}{2}+\bar{s}\right)\right\}\\
=&K\left\{\frac{N-1}{c}\frac{b^2}{12}-c\frac{1}{K}\sum_{j=1}^{K}\left[N\left(\frac{\bar{Z}_k}{2c}+\frac{1}{N}\right)^2-\sum_{i=1}^{N}\left(\frac{Z_{i,k}}{2c}\right)^2\right]-\left(\frac{b}{2}+\bar{s}\right)\right\}.
\end{aligned}
$$

Similarly, when bundling all K stocks together, the broker's payoff will be

$$
\begin{aligned}
&K\left\{-c\left[N\left(\frac{\bar{Z}}{2c}+\frac{1}{N}\right)^2-\sum_{i=1}^{N}\left(\frac{Z_i}{2c}\right)^2\right]+\frac{N-1}{c}\mathbb{E}\left[\hat{\epsilon}_i^2\right]-\left(\frac{b}{2K}+\bar{s}\right)\right\}\\
=&K\left\{\frac{N-1}{c}\frac{b^2}{12K^2}-c\left[N\left(\frac{\bar{Z}}{2c}+\frac{1}{N}\right)^2-\sum_{i=1}^{N}\left(\frac{Z_i}{2c}\right)^2\right]-\frac{b}{2K}-\bar{s}\right\}.
\end{aligned}
$$

Then unbundling dominates bundling if and only if

$$
\begin{aligned}
& -c\tfrac{1}{K}\textstyle\sum_{j=1}^{K}\left[N\left(\frac{\bar{Z}_k}{2c}+\frac{1}{N}\right)^2-\sum_{i=1}^{N}\left(\frac{Z_{i,k}}{2c}\right)^2\right]-\frac{b}{2} \;\overset{\frac{N-1}{c}\frac{b^2}{12}}{\geq}\; -c\left[N\left(\frac{\bar{Z}}{2c}+\frac{1}{N}\right)^2-\sum_{i=1}^{N}\left(\frac{Z_i}{2c}\right)^2\right]-\frac{b}{2K} \overset{\frac{N-1}{c}\frac{b^2}{12K^2}}{} \\
\iff & \frac{N-1}{c}\frac{b^2}{12}\left(1-\frac{1}{K^2}\right)-\frac{b}{2}\left(1-\frac{1}{K}\right)\geq \begin{array}{c} c\frac{1}{K}\sum_{j=1}^{K}\left[N\left(\frac{\bar{Z}_k}{2c}+\frac{1}{N}\right)^2-\sum_{i=1}^{N}\left(\frac{Z_{i,k}}{2c}\right)^2\right] \\ -c\left[N\left(\frac{\bar{Z}}{2c}+\frac{1}{N}\right)^2-\sum_{i=1}^{N}\left(\frac{Z_i}{2c}\right)^2\right] \end{array} \\
\iff & \frac{b}{2}\left(1-\frac{1}{K}\right)\left[\frac{N-1}{c}\frac{b}{6}\left(1+\frac{1}{K}\right)-1\right]\geq \begin{array}{c} c\frac{1}{K}\sum_{j=1}^{K}\left[\left(N\left(\frac{\bar{Z}_k}{2c}\right)^2+\frac{\bar{Z}_k}{c}+\frac{1}{N}\right)-\sum_{i=1}^{N}\left(\frac{Z_{i,k}}{2c}\right)^2\right] \\ -c\left[\left(N\left(\frac{\bar{Z}}{2c}\right)^2+\frac{\bar{Z}}{c}+\frac{1}{N}\right)-\sum_{i=1}^{N}\left(\frac{Z_i}{2c}\right)^2\right] \end{array} \\
\iff & \frac{b}{2}\left(1-\frac{1}{K}\right)\left[\frac{N-1}{c}\frac{b}{6}\left(1+\frac{1}{K}\right)-1\right]\geq\frac{1}{4c}\left\{\frac{1}{K}\sum_{j=1}^{K}\left[N\bar{Z}_k^2-\sum_{i=1}^{N}Z_{i,k}^2\right]-\left[N\bar{Z}^2-\sum_{i=1}^{N}Z_i^2\right]\right\} \\
\iff & \frac{b}{2}\left(1-\frac{1}{K}\right)\left[(N-1)\frac{b}{6}\left(1+\frac{1}{K}\right)-c\right]\geq\frac{N}{4}\left\{\left[\frac{1}{K}\sum_{j=1}^{K}\bar{Z}_k^2-\bar{Z}^2\right]-\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{K}\sum_{j=1}^{K}Z_{i,k}^2-Z_i^2\right)\right\}.
\end{aligned}
$$