What Does Best Execution Look Like?*

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

April 23, 2026

[Preliminary and Incomplete – Please Do Not Circulate]

Abstract

U.S. retail brokers provide "best execution" for client orders, though how this actually works 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 main 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 managing competition.

*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 that 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 proposed several new rules for equity trading, some recently adopted, motivated in no small part by concerns about the handling of retail orders.[2]

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 actually implement best execution and competition among wholesalers are comparatively unexplored. Our paper contributes to the literature by providing basic empirical insights into how best execution is implemented in practice, as well as basic theoretical results into how a broker with should design the order flow routing mechanism if its objective is to obtain the best expected price on a client's order

In practice, retail brokers work with several wholesalers. These wholesalers offer better prices for retail order flow than open market centers (such as exchanges) do,[4] due to lower adverse selection or correlation in the order flow.[5] Retail brokers do not communicate

[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 did not adopt, an order competition rule, which would mandate individual auctions for marketable retail orders, and a best-execution rule (currently, only FINRA has an explicit 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.

[5]Easley et al. (1996) provide empirical evidence suggesting that retail order flow is less informed, Battalio

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 develop a theoretical framework that studies best execution as 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 trades, 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 problem and show that it can be implemented via a performance-based order allocation rule that allocates the second period's order flow contingent on the price improvement wholesalers offered in the first period.

We then use this model to derive a number of implications, which we then examine in the data. In addition to the basic property that a wholesaler receives more order flow if it offers a lower historical execution price, we show that this sensitivity of routing to performance is greater when there is greater information asymmetry between wholesalers and brokers regarding their execution ability. In the model, wholesalers are rational and responsive to exactly how they are evaluated by the broker, and what matters for order allocation is the total price improvement across all trades on which the wholesaler is evaluated. As an intuitive prediction, the model suggests that the entry of a new wholesaler would lead to better prices for customers, particularly if the new entrant captures a greater share of the market (i.e., is more efficient). Finally, the model rationalizes why different brokers may optimally make different order bundling decisions, reflecting differences in the nature of the retail flow different brokers get.

Having derived this theoretical benchmark, we next examine how brokers implement a duty of best execution 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

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)).

insight into the specific practices of these brokers, including best-execution statistics on each of the wholesalers to whom they route. Based on this data, we establish the following empirical patterns.

Consistent with the most basic prediction of the model, brokers allocate more order flow to wholesalers with better past performance—and do so robustly across all three brokers. The key performance metric is the effective-over-quoted (EFQ) spread, which captures the fraction of the publicly available spread that retail investors pay; a lower EFQ means greater savings for investors. For every 1 percentage point increase in EFQ, wholesalers receive between 1.0 and 1.2 percentage points less order-flow allocation. If we consider the ordinal ranking of wholesalers, a worse ordinal ranking leads to 6 to 9 pp lower order flow allocation. The relationship is specific to the exact history each broker uses: evaluating wholesaler performance over alternative windows or order-size categories substantially reduces explanatory power, confirming that brokers are attuned to their own evaluation criteria rather than any generic measure of past performance.

Consistent with the prediction that the sensitivity of routing to performance is greater if wholesalers possess more private information about their cost of executing future orders, we find that routing is more sensitive to past performance in market segments where wholesaler performance varies more widely. For Broker A, which routes each stock symbol and order-size bin separately, a one-rank change for a higher-variance symbol (75th percentile of variance) is associated with 0.3 additional percentage points of lost order share compared to a lower-variance symbol (25th percentile). We further isolate an incentive channel distinct from pure Bayesian updating about which wholesalers are more efficient: even when the rank-1 wholesaler holds a large performance lead—making a change in ordinal ranking statistically very unlikely—routing remains highly sensitive to changes in the leader's margin of outperformance. This result supports the view that routing sensitivity serves as an incentive device.

To test the basic premise of the model that wholesalers are aware of how they get evaluated by brokers and respond to changes in incentives accordingly, we consider a natural experiment: on January 1, 2020, Broker B expanded its primary evaluation focus to include odd-lot orders (orders of fewer than 100 shares), which previously received minimal attention. Wholesalers respond rapidly and precisely as the theory predicts: EFQs on odd-lot orders fall from averages above 80% to below 50% within six months, and below 30% within a year. Simultaneously, and perhaps surprisingly, EFQ for large orders (2,000 or more shares) deteriorates substantially. These findings are, however, consistent with how the broker and wholesalers operate in the model: what matters for order allocation is the total price improvement across all order types that a wholesaler is evaluated on, so a change in the focus leads to a reallocation of a fixed pool of price improvement across order types. This finding also has implications for proposed changes to market structure: an attempt to improve execution along one dimension by altering routing criteria may trigger an unintended opposite response along other dimensions.

Additional evidence of strategic wholesaler behavior emerges from the timing of performance changes within the evaluation window. Two brokers in our sample evaluate and adjust routing allocations monthly, 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 variation in wholesaler performance increases toward the end of the window, consistent with wholesalers' strategic responses to offer price improvements closer to the end, when evaluation and re-adjustment are more time-constrained.

Our last set of analysis focuses on the competitive landscape for order routing. A large market maker, henceforth wholesaler A5, began working with Broker A in 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.

Taken together, our paper suggests that it is natural to interpret best execution as a mechanism design problem in which privately informed wholesalers compete for future order flow by offering price improvement on current orders, and the broker designs the allocation rule to manage this competition. The theory emphasizes that the price improvement wholesalers offer is an equilibrium outcome of the broker's routing policy, which can differ across brokers, depending on the nature of their order flow. Our empirical results provide a comprehensive picture of retail order routing, spanning broker responsiveness, wholesaler strategic behavior, and the competitive landscape, and offer direct evidence of the incentive channel emphasized in the model.

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 we know exactly which history (period, order sizes, stocks) each broker uses to evaluate wholesalers' performance. In contrast, Huang et al. (2023) is based only on small orders, and they do not observe how each broker evaluates the performance of wholesalers. At the same time, the advantage of Huang et al. (2023) over our data is that it includes data from six brokers, not three, and allows comparison across the brokers. 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 depending 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, despite the fact that the brokers differ in how they bundle the orders. 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 the historical window a broker uses 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 change in the order allocation process would likely cause a strategic response from wholesalers in what price improvement they are offering on different orders.

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. Retail liquidity programs on exchanges, which offer better prices only to retail investors.[6] NBBO liquidity is not all

[6]Jain, Linna, and McInish (2021) describes the NYSE RLP. Ernst, Sun, and Spatt (2024b) highlight general failures of these programs, while Chung, Rösch, and Zhang (2023) examine changes in RLP usage

equal–some exchanges charge take fees, while others pay rebates.[7] 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[8] for marketable orders.[9]

Several articles analyze the performance of wholesalers and market segmentation. Central to this segmentation is the idea that retail trades have lower adverse selection.[10] Cream-skimming can offer better prices for retail investors, but segmenting them can lead to worse prices on-exchange.[11] Reductions in on-exchange quotes can change the NBBO benchmark by which retail trades are measured.[12]. Off-exchange wholesalers tend to be concentrated, which has raised concerns both in equity and option markets.[13] Our contribution to this literature is a study of competition in the broker-wholesaler relationship, with brokers monitoring wholesaler performance and allocating order flow accordingly.

---

with market adverse selection.

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

[8]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.

[9]For limit orders, fill rates and implementation shortfall are the relevant metrics. Anand, Samadi, Sokobin, and Venkataraman (2025) present a novel analysis of retail limit orders, and compare their performance to retail market orders. Barardehi, Bogousslavsky, and Muravyev (2025) use trading journal data to analyze completed limit orders.

[10]Easley et al. (1996) model cream-skimming of retail trades. Battalio and Holden (2001) consider the extent to which segmentation benefits are passed on to retail investors. Baldauf et al. (2024) model retail traders as having lower correlation, rather than lower information, than professional traders. Battalio and Jennings (2022) empirically analyze proprietary wholesaler data which highlights the prices wholesalers give to retail investors as superior to those of public exchanges exchange, consistent with cream-skimming.

[11]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.

[12]van Kervel and Yueshen (2023) argue that when segmentation can widen on-exchange spreads, which reduces competitive pressure for off-exchange market makers who use exchange spreads as a benchmark.

[13]Hu and Murphy (2022) suggest concentration of off-exchange equity wholesaling leads to higher spreads. Bryzgalova, Pavlova, and Sikorskaya (2022) highlight concentration of option wholesalers, while Hendershott, Khan, and Riordan (2022) and Ernst and Spatt (2025) highlight specific option internalization rules which privilege large option DMMs.

III. Theoretical Framework

In this section, we develop a model in which a broker acting in the best interests of its retail clients designs an optimal mechanism to promote competition among privately informed wholesalers. We show that the optimal mechanism can be implemented by a performance-based order allocation rule that assigns future order flow as a function of past price improvement. We then derive comparative statics that generate testable predictions about broker responsiveness, wholesaler responsiveness, and competitive landscape. Finally, we present two extensions that rationalize cross-broker heterogeneity in bundling decisions and choices of window lengths.

A. Model Setup

We begin with the model primitives. There are two periods $t = 0, 1$. All players are risk neutral, and there is no discounting.

Players and payoffs. A broker (she) interacts with $N \geq 2$ wholesalers (he) in both periods. At the beginning of each period $t \in \{0, 1\}$, the broker receives a unit mass of divisible buy orders that she allocates between the wholesalers and the market. If wholesaler i executes an order mass $x_{i,t} \in [0, 1]$ at price (half bid-ask spread) $s_{i,t}$, 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 at time t. We assume that there is a publicly observable price $\bar{s}$, and any wholesaler i is restricted to choose $s_{i,t} \leq \bar{s}$ for any t. Here 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. The difference $\bar{s} - s_{i,t}$ is the price improvement offered by wholesaler i at time t. 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 her customers' (retail investors') payoff, which is equivalent to minimizing the spread paid by the customers in two periods. The broker's total payoff can thus be written as:

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

Period 0. At the beginning of period 0, the period 0 order allocation $\vec{x}_0 = \{x_{i,0}\}_{i=1}^N$ is exogenously given and publicly observable. For simplicity, we assume $x_{i,0} = 1/N$ for all wholesalers i.[14] The period-0 cost parameters $\vec{\zeta}_0 = \{\zeta_{i,0}\}_{i=1}^N$ also become public at the beginning of period 0, while period-1 cost parameter $\zeta_{i,1}$ is private information of wholesaler i.[15] It is common knowledge that

$$\zeta_{i,1} = \zeta_{i,0} + \epsilon_i, \tag{2}$$

where $\epsilon_i \sim \text{Unif}\left[-\frac{b}{2}, \frac{b}{2}\right]$ are mutually independent. We refer to ϵ_i as the (cost) type of wholesaler i. Here $b > 0$ measures the severity of information asymmetry between the broker and wholesalers. Outsiders know only the distribution of $\{\epsilon_i\}_{i=1}^N$. Without loss of generality, we assume $\zeta_{1,0} \geq \zeta_{2,0} \geq \cdots \geq \zeta_{N,0}$ and define $\bar{\zeta}_0 = \frac{1}{N} \sum_{i=1}^N \zeta_{i,0}$.

In period 0, wholesalers simultaneously execute period 0 orders received from the broker. Specifically, each wholesaler i chooses a price $s_{i,0}$ at which it executes his period 0 orders, without observing the prices chosen by other wholesalers. The entire price vector $\vec{s}_0 = (s_{1,0}, s_{2,0}, \ldots, s_{N,0})$ is then observed by the broker. Hence, the broker information set at the end of period 0 is

$$\mathcal{I}_0 = (\vec{x}_0, \vec{\zeta}_0, \vec{s}_0).$$

Period 1. The broker allocates the period 1 order flow across wholesalers based on $\mathcal{I}_0$, and wholesalers then choose their period 1 execution prices $\{s_{i,1}\}_{i=1}^N$. We assume that the

[14] Here we take the period 0 allocation as exogenous. In the baseline model, the specific value of the period 0 allocation is not important. What matters for the broker is the total price improvement across all orders that each wholesaler delivers at period 0, which in turn determines its share of period 1 order flow. A wholesaler who receives only a small fraction of orders at time 0 can still generate a large total price improvement by offering very low, possibly negative, prices to compete for future order flow. Hence the particular time 0 order share is not a key primitive in the baseline model.

In the Internet Appendix we present a dynamic model in which the allocation at the beginning of each period is determined endogenously. In that model, the initial order allocation in a period can matter, because it is correlated with wholesaler types from previous periods and, when types are persistent, also with price improvement and future order allocation.

[15] It is plausible that a wholesaler may not know its period 1 inventory cost with certainty. We can equivalently interpret $\zeta_{i,1}$ as wholesaler i best estimate in period 0 of its period 1 inventory cost. Since wholesalers are risk neutral and must execute any order flow they receive, the analysis is unchanged.

broker has full commitment power: before the game begins, the broker can commit to any period 1 allocation policy

$$\vec{x}_1(\mathcal{I}_0) = (x_{1,1}, x_{2,1}, \ldots, x_{N,1}),$$

which maps the time 0 information into the period 1 order shares. The period 1 order allocation must satisfy $x_{i,1} \geq 0$ and $\sum_{i=1}^{N} x_{i,1} \leq 1$ for all i. Since $\vec{x}_0$ and $\vec{\zeta}_0$ are exogenous, we sometimes explicitly write the allocation policy as $\vec{x}_1(\vec{s}_0)$ to emphasize its dependence on wholesalers' period 0 performance.[16]

B. *The Broker's Problem*

The broker's problem is to choose the period 1 order allocation policy $\vec{x}_1$ to maximize her ex ante payoff in Equation 1. The broker faces a fundamental tradeoff. Routing more orders to the lowest-cost wholesaler improves *allocative efficiency*. But privately informed wholesalers earn *information rents*: a low-cost wholesaler can mimic a higher-cost type and pocket the difference. The difficulty is that the broker does not observe wholesalers' period 1 costs directly and must design incentives for them to reveal this information. By the revelation principle (Myerson (1981)), we can without loss of generality restrict attention to direct mechanisms. A direct mechanism works in the following way: each wholesaler i simply reports his type $\hat{\epsilon}_i \in [-b/2, b/2]$ to the broker, and the broker specifies both the period 1 order allocation and monetary transfers from each wholesaler to the broker as functions of the reported types.[17] Formally, the broker commits to

$$\vec{X}_1(\cdot; \vec{\zeta}_0, \vec{x}_0) : \times_{i=1}^{N} \left[-\frac{b}{2}, \frac{b}{2}\right] \to \left\{ (X_{1,1}, X_{2,1}, ...X_{N,1}) \,|X_{i,1} \geq 0 \quad \forall i, \sum_{i=1}^{N} X_{i,1} \leq 1 \right\},$$

and

$$\vec{T}(\cdot; \vec{\zeta}_0, \vec{x}_0) : \times_{i=1}^{N} \left[-\frac{b}{2}, \frac{b}{2}\right] \to \left\{ (T_1, T_2, ...T_N) \in \mathbb{R}^N \right\},$$

[16] In fact, since $\vec{x}_0$ does not affect wholesalers' incentives, later we can show that the optimal period 1 allocation policy does not depend on $\vec{x}_0$.

[17] Later we can show that, in our setting, these transfers can be naturally implemented through period 0 price improvement.

where $X_{i,1}$ is wholesaler i's period 1 order share and T_i is the transfer from wholesaler i to the broker. The next proposition shows that the optimal mechanism maximizes a simple objective that nets out both inventory costs and information rents.

Proposition 1. *The period-1 allocation policy in the optimal direct mechanism, $\vec{X}_1^*(\vec{\epsilon})$, solves the following problem:*

$$\max_{\vec{X}_1} \sum_{i=1}^{N} \left[X_{i,1} \left(\bar{s} - (\zeta_{i,0} + \epsilon_i + cX_{i,1}) - \left(\epsilon_i + \frac{b}{2} \right) \right) \right] - 2\bar{s} \tag{3}$$

subject to constraints:

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

In period 1, because the game ends, all wholesalers execute at $\bar{s}$ and offer no price improvement. The broker's payoff therefore depends entirely on how much price improvement she extracts in period 0 through the incentives created by the period 1 allocation rule.

The bracketed term in Equation 3, $\bar{s} - (\zeta_{i,0} + \epsilon_i + c\, X_{i,1}) - (\epsilon_i + b/2)$, is the per-unit net surplus from allocating period 1 flow to wholesaler i of type ϵ_i. It has two cost components. First, $\zeta_{i,0} + \epsilon_i + c\, X_{i,1}$ is wholesaler i's average execution cost for period 1 orders. Second, $\epsilon_i + b/2$ is the information rent that the broker must concede to induce truthful reporting. The optimal mechanism chooses the allocation to maximize total net surplus across all wholesalers.

We introduce the following assumption on the cost c.

Assumption 1. $c \in \left(\frac{N}{2}\left(\zeta_{1,0} - \bar{\zeta}_0\right) + (N-1)b,\ \frac{N}{2}\left(\bar{s} - \bar{\zeta}_0 - \frac{3}{2}b\right)\right)$.

Assumption 1 ensures that under the optimal direct mechanism, the order allocation always satisfies $X_{i,1} > 0$ for all i and $\sum_{i=1}^{N} X_{i,1} = 1$. This appears to be the 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 direct mechanism under this assumption.

Proposition 2. *Under Assumption 1, the optimal direct mechanism is characterized by the period 1 allocation policy $\vec{X}_1^*(\vec{\epsilon})$ and the transfer schedule from wholesalers to the broker, $\vec{T}^*(\vec{\epsilon})$, given by*

$$X^*_{i,1}(\vec{\epsilon}) = A_i + \frac{1}{c}\left[\frac{1}{N}\sum_{j=1}^{N}\epsilon_j - \epsilon_i\right], \quad \forall i, \tag{4}$$

and

$$\begin{aligned} T^*_i(\epsilon_i) = &\left[A_i\left(\bar{s} - \zeta_{i,0}\right) - cA_i^2 - \frac{(N-1)\,b^2}{12N^2c} - A_i\frac{b}{2} + \frac{(N-1)\,b^2}{8Nc}\right] \\ &+ \left[2A_i - \frac{1}{c}\left(\bar{s} - \zeta_{i,0}\right)\right]\frac{N-1}{N}\epsilon_i - \left(\frac{N^2 - 3N + 2}{2N^2c}\right)\epsilon_i^2, \qquad \forall i. \end{aligned} \tag{5}$$

where $A_i = \frac{\bar{\zeta}_0 - \zeta_{i,0}}{2c} + \frac{1}{N}$.

The allocation policy in Equation 4 has two components. The first component, A_i, depends only on public information and captures ex ante heterogeneity in execution efficiency across wholesalers. The second component,

$$\frac{1}{c}\left[\frac{1}{N}\sum_{j=1}^{N}\epsilon_j - \epsilon_i\right],$$

reflects comparative advantage arising from the private cost types. A wholesaler receives a larger share when its type ϵ_i is lower.

The transfer from a wholesaler to the broker is also decreasing in its type ϵ_i, but in a nonlinear way. Hence, under the optimal direct mechanism, a wholesaler with a lower cost type makes a larger transfer to the broker and receives a larger share of period 1 order flow.

C. Implementation

A direct mechanism requires each wholesaler to report his type. In practice, brokers do not elicit type reports; instead, they observe execution prices and use them to set future allocations. We now show that the optimal direct mechanism in Proposition 2 can be implemented by a *performance-based* allocation policy—a rule $\vec{x}^*_1(\vec{s}_0) = (x_{1,1}(\vec{s}_0), x_{2,1}(\vec{s}_0), ...x_{N,1}(\vec{s}_0))$ that maps observed period 0 execution prices directly into period 1 order shares, without any explicit type report.[18]

[18] We use uppercase $\vec{X}^*_1(\vec{\epsilon})$ for the direct-mechanism allocation (indexed by types) and lowercase $\vec{x}^*_1(\vec{s}_0)$ for the performance-based allocation (indexed by prices).

Let $(T_i^*)^{-1}(\cdot)$ denote the inverse function of wholesaler i's transfer schedule in the optimal direct mechanism.

Proposition 3. *There exists a performance based period 1 allocation policy $\vec{x}_1^*(\vec{s}_0)$ that implements the optimal direct mechanism characterized in Proposition 2. In particular,*

$$\vec{x}_1^*(\vec{s}_0) = \vec{X}_1^*\Big((T_1^*)^{-1}\big((\bar{s} - s_{1,0})x_{1,0}\big), \ldots, (T_N^*)^{-1}\big((\bar{s} - s_{N,0})x_{N,0}\big)\Big).$$

The logic has three steps. First, $(\bar{s} - s_{i,0})\, x_{i,0}$ is the total price improvement that wholesaler i delivers in period 0; this is the observable counterpart of the transfer in the direct mechanism. Second, the inverse transfer schedule $(T_i^*)^{-1}$ maps this observed price improvement back into an implied type $\hat{\epsilon}_i$. Third, incentive compatibility ensures that each wholesaler optimally reveals his true type, so $\hat{\epsilon}_i = \epsilon_i$, and the direct-mechanism allocation $\vec{X}_1^*(\cdot)$ yields the correct period 1 order shares. In our Internet Appendix, we provide a numerical example about how optimal order allocation responds to performance.

D. Comparative Statics and Empirical Predictions

In this section, we examine the properties of the optimal order allocation policy $\vec{x}_1^*\,(\vec{s}_0)$.

Broker Responsiveness: Rewarding and Performance Slopes

First, we show under the optimal policy, a wholesaler that delivers more price improvement in period 0 receives a larger order share in period 1.

Lemma 1. *The optimal order allocation policy in Proposition 3 satisfies $\frac{dx_{i,1}^*}{ds_{i,0}} < 0$ for all i.*

This result is intuitive: more price improvement signals a lower cost type, so the broker rewards it with more future flow. Although intuitive, this result serves as a natural starting point for the subsequent analysis. It yields our first empirical prediction.

Prediction 1. *All else equal, a wholesaler receives more order flow if it offers a lower historical execution price.*

The performance slope, $\frac{dx_{i,1}^*}{ds_{i,0}} < 0$, is heterogeneous. Brokers in our sample differ in how they bundle orders when choosing their routing strategies. In particular, Broker A routes orders at the individual stock level, Broker B groups stocks into four security bins and routes at the bin level, and Broker C bundles all orders into a single pool. Because order characteristics differ across bins, the degree of information asymmetry between the broker and wholesalers can also differ. This can lead brokers to use different allocation rules across bins, including different responses of future order flow to past performance.

The key parameter governing the degree of information asymmetry is b, which captures the dispersion of wholesalers' private information. The next result shows that lower information asymmetry implies a less sensitive reward scheme.

Lemma 2. *The expected sensitivity of period 1 order flow to the price at which wholesaler i executes in period 0,*

$$\mathbb{E}\left[\left|dx^*_{i,1}/ds_{i,0}\right|\right],$$

is increasing in b.

The intuition is as follows. Making period 1 order flow sensitive to period 0 prices serves two purposes: efficiency and incentives. First, it improves allocative efficiency, because wholesalers with lower execution costs should receive more order flow. Since lower period 0 prices signal lower execution costs, they make it efficient to allocate more future order flow to those wholesalers. Second, it provides incentives by rewarding wholesalers for offering better prices in period 0. When private information is more severe, that is, when b is higher, the incentive motive becomes stronger, which makes the optimal reward scheme more sensitive to performance.

At the same time, we can show the following result.

Lemma 3. *The variation in wholesaler performance,* $\mathrm{Var}[T_i]$*, is higher when information asymmetry is more severe, that is, when b is higher.*

Lemma 2 and Lemma 3 jointly imply the following empirical prediction.

Prediction 2. *All else being equal, $\left|\frac{dx_i^*}{ds_i}\right|$ is higher for bins with greater variation in wholesaler performance.*

Next, we examine how the optimal allocation varies across three additional dimensions of our empirical setting: changes in broker focus, wholesaler entry, and order size.

Wholesaler Responsiveness: Effects of Focus Change

In our data, Broker B makes an adjustment to its order allocation policy. 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 study this question, we develop a simple extension of the baseline model that captures a focus change in the broker's allocation criterion. The Internet Appendix presents the full specification and derivations. Here we only provide the economic intuition and the final result.

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 the broker seeks to maximize total price improvement across all orders and that the routing decision is identical for both primary-focus and non-primary-focus stocks.[19] 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.

The key observation is that, because the period 1 allocation depends only on *total* price improvement, the optimal total price improvement is the same in both cases. What changes is its distribution across order types. When the primary focus expands (Case 1 to Case 2), wholesalers must offer more price improvement on newly included small orders to meet the threshold. Since the total is fixed, price improvement on large orders must decline. Based on this observation, we derive the following empirical prediction:

[19] This is consistent with practice, where Broker B applies the same routing decision regardless of order size.

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

Competitive Landscape: New Wholesaler Entry

In our data, a new wholesaler enters during the sample period, which may alter the competitive environment for retail order routing. This entry event provides a natural setting to study how an increase in competition affects market outcomes.

To capture this setting in the model, we compare the baseline model with N wholesalers to a model with an additional wholesaler, indexed $N+1$, with public period 0 cost $\zeta_{N+1,0}$. We do not impose restrictions on the value of $\zeta_{N+1,0}$, and it can be higher or lower than the period 0 costs of the incumbent wholesalers. Instead, we replace Assumption 1 with the following one in this extension.

Assumption 2. $c \in \left(\frac{N+1}{2}\left(\max\{\zeta_{1,0}, \zeta_{N+1,0}\} - \frac{1}{N+1}\sum_{i=1}^{N+1}\zeta_{i,0}\right) + Nb,\ \frac{N+1}{2}\left(\bar{s} - \frac{1}{N+1}\sum_{i=1}^{N+1}\zeta_{i,0} - \frac{3}{2}b\right)\right).$

This is the natural analogue of Assumption 1 for $N+1$ wholesalers and again ensures that all wholesalers receive positive allocations in equilibrium. Under Assumption 2, the model with wholesaler $N+1$ can be solved in the same way as the baseline model. For brevity, we report only the results that characterize the impact of entry.

Lemma 4. *In the model with the new wholesaler $N+1$,*

1. *the average execution price across all wholesalers is lower than in the model without wholesaler $N+1$, and it is increasing in $\zeta_{N+1,0}$;*

2. *wholesaler $N+1$'s expected market share in period 1 is decreasing in $\zeta_{N+1,0}$.*

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.*

Competitive Landscape: Order Size and Competition

Order characteristics shape the competitive landscape. In our data, Broker A conditions routing on order size within each symbol, providing a setting to study how size affects competition.

What distinguishes orders of different sizes is their adverse selection (Easley and O'hara, 1987; Barclay and Warner, 1993), with larger orders exhibiting more severe adverse selection. Accordingly, a wholesaler's execution cost reflects not only inventory considerations, but also its ability to manage adverse selection associated with future price movements. We capture this force by allowing the dispersion of wholesalers' private cost information, denoted by b, to be higher for larger orders. For tractability, in this extension we focus on the case in which wholesalers are ex ante identical, that is, $\zeta_{i,0} \equiv \bar{\zeta}_0$. The following result provides a complete characterization of how the equilibrium competitive landscape varies with the parameter b.

Lemma 5. *Assume $\zeta_{i,0} = \zeta_0$ for all $i \in \{1, \ldots, N\}$. Let $s_{(1),0} \leq s_{(2),0} \leq \cdots \leq s_{(N),0}$ denote the order statistics of the time 0 prices offered by wholesalers, and let*

$$\bar{s}_0 \coloneqq \frac{1}{N} \sum_{i=1}^{N} s_{i,0}$$

denote the average time 0 price. Define the Herfindahl Hirschman Index of wholesalers time 1 order shares by

$$\mathrm{HHI} \coloneqq \sum_{i=1}^{N} \big(x_{i,1}\big)^2.$$

Define the top wholesaler as the wholesaler offering the lowest time 0 price, and let its time 1 order share be

$$x_{(1),1} \coloneqq x_{j^*,1}, \quad \text{where } j^* \in \operatorname*{arg\,min}_{i \in \{1,\ldots,N\}} s_{i,0}.$$

Then $\mathbb{E}\big[\lvert s_{(1),0} - s_{(2),0} \rvert\big]$, $\mathbb{E}\big[\lvert s_{(1),0} - \bar{s}_0 \rvert\big]$, $\mathbb{E}[\mathrm{HHI}]$, $\mathbb{E}[\bar{s}_0]$, and $\mathbb{E}\big[x_{(1),1}\big]$ are all increasing in b.

Based on our earlier discussion, a higher b corresponds to larger orders with more severe adverse selection. This mapping yields the following empirical prediction.

Prediction 5. *All else being equal, the gap between the first-ranked and second-ranked wholesaler's EFQ, the gap between the first-ranked wholesaler's EFQ and the average EFQ across wholesalers, the HHI, the top wholesaler's order share, and the average EFQ across wholesalers are all higher for larger orders.*

E. Extensions: Bundling Choice and Window Length

So far, we have provided testable empirical predictions about the optimal order allocation rule that brokers should follow. While these properties apply broadly, the three brokers in our data exhibit meaningful heterogeneity in how they implement their allocation policies. Below, we discuss two key dimensions of this heterogeneity and present extensions of our baseline model that can rationalize them.

For these two extensions, due to data restrictions, we cannot test them in our setting. We therefore present them as theoretical results and leave the empirical test to future work.

Why Choose Bundling versus Unbundling?

Brokers differ in their bundling strategies, which raises a natural question: what determines whether a broker bundles or unbundles? To study this question, we extend the baseline model to multiple stocks and allow for richer private information about wholesalers' period 1 execution costs. This extension delivers theoretical predictions for the optimal choice of bundling.

In this extension, we assume there are 2^K stocks indexed by $j \in \{0, 1, 2, \ldots, 2^K - 1\}$, and in each period $t \in \{0, 1\}$, the broker receives one unit order in each of 2^K stocks. The broker allocates these orders across wholesalers. We compare two routing policies: (i) the broker chooses allocations stock by stock; and (ii) the broker bundles all stocks into one pool and chooses allocations for the pooled order flow. We introduce a new inventory cost structure: if wholesaler i executes $x_{i,j,t}$ units of stock j in period t, its total cost is

$$\sum_{j=0}^{2^K-1} \left(\zeta_{i,j,t} x_{i,j,t} + c x_{i,j,t}^2\right), \tag{6}$$

where $\zeta_{i,j,1} = \zeta_{i,j,0} + \epsilon_{i,j}$. As in the baseline model, $\zeta_{i,j,0}$ becomes observable by the end of period 0, so the remaining uncertainty about period 1 costs is captured by $\epsilon_{i,j}$.

The Internet Appendix describes the extension in full detail. A key new ingredient is that a given wholesaler's execution cost $\epsilon_{i,j}$ can vary across stocks j and exhibit cross-stock correlation. We summarize this dependence by the average pairwise correlation of wholesaler

i's cost component,

$$\bar{\rho}^{\epsilon} := \frac{2}{2^K\left(2^K-1\right)} \sum_{0 \leq j < k \leq 2^K-1} \mathrm{Corr}\big(\epsilon_{i,j}, \epsilon_{i,k}\big).$$

To keep the main text concise, we present only the equilibrium implications and direct readers to the Internet Appendix for full model details.

Lemma 6. *We have following results about the optimal choice of bundling:*

1. *there exists a threshold c^* such that unbundling dominates bundling if and only if $c < c^*$;*
2. *there exists a threshold N^* such that unbundling dominates bundling if and only if $N > N^*$;*
3. *there exists a threshold ρ^* such that unbundling dominates bundling if and only if $\bar{\rho}^{\epsilon} > \rho^*$;*
4. *there exists a threshold b^* such that unbundling dominates bundling if and only if $b > b^*$.*

The intuition of Lemma 6 is as follows. Relative to bundling, unbundling has two opposing effects. On the one hand, unbundling raises the information rents earned by wholesalers. On the other hand, it improves allocative efficiency because the broker can better match orders to the wholesalers with the lowest inventory costs. These lower inventory costs under unbundling arise from two sources. For a given stock, wholesalers may draw different realizations of the private cost shock $\epsilon_{i,j}$. In addition, their ex ante costs $\zeta_{i,j,0}$ may differ across stocks and across wholesalers.

Consider first the role of c. When c is large, the broker has strong incentives to smooth inventory and therefore allocates orders more uniformly across wholesalers. In that case, the potential gain in allocative efficiency from unbundling is small, while the increase in information rents remains. As a result, bundling is more attractive.

Next consider N. When there are many wholesalers, competition already drives information rents down. The additional increase in information rents caused by unbundling is then limited. At the same time, with more wholesalers the broker can more finely tailor allocations and thus obtain a larger efficiency gain from unbundling. In this case the efficiency effect dominates and unbundling tends to be better.

Next, consider $\bar{\rho}^{\epsilon}$. When average correlation of cost is high, the extra information rent created by unbundling is weaker. At the same time, allocative efficiency remains sensitive to cross stock differences in the cost terms $\zeta_{i,j,0}$. Hence, unbundling tends to be better for high $\bar{\rho}^{\epsilon}$.

Finally, consider b. The parameter b measures the dispersion of wholesalers' private information and also governs the scope for allocative efficiency gains from routing. When b is larger, both the potential efficiency gain from unbundling and the information rents earned by wholesalers increase. Lemma 6 shows that the efficiency effect dominates, and higher b makes unbundling more attractive.

Asymmetric Information and Window Length

When designing order allocation, brokers in our data use two windows (See Table I): they measure wholesalers' price improvement over an evaluation window and use it to determine order flow over the subsequent allocation window.

Window lengths vary substantially. Brokers B and C evaluate performance over 90 days and update allocations monthly, while Broker A evaluates over 30 days and updates daily. Thus, brokers differ in both evaluation and allocation window lengths, and shorter evaluation windows are paired with shorter allocation windows.

We rationalize these patterns with a simple extension of the baseline model. In the baseline model, each wholesaler i perfectly observes its period 1 cost component ϵ_i when choosing the period 0 execution price. We instead assume that wholesaler i observes only a noisy signal:

$$z_i = \begin{cases} \epsilon_i, & \text{with probability } p, \\ u_i, & \text{with probability } 1-p, \end{cases} \qquad u_i \sim \text{Unif}\left[-\tfrac{b}{2}, \tfrac{b}{2}\right],$$

where u_i is independent of ϵ_i and $\{u_i\}_{i=1}^{N}$ are independent. A longer evaluation or allocation window makes wholesalers less certain about future execution costs, corresponding to a lower p (noisier signal). At the extreme, $p = 1$ represents the shortest possible window (perfect information) and $p = 0$ the longest (no information beyond the prior).

This creates a tradeoff. A noisier signal reduces information asymmetry and therefore lowers information rents, but it can also reduce allocative efficiency because period 1 flow is more likely to reach higher-cost wholesalers.

The full solution to this extension is presented in the Internet Appendix. Here, we provide only the key implication.

Lemma 7. *The expected total price improvement received by broker is a U-shaped function of p. In particular, total price improvement is higher under $p = 1$ than under $p = 0$ if and only if $\frac{(N-1)b}{6c} \geq 1$.*

As discussed above, a noisier signal corresponds to a longer evaluation window or allocation window. Lemma 7 implies that the broker's objective can be maximized at either extreme, $p = 0$ or $p = 1$, depending on parameters. Therefore, optimal window choices can differ sharply across brokers, consistent with our data. The model also predicts that longer evaluation windows are paired with longer allocation windows, because both windows affect the precision of wholesalers' information about future execution costs in the same way.

Lemma 7 predicts that the broker is more likely to choose shorter windows when b is larger, where b measures the dispersion of wholesalers' private execution costs. Moreover, shorter windows are more likely when the number of wholesalers is larger or when the diseconomies of scale parameter c is lower.

Finally, Lemma 7 and our bundling results Lemma 6 deliver consistent comparative statics. The same parameters (like N, b and c) that make unbundling more attractive also tilt the broker toward shorter windows. This aligns with our data: the broker that unbundles (Broker A) also uses shorter windows, while the brokers that bundle (Brokers B and C) use longer windows.

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 document that brokers are selective in their choice of evaluation windows and allocation windows, in how they bundle orders across symbols and sizes when assessing performance, 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, as well as the data we obtain from each broker. 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.[20] 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.[21] 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

[20]Some brokers take payment-for-order flow. In guidance, FINRA recommends firms consider "potential conflicts of interest in order-routing decisions, including those related to its routing of orders to market centers that provide payment for order flow (PFOF) or other-routing inducements?" FINRA (2021) In practice, each broker chooses a PFOF rate, and charges that same rate to all wholesalers. This allows any decisions in routing between wholesalers to be made along the margin of price improvement, as each wholesaler pays the same PFOF.

[21]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.

likelihood, speed, and order fill size.

The brokers in our sample work with wholesalers who will always match or improve upon the public 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 (EFQ). 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. 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.[22]

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. In our model, the length of the evaluation window shapes the information environment in which wholesalers compete for future orders. In a longer window, wholesalers have less certainty about future execution costs, and therefore less certainty about marginal profit from additional order flow. This uncertainty leads to a tradeoff between information rents earned by wholesalers (which arise from wholesaler's private knowledge of future execution costs) and allocated efficiency (as less precise information about costs leads to less precise routing to the lowest cost wholesaler).

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

[22] 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 $\frac{3+3+4}{3+6+7} = \frac{10}{16} = 62.5\%$.

contrast, Broker A evaluates wholesaler performance for past orders in an individual symbol, further divided into five order 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. Within each category, Broker B measures performance for small (less than 100 shares), medium (100 to 2,000 shares), and large (over 2,000 shares) orders. Broker B routes all orders in a security bin in the same manner; when combining performance across order sizes within the bin, Broker B places a special focus on specific order sizes.

The bundling decision involves a tradeoff. When a broker bundles stocks together for evaluation, pooling reduces wholesalers' information rents: aggregating performance across stocks dilutes the informational advantage that each wholesaler holds about its idiosyncratic cost on any individual stock. At the same time, bundling requires the routing decision to be uniform across all stocks in a bundle, which reduces allocative efficiency. The most efficient wholesaler for one stock may not be the most efficient for another stock in the same bundle, and bundling prevents exploiting these differences.

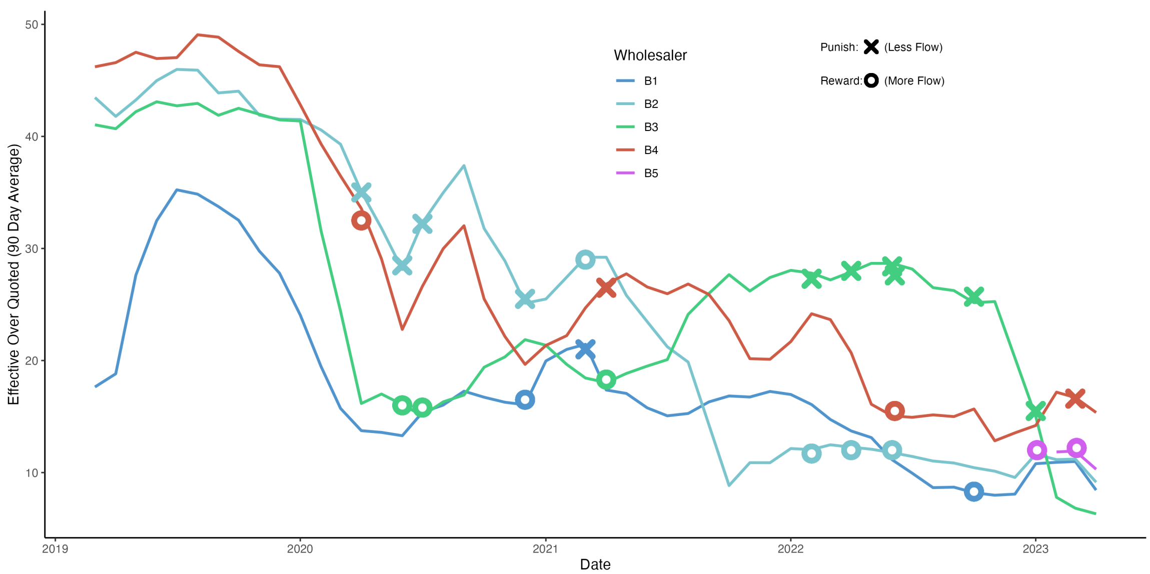
Figure 1 plots wholesaler performance over our sample period for the set of wholesalers who work with Brokers B and C. For Broker B, wholesalers who reduce their EFQ tend to obtain more order flow, while wholesalers who increase their effective-over-quoted spread tend to obtain less order flow. For Broker 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 tend to obtain more order flow while wholesalers with impaired scores tend to obtain less order flow.

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. Prediction 1 states that, under the optimal allocation rule, a wholesaler that offers more price improvement receives more future order flow. To evaluate the routing-performance relationship, we estimate the following regression:

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 tend to allocate more order flow to wholesalers charging lower EFQ spreads (adjustments depicted with circles) and 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

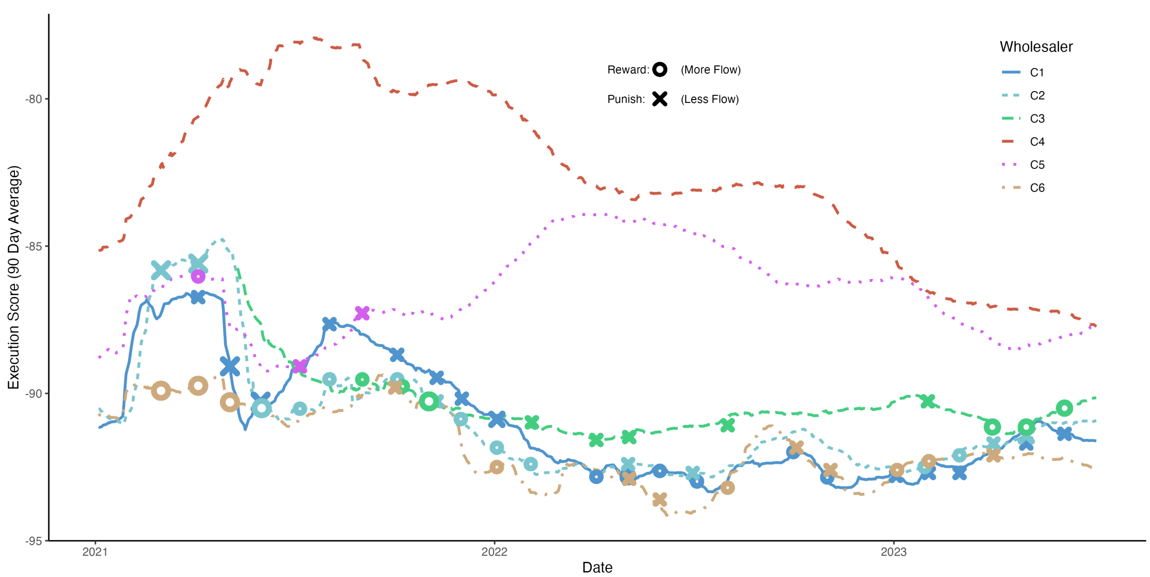


Table I: Best Execution Focus Points and Data. All brokers use a holistic process and consider all aspects of wholesaler performance, but have slightly different focuses in measuring wholesaler performance. We summarize some of the basic differences here. Category refers to the bin that a broker uses, 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. Panel B presents the data samples of proprietary data received from each wholesaler.

Panel A: Best Execution Focus Points

	Broker A	Broker B	Broker C
Category:	Symbol-Size Bins Individual Symbols Five Size Categories	Four Security Bins	One Bin
History:	30 Days	90 Days	90 Days
Decisions:	Daily (Rolling)	Monthly	Monthly

Panel B: Proprietary Data

	Observation Level	Date
Broker A	Trade-level data: 5 securities	March 13, 2023
	Wholesaler-Order-Size-Symbol-Day: EFQ 480 symbols - listed in Table XIV Size: (< 100), $[100, 500)$, $[500, 2000)$, $[2000, 5000)$, (> 5000)	Nov 15 - Dec 15 2021 & March 3 - April 3 & May 15 - June 15 & Aug - Nov 2022
Broker B	Wholesaler-Category-Size-Day EFQ Categories: Nasdaq 100, SP 500, All Nasdaq, All Listed Sizes: (< 100), *Focus**, > 2000	Jan 2019 - Jan 2023
	Limit Order Performance Categories: Nasdaq 100, SP 500, All Nasdaq, All Listed Sizes: All Size, (< 100) only, > 2000 only	Jan 2019 - May 2023
Broker C	Wholesaler-Day Score	Jan 2021 - July 2023
	Symbol-Day EFQ	Jan, Feb 2023

*Focus is 100-1999 prior to Jan 2020, and 1-1999 after.

REGRESSION 1: *For each broker, we estimate:*

$$OrderShare_{jkt} = \beta_0 + \beta_1 Prior\_EFQ_{jkt} + X_{kt} + \epsilon_{jkt}$$

for wholesaler j, time period t and security bin k and vector of controls X_{jk}.

We estimate Regression 1 separately using data from each broker.[23] We use two measurements of $Prior\_EFQ$. The first measure of $Prior\_EFQ$ is the prior period's effective-over-quoted ratio, which is the average effective spread paid divided by the average quoted spread, measured as a percentage. The second measure of $Prior\_EFQ$ is the wholesaler's ordinal rank (with the wholesaler having the best EFQ score in that period receiving a rank of one for that period, the wholesaler with the second-best EFQ score in that period receiving a rank 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 security bin. 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.[24]

Results of Regression 1 are presented in Table II. Across all three brokers, there is a strong relationship between performance and order share: wholesalers with worse EFQ obtain smaller order share allocations in future periods. This result is consistent with

[23]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.

[24]We present a broker routing with wholesaler fixed effects and broker routing based on changes in past performance in Appendix IA0.2. 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. 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.

Prediction 1, 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 percentage point increase in EFQ (i.e., charging a higher price, as a percentage of the quoted spread), we estimate that wholesalers receive 1.2 pp less order share from Broker A and 0.96 pp less order share from Broker B. For each lower rank, wholesalers receive 8.9 pp less order share from Broker A and 5.6 pp 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 pp (cardinal score measure) to 7.3 pp (ordinal score rank) less order flow from Broker C.

C. Performance Slopes

Our theory predicts that brokers should respond more strongly to wholesaler performance in environments with greater information asymmetry. In the model, the performance sensitivity of future order flow, $|dx_i^*/ds_i|$, is increasing in b, the dispersion of wholesalers' private information. We also show that greater information asymmetry is associated with greater variation in wholesaler performance. This yields a simple empirical implication within Broker A's security bins: changes in order flow should be more sensitive to changes in a wholesaler's performance when the underlying symbol-bin has more variable wholesaler performance. To test this prediction, we estimate the following specification.

REGRESSION 2: *Using observations at of wholesaler i symbol j and security bin k on date t, we estimate:*

$$\begin{aligned}\Delta(\mathit{Order\ Share})_{ijk(t+1)} = \quad & \beta_0 + \beta_1\Delta(EFQrank)_{ijkt} + \beta_2 SD_{jk} \\ & + \beta_3\Delta(EFQrank)_{ijkt} * SD_{jk} + \epsilon_{ijkt}\end{aligned}$$

Our interest is in β_3, the interaction term. Order flow adjustments to changes in EFQ ranking should be larger when the underlying bin has more variable performance of wholesalers. Results of Regression 2 are presented in Table III.[25] Changing rank for a higher-variance symbol (75^{th} percentile) is associated with a 0.3 additional percentage points of lost

[25]In the internet appendix, we consider variations of Regression 2 using changes in past EFQ, as well as a security-bin level estimation of incentive slopes via first-wholesaler order share across bins.

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 and date; observations are at the category-wholesaler-date level for Broker B with a fixed effect for each trade date and security bin, and we cluster standard errors by wholesaler and date; 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 and date.

	Dependent variable: OrderShare					
	Broker A Data		Broker B Data		Broker C Data	
	(1)	(2)	(3)	(4)	(5)	(6)
Prior EFQ	-1.230^{***}		-0.958^{***}			
	(0.130)		(0.104)			
Prior EFQ Rank		-8.882^{***}		-5.553^{***}		
		(0.746)		(0.757)		
Prior Score					-3.015^{***}	
					(0.583)	
Prior Score Rank						-7.294^{***}
						(0.577)
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		

order share compared to a lower-variance symbol (25^{th} percentile).

Table III: Order Flow Sensitivity and Rank. We estimate Regression 2 using Broker A data, as Broker A routes each symbol separately. $\Delta(OrderShare)$ measures daily changes in order share. We consider two measures of variability: SD_{WL}, where we measure daily EFQ for each individual wholesaler and then measure the standard deviation across wholesaler-day observations, and SD_{DL}, where we measure daily EFQ across all wholesalers, and then measure the standard deviation across daily observations. Δ_{rank} measures changes in ordinal wholesaler ranking (with the best, lowest EFQ wholesaler having rank 1). We fit a wholesaler fixed effect for each wholesaler (and cluster standard errors by wholesaler) in column (2) and (4).

	Dependent variable:			
	$\Delta(OrderShare)$			
	(1)	(2)	(3)	(4)
Δ_{rank}	-9.728^{***}	-9.749^{***}	-9.543^{***}	-9.564^{***}
	(0.071)	(1.328)	(0.071)	(1.243)
SD_{WL}	0.004^{***}	0.004		
	(0.001)	(0.002)		
$\Delta_{rank} * SD_{WL}$	-0.027^{***}	-0.027^{*}		
	(0.001)	(0.012)		
SD_{DL}			0.001	0.001
			(0.001)	(0.001)
$\Delta_{rank} * SD_{DL}$			-0.038^{***}	-0.038^{***}
			(0.001)	(0.005)
Constant	0.123^{**}		0.174^{***}	
	(0.049)		(0.049)	
Wholesaler FE		X		X
Observations	269,682	269,682	269,659	269,659
R^2	0.080	0.080	0.081	0.081
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$		

D. *Incentive Channel*

The routing-performance relationship documented above is consistent with both a informational motive (the broker updates its estimate of which wholesaler has the lowest cost) and an incentive motive (the broker adjusts flow to reward good performance and punish bad performance). To distinguish between these channels, we first establish that wholesaler performance is persistent, and then examine whether the broker continues to adjust order flow even when the leading wholesaler's ranking is unlikely to change—a setting in which learning alone would predict little reallocation.

Wholesaler performances are persistent. We estimate an autoregressive model of wholesaler EFQ against prior EFQ, and present results in Table IV. For Broker A, persistence of wholesaler EFQ is a statistically positive and significant 0.23 within each symbol and ordersize bin. For Broker B, persistence of wholesaler EFQ is estimated at 0.7, and for Broker C, persistence of wholesaler EFQ is 0.9. Wholesalers who outperform their competitors are likely to continue to outperform their competitors.

Because wholesaler performance is persistent, a purely informational routing rule would quickly identify the best wholesaler and allocate most flow accordingly, with little further adjustment once the ranking is well established. To test whether brokers also adjust flow for incentive reasons, we examine cases in which the leading wholesaler's advantage is large enough that the ordinal ranking is unlikely to change. If the broker still reallocates flow in response to marginal changes in the leader's performance in these settings, this suggests that the broker uses order flow not only to route to the best wholesaler, but also to incentivize continued effort. Rather than changes in EFQ, we consider changes in the lead of the rank-1 wholesaler in EFQ.[26]

Table V presents results of this estimation. We use data from Broker A as it has a large enough quantity of wholesaler-symbol-order size-date bins to divide symbols. A one percentage point decrease in EFQ lead of the rank-1 wholesaler is associated with a 0.4 percentage point decrease in order share received by the rank-1 wholesaler. This sensitivity holds true whether the lead of the rank-1 wholesaler is small (eg 5 to 10 pp- Column 2) or large (e.g., more than 20 pp - Column 4), and whether the lead is measured in the prior 15 days (Panel B) or in the prior 16 to 30 days (Panel C). Even in cases where the

[26] For example, if the first-ranked wholesaler charges an EFQ of 20 pp, and the second-ranked wholesaler charges an EFQ of 30 pp, the lead of the first-ranked wholesaler would be 10 pp.

Table IV: Persistence of Wholesaler Performance. We fit an autoregression model of order 1 for EFQ changes at the month-symbol-order-size level (Broker A), month-security-bin category (Broker B) or monthly level (Broker C). We use EFQ for Broker A and Broker B, and execution score for broker C. For Broker A, we cluster standard errors by market center, symbol, and order size (column 2) or date, symbol, and order size (column 3). For Broker B, we cluster standard errors by security bin and market center (column 2) or date and security bin (column 3). For Broker C we cluster standard errors by market center (column 2) or date (column 3).

Panel A: Broker A Data

	Dependent variable:		
	EFQ_t		
	(1)	(2)	(3)
EFQ_{t-1}	0.230***	0.230***	0.230***
	(0.023)	(0.035)	(0.043)
Date+Symbol+Ordersize FE	X	X	X
Observations	2,339	2,339	2,339
R^2	0.433	0.433	0.433

Panel B: Broker B Data

	Dependent variable: EFQ_t		
	(1)	(2)	(3)
EFQ_{t-1}	0.713***	0.713***	0.713***
	(0.025)	(0.020)	(0.047)
Security bin + Date FE	X	X	X
Observations	806	806	806
R^2	0.783	0.783	0.783

Panel C: Broker C Data

	Dependent variable: Execution $Score_t$		
	(1)	(2)	(3)
Execution $Score_{t-1}$	0.882***	0.882***	0.882***
	(0.037)	(0.080)	(0.055)
Date FE	X	X	X
Observations	176	176	176
R^2	0.800	0.800	0.800

ordinal ranking uncertainty is low, Broker A adjusts order flow in response to changes in past wholesaler performance.

E. Specificity and Self-Evaluation

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,[27] 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 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.[28] While the broker will

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

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

Table V: Routing Sensitivity and Wholesaler Lead. We estimate the routing-performance sensitivity using only the first-ranked wholesaler of Broker A, as Broker A adjusts routing daily. $\Delta(OrderShare)$ refers to the change in order share obtained by the first-ranked wholesaler comparing the last 30 days to 31 days. Δ EFQ Lead refers to the change in the lead of the rank 1 wholesaler comparing the last 32 to 31 days; lead is measured as the percentage point difference in thirty day average EFQ between the rank-1 and rank-2 wholesalers. Column 1 presents results over all security-symbol-order size bins, with a fixed effect for each wholesaler and standard errors clustered by wholesaler. Column 2 restricts to bins where the rank-1 wholesaler lead is at least 5 to 10 percentage points, Column 3 a lead of at least 10 to 20, and Column 4 a lead of 20 or more. Panel B uses an EFQ lead based on the prior 1 to 15 days, and Panel C uses an EFQ lead based on the prior 16 to 30 days.

Panel A:

	Dependent variable: $\Delta(OrderShare)$			
EFQ Lead:	All	$(5,10)$	$(10,20)$	>20
	(1)	(2)	(3)	(4)
$\Delta EFQ_{R1-lead}$	0.404***	0.428*	0.859***	0.253**
	(0.048)	(0.193)	(0.155)	(0.074)
Wholesaler FE	X	X	X	X
Observations	32,292	9,248	5,102	605
R^2	0.003	0.003	0.015	0.025

Panel B:

	Dependent variable: Δ Order Share			
EFQ Lead_{1-15}:	All	$(5,10)$	$(10,20)$	>20
	(1)	(2)	(3)	(4)
$\Delta EFQ_{R1-lead}$	0.404***	0.237*	0.784***	0.375**
	(0.048)	(0.087)	(0.127)	(0.091)
Wholesaler FE	X	X	X	X
Observations	32,292	9,362	6,129	1,021
R^2	0.003	0.001	0.010	0.019

Panel C:

	Dependent variable: Δ Order Share			
EFQ Lead_{16-30}:	All	$(5,10)$	$(10,20)$	>20
	(1)	(2)	(3)	(4)
$\Delta EFQ_{R1-lead}$	0.404***	0.360**	0.662**	0.344**
	(0.048)	(0.094)	(0.185)	(0.111)
Wholesaler FE	X	X	X	X
Observations	32,292	9,482	6,278	1,036
R^2	0.003	0.003	0.007	0.017

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.

In these specifications, we know the true routing history used by the broker. In Section IA0.3 of the internet appendix, we consider variations of the specific history with Broker A data (e.g., using shorter or longer histories, or different order-size bins), and this can reduce the predictive power of wholesaler prior EFQ in explaining wholesaler order share.[29] Our results underscore that for an individual retail investor to be able to understand broker routing decisions for their own trades, they may need an exceedingly large sample of trades as well as detailed knowledge of the specific focuses on which the broker places weight.

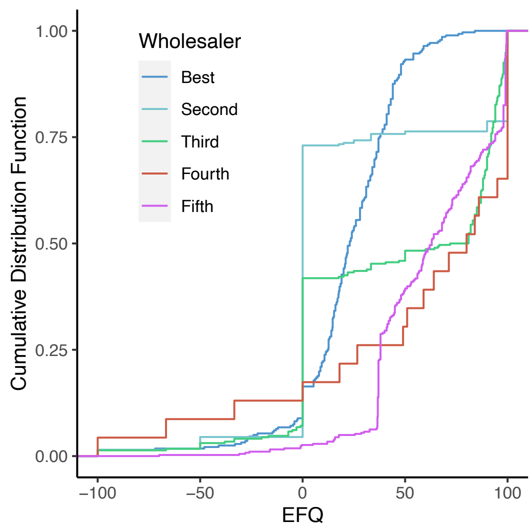
---

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

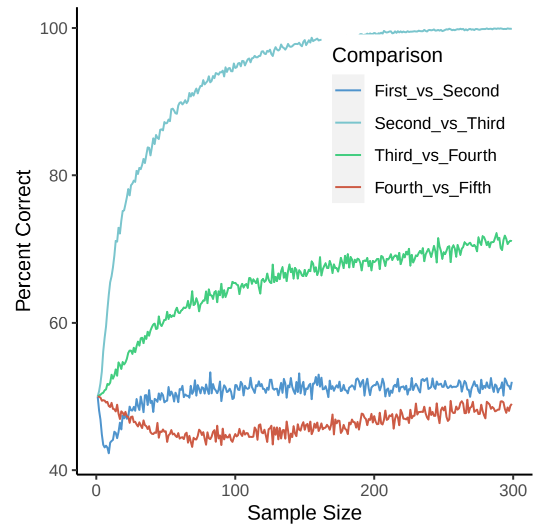
[29] We find that R^2 falls from 10% using a 30-day history window (the same as that used by Broker A) to 5% at shorter (5 days) or longer (45 days) time horizons. Across alternative sizes, we find that wholesaler performance in medium-orders predicts wholesaler order shares in medium orders with an R^2 of 12.5%; in contrast, small order performance explains 2.5% of the variation in medium-size order shares, while large order performance explains less than 1% of the variation in medium-size order shares.

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.

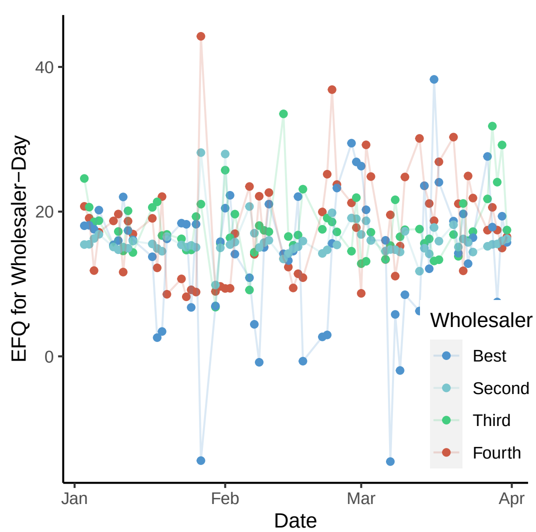
Panel A: Stock EFQ CDF (Broker A)



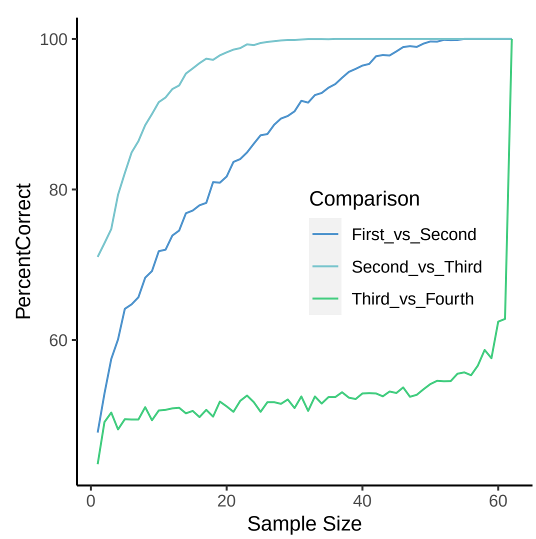
Panel B: Comparison Probability (Broker A)



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 evaluation method used by brokers. Specifically, we show that wholesalers are responsive to broker focuses, and that wholesalers change performance near the end of evaluation windows.

A. *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 a 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.[30] 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 of 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

[30]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.

closer to 150% of the average quoted spread six months after the focus change.[31] To confirm the intuition of Figure 3, we estimate the following regression:

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

$$\begin{aligned} EFQ_{jkt} = \quad & \beta_0 + \beta_1 Size\_Category_{jkt} + \beta_2 FocusChange_t \\ & + \beta_3 Size\_Category_{jkt} * FocusChange_t + X_{jk} + \epsilon_{jkt} \end{aligned}$$

$Size\_Category$ 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 security bin 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 pp better, while large orders obtain an EFQ spread which is 68 pp worse.[32] 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 pp better while large orders obtain an EFQ spread which is 85 pp worse.

The key conclusion of this analysis is that any substantial change in brokers' routing practices is likely to prompt a strategic response from wholesalers. In contemporaneous

[31] 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 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."

[32] 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.

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. Under the assumption that such hypothetical rerouting would not lead to different behavior by wholesalers, they find that it would significantly lower spreads. Our results in this section warn against making this assumption when studying counterfactual routing practices: we observed that wholesalers' behavior changed when Broker B implemented the focus change. In the focus change we analyze, there is a trade-off: While there are improvements along one dimension (execution quality of small orders), they are accompanied by dis-improvements along another dimension (execution quality of large orders). This suggests that the analysis of the costs and benefits of hypothetical changes in routing practices would likely require a structural model to account for wholesalers' strategic responses to those changes.

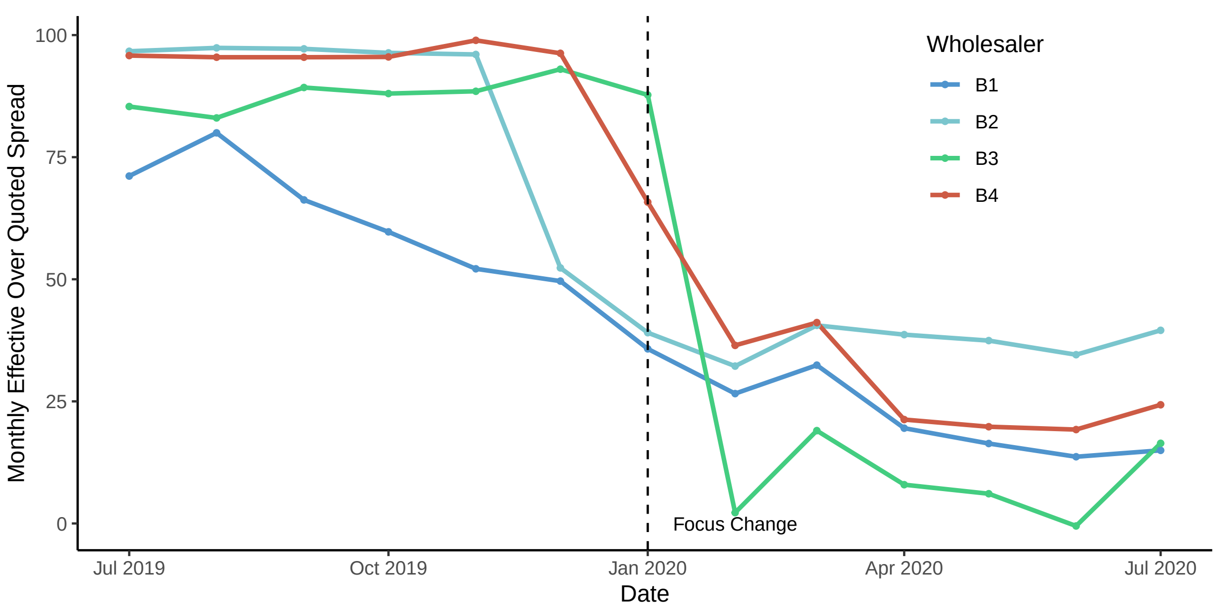
Any change to market structure will change the behavior of participants. Ernst, Spatt, and Sun (2024a) highlight how order-by-order auctions expose market makers to a winner's curse, and consequently the price improvement they provide in auctions can sometimes be worse than that obtained in the current 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 contrasts 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.

B. *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,

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

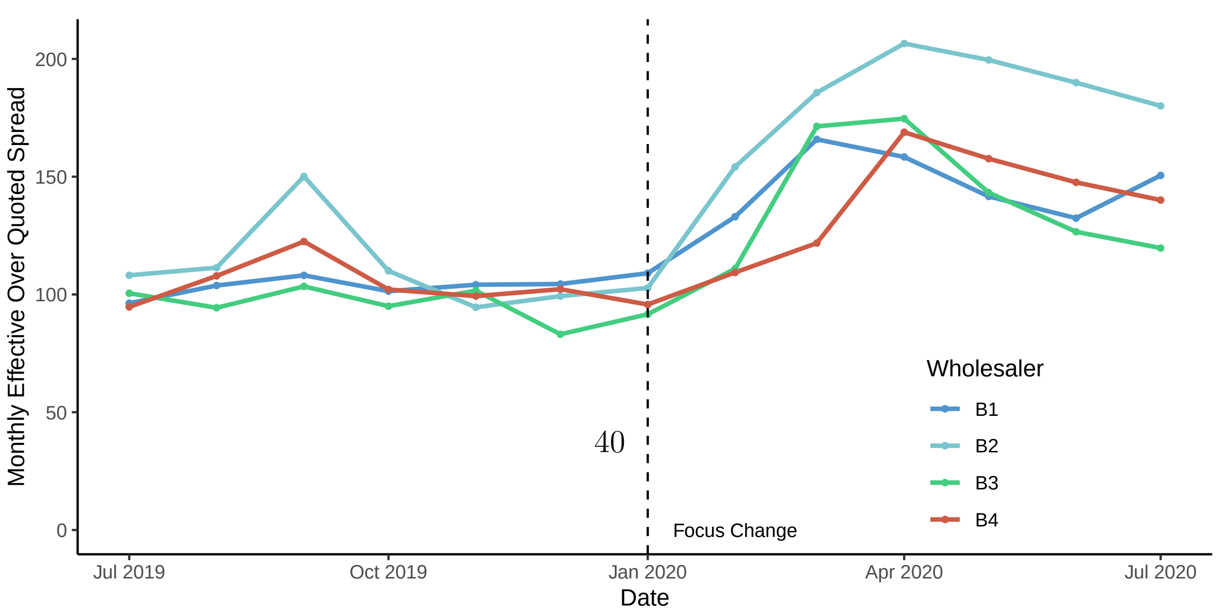


Table VI: Broker Focus-Change - Estimation of Regression 3. FocusChange is a variable which takes the value 1 after Broker B institutes a focus change on January 1, 2020. OddLot is a variable which takes the value 1 for orders of between 1 and 99 shares; LargeOrder is a variable which takes the value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

	Dependent variable: EFQ			
	Odd Lot		Large Orders	
	(1)	(2)	(3)	(4)
OddLot		44.685^{***}		
		(7.554)		
LargeOrder				66.695^{***}
				(5.093)
FocusChange	-61.821^{***}	-17.085^{***}	68.992^{***}	-16.847^{***}
	(7.984)	(1.646)	(4.252)	(0.995)
OddLot∗FocusChange		-44.790^{***}		
		(6.434)		
LargeOrder∗FocusChange				85.549^{***}
				(10.884)
Observations	820	1,650	820	1,650
R^2	0.841	0.809	0.544	0.886
Note:				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

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.

For each wholesaler j, we define $|(EFQ_{jt} - \mu_{jt})|$ as the absolute value of the difference between the EFQ given by wholesaler j on date t and their 90-day-average μ_{jt} from wholesaler j on date t. To capture the effect of the evaluation window drawing near, we define the indicator $LastFiveDays$ which takes the value 1 for the last five days of the month. For each possible wholesaler rank, we then regress $|(EFQ_{jt} - \mu_{jt})|$ on $LastFiveDays$ for each possible wholesaler rank.

Results are presented in Table VII. In the last five days of the month, wholesalers have larger absolute deviations from their 90-day average. Coefficients for the first-ranked wholesaler are not significant, but coefficients for the third and fourth-ranked wholesaler are highly significant for both Broker B and Broker C; over the last five days of the month, these wholesalers have an absolute deviation from average between 0.73 and 2.14 higher, consistent with these wholesalers offering much better or worse performance toward the end of the evaluation window. In the Internet Appendix Section IA0.7, we estimate an alternative specification using the number of days left in the evaluation window, and obtain similar results.

Table VII: Wholesaler Strategic Improvement. $LastFiveDays$ is an indicator which takes the value 1 if the date is in the last five days of the month. The dependent variable $|(EFQ_{jt} - \mu_{jt})|$ compares the EFQ (or score) given by wholesaler j on date t against the 90-day average EFQ μ_{jt} given by wholesaler j on date t. In column 1, we estimate for the 1st-Ranked (i.e., best) wholesaler. In column 2, for the 2nd-ranked, column 3 for the 3rd ranked, and column 4 for the 4th ranked. Panel A uses data from Broker B with observations are at the wholesaler-security bin-day level, so we fit a fixed effect for each security bin and cluster standard errors by security bin. Panel B uses data from Broker C with observations at the wholesaler-day level. We do not use data from Broker A has Broker A evaluates on a rolling basis.

Panel A: Broker B Data

	Dependent variable: $\|(EFQ_{jt} - \mu_{jt})\|$			
Rank:	One	Two	Three	Four
	(1)	(2)	(3)	(4)
$LastFiveDays$	0.434	0.723	1.331**	2.142***
	(0.418)	(0.723)	(0.315)	(0.294)
Security Bin FE	X	X	X	X
Observations	2,893	3,037	2,959	2,957
R^2	0.027	0.015	0.017	0.076

Panel B: Broker C Data

	Dependent variable: $\|(Score_{jt} - \mu_{jt})\|$			
Rank:	One	Two	Three	Four
	(1)	(2)	(3)	(4)
$LastFiveDays$	0.164	1.134***	0.729**	1.276***
	(0.182)	(0.292)	(0.356)	(0.271)
Constant	1.539***	1.820***	1.757***	1.553***
	(0.065)	(0.105)	(0.128)	(0.097)
Observations	628	628	628	628
R^2	0.001	0.024	0.007	0.034

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 first began working with a 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 4: *For observation symbol j in order size category k in time period t:*

$$A5\_Post\_Share_{jk(t+1)} = \beta_0 + \beta_1 Competition_t + X_{jk} + \epsilon_{ijkt}$$

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.[33]

Results of Regression 4 are presented in Table VIII. Wholesaler A5 enters categories with a larger gap between the first and second-ranked wholesalers; for every 1 pp increase

[33]Between December 15 and March 3, Broker A routed to wholesaler A5 on a provisional basis. In conversations with the broker, they highlighted March 3- April 3 as reflective of post-entry performance of wholesaler A5.

in $First - To - Second$ (the arithmetic difference in EFQ between the first-ranked to second-ranked wholesaler), wholesaler A5 obtains an additional 0.07 pp of order share in the post-entry period.[34] Similarly, when the top wholesaler has a 1 pp larger order share, wholesaler A5 has a 0.15 pp 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 5: *For each symbol j in order size category k in time period t:*

$$Competition_{jkt} = \beta_0 + \beta_1 Post_t + X_{jk} + \epsilon_{jkt}$$

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 March 3 to April 3, 2022 (after A5's entry). 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 pp, while the first-to-average gap narrows by 2 pp. EFQ decreases by 6 pp, 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 pp 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.

[34]In relative terms, a one standard deviation increase in $First - To - Second$ is associated with a 5 pp increase in the order share of wholesaler A5 in the post-entry period.

We also estimate Regression 5 with the level of order share (as opposed to a *Post* indicator) obtained by wholesaler A5 in the March 3 to April 3, 2022 time period. While category-specific order share obtained by wholesaler A5 is endogenous (depending on the EFQ provided by A5 and by competing wholesalers), in categories where wholesaler A5 obtains a larger proportion of the order flow, there are larger decreases in the $First-To-Average$ difference in EFQ, the HHI, and EFQ itself. The theoretical predictions in Lemma 4 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 4, 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 5, 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	3,157	2,106	3,467	3,467
R^2	0.294	0.197	0.418	0.458	0.548

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 6: *For each symbol j and order size category k:*

$$Competition_{jk} = \beta_0 + \beta_1 Order\_Size_k + \beta_2 Trading\_Volume\_Decile_j + Price_j + X_{jk} + \epsilon_{jk}$$

$Competition$ 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 the order size bin (1 - smallest to 5 - largest), $Trading\_Volume\_Decile$ is the decile of trading volume (0 - lowest to 9 - highest), and $Price$ is the average share price. X_{jk} is a fixed effect for the identity of the top wholesaler in symbol j and order size category k.

Results of Regression 6 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.

These empirical patterns map directly onto the model's comparative statics in the information-asymmetry parameter b. As discussed in the subsection "Competitive Landscape: Order Size and Competition," larger orders exhibit more severe adverse selection, and less liquid stocks likewise involve greater dispersion in wholesalers' private cost information; both correspond to a higher b in the model. Table X confirms each of these predictions in Lemma 5.

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 effective spreads in some stocks may be partially offset by quoted 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 6. $First-To-Second$ 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

VII. Conclusion

This paper studies how U.S. retail brokers implement best execution by designing performance-based rules that allocate order flow among competing, privately informed wholesalers. We develop a mechanism design model that balances allocative efficiency against information rents, and test its predictions using proprietary data from three large retail brokers.

Our empirical results support the model's predictions. Brokers allocate more future order flow to wholesalers with superior past performance, and this sensitivity is stronger where cross-wholesaler performance dispersion is higher. Wholesalers respond strategically: a shift in broker's evaluation focus improves execution on newly included orders at the expense of others, and lower-ranked wholesalers adjust pricing more aggressively near the end of the evaluation window. Entry of a new wholesaler reduces average spreads, and larger orders exhibit wider EFQ gaps and higher concentration, consistent with the model's comparative statics on competition and information asymmetry.

These findings suggest that brokers manage competition by designing reallocation rules that balance rewarding past price improvement with creating future incentives, and wholesalers optimize subject to these rules. For policy, a key implication is that changes in routing criteria trigger strategic responses by wholesalers, so the costs and benefits of proposed reforms cannot be evaluated under the assumption that wholesaler behavior remains unchanged.

Our model also delivers predictions about the optimal choice of stock bundling and evaluation window length that we leave to future empirical work. A natural theoretical extension is broker-side competition, which would endogenize wholesalers' outside options as they trade off flow across competing brokers. More broadly, the paper highlights that the design of best-execution algorithms is itself a mechanism design problem whose analysis requires accounting for the strategic interaction between broker routing rules and wholesaler incentives.

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.

Anand, Amber, Mehrdad Samadi, Jonathan Sokobin, and Kumar Venkataraman, 2025, Retail limit orders, *Review of Finance* rfaf049.

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.

Barardehi, Yashar H, Vincent Bogousslavsky, and Dmitriy Muravyev, 2025, Unpacking Retail Trading Costs: the Role of Options Trading and Limit Order Usage, *Available at SSRN 5979974* .

Barclay, Michael J, and Jerold B Warner, 1993, Stealth trading and volatility: Which trades move prices?, *Journal of financial Economics* 34, 281–305.

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* .

Bryzgalova, Svetlana, Anna Pavlova, and Taisiya Sikorskaya, 2022, Retail Trading in Options and the Rise of the Big Three Wholesalers, *Available at SSRN* .

Chung, Kee H, Dominik Rösch, and Yuhan Zhang, 2023, Competition and execution quality in the market for retail trading, *Available at SSRN 4669930* .

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.

Easley, David, and Maureen O'hara, 1987, Price, trade size, and information in securities markets, *Journal of Financial economics* 19, 69–90.

Ernst, Thomas, and Chester S Spatt, 2025, Payment for Order Flow and Asset Choice, *Review of Financial Studies, Forthcoming* .

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* .

FINRA, 2021, Best Execution, Financial Industry Regulatory Authority, Accessed: March 17, 2026.

Hendershott, Terrence, Saad Ali Khan, and Ryan Riordan, 2022, Option Auctions, *Working Paper* .

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* .

Jain, Pankaj K, Jared A Linna, and Thomas H McInish, 2021, An examination of the NYSE's retail liquidity program, *The Quarterly Review of Economics and Finance* 80, 367–373.

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.

Myerson, Roger B, 1981, Optimal auction design, *Mathematics of operations research* 6, 58–73.

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. Summary Statistics

Table XI: Summary Statistics on Order Share. We present summary statistics on the order share obtained by a wholesaler of each rank (EFQ for Broker A and B, score for Broker C). The first-ranked wholesaler has the best (i.e. lowest) effective-over-quoted spread or score. Panel A presents summary statistics for Broker A with observations at the symbol-order-size-day level. Panel B presents summary statistics for Broker B with observations at the asset-month level. Panel C presents summary statistics for Broker C with observations at the monthly level.

Panel A: Broker A Data

Wholesaler Rank	5th	25th	50th	75th	95th	Mean	SD
One	8	27	43	61	100	46	26
Two	4	15	26	39	66	29	19
Three	2	8	17	28	52	20	16
Four	1	4	9	18	39	13	13
Five or Six	0	1	4	10	30	8	10

Panel B: Broker B Data

Wholesaler Rank	5th	25th	50th	75th	95th	Mean	SD
One	5	25	35	40	45	32	11
Two	5	20	30	35	45	28	11
Three	5	20	25	35	40	25	10
Four	5	5	10	20	35	15	11
Five or Six	5	5	5	12	25	10	9

Panel C: Broker C Data

Wholesaler Rank	5th	25th	50th	75th	95th	Mean	SD
One	21	30	33	36	42	32	7
Two	9	24	27	34	45	28	9
Three	2	20	24	28	37	23	10
Four	2	9	12	17	28	13	7
Five or Six	1	1	1	1	3	2	5

Table XII Rank Share Statistics. For row i, we take wholesalers with order share ranking i (where 1 is the wholesaler receiving the most order flow), and calculate the share of these wholesalers with an EFQ or score rank j in column j; values sum to 100 for each row. Panel A presents results for Broker A with observations at the wholesaler-symbol-ordersize-date level. Panel B presents results for Broker B with observations at the wholesaler-asset category-month level. Panel C presents results at the wholesaler-month level.

Panel A: Broker A Data

		EFQ Rank				
		1	2	3	4	5
Share Rank	1	57.7	24.6	11.5	5.0	1.3
	2	31.6	36.2	20.8	9.1	2.4
	3	18.2	26.6	32.7	17.9	4.5
	4	12.3	15.5	25.9	36.5	9.9
	5	6.6	9.2	15.4	23.2	45.6

Panel B: Broker B Data

		EFQ Rank				
		1	2	3	4	5
Share Rank	1	56.3	13.1	18.8	8.2	3.7
	2	9.5	36.5	21.5	32.5	0.0
	3	17.1	34.6	27.6	20.7	0.0
	4	7.6	16.5	32.4	43.5	0.0
	5	10.0	0.0	40.0	40.0	10.0

Panel C: Broker C Data

		Score Rank					
		1	2	3	4	5	6
Share Rank	1	93.5	3.2	0.0	3.2	0.0	0.0
	2	6.7	80.0	10.0	3.3	0.0	0.0
	3	0.0	13.3	63.3	16.7	0.0	6.7
	4	0.0	0.0	13.3	50.0	33.3	3.3
	5	0.0	3.3	13.3	20.0	46.7	16.7
	6	0.0	0.0	0.0	8.0	24.0	68.0

Table XIII: Summary Statistics on Std. Dev. of Symbol-Day Volume. We present summary statistics on the standard deviation of volume at the symbol-day level for Broker A and Broker C. We do not have symbol-day level data from Broker B. We restrict analysis to the 400 symbols in Table XIV. For each symbol-day, volume is measured as the percentage of average daily volume in that symbol (i.e. 100% for a day with volume equal to average for that symbol, 200% for a day with volume double the average for that symbol). We then calculate the standard deviation across symbol-days.

Broker A

Decile	0	1	2	3	4	5	6	7	8	9
Standard Deviation	250	231	215	240	206	243	184	260	208	153

Broker C

Decile	0	1	2	3	4	5	6	7	8	9
Standard Deviation	168	154	153	190	142	160	173	153	171	142

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.

Appendix IA1. A Numerical Example about Optimal Order Allocation

Figure 4 provides one example of how order shares are split among wholesalers when $N = 4$. In this numerical example, Wholesalers 3 and 4 have lower ex ante execution costs, while Wholesalers 1 and 2 have higher ex ante execution costs. Holding fixed the period 0 execution costs of Wholesalers 1, 2, and 3, we increase Wholesaler 4's period 0 execution cost. As Wholesaler 4's time 0 execution price rises, its period 1 market share declines, and the displaced market share is reallocated to the other three wholesalers.

Appendix IA2. A Toy Model of Broker Focus Change

This appendix provides a simple model that illustrates the mechanism behind the *Effects of Focus Change* discussion in the main text. The key ingredients are: (i) the broker's future routing depends only on *total* price improvement in period 0, and (ii) "primary focus" is represented by a minimum per-order price-improvement requirement.[35]

Orders and price improvement. There are three size categories of orders:

$$k \in \{o, m, \ell\},$$

[35] This interpretation matches the main text: once an order type is labeled as primary focus, the broker requires the price improvement to exceed a fixed exogenous threshold.

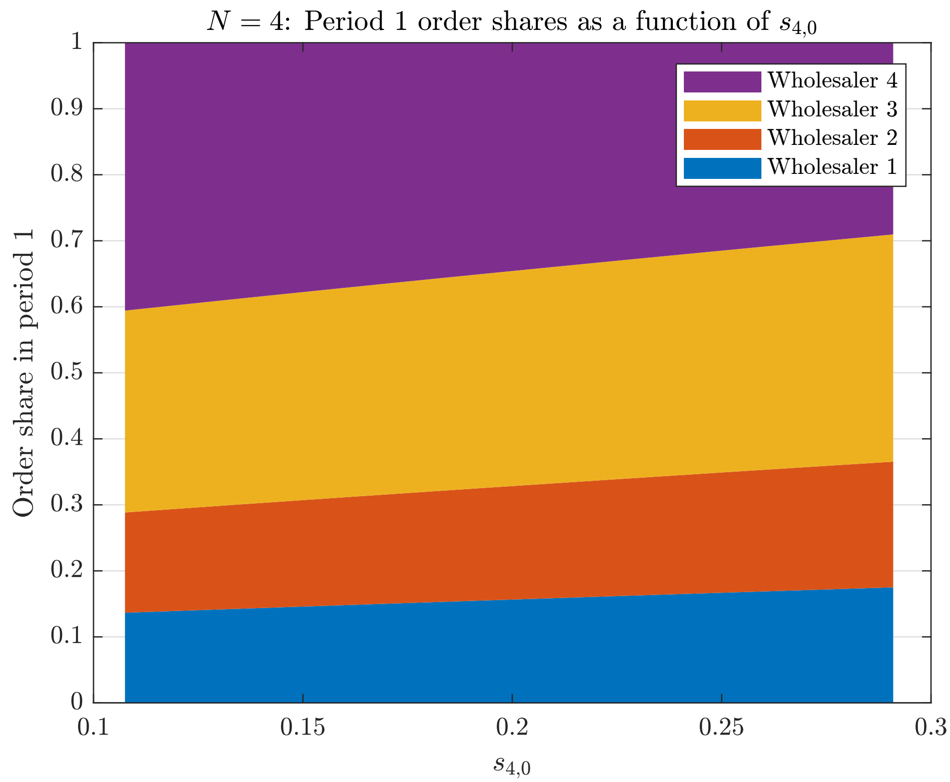


Figure 4. The figure plots the broker's optimal period 1 order allocation shares as functions of $s_{4,0}$ in the $N = 4$ case. Parameters are $\vec{\zeta}_0 = (0.40, 0.38, 0.18, 0.15)$, $\bar{s} = 1.00$, $b = 0.10$, and $c = 0.65$. We fix $s_{1,0} = 0.72$, $s_{2,0} = 0.68$, and $s_{3,0} = 0.27$.

where o denotes odd-lot orders (1–99 shares), m denotes medium orders (100–1,999 shares), and ℓ denotes large orders ($\geq$ 2,000 shares). Let their (exogenous) fractions be $\omega_o, \omega_m, \omega_\ell > 0$ with $\omega_o + \omega_m + \omega_\ell = 1$.

Execution in each category k occurs at a time-0 price $s_{k,0} \leq \bar{s}$. Define per-order *price improvement* by

$$\pi_k \;\coloneqq\; \bar{s} - s_{k,0} \;\geq 0.$$

Broker's state variable: total period-0 price improvement. Following the main text, when the period-1 allocation rule is the same for orders from all size categories, what matters is the total price improvement from all orders in period 0 (order type does not enter any primitives). Thus the wholesaler's relevant period-0 performance statistic is

$$T \;:=\; \omega_o \pi_o + \omega_m \pi_m + \omega_\ell \pi_\ell. \tag{IA21}$$

To keep the toy model minimal, consider the case where, for a given wholesaler type, the equilibrium pins down a fixed total price improvement level:

$$T = \tau, \tag{IA22}$$

and that τ does not change when the broker relabels which orders are in primary focus.

Primary-focus requirement. If an order category k is labeled *primary focus*, the broker imposes a minimum price improvement requirement:

$$\pi_k \;\geq\; \underline{\pi}, \qquad \underline{\pi} > 0. \tag{IA23}$$

If k is *non-primary focus*, the only requirement is $\pi_k \geq 0$.

Two cases (focus change). We compare two focus regimes.
Case 1 (pre-change). Only medium orders are primary focus:

$$m \text{ is focus}, \qquad o, \ell \text{ are non-focus}.$$

Thus $\pi_m \geq \underline{\pi}$ and $\pi_o, \pi_\ell \geq 0$.
Case 2 (post-change). Odd-lot orders become primary focus as well:

$$o, m \text{ are focus}, \qquad \ell \text{ is non-focus}.$$

Thus $\pi_o \geq \underline{\pi}$, $\pi_m \geq \underline{\pi}$, and $\pi_\ell \geq 0$.

Proposition IA2.1 (Mechanics of a focus expansion)**.** *Suppose the total price improvement is fixed at $T = \tau$ in both Case 1 and Case 2, and suppose $\pi_m = \underline{\pi}$ in both cases (the medium-order focus requirement binds). Let $\pi_o^{(1)}$ denote the odd-lot price improvement in Case 1 and $\pi_o^{(2)}$ denote the odd-lot price improvement in Case 2, with $\pi_o^{(2)} \geq \underline{\pi} > \pi_o^{(1)}$. Then the implied* *large-order price improvement satisfies*

$$\pi_\ell^{(2)} \;<\; \pi_\ell^{(1)}.$$

In particular, if odd-lot orders are executed at (or close to) NBBO pre-change so that $\pi_o^{(1)} \approx 0$ *and the new focus constraint binds so that* $\pi_o^{(2)} = \underline{\pi}$*, then*

$$\pi_\ell^{(2)} = \pi_\ell^{(1)} - \frac{\omega_o}{\omega_\ell}\underline{\pi}.$$

Proof. In either case, imposing $T = \tau$ and $\pi_m = \underline{\pi}$ in (IA21) implies that large-order price improvement must satisfy

$$\pi_\ell = \frac{\tau - \omega_m \underline{\pi} - \omega_o \pi_o}{\omega_\ell}.$$

Therefore

$$\pi_\ell^{(2)} - \pi_\ell^{(1)} = -\frac{\omega_o}{\omega_\ell}\big(\pi_o^{(2)} - \pi_o^{(1)}\big) \;<\; 0,$$

since $\omega_o, \omega_\ell > 0$ and $\pi_o^{(2)} > \pi_o^{(1)}$. The special case with $\pi_o^{(1)} \approx 0$ and $\pi_o^{(2)} = \underline{\pi}$ follows immediately. □

Interpretation. The proposition shows the core accounting logic: if the broker evaluates wholesalers based on total period-0 price improvement but imposes a stricter minimum requirement on the primary-focus orders, then expanding the focus set forces wholesalers to shift price improvement toward the newly focused orders. Holding total price improvement fixed, the remaining non-focus orders must receive less price improvement.

Appendix IA3. Wholesaler Responsiveness: Effects of Volatile Market

We examine how market conditions, specifically market 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 cost $\zeta_{i,0}$, 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: time 0 cost is now $\zeta_{i,0}(\sigma)$, scale cost $c(\sigma)$, information asymmetry $b(\sigma)$, and exchange spread $\bar{s}(\sigma)$, where

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

We do not impose specific functional forms and keep these mappings general. The average expected spread across all wholesalers is:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[s^*_{i,0}],$$

where $s^*_{i,0}$ is wholesaler i's period 0 execution price under the optimal allocation policy, and the expectation is taken over the vector $\vec{\epsilon}$. We establish the following results.

Lemma IA3.1. *When market volatility σ increases:*

1. *The average price offered by wholesalers, $\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[s^*_{i,0}]$, increases.*
2. *The change in the ratio $\frac{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[s^*_{i,0}]}{\bar{s}}$ is indeterminate.*

Proof. Recall that the period-0 allocation is uniform, $x_{i,0} = 1/N$ for all i, and that under the implementation in Proposition 3 the period-0 *total price improvement* delivered by wholesaler i,

$$(\bar{s} - s_{i,0})x_{i,0},$$

plays the role of the transfer to the broker in the optimal direct mechanism. Hence in equilibrium,

$$T^*_i(\epsilon_i) = (\bar{s} - s^*_{i,0}(\epsilon_i))\frac{1}{N} \quad \Longleftrightarrow \quad s^*_{i,0}(\epsilon_i) = \bar{s} - N\,T^*_i(\epsilon_i),$$

and therefore

$$\mathbb{E}[s^*_{i,0}] = \bar{s} - N\,\mathbb{E}[T^*_i(\epsilon_i)]. \tag{IA31}$$

Step 1: Compute $\mathbb{E}[T^*_i(\epsilon_i)]$. From Proposition 2 (Equation 5), the optimal transfer schedule is

$$T_i^*(\epsilon_i) = \left[A_i(\bar{s} - \zeta_{i,0}) - cA_i^2 - \frac{N-1}{12N^2c}b^2 - \frac{A_i b}{2} + \frac{N-1}{8Nc}b^2\right] + \left[2A_i - \frac{1}{c}(\bar{s} - \zeta_{i,0})\right]\frac{N-1}{N}\,\epsilon_i - \frac{N^2-3N+2}{2N^2c}\,\epsilon_i^2,$$

where $A_i = \frac{\bar{\zeta}_0 - \zeta_{i,0}}{2c} + \frac{1}{N}$ and $\bar{\zeta}_0 = \frac{1}{N}\sum_{i=1}^N \zeta_{i,0}$. Since $\epsilon_i \sim U[-b/2, b/2]$, we have $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i^2] = b^2/12$. Taking expectations and using $\mathbb{E}[\epsilon_i] = 0$ eliminates the linear term, yielding

$$\mathbb{E}[T_i^*(\epsilon_i)] = A_i(\bar{s} - \zeta_{i,0}) - cA_i^2 - \frac{A_i b}{2} + \frac{N-1}{12N}\frac{b^2}{c}. \tag{IA32}$$

Step 2: Express the average expected period-0 execution price. Averaging (IA31) across wholesalers and using (IA32),

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \mathbb{E}[s_{i,0}^*] &= \bar{s} - \sum_{i=1}^N \mathbb{E}[T_i^*(\epsilon_i)] \\ &= \bar{s} - \sum_{i=1}^N \left[A_i(\bar{s} - \zeta_{i,0}) - cA_i^2 - \frac{A_i b}{2} + \frac{N-1}{12N}\frac{b^2}{c}\right] \\ &= \sum_{i=1}^N \zeta_{i,0}A_i + \frac{b}{2} + c\sum_{i=1}^N A_i^2 - \frac{N-1}{12}\frac{b^2}{c},\end{aligned} \tag{IA33}$$

where we used $\sum_{i=1}^N A_i = 1$ (immediate from the definition of A_i).

Now substitute $A_i = \frac{\bar{\zeta}_0 - \zeta_{i,0}}{2c} + \frac{1}{N}$ into (IA33). A straightforward simplification gives the closed form

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}[s_{i,0}^*] = \bar{\zeta}_0 + \frac{b}{2} + \frac{c}{N} - \frac{N}{4c}\left(\frac{1}{N}\sum_{i=1}^N \zeta_{i,0}^2 - \bar{\zeta}_0^2\right) - \frac{N-1}{12}\frac{b^2}{c}. \tag{IA34}$$

(Recall that $\frac{1}{N}\sum_{i=1}^N \zeta_{i,0}^2 - \bar{\zeta}_0^2 \geq 0$ by Cauchy–Schwarz.)

Step 3: Comparative statics in volatility σ. Let $\zeta_{i,0} = \zeta_{i,0}(\sigma)$, $b = b(\sigma)$, $c = c(\sigma)$, and

$\bar{s} = \bar{s}(\sigma)$, with

$$\zeta'_{i,0}(\sigma) = \zeta'_{j,0}(\sigma) > 0 \; \forall i, j, \qquad b'(\sigma) > 0, \qquad c'(\sigma) > 0, \qquad \bar{s}'(\sigma) > 0.$$

Because all $\zeta_{i,0}(\sigma)$ shift by the same amount, the cross-sectional dispersion term $\frac{1}{N}\sum_{i=1}^{N}\zeta_{i,0}^2 - \bar{\zeta}_0^2$ is invariant to σ (a common shift does not change dispersion). Differentiating (IA34) therefore yields

$$\begin{aligned} \frac{d}{d\sigma}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[s_{i,0}^*]\right) &= \bar{\zeta}_0'(\sigma) + \left(\frac{1}{2} - \frac{N-1}{6}\frac{b(\sigma)}{c(\sigma)}\right)b'(\sigma) \\ &\quad + \left(\frac{1}{N} + \frac{N}{4c(\sigma)^2}\left(\frac{1}{N}\sum_{i=1}^{N}\zeta_{i,0}^2 - \bar{\zeta}_0^2\right) + \frac{N-1}{12}\frac{b(\sigma)^2}{c(\sigma)^2}\right)c'(\sigma). \qquad \text{(IA35)} \end{aligned}$$

The coefficient multiplying $c'(\sigma)$ is strictly positive. Also $\bar{\zeta}_0'(\sigma) > 0$ by assumption. Finally, Assumption 1 implies $c(\sigma) > (N-1)b(\sigma)$, hence

$$\frac{1}{2} - \frac{N-1}{6}\frac{b(\sigma)}{c(\sigma)} > \frac{1}{2} - \frac{1}{6} = \frac{1}{3} > 0.$$

Since $b'(\sigma) > 0$ and $c'(\sigma) > 0$, the right-hand side of (IA35) is strictly positive, proving part (1).

Step 4: The EFQ ratio is indeterminate. Let

$$P(\sigma) := \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[s_{i,0}^*], \qquad R(\sigma) := \frac{P(\sigma)}{\bar{s}(\sigma)}.$$

By the quotient rule,

$$R'(\sigma) = \frac{\bar{s}(\sigma)P'(\sigma) - P(\sigma)\bar{s}'(\sigma)}{\bar{s}(\sigma)^2}.$$

We have $P'(\sigma) > 0$ from part (1) and $\bar{s}'(\sigma) > 0$ by assumption, so the sign of $R'(\sigma)$ depends on the relative magnitudes of $P'(\sigma)$ and $\bar{s}'(\sigma)$ and can be either positive or negative. This proves part (2). □

Note that the ratio $\frac{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[s_{i,0}^*]}{\bar{s}}$ corresponds to the EFQ spread, thus we obtain the following empirical predictions from Lemma IA3.1.

Prediction 6. *When market volatility increases, the average spread offered by wholesalers increases, but the effect on the EFQ spread is ambiguous.*

Appendix IA4. Dynamic Extension with Limited Commitment

This appendix extends the baseline model to a dynamic environment in which the broker interacts with the same N wholesalers over multiple periods, but can commit only one period ahead. One period commitment means that, in period t, the broker can condition only next period allocations $\vec{x}_{t+1}$ on current period spreads $\vec{s}_t$. Allocations in periods $t+2, t+3, \ldots$ are chosen later as functions of future spreads. We show that the broker's problem reduces to a sequence of static mechanism design problems, so the baseline logic and comparative statics carry over period by period.

Appendix IA4.1. Environment, timing, and limited commitment

There is one broker (she) and N wholesalers (he) indexed by $i \in \{1, \ldots, N\}$. Time is discrete, $t \in \{0, 1, 2, \ldots, T\}$, and all agents discount by $\beta \in (0, 1)$. In each period t, the broker receives a unit mass of orders that she allocates across the N wholesalers and the market (the outside option). Let $x_{i,t} \in [0, 1]$ be the order mass executed by wholesaler i in period t, with feasibility $\sum_{i=1}^{N} x_{i,t} \leq 1$.

If wholesaler i executes $x_{i,t}$ at price $s_{i,t}$, his period t payoff is

$$x_{i,t}\big(s_{i,t} - \zeta_{i,t} - c_{i,t}(x_{i,t})\big) = x_{i,t}\big(\bar{s} - \zeta_{i,t} - c_{i,t}(x_{i,t})\big) - (\bar{s} - s_{i,t})x_{i,t}.$$

Here, $\zeta_{i,t}$ is the inventory cost of the first unit and $c_{i,t}(x)$ is the average inventory cost of executing x units. The term $(\bar{s} - s_{i,t})x_{i,t}$ is the total price improvement offered in period t. In the last period, wholesalers optimally choose $s_{i,T} = \bar{s}$.

The broker's period t payoff is

$$V_t = -\sum_{i=1}^{N} x_{i,t}s_{i,t} - \Big(1 - \sum_{i=1}^{N} x_{i,t}\Big)\bar{s} = -\bar{s} + \sum_{i=1}^{N} x_{i,t}(\bar{s} - s_{i,t}).$$

Thus, up to the constant $-\bar{s}$, the broker values total price improvement. Define the period t transfer from wholesaler i to the broker as

$$t_{i,t} = (\bar{s} - s_{i,t})x_{i,t}.$$

Information. In period t, wholesaler i privately observes his inventory cost $\zeta_{i,t}$. Types are independent across wholesalers. Each $\zeta_{i,t}$ follows a Markov process with conditional distribution $\zeta_{i,t+1} \sim F_i(\cdot \mid \zeta_{i,t})$. We assume $\zeta_{i,t}$ is revealed publicly at the beginning of period $t+1$, so private information is not persistent.

Assumption 3 (Cost structure)**.** *For any t,*

$$c_{i,t}(x) = cx, \qquad \zeta_{i,t+1} = \rho\zeta_{i,t} + (1-\rho)\epsilon_{i,t+1},$$

where $\rho \in [0,1]$ and the innovations $\epsilon_{i,t}$ are i.i.d. draws from a distribution F with density f on $[\underline{\epsilon}, \bar{\epsilon}]$.

$$\zeta_{i,t} = \rho\zeta_{i,t-1} + (1-\rho)\epsilon_{i,t}. \tag{IA41}$$

Timing within period $t < T$.

1. The vector of previous period types $\vec{\zeta}_{t-1} = (\zeta_{1,t-1}, \ldots, \zeta_{N,t-1})'$ becomes public.

2. Each wholesaler i privately learns $\zeta_{i,t}$ (equivalently $\epsilon_{i,t}$ via (IA41)).

3. The broker commits to a next period allocation rule $\vec{x}_{t+1}(\vec{s}_t) = \{x_{i,t+1}(\vec{s}_t)\}_{i=1}^N$, subject to $x_{i,t+1}(\vec{s}_t) \geq 0$ and $\sum_{i=1}^N x_{i,t+1}(\vec{s}_t) \leq 1$ for all $\vec{s}_t$.

4. Given current allocations $\vec{x}_t$ (chosen in period $t-1$) and the committed rule $\vec{x}_{t+1}(\cdot)$, wholesalers choose spreads $\vec{s}_t$ (equivalently transfers $\vec{t}_t$).

One period commitment. The broker can condition $\vec{x}_{t+1}$ on $\vec{s}_t$, but cannot condition $\vec{x}_{t+2}, \vec{x}_{t+3}, \ldots$ on $\vec{s}_t$. Moreover, since $\vec{\zeta}_t$ is publicly revealed at the beginning of $t+1$, period t private information affects incentives only through the promised period $t+1$ allocation.

Appendix IA4.2. Stage game: optimal direct mechanism

We characterize the broker's optimal period t mechanism, equivalently the optimal rule for $\vec{x}_{t+1}$ as a function of period t private information.

Proposition IA4.2. *Fix $t \in \{0, 1, \dots, T-1\}$ and the public state $\vec{\zeta}_{t-1}$. Under one period commitment, the broker's optimal period t direct mechanism induces a next period allocation rule $\vec{x}^*_{t+1}(\vec{\epsilon}_t; \vec{\zeta}_{t-1})$ that solves, for each realization $\vec{\epsilon}_t$,*

$$\max_{\vec{x}} \sum_{i=1}^{N} \left[x_i \left(\bar{s} - \rho^2 \zeta_{i,t-1} - \rho(1-\rho) \left(\epsilon_{i,t} + \frac{F(\epsilon_{i,t})}{f(\epsilon_{i,t})} \right) \right) - c x_i^2 \right] - \bar{s}, \tag{IA42}$$

subject to $x_i \geq 0$ for all i and $\sum_{i=1}^{N} x_i \leq 1$.

Remark. Relative to the baseline model, the broker again maximizes virtual surplus. The information rent term enters through the virtual cost component $\epsilon + F(\epsilon)/f(\epsilon)$. The dynamic features enter only through the public state $\zeta_{i,t-1}$ and the scaling $\rho(1-\rho)$ implied by Assumption 3 and one period commitment.

Proof. There are 4 steps in this proof:

Step 1. Direct mechanism and the role of one period commitment. By the revelation principle, it is without loss to restrict attention to direct mechanisms in which wholesaler i reports $\hat{\epsilon}_{i,t}$ and the mechanism specifies (i) a next period allocation $x_{i,t+1}(\hat{\vec{\epsilon}}_t)$ and (ii) a current transfer $t_{i,t}(\hat{\vec{\epsilon}}_t)$. Here the transfer is the total price improvement in period t, $t_{i,t} = (\bar{s} - s_{i,t})x_{i,t}$, and $x_{i,t}$ is predetermined when $s_{i,t}$ is chosen.

One period commitment is used in the continuation value. Conditional on the report in period t, the broker can commit only to $x_{i,t+1}$, and $\zeta_{i,t}$ becomes public at the beginning of $t+1$. Hence, the report affects wholesaler i only through the promised period $t+1$ allocation and the period t transfer. Future allocations beyond $t+1$ are chosen later as functions of future spreads and do not enter the period t incentive problem.

Step 2. Interim payoff from reporting. Suppress time subscripts and write ϵ_i for $\epsilon_{i,t}$. Let $x_i(\hat{\epsilon}_i, \epsilon_{-i})$ and $t_i(\hat{\epsilon}_i, \epsilon_{-i})$ denote the allocation in $t+1$ and the transfer in t induced by the direct mechanism when i reports $\hat{\epsilon}_i$ and others report truthfully. Define interim expectations

$$\bar{x}_i(\hat{\epsilon}_i) = \int x_i(\hat{\epsilon}_i, \epsilon_{-i})\, dF(\epsilon_{-i}), \qquad \bar{t}_i(\hat{\epsilon}_i) = \int t_i(\hat{\epsilon}_i, \epsilon_{-i})\, dF(\epsilon_{-i}).$$

Given a report $\hat{\epsilon}_i$ and true type ϵ_i, wholesaler i's relevant payoff in period t is the discounted maximum profit from executing the promised period $t+1$ order flow, minus the period t transfer:

$$U_i(\hat{\epsilon}_i, \epsilon_i) = \int\int \beta x_i(\hat{\epsilon}_i, \epsilon_{-i})\left(\bar{s} - \zeta_{i,t+1} - c\, x_i(\hat{\epsilon}_i, \epsilon_{-i})\right)\, dF(\epsilon_{-i})\, dF(\epsilon_{i,t+1}) - \bar{t}_i(\hat{\epsilon}_i). \tag{IA43}$$

Using $\zeta_{i,t+1} = \rho\zeta_{i,t} + (1-\rho)\epsilon_{i,t+1}$ and (IA41),

$$\zeta_{i,t+1} = \rho\big(\rho\zeta_{i,t-1} + (1-\rho)\epsilon_i\big) + (1-\rho)\epsilon_{i,t+1},$$

so

$$\mathbb{E}\big[\zeta_{i,t+1} \mid \zeta_{i,t-1}, \epsilon_i\big] = \rho^2\zeta_{i,t-1} + \rho(1-\rho)\epsilon_i + (1-\rho)\mathbb{E}[\epsilon].$$

Substituting into (IA43) and dropping the constant term $(1-\rho)\mathbb{E}[\epsilon]$ (which does not affect the optimal allocation) gives

$$U_i(\hat{\epsilon}_i, \epsilon_i) = \int \beta x_i(\hat{\epsilon}_i, \epsilon_{-i})\left(\bar{s} - \rho^2\zeta_{i,t-1} - \rho(1-\rho)\epsilon_i - c\, x_i(\hat{\epsilon}_i, \epsilon_{-i})\right)\, dF(\epsilon_{-i}) - \bar{t}_i(\hat{\epsilon}_i). \tag{IA44}$$

Step 3. Incentive compatibility and envelope. Incentive compatibility requires truthful reporting:

$$U_i(\epsilon_i) = \max_{\hat{\epsilon}_i} U_i(\hat{\epsilon}_i, \epsilon_i).$$

Since ϵ_i enters (IA44) only through the linear term $-\rho(1-\rho)\epsilon_i$, the generalized envelope theorem yields

$$U_i'(\epsilon_i) = -\beta\rho(1-\rho)\bar{x}_i(\epsilon_i). \tag{IA45}$$

Integrating (IA45) from ϵ_i to $\bar{\epsilon}$ gives

$$U_i(\epsilon_i) = U_i(\bar{\epsilon}) + \beta\rho(1-\rho)\int_{\epsilon_i}^{\bar{\epsilon}} \bar{x}_i(\tau)\, d\tau. \tag{IA46}$$

In an optimal mechanism, the broker minimizes information rents, so we impose the normalization $U_i(\bar{\epsilon}) = 0$.

Step 4. Transfers and virtual surplus. Evaluating (IA44) at truthful reporting and rearranging yields

$$\bar{t}_i(\epsilon_i) = \beta \bar{x}_i(\epsilon_i) \left(\bar{s} - \rho^2 \zeta_{i,t-1} - \rho(1-\rho)\epsilon_i\right) - \beta c \, \mathbb{E}\left[x_i(\vec{\epsilon})^2 \mid \epsilon_i\right] - U_i(\epsilon_i). \tag{IA47}$$

Taking expectations over ϵ_i and using (IA46) with $U_i(\bar{\epsilon}) = 0$,

$$\mathbb{E}[U_i(\epsilon_i)] = \beta\rho(1-\rho) \int_{\underline{\epsilon}}^{\bar{\epsilon}} \left(\int_{\epsilon_i}^{\bar{\epsilon}} \bar{x}_i(\tau)\, d\tau \right) f(\epsilon_i)\, d\epsilon_i.$$

Swap the order of integration:

$$\begin{aligned}
\int_{\underline{\epsilon}}^{\bar{\epsilon}} \left(\int_{\epsilon_i}^{\bar{\epsilon}} \bar{x}_i(\tau)\, d\tau \right) f(\epsilon_i)\, d\epsilon_i &= \int_{\underline{\epsilon}}^{\bar{\epsilon}} \left(\int_{\underline{\epsilon}}^{\tau} f(\epsilon_i)\, d\epsilon_i \right) \bar{x}_i(\tau)\, d\tau \\
&= \int_{\underline{\epsilon}}^{\bar{\epsilon}} \bar{x}_i(\tau) F(\tau)\, d\tau.
\end{aligned}$$

Therefore,

$$\mathbb{E}[U_i(\epsilon_i)] = \beta\rho(1-\rho) \int_{\underline{\epsilon}}^{\bar{\epsilon}} \bar{x}_i(\tau) F(\tau)\, d\tau.$$

The broker's period t expected payoff equals $\sum_{i=1}^{N} \mathbb{E}[\bar{t}_i(\epsilon_i)] - \bar{s}$. Substituting (IA47) and the expression for $\mathbb{E}[U_i(\epsilon_i)]$ gives

$$\begin{aligned}
\sum_{i=1}^{N} \mathbb{E}[\bar{t}_i(\epsilon_i)] - \bar{s} = \sum_{i=1}^{N} \mathbb{E}\Bigg[&\beta \bar{x}_i(\epsilon_i) \left(\bar{s} - \rho^2 \zeta_{i,t-1} - \rho(1-\rho)\epsilon_i\right) - \beta c \, \mathbb{E}\left[x_i(\vec{\epsilon})^2 \mid \epsilon_i\right] \\
&- \beta\rho(1-\rho) \bar{x}_i(\epsilon_i) \frac{F(\epsilon_i)}{f(\epsilon_i)} \Bigg] - \bar{s}.
\end{aligned}$$

Since β is a common factor on all choice relevant terms, the broker maximizes expected virtual surplus. The feasibility constraints bind pointwise in $\vec{\epsilon}$, so the problem separates by type realizations, yielding Equation IA42. □

Appendix IA4.3. Closed form allocations under uniform innovations

Assumption 4 (Uniform innovations and interiority). *The innovations are i.i.d. with $\epsilon_{i,t} \sim U[-b/2, b/2]$, where $b > 0$. Parameters are such that the solution to Equation IA42 is interior for all relevant public states and all realizations $\vec{\epsilon}_t$, so $x_i^* > 0$ for all i and $\sum_{i=1}^N x_i^* = 1$.*

Under Assumption 4,

$$\epsilon + \frac{F(\epsilon)}{f(\epsilon)} = \epsilon + \left(\epsilon + \frac{b}{2}\right) = 2\epsilon + \frac{b}{2},$$

which is increasing in ϵ.

Lemma IA4.2. *Under Assumption 4, the broker's optimal period t allocation rule is*

$$x_{i,t+1}^*(\vec{\epsilon}_t; \vec{\zeta}_{t-1}) = A_{i,t-1} + \frac{\rho(1-\rho)}{c}\left(\frac{1}{N}\sum_{j=1}^N \epsilon_{j,t} - \epsilon_{i,t}\right), \qquad i \in \{1, \dots, N\}, \tag{IA48}$$

where

$$A_{i,t-1} = \frac{1}{N} + \frac{\rho^2}{2c}\left(\frac{1}{N}\sum_{j=1}^N \zeta_{j,t-1} - \zeta_{i,t-1}\right). \tag{IA49}$$

Remark. Equation IA48 is the dynamic analog of the baseline linear allocation rule. Conditional on public costs, wholesalers with lower private innovations $\epsilon_{i,t}$ receive more next period order flow. Persistence affects only the coefficients through ρ^2 and $\rho(1-\rho)$.

Proof. Fix $t < T$ and $\vec{\zeta}_{t-1}$. Under Assumption 4, substitute $\epsilon_{i,t} + F(\epsilon_{i,t})/f(\epsilon_{i,t}) = 2\epsilon_{i,t} + b/2$ into Equation IA42. Dropping constants that do not depend on $\vec{x}$, for a given realization $\vec{\epsilon}_t$ the broker solves

$$\max_{\vec{x}} \sum_{i=1}^N \left[x_i\left(\bar{s} - \rho^2\zeta_{i,t-1} - \rho(1-\rho)\left(2\epsilon_{i,t} + \frac{b}{2}\right)\right) - cx_i^2\right] \quad \text{s.t.} \quad x_i \geq 0, \; \sum_{i=1}^N x_i \leq 1.$$

By interiority, $x_i > 0$ for all i and $\sum_{i=1}^N x_i = 1$ at the optimum. Let μ be the multiplier on $\sum_{i=1}^N x_i = 1$. The first order condition for x_i is

$$\bar{s} - \rho^2\zeta_{i,t-1} - \rho(1-\rho)\left(2\epsilon_{i,t} + \frac{b}{2}\right) - 2cx_i - \mu = 0,$$

so

$$x_i = \frac{1}{2c}\left(\bar{s} - \rho^2\zeta_{i,t-1} - \rho(1-\rho)\left(2\epsilon_{i,t} + \frac{b}{2}\right) - \mu\right). \tag{IA410}$$

Summing (IA410) over i and imposing $\sum_{i=1}^{N} x_i = 1$ gives

$$1 = \frac{1}{2c}\left(N\bar{s} - \rho^2\sum_{i=1}^{N}\zeta_{i,t-1} - \rho(1-\rho)\sum_{i=1}^{N}\left(2\epsilon_{i,t} + \frac{b}{2}\right) - N\mu\right),$$

hence

$$\mu = \bar{s} - \rho^2\frac{1}{N}\sum_{j=1}^{N}\zeta_{j,t-1} - \rho(1-\rho)\frac{1}{N}\sum_{j=1}^{N}\left(2\epsilon_{j,t} + \frac{b}{2}\right) - \frac{2c}{N}.$$

Substitute this expression for μ into (IA410):

$$\begin{aligned} x_i &= \frac{1}{2c}\left[\bar{s} - \rho^2\zeta_{i,t-1} - \rho(1-\rho)\left(2\epsilon_{i,t} + \frac{b}{2}\right) - \bar{s} + \rho^2\frac{1}{N}\sum_{j=1}^{N}\zeta_{j,t-1} + \rho(1-\rho)\frac{1}{N}\sum_{j=1}^{N}\left(2\epsilon_{j,t} + \frac{b}{2}\right) + \frac{2c}{N}\right] \\ &= \frac{1}{N} + \frac{\rho^2}{2c}\left(\frac{1}{N}\sum_{j=1}^{N}\zeta_{j,t-1} - \zeta_{i,t-1}\right) + \frac{\rho(1-\rho)}{2c}\left(\frac{1}{N}\sum_{j=1}^{N}\left(2\epsilon_{j,t} + \frac{b}{2}\right) - \left(2\epsilon_{i,t} + \frac{b}{2}\right)\right) \\ &= \frac{1}{N} + \frac{\rho^2}{2c}\left(\frac{1}{N}\sum_{j=1}^{N}\zeta_{j,t-1} - \zeta_{i,t-1}\right) + \frac{\rho(1-\rho)}{c}\left(\frac{1}{N}\sum_{j=1}^{N}\epsilon_{j,t} - \epsilon_{i,t}\right), \end{aligned}$$

which is Equation IA48 and Equation IA49. □

Implication. Lemma IA4.2 shows that the optimal allocation rule in the dynamic setting has the same structure as in the baseline static model. As a result, the implied incentives for price improvement and the associated equilibrium outcomes carry over. This extension therefore suggests that the baseline results are robust, provided the broker maintains commitment power over a single subsequent period.

Appendix IA5. Full Model: Asymmetric Information and Window Length

In this part, we provide the full model for the discussion of window length. We keep the baseline model unchanged except for wholesalers' observation of the period-1 shock. In particular, for each wholesaler i,

$$\zeta_{i,1} = \zeta_{i,0} + \epsilon_i, \qquad \epsilon_i \sim \text{Unif}\left[-\frac{b}{2}, \frac{b}{2}\right] \text{ i.i.d. across } i,$$

as in the baseline.

Signal and type. At the beginning of period 0, wholesaler i does *not* observe ϵ_i perfectly. Instead, he privately observes a signal z_i defined by:

$$z_i = \begin{cases} \epsilon_i, & \text{with probability } p, \\ u_i, & \text{with probability } 1-p, \end{cases} \qquad u_i \sim \text{Unif}\left[-\frac{b}{2}, \frac{b}{2}\right],$$

where u_i is independent of ϵ_i and $\{u_i\}_{i=1}^N$ are independent across wholesalers.

The wholesaler's type is z_i. The mechanism is direct in z_i: each wholesaler reports $\hat{z}_i \in [-b/2, b/2]$, and incentive compatibility is defined with respect to truthful reporting of z_i.

Distribution of z_i and posterior of $\epsilon_i \mid z_i$. Since both ϵ_i and u_i are uniform on the same support,

$$z_i \sim \text{Unif}\left[-\frac{b}{2}, \frac{b}{2}\right], \quad \text{independent across } i.$$

Moreover, conditional on $z_i = z$, the posterior distribution of ϵ_i is the mixture

$$\epsilon_i \mid z_i = z \;\sim\; p\,\delta_z + (1-p)\,\text{Unif}\left[-\frac{b}{2}, \frac{b}{2}\right],$$

where δ_z denotes a point mass at z. In particular,

$$\mathbb{E}[\epsilon_i \mid z_i = z] = p\,z. \tag{IA51}$$

Proposition IA5.3 (Optimal mechanism with noisy signals and expected total price improvement). *Maintain all baseline assumptions (including Assumption 1) and replace only the wholesaler's observation of ϵ_i by the signal structure above. Consider the optimal direct mechanism in reports $\hat{z}_i$.*

1. ***Allocation rule.*** *Under the optimal direct mechanism, the period-1 order share assigned to wholesaler i is*

$$X^*_{i,1}(\vec{z}) = A_i + \frac{p}{c}\left[\frac{1}{N}\sum_{j=1}^{N} z_j - z_i\right], \qquad \forall i, \tag{IA52}$$

where, as in the baseline,

$$\bar{\zeta}_0 := \frac{1}{N}\sum_{j=1}^{N}\zeta_{j,0}, \qquad A_i := \frac{\bar{\zeta}_0 - \zeta_{i,0}}{2c} + \frac{1}{N}.$$

2. ***Transfer rule.*** *There exists an optimal transfer schedule that depends only on i's report and is given by*

$$\begin{aligned}\widetilde{T}^*_i(z_i) = &\left[A_i(\bar{s} - \zeta_{i,0}) - cA_i^2 - \frac{(N-1)p^2b^2}{12N^2c} - A_i\frac{pb}{2} + \frac{(N-1)p^2b^2}{8Nc}\right] \\ &+ \left[2A_i - \frac{1}{c}(\bar{s} - \zeta_{i,0})\right]\frac{N-1}{N}\,p\,z_i - \left(\frac{N^2 - 3N + 2}{2N^2c}\right)p^2 z_i^2, \qquad \forall i.\end{aligned} \tag{IA53}$$

3. ***Expected total price improvement.*** *Let total period-0 price improvement from all wholesalers be the sum of transfers, $\sum_{i=1}^{N}\widetilde{T}^*_i(z_i)$ (equivalently, $\sum_{i=1}^{N}(\bar{s} - s_{i,0})x_{i,0}$ under implementation by period-0 price improvement, as in the baseline). Under the optimal mechanism,*

$$\mathbb{E}\left[\sum_{i=1}^{N}\widetilde{T}^*_i(z_i)\right] = (\bar{s} - \bar{\zeta}_0) - \frac{pb}{2} - \frac{c}{N} + \frac{1}{4c}\sum_{i=1}^{N}(\zeta_{i,0} - \bar{\zeta}_0)^2 + \frac{p^2(N-1)b^2}{12c}. \tag{IA54}$$

Proof. **Step 1: Direct mechanism and interim expected utility under type z_i.** As in the baseline, by the revelation principle we restrict attention to direct mechanisms. A direct mechanism specifies an allocation rule

$$\vec{X}_1(\cdot;\vec{\zeta}_0,\vec{x}_0):\ \times_{i=1}^{N}\left[-\frac{b}{2},\frac{b}{2}\right]\to\left\{(X_{1,1},\ldots,X_{N,1})\,\middle|\,X_{i,1}\geq 0,\ \sum_{i=1}^{N}X_{i,1}\leq 1\right\},$$

and a transfer rule

$$\vec{\widetilde{T}}(\cdot;\vec{\zeta}_0,\vec{x}_0):\ \times_{i=1}^{N}\left[-\frac{b}{2},\frac{b}{2}\right]\to\mathbb{R}^N,$$

where $\widetilde{T}_i$ is the transfer from wholesaler i to the broker (implemented in the baseline by period-0 total price improvement).

Fix a direct mechanism $(\vec{X}_1,\vec{\widetilde{T}})$. A wholesaler with signal z_i who reports $\hat{z}_i$ obtains expected utility

$$U_i(\hat{z}_i;z_i)=\mathbb{E}\Big[X_{i,1}(\hat{z}_i,\vec{z}_{-i})\big(\bar{s}-(\zeta_{i,0}+\epsilon_i+c\,X_{i,1}(\hat{z}_i,\vec{z}_{-i}))\big)-\widetilde{T}_i(\hat{z}_i,\vec{z}_{-i})\ \Big|\ z_i\Big],$$

where the expectation is taken over $\vec{z}_{-i}$ and ϵ_i conditional on z_i (and over any remaining uncertainty), using the same equilibrium concept as in the baseline.

Because ϵ_i enters linearly and wholesalers are risk neutral, the conditional expectation over $\epsilon_i \mid z_i$ depends on z_i only through $\mathbb{E}[\epsilon_i \mid z_i]$. By (IA51),

$$\mathbb{E}[\epsilon_i\mid z_i]=p\,z_i,$$

hence

$$U_i(\hat{z}_i;z_i)=\mathbb{E}\Big[X_{i,1}(\hat{z}_i,\vec{z}_{-i})\big(\bar{s}-(\zeta_{i,0}+pz_i+c\,X_{i,1}(\hat{z}_i,\vec{z}_{-i}))\big)-\widetilde{T}_i(\hat{z}_i,\vec{z}_{-i})\Big].\tag{IA55}$$

Define the truthful utility $U_i(z_i):=U_i(z_i;z_i)$.

Let F and f denote the CDF and PDF of $z_i\sim\text{Unif}[-b/2,b/2]$, so $f(z)=1/b$ and $F(z)=(z+b/2)/b$.

Step 2: Envelope condition and information rents. For each i, define the interim expected allocation and transfer as functions of own report:

$$\bar{X}_i(\hat{z}_i):=\int X_{i,1}(\hat{z}_i,\vec{z}_{-i})\,dF_{-i}(\vec{z}_{-i}),\qquad \bar{\widetilde{T}}_i(\hat{z}_i):=\int \widetilde{T}_i(\hat{z}_i,\vec{z}_{-i})\,dF_{-i}(\vec{z}_{-i}),$$

where F_{-i} is the product distribution of $\vec{z}_{-i}$ (signals are independent across wholesalers).

Incentive compatibility for type z_i means $\hat{z}_i = z_i$ maximizes (IA55). Since z_i enters (IA55) only via $-pz_i$ multiplied by the interim allocation, the envelope theorem yields

$$U_i'(z_i) = -p\,\bar{X}_i(z_i). \tag{IA56}$$

Integrating (IA56) gives

$$U_i(z_i) = U_i\left(\frac{b}{2}\right) + p\int_{z_i}^{b/2} \bar{X}_i(\tau)\,d\tau. \tag{IA57}$$

As in the baseline analysis, the broker minimizes information rents by setting the worst type's (interim) participation constraint to bind,

$$U_i\left(\frac{b}{2}\right) = 0, \tag{IA58}$$

so (IA57) becomes

$$U_i(z_i) = p\int_{z_i}^{b/2} \bar{X}_i(\tau)\,d\tau. \tag{IA59}$$

Step 3: Broker objective and virtual costs in the signal economy. Fix any IC mechanism $(\vec{X}_1, \vec{\bar{T}})$. Using (IA55) at truth and the definition of $U_i(z_i)$,

$$\bar{\bar{T}}_i(z_i) = \bar{X}_i(z_i)\big(\bar{s} - (\zeta_{i,0} + pz_i)\big) - c\,\mathbb{E}\big[X_{i,1}(\vec{z})^2 \mid z_i\big] - U_i(z_i).$$

Taking expectations and using (IA59),

$$\mathbb{E}\big[\bar{\bar{T}}_i(z_i)\big] = \mathbb{E}\big[X_{i,1}(\vec{z})\big(\bar{s} - (\zeta_{i,0} + pz_i)\big) - c\,X_{i,1}(\vec{z})^2\big] - p\,\mathbb{E}\left[\int_{z_i}^{b/2} \bar{X}_i(\tau)\,d\tau\right]. \tag{IA510}$$

Rewrite the last term by swapping the order of integration (integration by parts as in the baseline proof): since z_i has density f and CDF F,

$$\mathbb{E}\left[\int_{z_i}^{b/2} \bar{X}_i(\tau)\,d\tau\right] = \int_{-b/2}^{b/2}\left(\int_{z}^{b/2} \bar{X}_i(\tau)\,d\tau\right) f(z)\,dz = \int_{-b/2}^{b/2} \bar{X}_i(\tau)\,F(\tau)\,d\tau. \tag{IA511}$$

Using $F(\tau)/f(\tau) = \tau + b/2$ for the uniform distribution and substituting (IA511) into (IA510), we obtain

$$\mathbb{E}\big[\bar{\widetilde{T}}_i(z_i)\big] = \mathbb{E}\left[X_{i,1}(\vec{z})\left(\bar{s} - \zeta_{i,0} - pz_i - p\frac{F(z_i)}{f(z_i)}\right) - c\,X_{i,1}(\vec{z})^2\right]. \tag{IA512}$$

Summing (IA512) across i and noting that the broker's objective differs from $\sum_i \mathbb{E}[\widetilde{T}_i]$ only by the constant $-2\bar{s}$ (as in the baseline), we conclude that the optimal allocation rule solves the pointwise problem: for each realization $\vec{z}$,

$$\max_{\vec{X}_1} \sum_{i=1}^{N} \left[X_{i,1}\left(\bar{s} - \zeta_{i,0} - pz_i - p\frac{F(z_i)}{f(z_i)}\right) - cX_{i,1}^2\right] - 2\bar{s} \tag{IA513}$$

$$\text{s.t. } X_{i,1} \geq 0\ \forall i, \qquad \sum_{i=1}^{N} X_{i,1} \leq 1.$$

Since $F(z_i)/f(z_i) = z_i + b/2$, the term in parentheses in (IA513) becomes

$$\bar{s} - \zeta_{i,0} - pz_i - p(z_i + b/2) = \bar{s} - \zeta_{i,0} - 2pz_i - \frac{pb}{2}.$$

This establishes the signal-economy analogue of the baseline reduced-form problem.

Step 4: Solving (IA513) **yields the allocation rule** (IA52)**.** Fix $\vec{z}$. Under Assumption 1, the same interior conclusions as in the baseline apply here: $\sum_i X_{i,1} = 1$ and $X_{i,1} > 0$ for all i at the optimum.[36] Let μ be the multiplier on $\sum_i X_{i,1} \leq 1$. The first-order condition is, for each i,

$$\bar{s} - \zeta_{i,0} - 2pz_i - \frac{pb}{2} - 2cX_{i,1} - \mu = 0.$$

Solving and imposing $\sum_i X_{i,1} = 1$ gives

$$X_{i,1}^*(\vec{z}) = \frac{1}{N} + \frac{\bar{\zeta}_0 - \zeta_{i,0}}{2c} + \frac{p}{c}\left[\frac{1}{N}\sum_{j=1}^{N} z_j - z_i\right] = A_i + \frac{p}{c}\left[\frac{1}{N}\sum_{j=1}^{N} z_j - z_i\right],$$

[36] Because $p \in [0,1]$, the dispersion term coming from private information is weakly smaller than in the baseline ($pb \leq b$). Hence Assumption 1, which guarantees an interior solution in the baseline, also guarantees an interior solution here.

which is (IA52).

Step 5: Transfer rule (IA53)**.** Under the optimal allocation rule (IA52), the interim expected allocation given z_i is

$$\bar{X}_i(z_i) = \mathbb{E}_{\vec{z}_{-i}}\big[X^*_{i,1}(\vec{z}) \mid z_i\big] = A_i + \frac{p}{c}\left[\frac{1}{N}z_i - z_i\right] = A_i - \frac{p}{c}\frac{N-1}{N}z_i. \tag{IA514}$$

Using (IA59) and (IA514),

$$U_i(z_i) = p\int_{z_i}^{b/2}\left(A_i - \frac{p}{c}\frac{N-1}{N}\tau\right)d\tau = pA_i\left(\frac{b}{2} - z_i\right) - \frac{p^2(N-1)}{Nc}\left(\frac{b^2}{8} - \frac{z_i^2}{2}\right). \tag{IA515}$$

Next, compute the conditional second moment in the transfer identity. From (IA52),

$$X^*_{i,1}(\vec{z}) = A_i + \frac{p}{c}\left[-\frac{N-1}{N}z_i + \frac{1}{N}\sum_{j\neq i} z_j\right].$$

Conditional on z_i, the term $\frac{1}{N}\sum_{j\neq i} z_j$ has mean 0 and variance $\frac{N-1}{N^2}\cdot\frac{b^2}{12}$, so

$$\mathbb{E}\big[(X^*_{i,1}(\vec{z}))^2 \mid z_i\big] = \bar{X}_i(z_i)^2 + \frac{p^2}{c^2}\cdot\frac{N-1}{N^2}\cdot\frac{b^2}{12}. \tag{IA516}$$

Finally, combine (IA514), (IA515), and (IA516) in the identity

$$\widetilde{T}^*_i(z_i) = \bar{X}_i(z_i)\big(\bar{s} - \zeta_{i,0} - pz_i\big) - c\,\mathbb{E}\big[(X^*_{i,1}(\vec{z}))^2 \mid z_i\big] - U_i(z_i),$$

and expand. Collecting constants, linear terms in z_i, and quadratic terms in z_i^2 yields (IA53).

Step 6: Expected total price improvement (IA54)**.** Under the reduced-form problem (IA513), total period-0 price improvement (i.e., total transfers) at realization $\vec{z}$ equals

$$\sum_{i=1}^{N}\widetilde{T}^*_i(z_i) = \sum_{i=1}^{N} X^*_{i,1}(\vec{z})\left(\bar{s} - \zeta_{i,0} - \frac{pb}{2} - 2pz_i - cX^*_{i,1}(\vec{z})\right). \tag{IA517}$$

Taking expectations of (IA517), use $z_i \sim \mathrm{Unif}[-b/2, b/2]$ i.i.d., so $\mathbb{E}[z_i] = 0$, $\mathbb{E}[z_i z_j] = 0$ for

$i \neq j$, and $\mathbb{E}[z_i^2] = b^2/12$. Also, from (IA52), $\mathbb{E}[X_{i,1}^*(\vec{z})] = A_i$ and

$$\mathbb{E}\big[X_{i,1}^*(\vec{z})\, z_i\big] = \frac{p}{c}\left(\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} z_j\, z_i\right] - \mathbb{E}[z_i^2]\right) = -\frac{p}{c}\frac{N-1}{N}\cdot\frac{b^2}{12},$$

and

$$\mathbb{E}\big[(X_{i,1}^*(\vec{z}))^2\big] = A_i^2 + \frac{p^2}{c^2}\cdot\frac{N-1}{N}\cdot\frac{b^2}{12}.$$

Substituting these moments into (IA517) and simplifying gives

$$\mathbb{E}\left[\sum_{i=1}^{N}\widetilde{T}_i^*(z_i)\right] = \bar{s} - \frac{pb}{2} - \sum_{i=1}^{N} A_i\zeta_{i,0} - c\sum_{i=1}^{N} A_i^2 + \frac{p^2(N-1)b^2}{12c}. \tag{IA518}$$

It remains to express the two A_i-aggregates in baseline primitives. Using $A_i = \frac{1}{N} + \frac{\bar{\zeta}_0 - \zeta_{i,0}}{2c}$ and $\sum_i(\zeta_{i,0} - \bar{\zeta}_0) = 0$, one computes

$$\sum_{i=1}^{N} A_i\zeta_{i,0} = \bar{\zeta}_0 - \frac{1}{2c}\sum_{i=1}^{N}(\zeta_{i,0} - \bar{\zeta}_0)^2, \qquad \sum_{i=1}^{N} A_i^2 = \frac{1}{N} + \frac{1}{4c^2}\sum_{i=1}^{N}(\zeta_{i,0} - \bar{\zeta}_0)^2.$$

Substituting these identities into (IA518) yields (IA54), completing the proof. □

Now let's prove Lemma 7. From Equation IA54, the broker's expected total price improvement can be written as a function of p:

$$\Pi(p) \equiv \mathbb{E}\left[\sum_{i=1}^{N}\widetilde{T}_i^*(z_i)\right] = K - \frac{b}{2}\,p + \frac{(N-1)b^2}{12c}\,p^2, \tag{IA519}$$

where the p-independent constant is

$$K \equiv (\bar{s} - \bar{\zeta}_0) - \frac{c}{N} + \frac{1}{4c}\sum_{i=1}^{N}(\zeta_{i,0} - \bar{\zeta}_0)^2.$$

Since the coefficient on p^2 in (IA519) is $\frac{(N-1)b^2}{12c} > 0$, $\Pi(p)$ is a convex quadratic function of p on $[0,1]$. Therefore, its maximum over $p \in [0,1]$ is attained at an endpoint, i.e., at $p = 0$ or $p = 1$.

Evaluating (IA519) at the endpoints yields

$$\Pi(1) = K - \frac{b}{2} + \frac{(N-1)b^2}{12c} = (\bar{s} - \bar{\zeta}_0) - \frac{b}{2} - \frac{c}{N} + \frac{1}{4c}\sum_{i=1}^{N}(\zeta_{i,0} - \bar{\zeta}_0)^2 + \frac{(N-1)b^2}{12c},$$

and

$$\Pi(0) = K = (\bar{s} - \bar{\zeta}_0) - \frac{c}{N} + \frac{1}{4c}\sum_{i=1}^{N}(\zeta_{i,0} - \bar{\zeta}_0)^2.$$

So $\Pi(1) > \Pi(0)$ if and only if

$$-\frac{b}{2} + \frac{(N-1)b^2}{12c} > 0 \quad \Longleftrightarrow \quad \frac{(N-1)b^2}{12c} > \frac{b}{2} \quad \Longleftrightarrow \quad \frac{(N-1)b}{6c} > 1.$$

This establishes the stated condition and completes the proof.

Appendix IA6. Order Bundling Extension: Full Details and Proofs

This appendix provides the details and proofs for the multi-stock bundling extension. There are 2^K stocks indexed by $j \in \{0, 1, \dots, 2^K - 1\}$, and in each period $t \in \{0, 1\}$ the broker receives one unit order in each stock j. The broker allocates these orders across N wholesalers. We compare two routing policies: (i) *unbundling* (stock-by-stock allocation); and (ii) *bundling* (pooling all stocks into one order pool).

Multi-stock cost structure and the joint distribution of private costs

In the extension, if wholesaler i executes $x_{i,j,t}$ units of stock j in period t, its total inventory cost is

$$\sum_{j=0}^{2^K-1} \left(\zeta_{i,j,t}\, x_{i,j,t} + c\, x_{i,j,t}^2\right), \qquad \zeta_{i,j,1} = \zeta_{i,j,0} + \epsilon_{i,j}. \tag{IA61}$$

As in the baseline model, $\zeta_{i,j,0}$ becomes observable by the end of period 0, so the remaining uncertainty about period-1 costs is captured by $\epsilon_{i,j}$.

Cyclic Progression construction. Fix integers $S \geq K$ and let a satisfy $S-K \leq a \leq S$. For each wholesaler i, draw

$$W_i \sim \mathrm{Unif}\{0,1,\ldots,2^S-1\}, \qquad \theta_i \sim \mathrm{Unif}[0,1],$$

independently across i, and independently of each other. Define, for each stock index $j \in \{0,1,\ldots,2^K-1\}$,

$$\lambda_{i,j} \;:=\; \big(W_i + 2^a j\big) \bmod 2^S \;\in\; \{0,1,\ldots,2^S-1\}. \tag{IA62}$$

Finally define the private cost components by

$$\epsilon_{i,j} \;:=\; \Big(\theta_i + \lambda_{i,j}\Big)\frac{b}{2^S} - \frac{b}{2}. \tag{IA63}$$

The parameter a governs cross-stock dependence: larger a implies the progression in (IA62) cycles over fewer distinct values and hence generates stronger comovement across $\{\epsilon_{i,j}\}_j$.

Bundled type and average correlation. Define wholesaler i's *bundled* private cost by

$$\hat{\epsilon}_i \;:=\; \frac{1}{2^K}\sum_{j=0}^{2^K-1} \epsilon_{i,j},$$

and define the average pairwise cross-stock correlation (as in the main text) by

$$\bar{\rho}_\epsilon \;:=\; \frac{2}{2^K(2^K-1)} \sum_{0\leq j<k\leq 2^K-1} \mathrm{Corr}(\epsilon_{i,j},\epsilon_{i,k}).$$

Useful distributional facts under Cyclic Progression

Lemma IA6.3 (Marginals and bundled type under Cyclic Progression)**.** *Under* (IA62)–(IA63)*:*

1. *For every wholesaler i and stock j,* $\epsilon_{i,j} \sim \mathrm{Unif}[-b/2, b/2]$*.*
2. *Let* $L := 2^{S-a}$*. Then* $\hat{\epsilon}_i \sim \mathrm{Unif}\big[-\frac{b}{2L}, \frac{b}{2L}\big]$*.*

Proof. (1) Fix i and j. By construction, $\lambda_{i,j}$ is uniform on $\{0, 1, \ldots, 2^S - 1\}$, and independent of $\theta_i \sim \text{Unif}[0, 1]$. Conditional on $\lambda_{i,j} = \ell$, the random variable $(\theta_i + \ell)/2^S$ is uniform on the interval $[\ell/2^S, (\ell + 1)/2^S]$. Mixing uniformly over $\ell \in \{0, \ldots, 2^S - 1\}$ yields

$$\frac{\theta_i + \lambda_{i,j}}{2^S} \sim \text{Unif}[0, 1],$$

and the affine transformation (IA63) implies $\epsilon_{i,j} \sim \text{Unif}[-b/2, b/2]$.

(2) Write $W_i = s_i + 2^a r_i$ where $s_i \in \{0, \ldots, 2^a - 1\}$ collects the low a bits of W_i and $r_i \in \{0, \ldots, 2^{S-a} - 1\}$ collects the remaining bits. Adding $2^a j$ in (IA62) leaves the low a bits unchanged, so all $\lambda_{i,j}$ share the same s_i. Moreover, as j ranges over $\{0, \ldots, 2^K - 1\}$, the high-bit index cycles modulo $L = 2^{S-a}$ through each value in $\{0, \ldots, L - 1\}$ exactly $2^K/L$ times (since $S - K \leq a$ implies $L \leq 2^K$ and both are powers of two). Therefore,

$$\frac{1}{2^K} \sum_{j=0}^{2^K-1} \lambda_{i,j} \;=\; s_i \;+\; 2^a \cdot \frac{1}{L} \sum_{r=0}^{L-1} r \;=\; s_i \;+\; 2^a \cdot \frac{L-1}{2}.$$

Call the right-hand side m_i. As s_i is uniform on $\{0, \ldots, 2^a - 1\}$, m_i is uniform on a set of 2^a consecutive integers. Hence $(\theta_i + m_i)/2^S$ is uniform on an interval of length $2^a/2^S = 1/L$, centered at $1/2$. Applying (IA63) to the average yields

$$\hat{\epsilon}_i = \Big(\theta_i + m_i\Big) \frac{b}{2^S} - \frac{b}{2} \sim \text{Unif}\left[-\frac{b}{2L}, \frac{b}{2L}\right].$$

□

Lemma IA6.4 (Average cross-stock correlation under Cyclic Progression)**.** *Let* $L := 2^{S-a}$*. Under* (IA62)–(IA63)*, the average pairwise correlation is constant across wholesalers and equals*

$$\bar{\rho}_\epsilon \;=\; \frac{2^K/L^2 - 1}{2^K - 1} \;=\; \frac{2^{K+2a-2S} - 1}{2^K - 1}. \tag{IA64}$$

Equivalently,

$$\frac{1}{2^{S-a}} \;=\; \frac{1}{L} \;=\; \sqrt{\frac{1 + (2^K - 1)\bar{\rho}_\epsilon}{2^K}}. \tag{IA65}$$

Proof. Fix wholesaler i and suppress i. By Lemma IA6.3(1), $\text{Var}(\epsilon_j) = b^2/12$ for each j. By Lemma IA6.3(2), $\hat{\epsilon} = (1/2^K) \sum_{j=0}^{2^K-1} \epsilon_j$ is uniform on $[-b/(2L), b/(2L)]$, hence $\text{Var}(\hat{\epsilon}) =$

$b^2/(12L^2)$.

Using the variance decomposition for an average of identically distributed random variables,

$$\mathrm{Var}(\hat{\epsilon}) = \frac{1}{(2^K)^2}\left(2^K\,\mathrm{Var}(\epsilon_j) + 2^K(2^K-1)\,\mathrm{Var}(\epsilon_j)\,\bar{\rho}_\epsilon\right).$$

Substituting $\mathrm{Var}(\hat{\epsilon}) = b^2/(12L^2)$ and $\mathrm{Var}(\epsilon_j) = b^2/12$ and solving for $\bar{\rho}_\epsilon$ yields the first equality in (IA64); the second equality follows from $L = 2^{S-a}$. Rearranging the first equality gives (IA65). □

Proof of Lemma 6

Proof of Lemma 6. Let $M := 2^K$ denote the number of stocks. Write

$$\bar{\zeta}_{j,0} := \frac{1}{N}\sum_{i=1}^{N}\zeta_{i,j,0}, \qquad \zeta_i := \frac{1}{M}\sum_{j=0}^{M-1}\zeta_{i,j,0}, \qquad \bar{\zeta} := \frac{1}{N}\sum_{i=1}^{N}\zeta_i.$$

Let $\mathrm{Var}_j(\cdot)$ denote variance taken across stocks $j \in \{0,\ldots,M-1\}$.

Step 1: broker payoff under unbundling. Under unbundling, the broker solves M independent copies of the baseline problem, one per stock. For each stock j, the marginal distribution of $\epsilon_{i,j}$ matches the baseline (Lemma IA6.3(1)), so the optimal allocation and the implied expected broker payoff take the same functional form as in the baseline model, evaluated at $\{\zeta_{i,j,0}\}_{i=1}^{N}$ and b. Summing across stocks yields a total expected payoff of the form

$$\Pi^U = M\left(\frac{N-1}{c}\frac{b^2}{12} - c\cdot\frac{1}{M}\sum_{j=0}^{M-1}\left[N\left(\frac{\bar{\zeta}_{j,0}}{2c}+\frac{1}{N}\right)^2 - \sum_{i=1}^{N}\left(\frac{\zeta_{i,j,0}}{2c}\right)^2\right] - \left(\frac{b}{2}+\bar{s}\right)\right). \qquad \text{(IA66)}$$

Step 2: broker payoff under bundling. Under bundling, the broker pools all stocks and allocates the pooled order flow. The relevant private information for wholesaler i is its bundled type $\hat{\epsilon}_i$, whose distribution is uniform with tightened support $[-b/(2L), b/(2L)]$ where $L = 2^{S-a}$ (Lemma IA6.3(2)). Thus bundling is equivalent to a baseline problem with the same N, c, ex ante cost component ζ_i, and an effective information-asymmetry parameter

b/L. Therefore, the total expected payoff under bundling is

$$\Pi^B = M\left(\frac{N-1}{c}\frac{b^2}{12L^2} - c\cdot\left[N\Big(\frac{\bar{\zeta}}{2c}+\frac{1}{N}\Big)^2 - \sum_{i=1}^{N}\Big(\frac{\zeta_i}{2c}\Big)^2\right] - \Big(\frac{b}{2L}+\bar{s}\Big)\right). \tag{IA67}$$

Step 3: payoff comparison and the key inequality. Since (IA66)–(IA67) share the same factor $M>0$ and the same $-\bar{s}$ term, unbundling dominates bundling if and only if

$$\frac{N-1}{c}\frac{b^2}{12}\Big(1-\frac{1}{L^2}\Big)-\frac{b}{2}\Big(1-\frac{1}{L}\Big) \;\geq\; c\cdot\left(\frac{1}{M}\sum_{j=0}^{M-1}\left[N\Big(\frac{\bar{\zeta}_{j,0}}{2c}+\frac{1}{N}\Big)^2-\sum_{i=1}^{N}\Big(\frac{\zeta_{i,j,0}}{2c}\Big)^2\right]-\left[N\Big(\frac{\bar{\zeta}}{2c}+\frac{1}{N}\Big)^2-\sum_{i=1}^{N}\Big(\frac{\zeta_i}{2c}\Big)^2\right.\right. \tag{IA68}$$

Multiplying (IA68) by c and rearranging yields the equivalent condition

$$\frac{b}{2}\Big(1-\frac{1}{L}\Big)\left[(N-1)\frac{b}{6}\Big(1+\frac{1}{L}\Big)-c\right] \;\geq\; \Delta, \tag{IA69}$$

where

$$\Delta \;:=\; \frac{N}{4}\left(\mathrm{Var}_j(\bar{\zeta}_{j,0})-\frac{1}{N}\sum_{i=1}^{N}\mathrm{Var}_j(\zeta_{i,j,0})\right) \;\leq\; 0. \tag{IA610}$$

(The inequality $\Delta \leq 0$ follows because $\bar{\zeta}_{j,0} = (1/N)\sum_i \zeta_{i,j,0}$ and variance is convex, so averaging across i weakly reduces variance across j.)

Step 4: threshold comparative statics. Fix all primitives other than one of $\{c, N, a\}$ (equivalently, other than $\{c, N, \bar{\rho}_\epsilon\}$). Condition (IA69) characterizes unbundling vs bundling.

(i) Threshold in c. For fixed (N, L) and fixed Δ, the left-hand side of (IA69) is affine and strictly decreasing in c. Hence there exists a (possibly boundary) threshold c^* such that (IA69) holds if and only if $c < c^*$.

(ii) Threshold in N. Holding (c, L) and Δ fixed, the left-hand side of (IA69) is strictly increasing in N (it depends on N only through the factor $N-1$ inside the bracket). Hence there exists a (possibly boundary) threshold N^* such that (IA69) holds if and only if $N > N^*$.

(iii) Threshold in dependence (equivalently $\bar{\rho}_\epsilon$). Write $\kappa := 1/L = 2^{a-S} \in (0, 1]$. The

left-hand side of (IA69) can be written as

$$\frac{b}{2}(1-\kappa)\left[(N-1)\frac{b}{6}(1+\kappa)-c\right],$$

which is continuous in κ. Under the maintained interiority restrictions of the baseline model (which imply c is sufficiently large relative to $(N-1)b$), this expression is monotone increasing in κ on the admissible range, so higher dependence (larger κ) makes unbundling more attractive. By Lemma IA6.4, κ is strictly increasing in the average correlation $\bar{\rho}_\epsilon$ via (IA65). Therefore there exists a (possibly boundary) threshold ρ^* such that unbundling dominates bundling if and only if $\bar{\rho}_\epsilon > \rho^*$.

(iv) Threshold in b. The left-hand side of (IA69) is increasing in b. Hence there exists a threshold b^* such that (IA69) holds if and only if $b > b^*$.

This establishes the statements in Lemma 6. □

Appendix IA7. Omitted Proofs for Theoretical Framework

Proof of Proposition 1

We first rewrite payoffs in a form that makes clear what the broker can affect. Each wholesaler i's total payoff is

$$U_i \;=\; \sum_{t=0}^{1} x_{i,t}\big[s_{i,t}-(\zeta_{i,t}+cx_{i,t})\big].$$

Since period 1 is the final stage and wholesalers choose $s_{i,1}\le\bar{s}$, we have $s_{i,1}=\bar{s}$ for all i. In period 0, $(\vec{x}_0,\vec{\zeta}_0)$ is exogenous, so $\sum_i x_{i,0}[\bar{s}-(\zeta_{i,0}+cx_{i,0})]$ is a constant from the perspective of the mechanism design problem. Thus, up to an additive constant, wholesaler i's payoff can be written as

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

Similarly, the broker's payoff is

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

Because $\sum_i x_{i,0} = 1$ and $s_{i,1} = \bar{s}$, the broker's payoff simplifies to

$$V \;=\; \sum_{i=1}^{N} x_{i,0}(\bar{s} - s_{i,0}) \;-\; 2\bar{s}.$$

Hence the broker's objective is to maximize total period-0 price improvement, i.e. total transfers from wholesalers, up to the constant $-2\bar{s}$.

Step 1: Direct mechanism and interim objects. By the revelation principle (Myerson (1981)), we can restrict attention to direct revelation mechanisms. Each wholesaler i reports a type $\hat{\epsilon}_i \in [-b/2, b/2]$, and the broker chooses an allocation $\vec{X}_1(\hat{\vec{\epsilon}})$ and transfers $\vec{T}(\hat{\vec{\epsilon}})$, where $X_{i,1}$ is wholesaler i's period-1 order share and T_i is the transfer from wholesaler i to the broker.

Fix a mechanism $(\vec{X}_1, \vec{T})$. Define the interim (expected) allocation and transfer for wholesaler i by

$$\bar{X}_i(\hat{\epsilon}_i) \;:=\; \int X_{i,1}(\hat{\epsilon}_i, \vec{\epsilon}_{-i})\, dF_{-i}(\vec{\epsilon}_{-i}), \qquad \bar{T}_i(\hat{\epsilon}_i) \;:=\; \int T_i(\hat{\epsilon}_i, \vec{\epsilon}_{-i})\, dF_{-i}(\vec{\epsilon}_{-i}),$$

where F_{-i} is the joint distribution of $\vec{\epsilon}_{-i}$.

Using $\zeta_{i,1} = \zeta_{i,0} + \epsilon_i$, wholesaler i's interim expected payoff with true type ϵ_i and report $\hat{\epsilon}_i$ is

$$\begin{aligned} U_i(\hat{\epsilon}_i, \epsilon_i) &= \int \Big[X_{i,1}(\hat{\epsilon}_i, \vec{\epsilon}_{-i})\big(\bar{s} - (\zeta_{i,0} + \epsilon_i) - c\, X_{i,1}(\hat{\epsilon}_i, \vec{\epsilon}_{-i})\big) - T_i(\hat{\epsilon}_i, \vec{\epsilon}_{-i}) \Big]\, dF_{-i}(\vec{\epsilon}_{-i}) \\ &= \bar{X}_i(\hat{\epsilon}_i)\big(\bar{s} - \zeta_{i,0} - \epsilon_i\big) \;-\; c\,\mathbb{E}\big[X_{i,1}(\hat{\epsilon}_i, \vec{\epsilon}_{-i})^2 \mid \hat{\epsilon}_i\big] \;-\; \bar{T}_i(\hat{\epsilon}_i). \end{aligned}$$

Step 2: Incentive compatibility and the envelope formula. Let $U_i(\epsilon_i) := U_i(\epsilon_i, \epsilon_i)$ denote equilibrium interim utility under truthful reporting. Incentive compatibility (IC) requires

$$U_i(\epsilon_i) \;\geq\; U_i(\hat{\epsilon}_i, \epsilon_i) \qquad \forall \hat{\epsilon}_i \in [-b/2, b/2].$$

Because ϵ_i enters $U_i(\hat{\epsilon}_i, \epsilon_i)$ only through the linear term $-\epsilon_i \bar{X}_i(\hat{\epsilon}_i)$, the envelope theorem implies

$$U_i'(\epsilon_i) \;=\; -\bar{X}_i(\epsilon_i) \qquad \text{for a.e. } \epsilon_i \in [-b/2, b/2].$$

As standard in the literature, we also need the following implementability constraint

$$\bar{X}_i(\epsilon) \text{ is weakly decreasing in } \epsilon \in [-b/2, b/2].$$

Integrating from ϵ_i up to the highest cost type $b/2$ gives

$$U_i(\epsilon_i) \ = \ U_i(b/2) \ + \ \int_{\epsilon_i}^{b/2} \bar{X}_i(\tau)\, d\tau.$$

Imposing interim individual rationality (IR), $U_i(\epsilon_i) \geq 0$ for all ϵ_i, the broker weakly prefers to set the worst type's rent to zero (otherwise she can increase $\bar{T}_i$ by a constant without violating IC). Hence we take

$$U_i(b/2) = 0, \qquad \Rightarrow \qquad U_i(\epsilon_i) = \int_{\epsilon_i}^{b/2} \bar{X}_i(\tau)\, d\tau.$$

Step 3: Expected transfers and virtual costs. Under truthful reporting, rearranging the definition of $U_i(\epsilon_i)$ yields

$$\bar{T}_i(\epsilon_i) = \bar{X}_i(\epsilon_i)\big(\bar{s} - \zeta_{i,0} - \epsilon_i\big) - c\,\mathbb{E}\big[X_{i,1}(\vec{\epsilon})^2 \mid \epsilon_i\big] - U_i(\epsilon_i).$$

Taking expectations over ϵ_i and using the law of iterated expectations, $\mathbb{E}[\mathbb{E}[X_{i,1}(\vec{\epsilon})^2 \mid \epsilon_i]] = \mathbb{E}[X_{i,1}(\vec{\epsilon})^2]$. Moreover, using $U_i(b/2) = 0$ and Fubini's theorem,

$$\mathbb{E}[U_i(\epsilon_i)] = \int_{-b/2}^{b/2} \left(\int_{\epsilon_i}^{b/2} \bar{X}_i(\tau)\, d\tau \right) f(\epsilon_i)\, d\epsilon_i = \int_{-b/2}^{b/2} \bar{X}_i(\tau) F(\tau)\, d\tau = \mathbb{E}\left[\bar{X}_i(\epsilon_i) \frac{F(\epsilon_i)}{f(\epsilon_i)} \right].$$

Therefore,

$$\mathbb{E}[\bar{T}_i(\epsilon_i)] = \mathbb{E}\left[X_{i,1}(\vec{\epsilon}) \left(\bar{s} - \zeta_{i,0} - \epsilon_i - \frac{F(\epsilon_i)}{f(\epsilon_i)} \right) - c\, X_{i,1}(\vec{\epsilon})^2 \right].$$

Step 4: Pointwise maximization. The broker's expected payoff is $\mathbb{E}\big[\sum_i T_i(\vec{\epsilon})\big] - 2\bar{s}$. Combining the expression above across i gives

$$\max_{\vec{X}_1(\cdot)} \ \mathbb{E}\left[\sum_{i=1}^{N} \left\{ X_{i,1}(\vec{\epsilon}) \left(\bar{s} - \zeta_{i,0} - \epsilon_i - \frac{F(\epsilon_i)}{f(\epsilon_i)} \right) - c\, X_{i,1}(\vec{\epsilon})^2 \right\} \right] - 2\bar{s},$$

subject to feasibility $X_{i,1}(\vec{\epsilon}) \geq 0$ for all i and $\sum_i X_{i,1}(\vec{\epsilon}) \leq 1$ for all $\vec{\epsilon}$.

Because the expectation is linear and the feasibility constraints hold pointwise in $\vec{\epsilon}$, the problem decomposes: for each realization $\vec{\epsilon}$ the optimal allocation $\vec{X}_1^*(\vec{\epsilon})$ solves the corresponding pointwise maximization problem.

Finally, since $\epsilon_i \sim U[-b/2, b/2]$, we have $\frac{F(\epsilon_i)}{f(\epsilon_i)} = \epsilon_i + \frac{b}{2}$. Substituting into the pointwise objective and rewriting yields exactly Equation 3 in Proposition 1:

$$\max_{\vec{X}_1} \sum_{i=1}^{N} \left[X_{i,1} \Big(\bar{s} - (\zeta_{i,0} + \epsilon_i + cX_{i,1}) - (\epsilon_i + \tfrac{b}{2}) \Big) \right] - 2\bar{s},$$

subject to $X_{i,1} \geq 0$ and $\sum_i X_{i,1} \leq 1$.

The last step is to check the monotonicity constraint. The problem is concave and the feasible set does not depend on $\vec{\epsilon}$. Moreover, for each i, ϵ_i appears only in $-2\epsilon_i X_{i,1}$, so the objective has decreasing differences in $(X_{i,1}, \epsilon_i)$. Therefore $X_{i,1}^*(\epsilon_i, \vec{\epsilon}_{-i})$ is weakly decreasing in ϵ_i, and hence the interim allocation $\bar{X}_i(\epsilon_i)$ is weakly decreasing as well. Thus the IC monotonicity constraint is automatically satisfied.

Proof of Proposition 2

Throughout, fix public $(\vec{\zeta}_0, \bar{s}, b, c)$ and a realization of types $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_N) \in [-\frac{b}{2}, \frac{b}{2}]^N$. Recall that $\zeta_{1,0} = \max_i \zeta_{i,0}$ and $\bar{\zeta}_0 = \frac{1}{N} \sum_{i=1}^{N} \zeta_{i,0}$.

Step 1: Solve for the optimal allocation rule. From Proposition 1, the broker's period-1 allocation rule in the optimal direct mechanism solves, pointwise in $\vec{\epsilon}$,

$$\begin{aligned} \max_{\{X_{i,1}\}_{i=1}^N} \quad & \sum_{i=1}^{N} \Big(X_{i,1} \big(\bar{s} - \zeta_{i,0} - 2\epsilon_i - \tfrac{b}{2} \big) - cX_{i,1}^2 \Big) \ - \ 2\bar{s} \\ \text{s.t.} \quad & X_{i,1} \geq 0 \ \forall i, \qquad \sum_{i=1}^{N} X_{i,1} \leq 1. \end{aligned} \tag{IA71}$$

Dropping the constant $-2\bar{s}$, the objective is strictly concave in $(X_{1,1}, \ldots, X_{N,1})$, so KKT conditions are sufficient.

Let $\mu \geq 0$ be the multiplier on $\sum_i X_{i,1} \leq 1$ and $\lambda_i \geq 0$ be the multiplier on $X_{i,1} \geq 0$.

The Lagrangian is

$$\mathcal{L} = \sum_{i=1}^{N} \Big(X_{i,1}\big(\bar{s} - \zeta_{i,0} - 2\epsilon_i - \tfrac{b}{2}\big) - cX_{i,1}^2 \Big) + \mu\Big(1 - \sum_{i=1}^{N} X_{i,1}\Big) + \sum_{i=1}^{N} \lambda_i X_{i,1}.$$

The KKT stationarity condition for each i is

$$\bar{s} - \zeta_{i,0} - 2\epsilon_i - \tfrac{b}{2} \; - \; 2cX_{i,1} \; - \; \mu \; + \; \lambda_i \; = \; 0. \tag{IA72}$$

We now use Assumption 1 to verify that the solution is interior: $X_{i,1}^*(\vec{\epsilon}) > 0$ for all i and $\sum_i X_{i,1}^*(\vec{\epsilon}) = 1$ for all $\vec{\epsilon}$. Hence $\lambda_i = 0$ for all i and $\mu > 0$. Under these conditions, (IA72) becomes

$$2cX_{i,1}^*(\vec{\epsilon}) = \bar{s} - \zeta_{i,0} - 2\epsilon_i - \tfrac{b}{2} - \mu.$$

Summing over i and imposing $\sum_i X_{i,1}^*(\vec{\epsilon}) = 1$ yields

$$2c = \sum_{i=1}^{N} \Big(\bar{s} - \zeta_{i,0} - 2\epsilon_i - \tfrac{b}{2} - \mu \Big) = N\Big(\bar{s} - \bar{\zeta}_0 - \tfrac{b}{2} - \mu \Big) - 2N\bar{\epsilon},$$

where $\bar{\epsilon} = \frac{1}{N}\sum_{i=1}^{N} \epsilon_i$. Therefore

$$\mu = \bar{s} - \bar{\zeta}_0 - \tfrac{b}{2} - 2\bar{\epsilon} - \frac{2c}{N}. \tag{IA73}$$

Substituting (IA73) back gives

$$\begin{aligned} X_{i,1}^*(\vec{\epsilon}) &= \frac{1}{2c}\Big(\bar{s} - \zeta_{i,0} - 2\epsilon_i - \tfrac{b}{2} - \mu \Big) \\ &= \frac{1}{2c}\Big(\bar{\zeta}_0 - \zeta_{i,0} + \frac{2c}{N} + 2\bar{\epsilon} - 2\epsilon_i \Big) \\ &= \underbrace{\Big(\frac{1}{N} + \frac{\bar{\zeta}_0 - \zeta_{i,0}}{2c} \Big)}_{=:A_i} + \frac{1}{c}\Big(\bar{\epsilon} - \epsilon_i \Big), \end{aligned} \tag{IA74}$$

which is exactly (IA74).

Step 2: Verify interiority under Assumption 1. First, we show $X_{i,1}^*(\vec{\epsilon}) > 0$ for all i and all $\vec{\epsilon}$. The smallest value of $X_{i,1}^*$ occurs for the wholesaler with the largest public cost,

$i = 1$, and for the most unfavorable type configuration for $X^*_{1,1}$, namely $\epsilon_1 = \frac{b}{2}$ and $\epsilon_j = -\frac{b}{2}$ for all $j \neq 1$, which implies $\bar{\epsilon} - \epsilon_1 = -\frac{N-1}{N} b$. Thus

$$\min_{\vec{\epsilon}} X^*_{1,1}(\vec{\epsilon}) = \frac{1}{N} + \frac{\bar{\zeta}_0 - \zeta_{1,0}}{2c} - \frac{N-1}{Nc} b.$$

This is strictly positive if and only if

$$c > \frac{N}{2}(\zeta_{1,0} - \bar{\zeta}_0) + (N-1)b,$$

which is exactly the lower bound in Assumption 1. Hence $X^*_{i,1} > 0$ for all i.

Second, we show the capacity constraint binds, i.e. $\sum_i X^*_{i,1}(\vec{\epsilon}) = 1$. From (IA73), the smallest value of μ arises at the largest $\bar{\epsilon}$, i.e. $\bar{\epsilon} = \frac{b}{2}$, which yields

$$\min_{\vec{\epsilon}} \mu = \bar{s} - \bar{\zeta}_0 - \tfrac{b}{2} - 2 \cdot \tfrac{b}{2} - \frac{2c}{N} = \bar{s} - \bar{\zeta}_0 - \frac{3b}{2} - \frac{2c}{N}.$$

The upper bound in Assumption 1 is precisely $c < \frac{N}{2}\left(\bar{s} - \bar{\zeta}_0 - \frac{3b}{2}\right)$, which implies $\min_{\vec{\epsilon}} \mu > 0$. Therefore the constraint binds for all $\vec{\epsilon}$, so $\sum_i X^*_{i,1}(\vec{\epsilon}) = 1$.

Step 3: Derive the (interim) transfer schedule. Fix wholesaler i. Under the direct mechanism, conditional on type ϵ_i, his interim (expected) period-1 allocation is

$$\begin{aligned} \bar{X}_{i,1}(\epsilon_i) := \mathbb{E}_{\epsilon_{-i}}\Big[X^*_{i,1}(\epsilon_i, \epsilon_{-i})\Big] &= A_i + \frac{1}{c}\Big(\mathbb{E}[\bar{\epsilon} \mid \epsilon_i] - \epsilon_i\Big) \\ &= A_i + \frac{1}{c}\Big(\frac{\epsilon_i}{N} - \epsilon_i\Big) = A_i - \frac{N-1}{Nc}\epsilon_i, \end{aligned} \tag{IA75}$$

since $\mathbb{E}[\epsilon_j] = 0$ for $j \neq i$.

Let $U_i(\epsilon_i)$ denote wholesaler i's interim utility *net of constants that do not depend on his report* (equivalently, focusing on the period-0 price improvement and the period-1 continuation payoff). As usual in one-dimensional quasi-linear environments, Bayesian incentive compatibility implies the envelope condition

$$\frac{d}{d\epsilon_i} U_i(\epsilon_i) = -\bar{X}_{i,1}(\epsilon_i). \tag{IA76}$$

The broker minimizes information rents by setting the interim utility of the worst type $\epsilon_i = \frac{b}{2}$

to zero, i.e. $U_i(\frac{b}{2}) = 0$. Integrating (IA76) gives

$$U_i(\epsilon_i) = \int_{\epsilon_i}^{b/2} \bar{X}_{i,1}(\tau)\, d\tau = \int_{\epsilon_i}^{b/2} \Big(A_i - \frac{N-1}{Nc}\tau\Big)\, d\tau$$
$$= A_i\Big(\frac{b}{2} - \epsilon_i\Big) - \frac{N-1}{2Nc}\Big(\frac{b^2}{4} - \epsilon_i^2\Big). \tag{IA77}$$

Let $T_i^*(\epsilon_i)$ denote the interim expected transfer *from wholesaler i to the broker* under the optimal direct mechanism. By definition of utility,

$$U_i(\epsilon_i) = \mathbb{E}_{\epsilon_{-i}}\Big[X_{i,1}^*(\vec{\epsilon})\big(\bar{s} - \zeta_{i,0} - \epsilon_i - cX_{i,1}^*(\vec{\epsilon})\big)\Big] \;-\; T_i^*(\epsilon_i), \tag{IA78}$$

so

$$T_i^*(\epsilon_i) = \big(\bar{s} - \zeta_{i,0} - \epsilon_i\big)\bar{X}_{i,1}(\epsilon_i) - c\,\mathbb{E}_{\epsilon_{-i}}\Big[\big(X_{i,1}^*(\vec{\epsilon})\big)^2\Big] - U_i(\epsilon_i). \tag{IA79}$$

To compute $\mathbb{E}[(X_{i,1}^*)^2 \mid \epsilon_i]$, write (from (IA74))

$$X_{i,1}^*(\vec{\epsilon}) = \bar{X}_{i,1}(\epsilon_i) + \frac{1}{Nc}\sum_{j\neq i}\epsilon_j.$$

Because $\{\epsilon_j\}_{j\neq i}$ are independent and mean zero with $\mathrm{Var}(\epsilon_j) = b^2/12$, we obtain

$$\mathbb{E}_{\epsilon_{-i}}\Big[\big(X_{i,1}^*(\vec{\epsilon})\big)^2\Big] = \bar{X}_{i,1}(\epsilon_i)^2 + \frac{1}{N^2c^2}\mathrm{Var}\Big(\sum_{j\neq i}\epsilon_j\Big)$$
$$= \bar{X}_{i,1}(\epsilon_i)^2 + \frac{N-1}{N^2c^2}\cdot\frac{b^2}{12}. \tag{IA710}$$

Plugging (IA75), (IA77), and (IA710) into (IA79) and simplifying yields the closed form:

$$T_i^*(\epsilon_i) = \Big[A_i(\bar{s} - \zeta_{i,0}) - cA_i^2 - \frac{(N-1)b^2}{12N^2c} - \frac{A_ib}{2} + \frac{(N-1)b^2}{8Nc}\Big]$$
$$+ \Big[2A_i - \frac{1}{c}(\bar{s} - \zeta_{i,0})\Big]\frac{N-1}{N}\epsilon_i - \Big(\frac{N^2-3N+2}{2N^2c}\Big)\epsilon_i^2, \tag{IA711}$$

which matches (IA711) and completes the proof.

Proof of Proposition 3

Fix the optimal direct mechanism from Proposition 2, with allocation rule $\vec{X}_1^*(\vec{\epsilon})$ and (interim) transfer schedules $\{T_i^*(\epsilon_i)\}_{i=1}^N$.

Step 1: Transfers can be implemented by period-0 price improvement. Given the exogenous period-0 allocation $x_{i,0}$, define wholesaler i's *total* period-0 price improvement (i.e. the transfer to the broker) as

$$t_i := (\bar{s} - s_{i,0})x_{i,0}.$$

Because $x_{i,0}$ is fixed and publicly observed, choosing $s_{i,0}$ is equivalent to choosing t_i (via $s_{i,0} = \bar{s} - t_i/x_{i,0}$). Thus the broker can treat the observed t_i as a monetary transfer.

Step 2: Invertibility. Under Assumption 1, $T_i^*(\cdot)$ is strictly monotone on $[-\frac{b}{2}, \frac{b}{2}]$ and hence invertible.[37] Let $(T_i^*)^{-1}$ denote its inverse.

Step 3: Define the performance-based allocation rule. Consider the following period-1 routing rule based on the observed period-0 execution prices $\vec{s}_0 = (s_{1,0}, \dots, s_{N,0})$:

$$\vec{x}_1^*(\vec{s}_0) = \vec{X}_1^*\big((T_1^*)^{-1}\big((\bar{s} - s_{1,0})x_{1,0}\big), \dots, (T_N^*)^{-1}\big((\bar{s} - s_{N,0})x_{N,0}\big)\big)\,,$$

which is exactly the rule stated in the proposition.

Step 4: Incentives coincide with the direct mechanism. Fix wholesaler i with type ϵ_i. Given any choice of period-0 price $s_{i,0}$ (equivalently, any transfer t_i), the broker forms the "implied report"

$$\hat{\epsilon}_i := (T_i^*)^{-1}(t_i) = (T_i^*)^{-1}\big((\bar{s} - s_{i,0})x_{i,0}\big),$$

and then assigns period-1 order flow according to $\vec{X}_1^*(\hat{\epsilon}_i, \hat{\epsilon}_{-i})$. Wholesaler i's period-0 payoff contribution is $-t_i$ (up to constants), and his period-1 payoff contribution is

$$X_{i,1}^*(\hat{\epsilon}_i, \hat{\epsilon}_{-i})\Big(\bar{s} - (\zeta_{i,0} + \epsilon_i + cX_{i,1}^*(\hat{\epsilon}_i, \hat{\epsilon}_{-i}))\Big),$$

since in period 1 (the terminal period) wholesalers optimally set $s_{i,1} = \bar{s}$.

[37] One way to see this is to differentiate (IA711) and use the upper bound in Assumption 1 to ensure the derivative is strictly negative over the whole type support.

Therefore, for any deviation that induces some $\hat{\epsilon}_i$, wholesaler i's expected payoff in the original game (under the rule $\vec{x}_1^*$) coincides with his expected payoff in the direct mechanism if he were to *report* $\hat{\epsilon}_i$ and make transfer $T_i^*(\hat{\epsilon}_i) = t_i$. Since the direct mechanism is incentive compatible, truthful reporting is optimal: the best response is $\hat{\epsilon}_i = \epsilon_i$, equivalently $t_i = T_i^*(\epsilon_i)$.

Step 5: Implementation. When each wholesaler chooses $t_i = T_i^*(\epsilon_i)$ in period 0 (equivalently, $s_{i,0} = \bar{s} - T_i^*(\epsilon_i)/x_{i,0}$), the broker recovers $\hat{\epsilon}_i = \epsilon_i$ for all i and assigns $\vec{x}_1^*(\vec{s}_0) = \vec{X}_1^*(\vec{\epsilon})$. Thus the induced equilibrium allocation coincides with the optimal direct mechanism's allocation, and the period-0 price improvements coincide with the direct mechanism's transfers. Hence the performance-based rule implements the optimal direct mechanism.

Proof of Lemma 1

Fix a wholesaler i and hold $(s_{j,0})_{j \neq i}$ fixed. Under the performance-based implementation in Proposition 3, define wholesaler j's *observed* period-0 total price improvement (i.e., the implemented transfer) by

$$t_j(\vec{s}_0) := (\bar{s} - s_{j,0})x_{j,0}.$$

The induced (reported/implied) type is

$$\hat{\epsilon}_j(\vec{s}_0) := (T_j^*)^{-1}\big(t_j(\vec{s}_0)\big),$$

and the implemented period-1 allocation satisfies

$$x_{i,1}^*(\vec{s}_0) = X_{i,1}^*\big(\hat{\epsilon}_1(\vec{s}_0), \ldots, \hat{\epsilon}_N(\vec{s}_0)\big).$$

Since only $t_i(\vec{s}_0)$ depends on $s_{i,0}$, only $\hat{\epsilon}_i(\vec{s}_0)$ varies with $s_{i,0}$, and hence by the chain rule,

$$\frac{dx_{i,1}^*(\vec{s}_0)}{ds_{i,0}} = \frac{\partial X_{i,1}^*(\vec{\epsilon})}{\partial \epsilon_i}\bigg|_{\vec{\epsilon}=\hat{\vec{\epsilon}}(\vec{s}_0)} \cdot \frac{d\hat{\epsilon}_i(\vec{s}_0)}{ds_{i,0}}. \tag{IA712}$$

Step 1: $\partial X^*_{i,1}/\partial \epsilon_i < 0$. From Equation 4,

$$X^*_{i,1}(\vec{\epsilon}) = A_i + \frac{1}{c}\Big(\bar{\epsilon} - \epsilon_i\Big), \qquad \bar{\epsilon} := \frac{1}{N}\sum_{j=1}^{N} \epsilon_j.$$

Therefore, holding ϵ_{-i} fixed,

$$\frac{\partial X^*_{i,1}(\vec{\epsilon})}{\partial \epsilon_i} = \frac{1}{c}\Big(\frac{1}{N} - 1\Big) = -\frac{N-1}{Nc} < 0.$$

Step 2: $d\hat{\epsilon}_i/ds_{i,0} > 0$. Differentiate $t_i(\vec{s}_0) = (\bar{s} - s_{i,0})x_{i,0}$:

$$\frac{dt_i(\vec{s}_0)}{ds_{i,0}} = -x_{i,0} < 0,$$

where $x_{i,0} > 0$ by assumption (the broker routes a positive period-0 share to each wholesaler). Using $\hat{\epsilon}_i(\vec{s}_0) = (T^*_i)^{-1}(t_i(\vec{s}_0))$,

$$\frac{d\hat{\epsilon}_i(\vec{s}_0)}{ds_{i,0}} = \big((T^*_i)^{-1}\big)'\big(t_i(\vec{s}_0)\big) \cdot \frac{dt_i(\vec{s}_0)}{ds_{i,0}}. \tag{IA713}$$

Thus it suffices to show that $\big((T^*_i)^{-1}\big)' < 0$, i.e., T^*_i is strictly decreasing.

From Equation 5, T^*_i is a quadratic in ϵ_i:

$$T^*_i(\epsilon_i) = K_i + L\,\epsilon_i - Q\,\epsilon_i^2,$$

where K_i collects the terms constant in ϵ_i,

$$L = \frac{N-1}{N}\Big(2A_i - \frac{\bar{s} - \zeta_{i,0}}{c}\Big) = \frac{N-1}{N}\Big(\frac{\bar{\zeta}_0 - \bar{s}}{c} + \frac{2}{N}\Big), \qquad Q = \frac{N^2 - 3N + 2}{2N^2 c} \geq 0,$$

and $\bar{\zeta}_0 := \frac{1}{N}\sum_{j=1}^{N} \zeta_{j,0}$. Hence

$$\frac{dT^*_i(\epsilon_i)}{d\epsilon_i} = L - 2Q\,\epsilon_i.$$

Over $\epsilon_i \in [-b/2, b/2]$, this derivative is maximized at $\epsilon_i = -b/2$, so

$$\max_{\epsilon_i \in [-b/2, b/2]} \frac{dT_i^*(\epsilon_i)}{d\epsilon_i} = L + Qb.$$

Using the *upper bound* in Assumption 1,

$$c < \frac{N}{2}\Big(\bar{s} - \bar{\zeta}_0 - \frac{3}{2}b\Big) \quad \Longrightarrow \quad \frac{\bar{\zeta}_0 - \bar{s}}{c} + \frac{2}{N} < -\frac{3b}{2c},$$

so

$$L < -\frac{N-1}{N} \cdot \frac{3b}{2c}.$$

Therefore,

$$L + Qb < -\frac{N-1}{N} \cdot \frac{3b}{2c} + \frac{N^2 - 3N + 2}{2N^2 c} b = -\frac{(N-1)(N+1)}{N^2 c} b < 0.$$

Hence $dT_i^*(\epsilon_i)/d\epsilon_i < 0$ for all $\epsilon_i \in [-b/2, b/2]$, so T_i^* is strictly decreasing and invertible, and thus $\big((T_i^*)^{-1}\big)'(t) < 0$ for all t in its range. Combining this with $dt_i/ds_{i,0} = -x_{i,0} < 0$ in (IA713) yields

$$\frac{d\hat{\epsilon}_i(\vec{s}_0)}{ds_{i,0}} > 0.$$

Step 3: Conclude $dx_{i,1}^*/ds_{i,0} < 0$. Plugging Step 1 and Step 2 into (IA712) implies

$$\frac{dx_{i,1}^*(\vec{s}_0)}{ds_{i,0}} < 0,$$

which proves the lemma.

Proof of Lemma 2

Fix wholesaler i. In the implementation described in the main text, period-0 performance is summarized by a transfer (price-improvement payment)

$$t_i \;=\; x_{i,0}\big(\bar{s} - s_{i,0}\big),$$

and the broker uses t_i to infer wholesaler i's private cost component ϵ_i. Write the equilibrium transfer schedule as $\bar{t}_i(\epsilon_i)$, which is strictly decreasing in ϵ_i under the maintained interiority conditions, hence invertible.

The period-1 allocation rule is linear in types:

$$x^*_{i,1}(\epsilon) \;=\; A_i + \frac{1}{c}\big(\bar{\epsilon} - \epsilon_i\big), \qquad \bar{\epsilon} := \frac{1}{N}\sum_{r=1}^{N} \epsilon_r,$$

so $\partial x^*_{i,1}/\partial \epsilon_i = -(N-1)/(Nc)$. By the chain rule,

$$\frac{dx^*_{i,1}}{dt_i} \;=\; \frac{\partial x^*_{i,1}}{\partial \epsilon_i} \cdot \frac{d\epsilon_i}{dt_i} \;=\; -\frac{N-1}{Nc} \cdot \frac{1}{\bar{t}'_i(\epsilon_i)}.$$

From the IC/envelope characterization of the transfer schedule in the baseline model, $\bar{t}'_i(\epsilon_i)$ is affine in ϵ_i and can be written (up to a common multiplicative constant that cancels above) as

$$\bar{t}'_i(\epsilon_i) \;=\; -\frac{N-1}{Nc}\Big[\Big(1 - \frac{2}{N}\Big)\epsilon_i + \big(\bar{s} - \bar{\zeta}_0\big) - \frac{2c}{N}\Big], \qquad \bar{\zeta}_0 := \frac{1}{N}\sum_{r=1}^{N} \zeta_{r,0}.$$

Substituting yields

$$\frac{dx^*_{i,1}}{dt_i} \;=\; \frac{1}{\Big(1 - \frac{2}{N}\Big)\epsilon_i + \big(\bar{s} - \bar{\zeta}_0\big) - \frac{2c}{N}} \;>\; 0,$$

where strict positivity holds under the maintained parameter restriction ensuring the denominator is positive for all $\epsilon_i \in [-b/2, b/2]$.

Because $t_i = x_{i,0}(\bar{s} - s_{i,0})$, we have

$$\frac{dx^*_{i,1}}{ds_{i,0}} \;=\; \frac{dx^*_{i,1}}{dt_i} \cdot \frac{dt_i}{ds_{i,0}} \;=\; -x_{i,0}\frac{dx^*_{i,1}}{dt_i} \;<\; 0,$$

so the (absolute) sensitivity satisfies

$$\Big|\frac{dx^*_{i,1}}{ds_{i,0}}\Big| \;=\; x_{i,0}\frac{dx^*_{i,1}}{dt_i}.$$

Taking expectations over $\epsilon_i \sim \text{Unif}[-b/2, b/2]$ gives

$$\mathbb{E}\left[\frac{dx_{i,1}^*}{dt_i}\right] = \int_{-b/2}^{b/2} \frac{1}{b} \cdot \frac{1}{\left(1-\frac{2}{N}\right)\epsilon + \left(\bar{s} - \bar{\zeta}_0\right) - \frac{2c}{N}} \, d\epsilon.$$

The integrand is a *positive, decreasing, convex* function of ϵ (it is the reciprocal of an affine function that stays positive on the support). Increasing b induces a mean-preserving spread of ϵ_i, and by Jensen's inequality the expectation of any convex function of ϵ_i weakly increases with b. Hence $\mathbb{E}\left[dx_{i,1}^*/dt_i\right]$ is increasing in b, and therefore

$$\mathbb{E}\left[\left|\frac{dx_{i,1}^*}{ds_{i,0}}\right|\right] = x_{i,0}\, \mathbb{E}\left[\frac{dx_{i,1}^*}{dt_i}\right]$$

is increasing in b as well. (Equivalently, since $dx_{i,1}^*/ds_{i,0} < 0$ pointwise, the derivative becomes more negative as b increases.)

Proof of Lemma 3

Recall that, in the Theoretical Framework, the optimal transfer can be written as

$$T_i^*(\epsilon_i) = K_i(b) + \alpha_i \epsilon_i + \delta \epsilon_i^2,$$

where $K_i(b)$ is constant with respect to ϵ_i,

$$\alpha_i = \left[2A_i - \frac{1}{c}\left(\bar{s} - \zeta_{i,0}\right)\right] \frac{N-1}{N},$$

and

$$\delta = -\frac{N^2 - 3N + 2}{2N^2 c}.$$

Also, $\epsilon_i \sim \text{Unif}\left[-\frac{b}{2}, \frac{b}{2}\right]$.

Since $K_i(b)$ does not vary with ϵ_i, it does not affect the variance. Hence,

$$\text{Var}(T_i^*(\epsilon_i)) = \text{Var}\left(\alpha_i \epsilon_i + \delta \epsilon_i^2\right).$$

Let

$$Y = \alpha_i \epsilon_i + \delta \epsilon_i^2.$$

Then

$$\mathrm{Var}(T_i^*(\epsilon_i)) = \mathrm{Var}(Y).$$

We first compute the relevant moments of ϵ_i. Because $\epsilon_i \sim \mathrm{Unif}\left[-\frac{b}{2}, \frac{b}{2}\right]$,

$$\mathbb{E}[\epsilon_i] = 0,$$

and

$$\mathbb{E}[\epsilon_i^2] = \frac{1}{b}\int_{-b/2}^{b/2} x^2\,dx = \frac{1}{b}\left[\frac{x^3}{3}\right]_{-b/2}^{b/2} = \frac{1}{b} \cdot \frac{2(b/2)^3}{3} = \frac{b^2}{12}.$$

By symmetry,

$$\mathbb{E}[\epsilon_i^3] = 0.$$

Similarly,

$$\mathbb{E}[\epsilon_i^4] = \frac{1}{b}\int_{-b/2}^{b/2} x^4\,dx = \frac{1}{b}\left[\frac{x^5}{5}\right]_{-b/2}^{b/2} = \frac{1}{b} \cdot \frac{2(b/2)^5}{5} = \frac{b^4}{80}.$$

Next,

$$\mathbb{E}[Y] = \alpha_i \mathbb{E}[\epsilon_i] + \delta \mathbb{E}[\epsilon_i^2] = \delta \frac{b^2}{12}.$$

Moreover,

$$Y^2 = \alpha_i^2 \epsilon_i^2 + 2\alpha_i \delta \epsilon_i^3 + \delta^2 \epsilon_i^4.$$

Taking expectations gives

$$\mathbb{E}[Y^2] = \alpha_i^2 \mathbb{E}[\epsilon_i^2] + 2\alpha_i \delta \mathbb{E}[\epsilon_i^3] + \delta^2 \mathbb{E}[\epsilon_i^4] = \alpha_i^2 \frac{b^2}{12} + \delta^2 \frac{b^4}{80}.$$

Therefore,

$$\begin{aligned}\operatorname{Var}(Y) &= \mathbb{E}[Y^2] - \big(\mathbb{E}[Y]\big)^2 \\ &= \alpha_i^2 \frac{b^2}{12} + \delta^2 \frac{b^4}{80} - \left(\delta \frac{b^2}{12}\right)^2 \\ &= \alpha_i^2 \frac{b^2}{12} + \delta^2 b^4 \left(\frac{1}{80} - \frac{1}{144}\right).\end{aligned}$$

Since

$$\frac{1}{80} - \frac{1}{144} = \frac{9-5}{720} = \frac{1}{180},$$

we obtain

$$\operatorname{Var}(T_i^*(\epsilon_i)) = \frac{\alpha_i^2}{12} b^2 + \frac{\delta^2}{180} b^4.$$

Finally, differentiate with respect to b:

$$\frac{d}{db}\operatorname{Var}(T_i^*(\epsilon_i)) = \frac{\alpha_i^2}{6} b + \frac{\delta^2}{45} b^3.$$

For every $b > 0$, the right hand side is weakly positive, and it is strictly positive whenever $\alpha_i \neq 0$ or $\delta \neq 0$. In particular, in the present model $\delta^2 \geq 0$, so

$$\operatorname{Var}(T_i^*(\epsilon_i))$$

is increasing in b, and strictly increasing for $b > 0$ whenever the transfer rule is nondegenerate.

Proof of Lemma 4

We prove the two parts in turn.

Part (1). Recall that the broker maximizes customers' execution quality, and the broker payoff V in Equation 1 is the negative of the (order-flow weighted) execution price. Hence, *a higher optimal broker payoff is equivalent to a lower average execution price.*

Consider the entry model with wholesalers $i \in \{1, \ldots, N+1\}$. As in Proposition 1, for any (feasible) period-1 allocation policy $\vec{X}_1(\cdot)$ satisfying $X_{i,1}(\vec{\epsilon}) \geq 0$ for all i and $\sum_{i=1}^{N+1} X_{i,1}(\vec{\epsilon}) \leq 1$ for all $\vec{\epsilon}$, define the broker's objective value (the pointwise objective in Proposition 1, integrated over types) by

$$\mathcal{V}(\zeta_{N+1,0};\vec{X}_1) := \mathbb{E}_{\vec{\epsilon}}\left[\sum_{i=1}^{N+1} X_{i,1}(\vec{\epsilon})\Big(\bar{s} - (\zeta_{i,0} + \epsilon_i + cX_{i,1}(\vec{\epsilon})) - (\epsilon_i + \tfrac{b}{2})\Big)\right] - 2\bar{s}.$$

Let $\mathcal{V}^*(\zeta_{N+1,0}) := \max_{\vec{X}_1} \mathcal{V}(\zeta_{N+1,0};\vec{X}_1)$.

Monotonicity in $\zeta_{N+1,0}$. For any fixed feasible $\vec{X}_1$, the parameter $\zeta_{N+1,0}$ enters $\mathcal{V}$ only through the term $-\zeta_{N+1,0}\, X_{N+1,1}(\vec{\epsilon})$. Therefore, for any $\zeta' < \zeta$,

$$\mathcal{V}(\zeta';\vec{X}_1) = \mathcal{V}(\zeta;\vec{X}_1) + (\zeta-\zeta')\,\mathbb{E}_{\vec{\epsilon}}\big[X_{N+1,1}(\vec{\epsilon})\big] \;\; \geq \;\; \mathcal{V}(\zeta;\vec{X}_1),$$

since $X_{N+1,1}(\vec{\epsilon}) \geq 0$. Taking $\vec{X}_1 = \vec{X}_1^*(\zeta)$ (an optimizer at ζ) gives

$$\mathcal{V}^*(\zeta') \;\geq\; \mathcal{V}(\zeta';\vec{X}_1^*(\zeta)) \;\geq\; \mathcal{V}(\zeta;\vec{X}_1^*(\zeta)) \;=\; \mathcal{V}^*(\zeta),$$

so $\mathcal{V}^*(\zeta_{N+1,0})$ is weakly decreasing in $\zeta_{N+1,0}$. Equivalently, the average execution price is weakly increasing in $\zeta_{N+1,0}$. Under Assumption 2, the optimal allocation is interior (in particular $X^*_{N+1,1}(\vec{\epsilon}) > 0$ for all $\vec{\epsilon}$), so the inequality is strict and the average execution price is strictly increasing in $\zeta_{N+1,0}$.

Entry lowers average execution price. Let $\mathcal{V}^*_N$ denote the optimal value in the baseline model with only N wholesalers. In the $(N+1)$-wholesaler model, the broker can always *ignore the entrant* by setting $X_{N+1,1}(\vec{\epsilon}) \equiv 0$ and using the baseline-optimal policy among incumbents. This is feasible, hence $\mathcal{V}^*(\zeta_{N+1,0}) \geq \mathcal{V}^*_N$. Under Assumption 2, the optimal solution assigns strictly positive order flow to wholesaler $N+1$, so the broker strictly improves relative to ignoring the entrant, implying $\mathcal{V}^*(\zeta_{N+1,0}) > \mathcal{V}^*_N$. Since higher broker payoff corresponds to lower average execution price, the average execution price is lower with entry.

Part (2). Under Assumption 2, the entry model is solved exactly as in Proposition 2, with N replaced by $N+1$. Define the $(N+1)$-wholesaler averages

$$\bar{\zeta}_0^+ := \frac{1}{N+1}\sum_{j=1}^{N+1}\zeta_{j,0}, \qquad \bar{\epsilon}^+ := \frac{1}{N+1}\sum_{j=1}^{N+1}\epsilon_j,$$

and

$$A_i^+ := \frac{\bar{\zeta}_0^+ - \zeta_{i,0}}{2c} + \frac{1}{N+1}.$$

Then the optimal period-1 allocation takes the form

$$X_{i,1}^*(\vec{\epsilon}) = A_i^+ + \frac{1}{c}\big(\bar{\epsilon}^+ - \epsilon_i\big), \qquad i = 1, \ldots, N+1.$$

Taking expectations and using $\mathbb{E}[\epsilon_i] = 0$ yields

$$\mathbb{E}\big[X_{N+1,1}^*(\vec{\epsilon})\big] = A_{N+1}^+ = \frac{\bar{\zeta}_0^+ - \zeta_{N+1,0}}{2c} + \frac{1}{N+1}.$$

Because $\bar{\zeta}_0^+$ depends on $\zeta_{N+1,0}$ with weight $1/(N+1)$, we have

$$\frac{\partial}{\partial \zeta_{N+1,0}} \mathbb{E}\big[X_{N+1,1}^*(\vec{\epsilon})\big] = \frac{1}{2c}\left(\frac{1}{N+1} - 1\right) = -\frac{N}{2c(N+1)} < 0.$$

Thus wholesaler $N+1$'s expected period-1 market share is decreasing in $\zeta_{N+1,0}$.

Proof of Lemma 5

Step 1: Equilibrium allocation and the monotone time-0 pricing function. In the homogeneous case, Proposition 2 implies $A_i = 1/N$ for all i, hence the optimal period-1 allocation rule simplifies to

$$x_{i,1}^*(\vec{\epsilon}) = \frac{1}{N} + \frac{1}{c}\big(\bar{\epsilon} - \epsilon_i\big), \qquad \bar{\epsilon} := \frac{1}{N}\sum_{k=1}^{N} \epsilon_k. \tag{IA714}$$

Under the implementation in Proposition 3 with $x_{i,0} = 1/N$, the time-0 price offered by a wholesaler with type ϵ is

$$s^*(\epsilon) = \bar{s} - N\,T^*(\epsilon), \tag{IA715}$$

where $T^*(\cdot)$ is the common interim transfer schedule in the homogeneous case. Specializing the transfer formula from Proposition 2 yields the quadratic form

$$T^*(\epsilon) = K + L\,\epsilon - Q\,\epsilon^2, \tag{IA716}$$

where K is constant in ϵ and

$$L = \frac{N-1}{N}\left(\frac{2}{N} + \frac{\zeta_0 - \bar{s}}{c}\right), \qquad Q = \frac{(N-1)(N-2)}{2N^2 c} \geq 0. \tag{IA717}$$

Assumption 1 implies $T^{*\prime}(\epsilon) = L - 2Q\epsilon < 0$ for all $\epsilon \in [-b/2, b/2]$, so $T^*(\cdot)$ is strictly decreasing and therefore $s^*(\cdot)$ is strictly increasing:

$$\frac{d}{d\epsilon} s^*(\epsilon) = -N\,T^{*\prime}(\epsilon) > 0.$$

Let $\epsilon_{(1)} \leq \cdots \leq \epsilon_{(N)}$ be the order statistics of $\{\epsilon_i\}_{i=1}^N$ and $s_{(1),0} \leq \cdots \leq s_{(N),0}$ be the order statistics of $\{s_{i,0}\}_{i=1}^N$. Since s^* is strictly increasing, we have the mapping

$$s_{(k),0} = s^*\big(\epsilon_{(k)}\big) \qquad \text{for } k = 1, \ldots, N. \tag{IA718}$$

In particular, $j^* \in \arg\min_i s_{i,0}$ corresponds to $\epsilon_{(1)}$. Because $x^*_{i,1}(\vec{\epsilon})$ is strictly decreasing in ϵ_i holding $\bar{\epsilon}$ fixed (cf. (IA714)), the wholesaler with $\epsilon_{(1)}$ has the largest period-1 share, and

$$x_{(1),1} = x^*_{j^*,1}(\vec{\epsilon}) = \frac{1}{N} + \frac{1}{c}\big(\bar{\epsilon} - \epsilon_{(1)}\big). \tag{IA719}$$

For later use, define the convenient coefficients

$$\alpha := -NL = (N-1)\left(\frac{\bar{s} - \zeta_0}{c} - \frac{2}{N}\right) > 0, \qquad \beta := NQ = \frac{(N-1)(N-2)}{2Nc} \geq 0, \tag{IA720}$$

so that from (IA715)–(IA716),

$$s^*(\epsilon) = s_0 + \alpha\,\epsilon + \beta\,\epsilon^2, \qquad s_0 := \bar{s} - NK. \tag{IA721}$$

Step 2: Order-statistic moments under uniform types. Let $U_i := (\epsilon_i + b/2)/b \sim \text{Unif}[0,1]$ and let $U_{(k)}$ denote its order statistics. It is standard that $U_{(k)} \sim \text{Beta}(k, N+1-k)$, hence

$$\mathbb{E}[U_{(k)}] = \frac{k}{N+1}, \qquad \mathbb{E}[U_{(k)}^2] = \frac{k(k+1)}{(N+1)(N+2)}.$$

Since $\epsilon_{(k)} = b(U_{(k)} - \frac{1}{2})$, it follows that

$$\mathbb{E}\big[\epsilon_{(2)} - \epsilon_{(1)}\big] = \frac{b}{N+1}, \tag{IA722}$$

$$\mathbb{E}\big[\epsilon_{(2)}^2 - \epsilon_{(1)}^2\big] = -\frac{N-2}{(N+1)(N+2)}\, b^2, \tag{IA723}$$

$$\mathbb{E}[\epsilon_{(1)}] = -\frac{N-1}{2(N+1)}\, b, \tag{IA724}$$

$$\mathbb{E}[\epsilon_{(1)}^2] - \mathbb{E}[\epsilon^2] = \frac{(N-1)(N-2)}{6(N+1)(N+2)}\, b^2, \tag{IA725}$$

where $\mathbb{E}[\epsilon^2] = b^2/12$ for $\epsilon \sim \mathrm{Unif}[-b/2, b/2]$.

Step 3: $\mathbb{E}\big[\big|s_{(1),0} - s_{(2),0}\big|\big]$ is increasing in b. Because $s_{(1),0} \leq s_{(2),0}$, we have

$$\big|s_{(1),0} - s_{(2),0}\big| = s_{(2),0} - s_{(1),0} = s^*(\epsilon_{(2)}) - s^*(\epsilon_{(1)}).$$

Using (IA721),

$$s_{(2),0} - s_{(1),0} = \alpha\big(\epsilon_{(2)} - \epsilon_{(1)}\big) + \beta\big(\epsilon_{(2)}^2 - \epsilon_{(1)}^2\big).$$

Taking expectations and applying (IA722)–(IA723) yields

$$\mathbb{E}\big[\big|s_{(1),0} - s_{(2),0}\big|\big] = \frac{\alpha}{N+1}\, b - \frac{\beta(N-2)}{(N+1)(N+2)}\, b^2. \tag{IA726}$$

Differentiate with respect to b:

$$\frac{\partial}{\partial b}\mathbb{E}\big[\big|s_{(1),0} - s_{(2),0}\big|\big] = \frac{\alpha}{N+1} - \frac{2\beta(N-2)}{(N+1)(N+2)}\, b. \tag{IA727}$$

Using (IA720), the derivative in (IA727) is strictly positive whenever

$$(\bar{s} - \zeta_0) - \frac{2c}{N} > \frac{(N-2)^2}{N(N+2)}\, b.$$

Assumption 1 implies $c < \frac{N}{2}(\bar{s} - \zeta_0 - \frac{3}{2}b)$, i.e. $(\bar{s} - \zeta_0) - \frac{2c}{N} > \frac{3}{2}b$, and $\frac{3}{2} > \frac{(N-2)^2}{N(N+2)}$ for all $N \geq 2$. Hence (IA727) is strictly positive, proving the claim.

Step 4: $\mathbb{E}\big[|s_{(1),0} - \bar{s}_0|\big]$ is increasing in b. Since $s_{(1),0}$ is the minimum, $s_{(1),0} \leq \bar{s}_0$, so $|s_{(1),0} - \bar{s}_0| = \bar{s}_0 - s_{(1),0}$. Using (IA721),

$$\bar{s}_0 = \frac{1}{N}\sum_{i=1}^{N} s^*(\epsilon_i) = s_0 + \alpha\,\bar{\epsilon} + \beta\,\overline{\epsilon^2}, \qquad \overline{\epsilon^2} := \frac{1}{N}\sum_{i=1}^{N}\epsilon_i^2.$$

Therefore

$$\bar{s}_0 - s_{(1),0} = \alpha(\bar{\epsilon} - \epsilon_{(1)}) + \beta(\overline{\epsilon^2} - \epsilon_{(1)}^2).$$

Taking expectations and using $\mathbb{E}[\bar{\epsilon}] = 0$ and $\mathbb{E}[\overline{\epsilon^2}] = \mathbb{E}[\epsilon^2] = b^2/12$ gives

$$\mathbb{E}[\bar{s}_0 - s_{(1),0}] = \alpha\big(-\mathbb{E}[\epsilon_{(1)}]\big) + \beta\left(\frac{b^2}{12} - \mathbb{E}[\epsilon_{(1)}^2]\right).$$

Applying (IA724)–(IA725) yields

$$\mathbb{E}\big[|s_{(1),0} - \bar{s}_0|\big] = \frac{\alpha(N-1)}{2(N+1)}\,b - \frac{\beta(N-1)(N-2)}{6(N+1)(N+2)}\,b^2. \tag{IA728}$$

Differentiating and using the same implication of Assumption 1 as in Step 3 shows the derivative is strictly positive, so $\mathbb{E}\big[|s_{(1),0} - \bar{s}_0|\big]$ is increasing in b.

Step 5: $\mathbb{E}[\text{HHI}]$ is increasing in b. From (IA714),

$$\text{HHI} = \sum_{i=1}^{N}\left(\frac{1}{N} + \frac{1}{c}(\bar{\epsilon} - \epsilon_i)\right)^2 = \frac{1}{N} + \frac{1}{c^2}\sum_{i=1}^{N}(\epsilon_i - \bar{\epsilon})^2,$$

because $\sum_i(\epsilon_i - \bar{\epsilon}) = 0$ eliminates the cross term. Taking expectations and using $\mathbb{E}[\sum_i(\epsilon_i - \bar{\epsilon})^2] = (N-1)\text{Var}(\epsilon)$ with $\text{Var}(\epsilon) = b^2/12$ yields

$$\mathbb{E}[\text{HHI}] = \frac{1}{N} + \frac{N-1}{12c^2}\,b^2, \tag{IA729}$$

which is strictly increasing in b.

Step 6: $\mathbb{E}[\bar{s}_0]$ is increasing in b. Using $s^*_{i,0}(\epsilon_i) = \bar{s} - NT^*(\epsilon_i)$ and symmetry, $\mathbb{E}[\bar{s}_0] = \bar{s} - N\mathbb{E}[T^*(\epsilon)]$. From Proposition 2 specialized to $A_i = 1/N$ and $\zeta_{i,0} = \zeta_0$,

$$\mathbb{E}[T^*(\epsilon)] = \frac{\bar{s} - \zeta_0}{N} - \frac{c}{N^2} - \frac{b}{2N} + \frac{N-1}{12Nc}\,b^2,$$

so

$$\mathbb{E}[\bar{s}_0] = \zeta_0 + \frac{c}{N} + \frac{b}{2} - \frac{N-1}{12c}\,b^2. \tag{IA730}$$

Hence

$$\frac{\partial}{\partial b}\mathbb{E}[\bar{s}_0] = \frac{1}{2} - \frac{N-1}{6c}\,b.$$

Assumption 1 implies $c > (N-1)b$, so $\frac{N-1}{6c}b < \frac{1}{6}$ and the derivative is strictly positive.

Step 7: $\mathbb{E}[x_{(1),1}]$ is increasing in b. By (IA719),

$$x_{(1),1} = \frac{1}{N} + \frac{1}{c}\big(\bar{\epsilon} - \epsilon_{(1)}\big).$$

Taking expectations and using $\mathbb{E}[\bar{\epsilon}] = 0$ and (IA724) gives

$$\mathbb{E}[x_{(1),1}] = \frac{1}{N} - \frac{1}{c}\mathbb{E}[\epsilon_{(1)}] = \frac{1}{N} + \frac{N-1}{2c(N+1)}\,b,$$

which is strictly increasing in b.

Internet Appendix - Empirics

Appendix IA0.1. Additional Details and Summary Statistics

We present the list of symbols provided by Broker A in Table XIV.

Appendix IA0.2. Routing Alternative Specifications

We consider a version of Regression 1 with fixed effects for each market center in Table XV. We also consider a version of Regression 1 with changes in EFQ and order share rather than levels. Estimates are presented in Table XVI. All estimates of β_1 are negative. A decrease in EFQ of 1 percentage point is associated with a 0.2 percentage point decline in next-day order share at Broker A and a 0.03 percentage point decline in next-month order share at Broker B. A decrease of 1-percentage point in wholesaler execution score is associated with a 0.4 percentage point drop in subsequent order share at Broker C.

Appendix IA0.3. History Specificity

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 XVII. 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: if a broker evaluates 30 days of history, a longer or shorter window will explain much less of the variation in broker routing decisions. EFQ spreads for orders of one size may offer less, or potentially zero, explanatory power for routing decisions for orders of another size.

Appendix IA0.4. Performance Sensitivity

Table XIX presents summary statistics on key variables for performance sensitivity. We consider several possible alternative measures of incentive slopes. As a symbol-security-bin level measure of incentive slope, we use $FirstRankShare_{jk}$, which measures the average share of order flow allocated to the first-ranked wholesaler in symbol j and security size bin k.

REGRESSION 7: *Using observations of symbol j and order size bin k, we estimate:*

$$FirstRankShare_{jk} = \beta_0 + \beta_1 SD_{jk} + \beta_2 EFQ\_Lead_{jk} + \epsilon_{jk}$$

Higher first-rank share is equivalent to a steeper incentive slope (as the first-ranked firm captures a larger percentage of order flow). We also include a control EFQ_Lead, which measures the average lead of the first-ranked wholesaler. Estimates of Regression 7 are presented in Table XVIII. Symbols with more variation in EFQ are associated with a larger share of orders allocated to the first-ranked wholesaler. Moving from the 25^{th} percentile of symbol-ordersize level standard deviation to the 75^{th} percentile of symbol-ordersize level standard deviation is associated with an increase in Rank-1 wholesaler order allocation of 0.8 percentage points.[38] Regression 7 uses security-ordersize level observations to measure incentive slopes. We can also measure the sensitivity of the broker responses to changes in wholesaler performance at the daily level, and compare that sensitivity across security-ordersize bins based on their variability.

We consider a variation of 2 using changes in EFQ rather than changes in EFQ rank. One measurement challenge is the inherent conflation between sensitivity to changes in EFQ and the greater or lesser raw variation in EFQ across order-size bins. As one method of avoiding

[38]We use Column (2) of Table XVIII in this calculation. Table XIX in the internet appendix provides the necessary summary statistics.

this conflation, we replace EFQ with a standardized measure of EFQ.[39] Results are presented in Table XXII. Moving from the 25^{th} percentile of symbol-order size level standard deviation to the 75^{th} percentile of symbol-order size level standard deviation is associated with an additional 0.2 to 0.3 percentage points of order flow adjustment to changes in EFQ. Table XXIII is a robustness check of Table III, and includes an additional control of wholesaler lead for the adjustment process.

[39] We consider several possible standardization schemes. In the simplest, we split symbol-size bins into below-median or above-median, and standardize EFQ changes within each group (Table 2).

Table XIV: List of Broker A Symbols. From Broker A we obtain a sample of 480 stocks. All symbols are divided into ten deciles based on trading volume; from each decile, we obtain 50 symbols drawn at random (with replacement, thus some categories have slightly less than 50 symbols).

Decile	Symbols
0	AAWW, ASHR, BHF, CTAS, GLDD, KNTK, SCHL, SWIR, ZSL, ABIO, ALHC, AROC, BC, BTRS, CBOE, CVGW, FGF, INFL, MMMB, NUVL, PH, RGEN, STCN, ZBH, FCFS, IYY, LAMR, MJXL, MNDO, NEWR, NPTN, PAYA, PRSO, RMD, UGP, VGSH, WFRD, IMRX, SEER, SPBC, WEST, RTO
1	ARDS, ASPC, CCRV, DK, ESS, FOXA, FSP, OPRT, PAGP, RYI, SBIG, SFYX, SPYX, SRVR, TSEM, UNM, USRT, AGS, CCK, CDZI, CNEY, FWRG, NTCO, OMIC, SBTX, SELF, SPB, TPOR, VCIT, AMPH, AMX, ASPU, ATCO, ATY, EHAB, MBUU, PFHC, PUMP, TOKE, AIP, AXL, BSGA, SCPH, CKX, ROG, KNW
2	AACG, AJRD, ATEN, BHSE, BNTC, BOLT, BURL, BXP, CGNT, CQP, FDN, FDVV, HESM, HYDR, IHI, LND, PFFA, REK, ROBO, THO, TRVI, UNIT, VNQI, VVV, BIPC, CANF, CSGP, GPC, HWM, MFGP, MIST, NAPA, NSC, SDEM, VGK, YMAB, AIG, AIRI, AY, ELP, LIAN, LQD, NCR, OOTO, RVAC, EDTK
3	AMPS, BAK, BAX, CHWA, CTMX, HACK, IEA, MBRX, MDVL, NCNO, QURE, SNCR, TWI, TWNK, TX, WCLD, WISA, XLC, AAAU, ACON, AGG, CTLT, DLPN, ICE, LPG, ONDS, SHC, SHY, SNPS, TDCX, TRP, XLB, XNET, APTV, ARWR, HTOO, IMAQ, KNOP, SANW, UVV, WES, TAOP, CGTX, UPXI, VANI, NYAX
4	ALBO, BNED, BWA, FSK, HASI, HCTI, HES, IGV, MCW, NLS, OILU, OM, PIPP, PRNT, RF, SKIN, SLCA, WW, YPF, ZLAB, ALEC, AMBC, ATOM, CMC, DBA, INDP, IPOD, LPL, MIGI, TIP, TOL, TROX, USFD, WIT, WKEY, ZYME, DOMO, DRI, ERJ, FPAY, NAUT, REAX, UCL, AAPD, INLX
5	ALRN, ASM, ATRA, BETZ, CCO, CNSP, COWZ, CRCT, CVAC, GRMN, HDSN, LCTX, LSPD, MAXR, OCX, OIH, PD, PLNT, PRTS, TRT, TWOU, USD, VTIP, ARHS, DISH, DPZ, EC, FIS, HRL, KDP, MUX, NMTC, NRG, PAVE, PMT, TSCO, CARG, CMCM, DAWN, LADR, PAYO, SQSP, SST, BNSO, RNXT, CNTQ
6	ADMA, ANET, AREC, AVAH, AVCO, BIG, BLOK, COUR, CURV, CWBR, DFS, DOGZ, FLWS, GMDA, HOOK, JKS, MAXN, MCHP, NEX, POLA, PSTG, QFIN, QNCX, SIEN, SKYX, TROW, WMB, XLU, XPO, ACMR, AMWL, BQ, EQNR, HCP, JOB, KRBN, MTNB, NOBL, RKLY, RSI, VORB, WMC, EA, GBS, ROST, SDGR, TASK
7	ALT, AREB, ATNX, CBAT, CMRA, COOK, CREX, CRTD, DDD, DQ, DS, EXPI, GLBE, HOUR, KSPN, LAW, LFLY, MDT, MTCH, NCTY, NMTR, NUSI, OCFT, SMRT, TAN, VVNT, VXUS, BARK, CEMI, CSX, DLO, GFI, IMGN, IMUX, INVZ, IRM, NVIV, PM, RAAS, SMFR, SQM, STWD, YCBD, AR, CIM, FRGT, ICU, ACRV
8	APA, AQB, ARKW, BEST, BMBL, BOXD, BRDS, CBAY, CELZ, CIDM, CTRA, CVE, CYRN, DLTR, DX, EFOI, EJH, EOSE, JDST, KMI, LDI, LLY, LOTZ, LUMN, ON, PFLT, PTRA, PUBM, QLD, RTX, SHIP, SLB, SPYD, TUEM, UNH, UPS, WTRH, Z, BRMK, ERX, ITRM, PIK, PIXY, SEAC, UONE, BCAN, JUPW, SLNA, NWTN
9	AI, AMD, BILI, BITF, BROS, BTBT, BTU, C, CHPT, COMS, DNA, EMBK, ENDP, FNGU, GM, GOVX, HKD, HLBZ, HOOD, HUSA, IMPP, INO, JOBY, MARA, MMAT, MRO, NNDM, PSFE, PSNY, PYPL, QCOM, QS, REV, RKLB, ROKU, RUN, SDIG, SIGA, SOFI, SPYG, STEM, TZA, USOI, UWMC, VAXX, VLTA, RDBX

Table XV: Order Share and Execution Quality - Wholesaler Fixed Effects. We estimate Regression 1 which regresses the $OrderShare$ allocated to each wholesaler on prior market performance. Relative to Table II we include a fixed effect for each wholesaler. $OrderShare$ 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, wholesaler, and order size and we cluster standard errors by wholesaler and date; observations are at the category-wholesaler-date level for Broker B with a fixed effect for each trade date, wholesaler, and security bin, and we cluster standard errors by wholesaler and date; and observations are at the wholesaler-date level for Broker C, with a fixed effect for each date and wholesaler, and we cluster standard errors by wholesaler and date.

	Dependent variable: OrderShare					
	Broker A Data		Broker B Data		Broker C Data	
	(1)	(2)	(3)	(4)	(5)	(6)
Prior EFQ	-1.078^{***}		-0.261^{*}			
	(0.140)		(0.114)			
Prior EFQ Rank		-8.557^{***}		-2.265^{**}		
		(1.141)		(0.513)		
Prior Score					-0.416	
					(0.395)	
Prior Score Rank						-1.731^{***}
						(0.224)
Observations	129,526	129,526	806	806	170	176
R^2	0.332	0.342	0.684	0.698	0.916	0.917
Note:						*p<0.1; **p<0.05; ***p<0.01

Table XVI: Order Flow Changes. $\Delta OrderShare$ measures the change in order share at the daily (Broker A) or monthly (Broker B and C) horizon. ΔEFQ measures prior changes in EFQ at the daily (Broker A) or monthly (Broker B and C) observation level; for Broker C we use changes in a proprietary execution score metric of which EFQ is an important component. For Broker A, we cluster standard errors by market center, symbol, and order size (column 2) or date, symbol, and order-size (Column 3). For Broker B, we cluster standard errors by security bin and market center (column 2) or date and security bin (column 3). For Broker C we cluster standard errors by market center (column 2) or date (column 3).

Panel A: Broker A Data

	Dependent variable: $\Delta(OrderShare)_{30vs31}$		
	(1)	(2)	(3)
ΔEFQ_{32vs31}	-0.191^{***}	-0.191^{**}	-0.191^{***}
	(0.051)	(0.056)	(0.031)
Date+Symbol+Ordersize FE	X	X	X
Observations	93,574	93,574	93,574
R^2	0.035	0.035	0.035

Panel B: Broker B Data

	Dependent variable: $\Delta OrderShare_t$		
	(1)	(2)	(3)
ΔEFQ_{t-1}	-0.027^{***}	-0.027^{*}	-0.027^{**}
	(0.010)	(0.009)	(0.008)
Security Bin+Date FE	X	X	X
Observations	762	762	762
R^2	0.011	0.011	0.011

Panel C: Broker C Data

	Dependent variable: $\Delta OrderShare_t$		
	(1)	(2)	(3)
$\Delta Score_{t-1}$	-0.397^{*}	-0.397	-0.397
	(0.210)	(0.272)	(0.284)
Date FE	X	X	X
Observations	164	164	164
R^2	0.026	0.026	0.026
Note:	*p<0.1; **p<0.05; ***p<0.01		

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

Panel A: Different History Windows

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

Panel B: Different Order Sizes

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

Table XVIII: Incentive Slopes at Symbol Level. We estimate Regression 7 using observations at the symbol-security size bin level using Broker A data. Incentive slopes are measured as the average (across days) order share allocation to the first-ranked wholesaler. We consider two measures of variability: SD_{WL}, where we measure daily EFQ for each individual wholesaler and then measure the standard deviation across wholesaler-day observations, and SD_{DL}, where we measure daily EFQ across all wholesalers, and then measure the standard deviation across daily observations.

	Dependent variable: FirstRankShare			
	(1)	(2)	(3)	(4)
SD_{WL}	0.004***	0.077***		
	(0.002)	(0.027)		
EFQ_Lead_{Rank1}		0.508***		0.541***
		(0.053)		(0.050)
SD_{DL}			0.008***	0.040***
			(0.002)	(0.011)
Constant	69.325***	50.846***	69.204***	51.155***
	(0.390)	(0.489)	(0.389)	(0.433)
Observations	1,969	884	1,968	884
R^2	0.004	0.131	0.010	0.136
Note:				*p<0.1; **p<0.05; ***p<0.01

Table XIX: Summary Statistics on Performance Sensitivity Variables - Broker A Data. Panel A presents summary statistics at the symbol-order size bin level. Rank 1 Mean order share is the mean order share obtained by the rank 1 wholesaler. SD_{WL} measures the standard deviation of EFQ measured at the wholesaler-day level. SD_{DL} measures the standard deviation of EFQ measured at the day level. Panel B presents summary statistics at the wholesaler-day-symbol-order size bin level. $\Delta_{ordershare}$ is the change in order share to the day prior, and $\Delta_{ordershare}$ - R1 only is this value restricted to only the first-ranked wholesaler. EFQ scaled is the z-score of EFQ, split into observations with either SD_{WL} below or above the median SD_{WL}, or SD_{DL} below or above the median SD_{DL}.

Panel A: Symbol-Ordersize Bin Level Statistics

Statistic	5	25	50	75	95	Mean	SD
Rank 1 Mean Order Share	45.9	55.0	68.8	85.2	100.0	70.8	17.9
SD_{WL}	7.3	11.9	15.7	22.7	56.4	23.3	37.6
SD_{DL}	4.1	10.0	15.0	21.8	50.2	24.0	67.3
Rank 1 Mean EFQ Lead	1.7	3.4	5.2	8.2	15.5	6.6	5.1

Panel B: Wholesaler-Day-Symbol-Ordersize Bin Level Statistics

Statistic	5	25	50	75	95	Mean	SD
$\Delta_{ordershare}$	-44.2	-9.9	0.0	10.1	44.8	0.1	25.3
$\Delta_{ordershare}$ - R1 Only	-46.667	-8.633	0.000	16.701	62.500	4.365	29.963
EFQ Scaled - WL	-1.448	-0.008	-0.008	-0.008	1.433	-0.000	1.000
SD Scaled - DL	-1.448	-0.008	-0.008	-0.008	1.433	-0.000	1.000

Table XX Order Flow Sensitivity Within Bins. We estimate Regression 2. $\Delta(OrderShare)$ measures daily changes in order share. Variability for the symbol-ordersize bin is measured with SD_{DL}, where we measure daily EFQ for each day and then measure the standard deviation across day observations. ΔEFQ measures changes in EFQ, standardized above or below median values of SD_{DL}. (We consider alternative schemes in the Internet Appendix). We fit a wholesaler fixed effect for each wholesaler (and cluster standard errors by wholesaler) in columns (2) and (4). Columns (3) and (4) are restricted to the first-ranked wholesaler only. Relative to Table XXII, we use SD_{DL} (day-level) in place of SD_{WL} (wholesaler-day level).

	Dependent variable: $\Delta(OrderShare)$			
	All Wholesalers		Rank 1 Only	
	(1)	(2)	(3)	(4)
ΔEFQ (Standardized)	-6.665^{***}	-6.679^{***}	-11.080^{***}	-10.988^{***}
	(0.049)	(0.858)	(0.132)	(1.294)
SD_{DL}	-0.0001	0.0001	-0.002	-0.002
	(0.001)	(0.002)	(0.002)	(0.003)
Δ EFQ (Standardized)$*SD_{DL}$	-0.024^{***}	-0.024^{***}	-0.024^{***}	-0.023^{**}
	(0.001)	(0.003)	(0.002)	(0.006)
Constant	0.130^{***}		0.640^{***}	
	(0.049)		(0.109)	
Wholesaler FE		X		X
Observations	269,657	269,657	86,154	86,154
R^2	0.081	0.081	0.092	0.094
Note:				*p<0.1; **p<0.05; ***p<0.01

Table XXI Order Flow Sensitivity Within Bins. We estimate Regression 2. $\Delta(OrderShare)$ measures daily changes in order share. Variability in each symbol-order size bin is measured with SD_{WL}, where we measure daily EFQ for each individual wholesaler and then measure the standard deviation across wholesaler-day observations. ΔEFQ measures changes in EFQ, standardized above or below median values of SD_{WL}. (We consider alternative schemes in the Internet Appendix). We fit a wholesaler fixed effect for each wholesaler (and cluster standard errors by wholesaler) in columns (2) and (4). Columns (3) and (4) are restricted to the first-ranked wholesaler only. Relative to Table XXII, which standardizes EFQ across the median SD_{WL}, here we split symbol-ordersize into SD_{WL} bins with the following cuts: (0,5,10,20,50).

	Dependent variable: DeltaOrderShare			
	All Wholesalers		Rank 1 Only	
	(1)	(2)	(3)	(4)
EFQ (Standardized)	-6.129^{***}	-6.144^{***}	-10.095^{***}	-9.984^{***}
	(0.051)	(1.011)	(0.141)	(1.476)
SD_{WL}	0.005^{***}	0.006	0.0004	0.001
	(0.001)	(0.003)	(0.001)	(0.001)
EFQ (Standardized)$*SD_{WL}$	-0.034^{***}	-0.034^{*}	-0.059^{***}	-0.058^{***}
	(0.001)	(0.014)	(0.003)	(0.010)
Constant	0.048		0.841^{***}	
	(0.049)		(0.108)	
Wholesaler FE		X		X
Observations	269,682	269,682	86,178	86,178
R^2	0.072	0.072	0.083	0.085
Note:				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table XXII Order Flow Sensitivity Within Bins. We estimate Regression 2 using Broker A data. $\Delta(OrderShare)$ measures daily changes in order share. Variability in each symbol-order size bin is measured with SD_{WL}, where we measure daily EFQ for each individual wholesaler and then measure the standard deviation across wholesaler-day observations. ΔEFQ measures changes in EFQ, standardized above or below median values of SD_{WL}. (We consider alternative schemes in the Internet Appendix). We fit a wholesaler fixed effect for each wholesaler (and cluster standard errors by wholesaler) in columns (2) and (4). Columns (3) and (4) are restricted to the first-ranked wholesaler only.

	Dependent variable: $\Delta(OrderShare)$			
	All Wholesalers		Rank 1 Only	
	(1)	(2)	(3)	(4)
ΔEFQ (Standardized)	-6.685^{***}	-6.699^{***}	-10.935^{***}	-10.840^{***}
	(0.050)	(0.904)	(0.137)	(1.348)
SD_{WL}	-0.0001	-0.00001	-0.006^{***}	-0.006^{**}
	(0.001)	(0.003)	(0.002)	(0.001)
Δ EFQ (Standardized)$*SD_{WL}$	-0.023^{***}	-0.023^{**}	-0.033^{***}	-0.032^{**}
	(0.001)	(0.006)	(0.003)	(0.008)
Constant	X		X	
Wholesaler FE		X		X
Observations	269,680	269,680	86,176	86,176
R^2	0.080	0.080	0.092	0.094
Note:				*p<0.1; **p<0.05; ***p<0.01

Table XXIII: Order Flow Sensitivity and Rank - Additional Control. Relative to Table III, we include a control for the lead of the first-ranked wholesaler in each symbol-ordersize bin. As before, we estimate Regression 2. $\Delta(OrderShare)$ measures daily changes in order share. Variability is measured with SD_{WL}, where we measure daily EFQ for each individual wholesaler and then measure the standard deviation across wholesaler-day observations. We consider two measures of variability: SD_{WL}, where we measure daily EFQ for each individual wholesaler and then measure the standard deviation across wholesaler-day observations, and SD_{DL}, where we measure daily EFQ across all wholesalers, and then measure the standard deviation across daily observations. Δ_{rank} measures changes in ordinal wholesaler ranking (with the best, lowest EFQ wholesaler having rank 1). We fit a wholesaler fixed effect for each wholesaler (and cluster standard errors by wholesaler) in columns (2) and (4).

	Dependent variable:			
	DeltaOrderShare			
	(1)	(2)	(3)	(4)
δ_{EFQ}	-2.695^{***}	-2.754^{***}	-3.561^{***}	-3.619^{***}
	(0.163)	(0.349)	(0.119)	(0.440)
SD_{WL}	-0.068^{***}	-0.067^{**}		
	(0.010)	(0.022)		
SD_{DL}			-0.028^{***}	-0.027^{*}
			(0.005)	(0.010)
$EFQ_{R1-Lead}$	0.008	0.007	-0.009	-0.009
	(0.011)	(0.009)	(0.011)	(0.011)
$\Delta_{rank} * SD_{WL}$	-0.106^{***}	-0.106^{***}		
	(0.010)	(0.020)		
$\Delta_{rank} * SD_{DL}$			-0.044^{***}	-0.044^{**}
			(0.006)	(0.011)
Constant	0.116		-0.331^{***}	
	(0.129)		(0.096)	
Wholesaler FE		X		X
Observations	134,579	134,579	134,579	134,579
R^2	0.014	0.014	0.013	0.014
Note:				*p<0.1; **p<0.05; ***p<0.01

Appendix IA0.5. Routing and Stock 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 (Easley et al. (1996)). Days with higher absolute log returns 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 8: *For each broker we estimate:*

$$Spread_{jkt} = \quad \beta_0 + \beta_1 |LogReturn|_{jkt} + X_{jkt} + \epsilon_{ijkt}$$

for wholesaler j, time period t, and security bin k.

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. $|LogReturn|\_jkt$ 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 8 are presented in Table XXIV (for Broker C) and Table **??** (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 IA3.1.

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.[40] 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.

Appendix IA0.6. Wholesaler Entry

In this section, we consider alternative measures of competition including the order share of the top two firms, or the order share of the top three firms.

Appendix IA0.7. Strategic Wholesaler Behavior

Figure 5 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

[40] As EFQ is calculated as the total effective spread over the total quoted spread,i.e. a ratio of averages rather than an average ratio.

Table XXIV: Market Volatility and Retail Outcomes. This table presents results of Regression 8. $|LogReturn|$ is the absolute log return (measured intrday). 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 date, and cluster by symbol and dates. Panel A presents results for Broker A. Panel B presents results for Broker C, restricted to the set of symbols in XIV. We do not have symbol-level data for Broker B.

Panel A: Broker A Data

	Dependent variable:		
	EFQ	Effective	Quoted
	(1)	(2)	(3)
Volume-Weighted Average Price	0.004	0.143***	0.177***
	(0.009)	(0.038)	(0.045)
Total Volume	0.966**	−2.829	3.970
	(0.468)	(1.857)	(6.451)
$\|LogReturn\|$	14.099	50.366*	65.333*
	(10.072)	(26.259)	(37.966)
Observations	53,855	53,855	53,855
R^2	0.115	0.614	0.591

Panel B: Broker C Data

	Dependent variable:		
	EFQ	Effective	Quoted
	(1)	(2)	(3)
Volume-Weighted Average Price	0.008	0.249	0.508***
	(0.100)	(0.208)	(0.158)
Total Volume	−0.002	−0.085	−0.117**
	(0.030)	(0.063)	(0.048)
$\|LogReturn\|$	13.157	67.190***	42.095**
	(10.788)	(22.473)	(17.107)
Observations	12,340	12,340	12,341
R^2	0.166	0.380	0.817
Note:			*p<0.1; **p<0.05; ***p<0.01

Table XXV: Wholesaler Entry Decision. We estimate Regression 4, 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. We use a different set of measures than Table VIII. Within each category, we measure $First-To-Average$ (the difference between the EFQ of the first wholesaler vs. volume-weighted average), and $TopTwoOrderShare$ (the share of orders obtained by the top two wholesalers in that category),$TopThreeOrderShare$ (the share of orders obtained by the top three wholesalers in that category), EFQ is the effective-over-quoted spread. 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)
First-To-Avg	0.054			0.042
	(0.086)			(0.073)
Top Two Order Share		0.110**		0.086
		(0.056)		(0.055)
Top Three Order Share			0.063	
			(0.104)	
EFQ				0.145**
				(0.073)
Observations	1,461	1,464	1,297	1,461
R^2	0.442	0.444	0.509	0.449
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table XXVI: Wholesaler Entry and Competition. We estimate Regression 5, 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. Relative to Table IX, we use a different set of measures: First-to-Average, the absolute difference in EFQ between the first-ranked wholesaler and the average, Top Two Order Share, the total share of orders allocated to the top two wholesalers, and Top Three Order Share, the total share of orders allocated to the top three wholesalers.

Panel A: Excluding Wholesaler A5

	Dependent variable:		
	First-To-Avg	Top Two Order Share	Top Three Order Share
	(1)	(2)	(3)
Post	-2.003^{***}	-14.494^{***}	-17.736^{***}
	(0.421)	(0.565)	(0.471)
Observations	3,133	3,093	2,683
R^2	0.232	0.504	0.594

Panel B: Including Wholesaler A5

	Dependent variable:		
	First-To-Avg	Top Two Order Share	Top Three Order Share
	(1)	(2)	(3)
Post	-2.963^{***}	-7.738^{***}	-6.435^{***}
	(0.432)	(0.489)	(0.296)
Observations	3,157	3,162	2,792
R^2	0.205	0.455	0.477

Panel C: Wholesaler A Entry As Independent Variable

	Dependent variable:		
	First-To-Avg	Top Two Order Share	Top Three Order Share
A5_Share	-0.068^{***}	-0.278^{***}	-0.207^{***}
	(0.016)	(0.021)	(0.013)
Observations	3,157	3,162	2,792
R^2	0.197	0.460	0.445
Note:			*p<0.1; **p<0.05; ***p<0.01

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 examine changes in wholesaler performance relative to their average.

We consider an alternative measure of strategic behavior, using the continuous number of days until adjustment rather than an indicator for the last five days of the month. Note that while LastFiveDays was defined as an indicator which took the value 1 in the last five days of the month, and zero otherwise. Here, Days Until Evaluation takes the value 1 at the very last day of the month, 2 at the second-last day of the month, and so on, so higher values of days until evaluation are associated with larger distance to the end of the month. Consequently, the negative coefficients here indicate volatility is lower toward the beginning of the month, consistent with our earlier estimates that volatility is higher toward the end of the month.

Table XXVII: Wholesaler Strategic Improvement. : **Wholesaler Strategic Improvement**. In contrast to Table VII, we use days until evaluation rather than an indicator for the last five days of the month; dates closer to the next order allocation will have a lower value for days until evaluation. In column 1, uses the 1st-Ranked (i.e. best) wholesaler. In column 2, the 2nd-ranked, column 3 the 3rd ranked, and column 4 the 4th ranked wholesaler. Panel A uses data from Broker B, with the dependent variable being absolute difference in wholesaler EFQ from the 90 day average of wholesaler EFQ; we also fit a fixed effect for each asset category and cluster standard errors by asset category. Panel B uses data from Broker C, with the dependent variable being the absolute value deviation of the day wholesaler score from the 90 day average wholesaler score. We do not use data from Broker A has Broker A evaluates on a rolling basis.

Panel A: Broker B Data

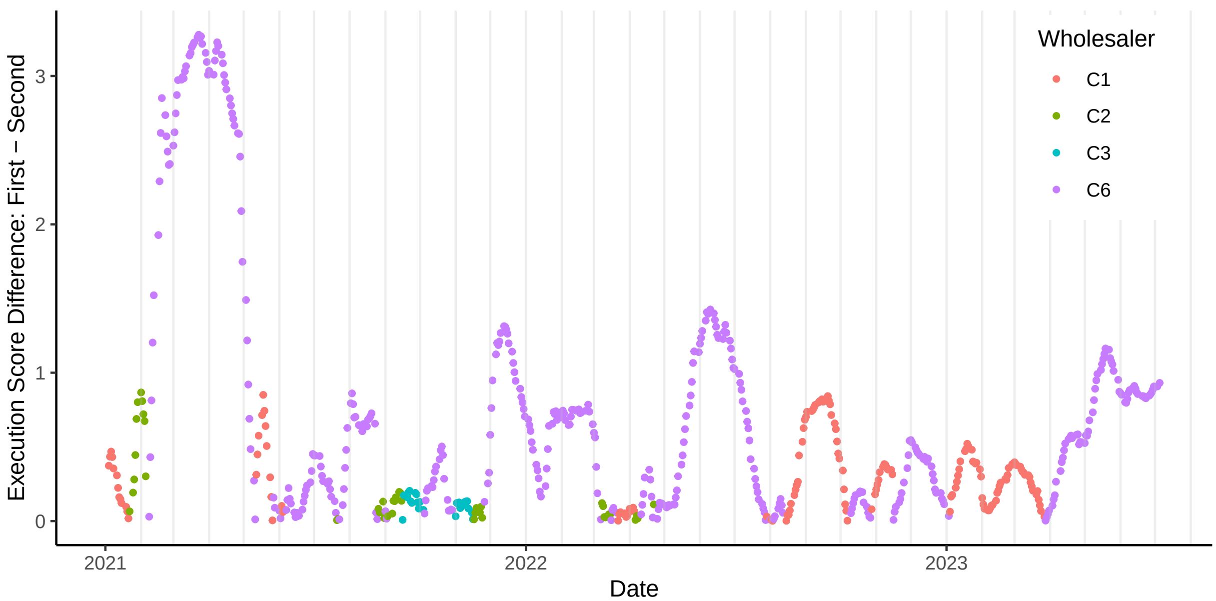
	Dependent variable: $\lvert\Delta EFQ\rvert$			
	(1)	(2)	(3)	(4)
Days Until Evaluation	−0.008	−0.029	−0.049**	−0.045**
	(0.010)	(0.031)	(0.014)	(0.008)
Observations	2,893	3,037	2,959	2,957
R^2	0.026	0.015	0.017	0.072

Panel B: Broker C Data

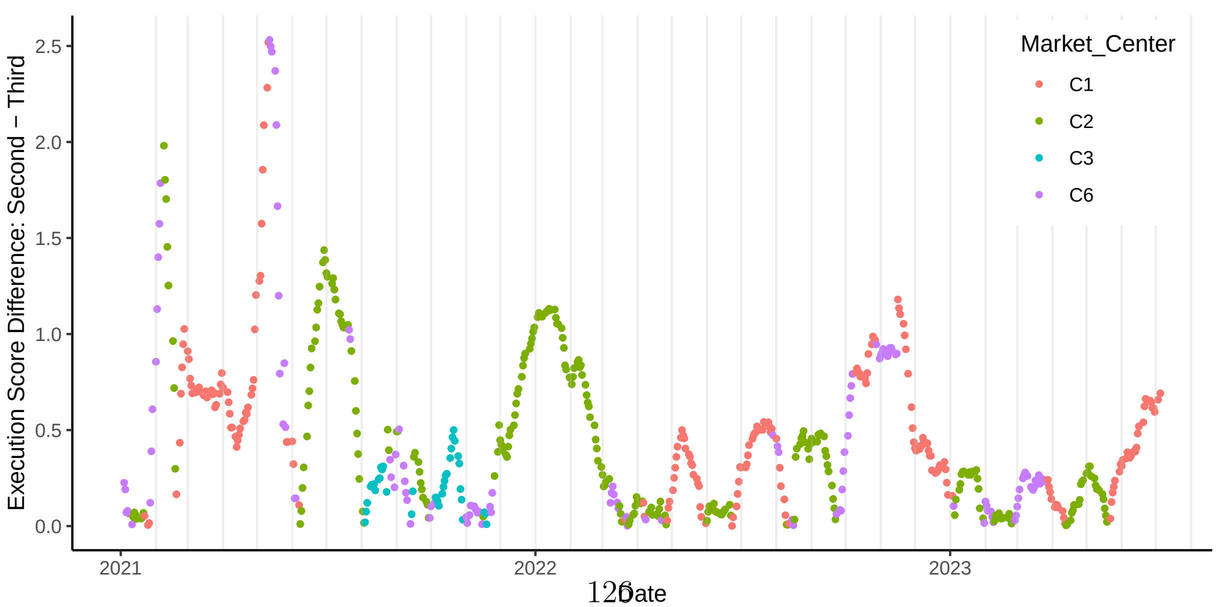
	Dependent variable: $\lvert\Delta Score\rvert$			
	(1)	(2)	(3)	(4)
Days Until Evaluation	−0.028**	−0.040*	−0.042	−0.096***
	(0.014)	(0.022)	(0.027)	(0.021)
Constant	1.772***	2.269***	2.167***	2.447***
	(0.121)	(0.197)	(0.238)	(0.181)
Observations	628	628	628	628
R^2	0.006	0.005	0.004	0.033
Note:				*p<0.1; **p<0.05; ***p<0.01

Figure 5. 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



Appendix IA0.8. 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.[41] Any non-marketable limit orders which a wholesaler receives and does not immediately fill must be displayed on an exchange. Effectively, the wholesalers serve largely as a pass-through, receiving limit orders from brokers and posting them to exchanges. Spatt (2019) highlights an important reason for this pass-through behavior: exchange pricing is often volume-tiered. Limit orders are eligible for a market-making rebate, and the more a market participant trades, the larger the rebate they earn. Market makers who are active in the market may qualify for the best rebate tier, while a small or even medium broker may not qualify for the best rebate tier. We compare the persistence of performance between market and limit orders with the following regression:

REGRESSION 9: *For each wholesaler j, time period t and security bin k, we estimate:*

$$\mathit{RelativePerformance}_{jkt} = \beta_0 + \beta_1 Prior\_Relative\_Performance_{jkt} + X_{jk} + \epsilon_{jkt}$$

$RelativePerformance$ 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 9 are presented in Table XXVIII. 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

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

Table XXVIII: Limit Order Persistence. We estimate Regression 9, 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

wholesalers will obtain different financial rewards from posting limit orders to the exchange. Even if the performance, as measured by fill rates obtained by customers, is the same, wholesaler profits will not be.

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