# Why Did Retail Liquidity Programs Fail?*

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

August 15, 2026

**Abstract**

Retail Liquidity Programs (RLPs) allow on-exchange, sub-penny price improvement for retail orders, yet capture only a negligible fraction of retail flow. We find RLPs deliver less price improvement than off-exchange wholesalers, wholesalers route more orders to RLPs when market conditions deteriorate, and wholesalers internalize even when an RLP offers a superior price. Evidence from the SEC's Tick Size Pilot further shows that off-exchange effective spreads are insensitive to wider quoting increments, whereas RLP effective spreads narrow. We rationalize these findings with a model of wholesaler routing built around three frictions: non-displayed prices, absence of order protection, and wholesaler routing discretion.

*Some analysis in this paper was previously circulated as part of the paper *Would Order-by-Order Auctions Be Competitive?* For helpful comments and feedback we are thankful to Yashar Barardehi, Robert Battalio, Larry Glosten, Terrence Hendershott, Pete Kyle, Mark Loewenstein, Josh Mollner, Alberto Teguia, Vincent Van Kervel, and Bart Zhou Yueshen, along with seminar and conference participants at Arizona State University, Carnegie Mellon–Penn State–Pittsburgh joint conference, China International Conference in Finance, Eastern Finance Association, HEC (Paris), Imperial College, London Business School, University of Maryland, the Microstructure Exchange, New York University Stern Market Microstructure Conference, Securities and Exchange Commission Economics Conference, Society for Financial Studies Cavalcade, Western Finance Association, and numerous anonymous industry participants.

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

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

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

# I. Introduction

Retail marketable orders are central to modern U.S. equity trading, yet the vast majority are executed off exchange by a small number of wholesalers rather than on transparent exchange order books. Beginning with the NYSE in 2012, several U.S. exchanges introduced Retail Liquidity Programs (RLPs): on-exchange mechanisms that segment retail order flow and allow price improvement in sub-penny increments. RLPs rest on the premise that retail orders carry lower adverse selection risk, enabling liquidity providers to quote tighter prices than those available in the standard displayed book. While RLPs have at least one-sided liquidity more than one-fifth of the day in large-cap stocks, they execute trivial volume: just 0.1% of U.S. equity trading volume. This paper analyzes the economics of RLPs and extends empirical analysis with a simple theoretical framework that captures the central features of RLPs to develop an explanation for why on-exchange competition for retail order flow has failed to gain traction and what this failure reveals about retail execution and venue design.

There are three institutional features that are central to RLPs. First, RLP liquidity is non-displayed: although market participants can observe signals about the presence of RLP interest, the price and depth of resting RLP liquidity are not observable to wholesalers. Second, RLP prices are not protected quotations under Regulation NMS, so there is no legal requirement that retail orders execute against superior RLP liquidity even when it is available. For example, the IEX RLP requires the RLP price to equal the mid-quote, making it publicly observable, yet wholesalers are not obligated to route retail orders to this superior price. Third, wholesalers exercise full routing discretion over the retail orders they receive, choosing when, how much, and which orders to internalize versus route to an RLP. We investigate each of these features empirically, and develop a model to understand how they interact.

We begin our empirical analysis with the display and price improvement of RLPs. Using NYSE TAQ data from January 2019 to May 2022 for U.S. equities of the Russell 3000 stocks, along with the 100 most frequently traded ETFs, we establish a link between RLP availability and market conditions. RLP liquidity is available when market imbalances are low, but disappears when orders are predominantly buys or predominantly sells. The price and size of RLP liquidity is non-displayed: we document that orders executed in RLPs rarely exceed 100 shares. Crucially, we note that RLPs deliver substantially less price improvement than off-exchange wholesalers—11 mils on average compared to 19 mils from wholesalers.

RLP volume share rises with volatility and wider spreads, consistent with wholesalers routing more orders to RLPs when market conditions deteriorate. At the same time, RLP liquidity providers concentrate their quotes in calm periods with balanced order flow, withdrawing when conditions deteriorate. These demand and supply forces work at cross purposes: wholesalers most want to route to RLPs precisely when RLP liquidity is least available.

RLP liquidity, by design, is non-displayed and therefore does not receive order protection. This weakens both routing incentives and liquidity provision incentives: even when retail only liquidity is available at a superior price, there is no legal requirement that a retail order execute against it. We illustrate the role of protection using the IEX RLP, which differs from other RLP programs by only permitting midpoint pricing. Because midpoint is a specific and economically salient benchmark, the presence of IEX's RLP indicator provides a particularly clean environment to study whether retail executions reliably reflect contemporaneous retail only liquidity. We find that wholesalers sometimes internalize orders at prices inferior to the mid-quote available on the RLP: 75% of retail orders receive mid-quote or better when the IEX RLP flag is active, compared to 60% when it is not.[1] These observations are not legal trade-throughs, as the IEX RLP is not a protected quote, but they highlight how the absence of order protection can limit the competitive bite of a retail-only venue.[2] We also note that while some trades execute at worse than the midpoint, wholesalers offer *average* price improvement which is better than the midquote when the IEX RLP is active.

We exploit quasi-experimental variation from the SEC Tick Size Pilot to study how RLP usage changes with an exogenous change to market conditions. When trading increments are restricted, RLP sub-penny volume drops to near zero, confirming that the mechanism depends on sub-penny pricing flexibility. When only quoting increments are widened, however, RLP execution quality improves significantly—effective spreads decrease and price impact declines—while off-exchange effective spreads are unchanged. In the language of liquidity pecking orders (Menkveld, Yueshen, and Zhu (2017)), RLPs behave like an intermediate venue, having less adverse selection (and lower effective spreads) than exchange trades, yet more adverse selection (and higher effective spreads) than off-exchange trades.

To rationalize these findings, we develop a simple model of wholesaler routing. A wholesaler chooses between internalizing a retail order, earning spread revenue that depends on a privately observed execution cost, and routing it to the RLP, forgoing current revenue. Brokers evaluate wholesalers on execution quality and reallocate future order flow accordingly, creating a dynamic tradeoff between current spread revenue and the continuation value of future order flow. The optimal strategy takes a cutoff form: the wholesaler internalizes low-cost orders and routes high-cost orders to the RLP. This framework generates three predictions that match the data. First, internalized orders receive greater price improvement than RLP orders, because wholesalers selectively

[1]We focus on orders of 100 shares or less (which ensures the order size is less than the minimum IEX quantity) and confirm these results hold under the restriction that the IEX flag is active for at least one second prior to the trade, to account for potential latencies in order transmission and execution.

[2]These concerns may become even more pressing if the SEC adopts a recent proposal to eliminate order protection.

internalize when they can offer attractive prices. Second, worsening market conditions shift the cost distribution rightward, increasing the share of orders routed to the RLP. Third, the wholesaler may internalize even when the RLP offers a known superior price (such as the IEX mid-quote), because current-period profit can exceed the future order-flow gain from routing.

The insensitivity of off-exchange spreads to wider quoting increments has implications for the broader debate on retail execution quality. It indicates that wholesaler pricing reflects underlying execution costs rather than a mechanical markdown from the quoted spread, consistent with a theoretical framework with competitive wholesaler market in which spreads are determined by cost structure (Ernst, Malenko, Spatt, and Sun (2026)).

Taken together, our results have direct implications for market design proposals aimed at increasing competition for retail order flow. Nothing in the evidence suggests that on-exchange mechanisms, including a continuous double auction, are inherently unable to compete for segmented retail flow. Rather, our findings show that simply creating a venue is insufficient when the best retail-only prices are not observable, those prices are not protected, and the routing intermediary controls order composition. Evidence from the Tick Size Pilot demonstrates that when wider quoting increments expand the pricing space for sub-penny improvement, RLPs can deliver meaningfully better execution quality relative to off-exchange mechanisms. Venue design reforms that address the three frictions we identify could therefore make on-exchange competition for retail order flow viable.

The remainder of the paper is organized as follows. Section II reviews the related literature. Section III describes RLP institutional details and the data. Section IV develops the theoretical framework. Section V presents the empirical results in four subsections: price improvement and trade size (Section V.A), RLPs and market conditions (Section V.B), the IEX RLP and price visibility (Section V.C), and the Tick Size Pilot (Section V.D). Section VI concludes.

## II. Prior Literature and Contribution

Our paper contributes to the literature on market design for retail order flow. Comerton-Forde, Malinova, and Park (2018) show that Canada's trade-at rule increased exchange liquidity but reduced retail price improvement, while Foley and Putniņš (2016) find that trade-at rules in Canada and Australia effectively increased tick sizes without improving market quality—a finding that foreshadows the tick-size channel we exploit in our Tick Size Pilot analysis. Kwan, Masulis, and McInish (2015) argue that off-exchange sub-penny pricing undermines exchange competitiveness; RLPs address this by permitting sub-penny pricing within an exchange-based framework. A more sweeping proposal, the SEC's Order Competition Rule (Securities and Exchange Commission (2022)), would have replaced wholesaler internalization with order-by-order auctions but was formally withdrawn in June 2025; Battalio and Jennings (2023) identify its susceptibility to quote fading, and Ernst, Spatt, and Sun (2025) highlight the trade-off between allocative efficiency and competition.[3] Our paper uses the decade-long operating history of RLPs to study whether on-exchange competition for retail flow can succeed, and to identify the specific conditions under which it does.

A rapidly growing literature documents the economics of retail wholesaling. Dyhrberg, Shkilko, and Werner (2025) and Battalio and Jennings (2022) provide comprehensive evidence that wholesalers deliver substantial price improvement and compete actively for order flow, while Schwarz, Barber, Huang, Jorion, and Odean (2022) estimate that retail investors receive an average of 60% of the bid-

[3]As noted by Bishop (2022), RLPs offered a close existing parallel to the proposed auctions: both provided a mechanism for liquidity providers to compete for retail order flow, and in both cases wholesalers could bypass the mechanism by matching or improving the available price.

ask spread.[4] Earlier work identified RLP underperformance as an open question: O'Hara (2015) and Jain, Linna, and McInish (2021) highlighted the NYSE RLP as a potential venue for retail investors, and analogous programs have become important in Europe (Aramian and Comerton-Forde (2023)). Our paper provides the first systematic analysis of why U.S. RLPs have failed to compete, tracing the failure to specific, identifiable frictions rather than treating it as a single phenomenon. Complementary to our work, Chung, Rösch, and Zhang (2023) focus on the specific instances in which RLP trades co-occur alongside off-exchange trades, and present empirical findings that RLP trades occur between off-exchange midquote and off-exchange subpenny prices.

Our theoretical framework connects to models of order flow segmentation and the debate on wholesaler pricing. Easley, Kiefer, and O'Hara (1996) formalize cream-skimming through purchased order flow, and Battalio and Holden (2001) model internalization as profitable precisely because it segments retail from institutional flow. Baldauf, Mollner, and Yueshen (2022) analyze the welfare implications of segmentation from a portfolio perspective. Menkveld et al. (2017) develop a theory of venue hierarchy; our finding that RLPs occupy an intermediate position between wholesalers and exchanges in the price improvement ranking is consistent with this framework. On wholesaler pricing, van Kervel and Yueshen (2023) argue that wholesalers engage in anti-competitive reference pricing, benchmarking execution prices to the quoted spread. By contrast, Ernst et al. (2026) provide evidence that the wholesaler market is sufficiently competitive that execution quality reflects the underlying cost structure. Our Tick Size Pilot results support the competitive interpretation and extend it to the RLP context. The model formalizes the channel

[4] Related work includes Eaton, Green, Roseman, and Wu (2022) on how retail participation affects market quality, Parlour and Rajan (2003) on the welfare implications of payment for order flow, Hu and Murphy (2022) and Jain, Mishra, O'Donoghue, and Zhao (2020) on variation in execution quality across brokers, Bartlett, McCrary, and O'Hara (2022b) on odd-lot quote improvement, and Bartlett, McCrary, and O'Hara (2022a) on fractional share execution at mid-quote.

through which competitive pressure operates: brokers evaluate wholesalers on execution quality and reallocate future order flow accordingly (Ernst et al. (2026)), creating the incentive structure that generates both the price improvement ranking between internalized and RLP executions and the selective routing patterns we document.

# III. RLP Background and Summary Statistics

Retail Liquidity Programs allow segmentation of retail orders in a competitive limit order book setting. In RLPs, any market participant can submit limit orders, but these orders are available only to marketable orders from retail investors—defined as orders originating from a "natural person" and not from a trading algorithm or computerized methodology (SEC (2013)). RLPs are motivated by the premise that if retail order flow carries lower adverse selection risk, liquidity providers can price RLP orders more favorably than standard displayed quotes.

Orders in the RLP differ from standard exchange trading along four dimensions: access, pricing, visibility, and trading fees. For *access*, any exchange member can submit a Retail Price Improvement Order (RPI), similar to a standard limit order, but these orders are accessible only to incoming retail marketable orders. For *pricing*, RPI orders can be priced in tenth-of-a-penny increments, provided the order improves on the exchange's best bid or offer—a much finer grid than the one-penny minimum tick that governs standard exchange quotes. For *visibility*, RLP orders are non-displayed: both the price and size are hidden from market participants, with exchanges disseminating only an indicative flag showing whether at least one round lot of interest is available on each side of the book. The IEX RLP is the exception: its midpoint-only pricing renders the execution price observable, leaving only the available size hidden (Section V.C). Finally, *trading* *fees* for RLP transactions are substantially lower than for standard exchange trades, with some programs charging no fees to either side. As we discuss in Section III.A, the non-displayed status of RLP orders is a regulatory necessity with important consequences for program performance.

Incoming marketable retail orders trade against the best-priced resting interest: first against any mid-quote liquidity—whether regular hidden mid-quote interest or RLP mid-quote interest—then against RLP interest priced between the mid-quote and the displayed quote. Retail orders can be designated to interact only with RPI and non-displayed orders, or to sweep through RPI and non-displayed interest and then execute against displayed quotes up to the limit price.

## A. *Individual Program Details*

Four U.S. exchanges have launched Retail Liquidity Programs. The NYSE introduced the first program on August 1, 2012.[5] BATS and NASDAQ followed shortly after, with pilot programs approved on November 27, 2012 and February 15, 2013, respectively.[6] The Investors Exchange (IEX) launched its program on October 1, 2019. A fifth exchange, the Members Exchange (MEMX), applied to create an RLP but was denied by the SEC, as discussed below. While initially temporary pilot programs, the NYSE RLP received permanent approval on February 15, 2019; NASDAQ and BATS received similar permanent approvals.

All approved programs share two regulatory constraints that are central to our analysis. First, RLP orders are *non-displayed*: both the price and size of resting RLP interest are hidden from

[5]The introduction of the data field for the RLP led to the $400 million trading glitch at Knight Capital Group on the first day that the new data field was active.

[6]The surviving NASDAQ-family program operates on Nasdaq BX rather than the flagship venue: NASDAQ retired the Retail Price Improvement program on The Nasdaq Stock Market in December 2014, and on February 13, 2026 Nasdaq BX was reorganized as Nasdaq Texas, LLC, which now hosts the program. The BATS program is now operated by Cboe, which should be distinguished from Cboe's separate "Retail Priority" mechanism—a displayed-queue-priority model launched on Cboe EDGX in 2019—and from the full sub-penny Retail Price Improvement program that Cboe adopted on EDGX in April 2026.

market participants. This is not a design choice by the exchanges but a regulatory necessity. Regulation NMS requires all displayed quotes to be immediately accessible to all market participants; because RLP quotes are available only to retail orders, exchanges must hide them to satisfy the fair-access requirements of Rule 610.[7] Second, because RLP quotes are non-displayed, they receive *no trade-through protection* under Rule 611 of Regulation NMS. The Order Protection Rule prohibits trade-throughs only of protected quotations, which must be both displayed and immediately accessible. A wholesaler may therefore execute a retail order at a price inferior to available RLP interest without violating any routing obligation. These two constraints—non-display and the absence of trade-through protection—are regulatory consequences of the segmented-access model and have direct implications for how wholesalers interact with RLP liquidity.

The programs differ along several other dimensions. The NYSE, NASDAQ, and BATS programs allow Retail Price Improvement (RPI) orders to be priced in tenth-of-a-penny increments, provided the order improves on the exchange's best bid or offer. Trading fees are substantially lower than for standard exchange trades. The NYSE RLP charges no fee to qualifying retail market orders and pays a \$0.0003 per-share credit to Retail Liquidity Providers whose orders fill retail flow, provided the liquidity provider maintains resting RPI orders that improve the best bid or offer for at least 5% of the trading day; if a firm fails to meet the 5% threshold, it can still post RPI orders but forgoes the rebate.[8] The 5% quoting requirement distinguishes the NYSE program from those

[7]Exchanges do disseminate a limited indicative message through the SIP indicating whether there is at least one round lot of RLP interest on the bid side, the ask side, or both. This flag reveals the existence of RLP liquidity but neither its price nor its size.

[8]These two defining features—the \$0.0003 credit and the 5% quoting requirement—remain in force through 2026. The set of NYSE Group exchanges operating an RLP did shift after our sample period, however. NYSE Arca discontinued its Retail Liquidity Program (Rule 7.44-E) in tandem with the launch of a similarly structured program on NYSE National (Rule 7.44) on September 12, 2023. The NYSE National program follows the IEX design described below, accepting only midpoint interest rather than sub-penny-priced orders, and the SEC granted it a limited exemption from Rule 602 (Release No. 34-98206) to disseminate its retail liquidity identifier. NYSE Group later introduced a separate "Retail Attested Order" classification across its exchanges.

offered by NASDAQ and BATS, which impose no minimum quoting obligation and are therefore more accessible to firms that do not continuously make markets. The IEX RLP charges no trading fees to either side of the trade.

In approving the NASDAQ RLP, the SEC noted that "the Program might also create a desirable opportunity for institutional investors to interact with retail order flow that they are not able to reach currently. Today, institutional investors often do not have the chance to interact with marketable retail orders that are executed pursuant to internalization arrangements. Thus, by submitting RPI Orders, institutional investors may be able to reduce their possible adverse selection costs by interacting with retail order flow" (SEC (2013)). In practice, however, the real-time pricing, sub-penny quoting infrastructure, and adverse-selection management required to provide competitive RLP liquidity have limited participation largely to professional market-making firms. This concentration is not mandated by the program rules—any exchange member may submit RPI orders—but arises endogenously from the demands of the activity.

The IEX RLP differs from all other programs in a crucial respect: it accepts only midpoint-peg orders. Because the midpoint of the NBBO is publicly observable, the execution price of an IEX RLP order is effectively known even though the order itself is non-displayed—only the available size remains hidden. All midpoint-peg orders, whether or not they are designated as retail-only, enter the same time-priority queue.[9] IEX initially operated without RLP indicators but added indicative flags on October 13, 2021. Disseminating these flags required an approved exemption from SEC Rule 242.602, as the indicator reveals a specific price (the midpoint) and a minimum quantity (one round lot), yet the underlying interest is not accessible to non-retail orders.[10] The

[9]Midpoint orders, whether RLP or not, have queue priority over the IEX D-limit order, a discretionary limit order whose price IEX's proprietary algorithm adjusts whenever it forecasts a potentially disadvantageous change in quotes.

[10]This midpoint-only design is unchanged as of 2026; a 2023 revision granted retail interest orders priority over other non-displayed midpoint orders when a retail counterparty is present, and formalized the retail liquidity identifier

IEX program provides a natural setting to test whether greater price observability improves RLP outcomes, and we return to this analysis in Section V.

The MEMX application illustrates the regulatory constraints that bind these programs. Under the MEMX proposal, incoming retail market orders would interact with hidden RPI orders before hidden non-RPI orders at the same price level, even if the non-RPI orders had time priority. MEMX argued that because hidden RPI orders contribute to the dissemination of the indicative interest indicator, they should receive priority analogous to that of displayed orders over hidden orders at the same price. The SEC disagreed, ruling that the proposed change would violate Section 6(b)(5) and Section 11A of the Exchange Act.[11]

Several regulatory developments since our sample period bear on the economics of RLPs. The SEC's proposed Order Competition Rule (Rule 615), which would have routed marketable retail orders into order-by-order auctions, was formally withdrawn on June 12, 2025, together with the companion Regulation Best Execution proposal.[12] The SEC instead finalized two narrower reforms from its December 2022 market-structure package. First, amendments to Rule 605 (adopted March 6, 2024; compliance date August 1, 2026) substantially expand the order-execution-quality statistics that market centers and large broker-dealers must disclose, including odd-lot- and fractional-share-aware effective and realized spreads. Second, amendments to Regulation NMS (adopted September 18, 2024; Release No. 34-101070) introduce a half-penny ($0.005) minimum quoting

that IEX disseminates when at least one round lot of midpoint interest is available.

[11] We previously dated this denial to February 14, 2022, which was in fact a procedural notice designating a longer period to act on Amendment No. 1. The operative disapproval was issued under delegated authority by the Division of Trading and Markets on May 6, 2022 (Release No. 34-94866). The Commission then noticed the matter for its own review in February 2023, and as of mid-2026 had issued no final decision; MEMX has not re-filed a retail program in the interim.

[12] The withdrawn Rule 615 would have required exchanges to give auction responses priority over hidden limit orders—the same priority structure that the SEC denied to MEMX in the RLP context (Securities and Exchange Commission (2022), Proposed Rule 615 Section IV C.5.).

increment for the most tick-constrained stocks and reduce the access-fee cap from thirty mils to ten mils per share.[13] The tick-size change compresses the sub-penny price-improvement gap that RLPs monetize in exactly the tick-constrained securities where RLP activity concentrates, while the lower fee cap pressures the rebates that fund RLP liquidity provision. Importantly, the final rule did not adopt the proposed prohibition on sub-penny price improvement, so the RLPs retain their Rule 612 exemptions and continue to operate.[14]

## B. *Data and Summary Statistics*

We use NYSE TAQ data from January 1, 2019 to May 30, 2022, covering all securities in the Russell 3000 index and the 100 most actively traded ETFs with prices above $1 per share. RLP indicators are distributed through the SIP and recorded in TAQ. Because an exchange may disseminate an indicator even when its posted best bid or offer does not coincide with the NBBO, we obtain indicator flags from the TAQ Quotes file.

Despite their non-displayed nature, RLP interest is available for a substantial portion of the trading day. Figure 1 plots the percentage of time, by asset class, that at least one-sided RLP interest is posted. For ETFs, resting RLP orders are present 50 to 75% of the trading day. For Russell 1000 stocks, NYSE, CBOE, and NASDAQ each have RPI interest for over 20% of the day; for Russell 2000 stocks, CBOE and NASDAQ exceed 20%. Differences across assets partly reflect program rules but also listing-exchange advantages: NYSE Arca, for example, has RPI

[13]These amendments were upheld by the D.C. Circuit on October 14, 2025. Compliance with the tick-size and access-fee provisions was subsequently deferred to the first business day of November 2026.

[14]Among newer venues, the SEC approved the 24X National Exchange, the Texas Stock Exchange, and the Green Impact Exchange in 2024–2025; of these, only 24X has so far introduced retail-order pricing aimed at on-exchange price improvement. With Rule 615 withdrawn, the live regulatory question for these programs is the SEC's reconsideration of the order-protection (trade-through) rule, Rule 611, which was the subject of Commission roundtables in September and December 2025.

interest less than 20% of the day for individual stocks but over 75% for ETFs, consistent with Arca's strength in ETF listings.

**Figure 1. Time Share of Retail Liquidity Programs**. We plot the average time that an RLP indicator is active, measured as percentage of time active out of the total trading day. Our sample can be divided into three groups: stocks in the Russell 1000 index, stocks in the Russell 2000 index, and ETFs from our sample. We provide a fourth (overlapping) group, "Tick Constrained", comprised of any stock or ETF which meets the criteria of having a quoted bid-ask spread of one penny at least 50% of the day for at least one-third of the days of our sample.

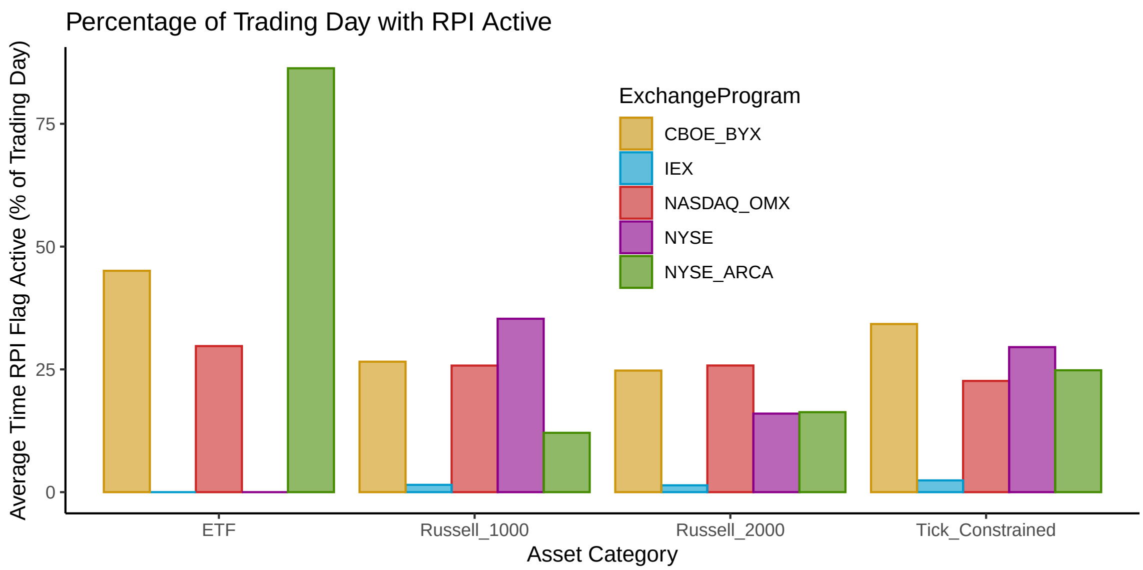
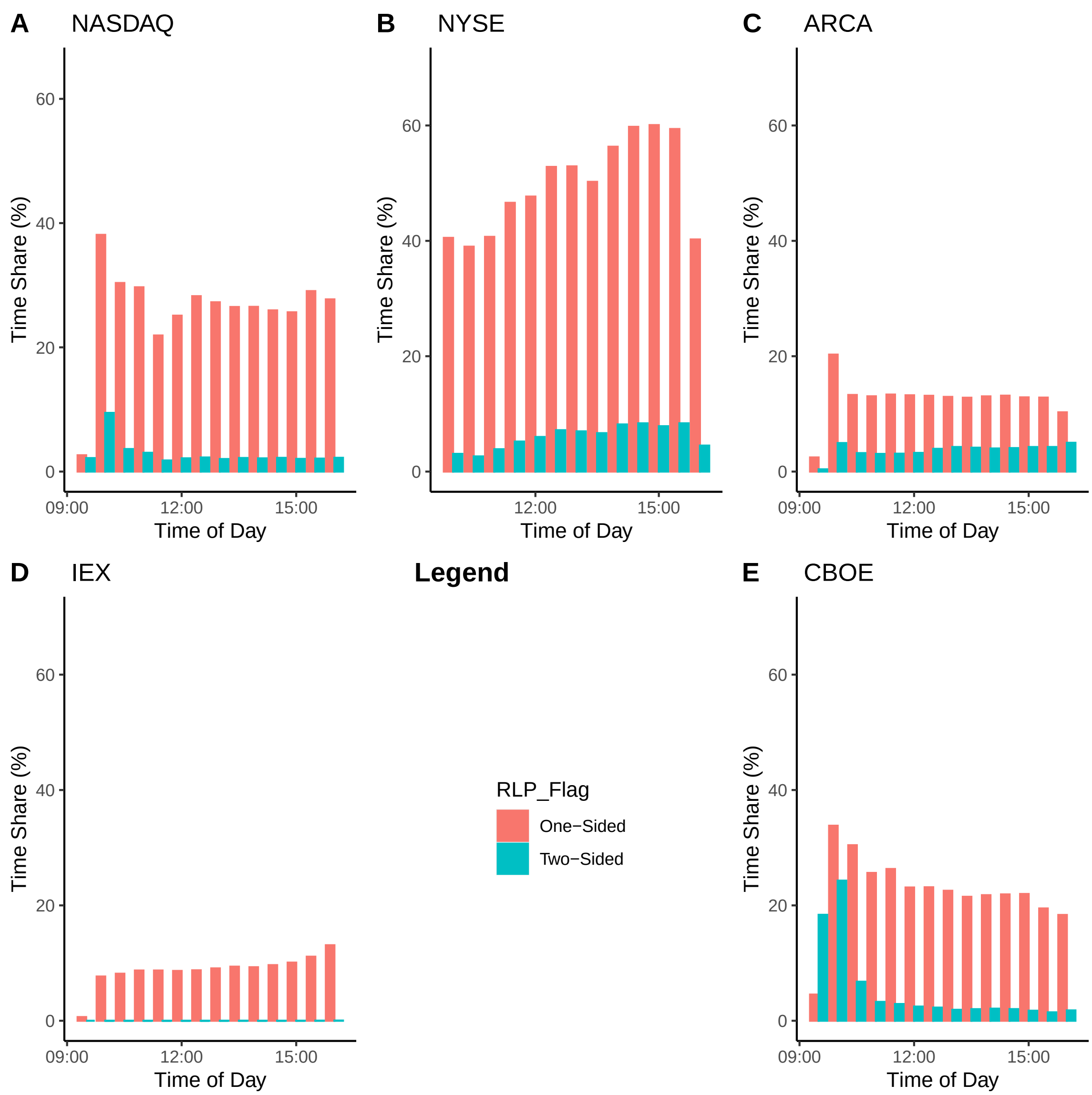



Figure 2 highlights variation across programs throughout the trading day, with some programs showing a gradual increase in liquidity throughout the day while others are more active in the morning, and have less liquidity later in the day. Across each time interval, the IEX RLP is active for a notably smaller percentage of time relative to any competing RLPs, as the IEX RLP requires orders to be placed at mid-quote, while competitor programs only require a minimum of 10 mils of improvement relative to the NBBO. The RLP flags do not display the exact amount of price improvement available, nor do they display any indication of the size available, with the flag only indicating whether there is at least one round lot. The IEX RLP flag is an exception on price: because IEX accepts only mid-quote orders, an active flag implies the available price is the midpoint, though the size remains undisclosed.

**Figure 2. Intra-day Time Share**. We plot the average percent of time that the RLP Flag is active throughout the trading day on January 3, 2022. For each exchange, we divide the trading day into 30-minute intervals and calculate the average across stocks of the percentage of time for which the RLP Flag is active. One-sided liquidity is the percentage of time for which there is a quote on either the bid, the ask, or both, and therefore includes the time for which there is two-sided liquidity (i.e., a flag indicating RLP interest on both the bid and ask at the same time).




Despite this extensive availability, very little volume actually executes through RLPs. Figure 3 depicts the trading volume split when the exchange's RPI Flag is active and when it is not. Sub-penny executions account for less than 1% of total exchange volume even when RLP interest is available.[15] Exchange RLP trades comprise just 0.3% of all ETF trading volume and 0.1% of trading in Russell 1000 or Russell 2000 stocks. Table I presents total volumes executed when RLP programs are active and when they are not.[16]

On-exchange mid-quote volume is considerably higher when the RPI Flag is active but still represents less than 2% of total trading volume. This mid-quote volume mixes retail RLP interest—including from the IEX RLP, which accepts only mid-quote orders—with non-RLP hidden mid-quote liquidity. The vast majority of mid-quote and sub-penny trading occurs off-exchange; when RPI Flags are active, a larger share of off-exchange volume executes at sub-penny or mid-quote prices.

The total volume share of RLP trades is stable over our sample period. As Panel A of Figure 4 shows, on-exchange sub-penny retail trades average below 0.2% of total volume for Russell 1000 and Russell 2000 stocks. The volume share is slightly higher for ETFs and tick-constrained stocks, fluctuating between 0.2% and 0.5%.[17] Panel B of Figure 4 broadens the measure to include all exchange sub-penny or mid-quote executions while the RPI Flag is active. This captures both non-retail hidden mid-quote liquidity and retail trading through the IEX RLP, which accepts only mid-quote orders. Under this broader measure, ETFs and tick-constrained stocks reach 0.5% to 1.0% of total trading volume. While small in absolute terms, these figures represent a meaningful fraction of retail-only volume: Morgan Stanley estimates retail trading at approximately 10% of

[15] Exchange sub-penny volume when there is no RPI Flag is small but non-zero. This arises when an odd-lot limit order rests in the RLP program: the flag is not disseminated (as there are fewer than 100 shares of interest) but retail market orders can still execute against this odd-lot volume. Another possible explanation is inaccuracy in timestamp-based matching of the sort described by Schwenk-Nebbe (2021), who show that the exchange processing and dissemination of quotes is typically several microseconds faster than that of trades.

[16] To exclude biases from incorrect interpretation of fractional shares, as documented by Bartlett et al. (2022a), we exclude any orders for exactly 1 share, as these may be orders for a fractional share reported as 1 share.

[17] We define a stock as tick-constrained if it has a one-penny bid-ask spread for at least 50% of the trading day on at least one-third of the trading days in our sample. For these stocks, competition for marketable orders is heightened by the tick constraint, increasing the incentive to provide liquidity in an RLP.

**Table I: Summary Volumes By Each Price Increment.** This table presents summary total trading volume (in billions of dollars) in our sample for each sub-penny category of trade: at-quote, mid-quote, and sub-penny. Our sample can be divided into three asset groups: stocks in the Russell 1000 index, stocks in the Russell 2000 index, and ETFs from our sample. We provide a fourth (overlapping) asset group, "Tick Constrained," comprised of any stock or ETF which meets the criteria of having a quoted bid-ask spread of one penny at least 50% of the day for at least one-third of the days of our sample.

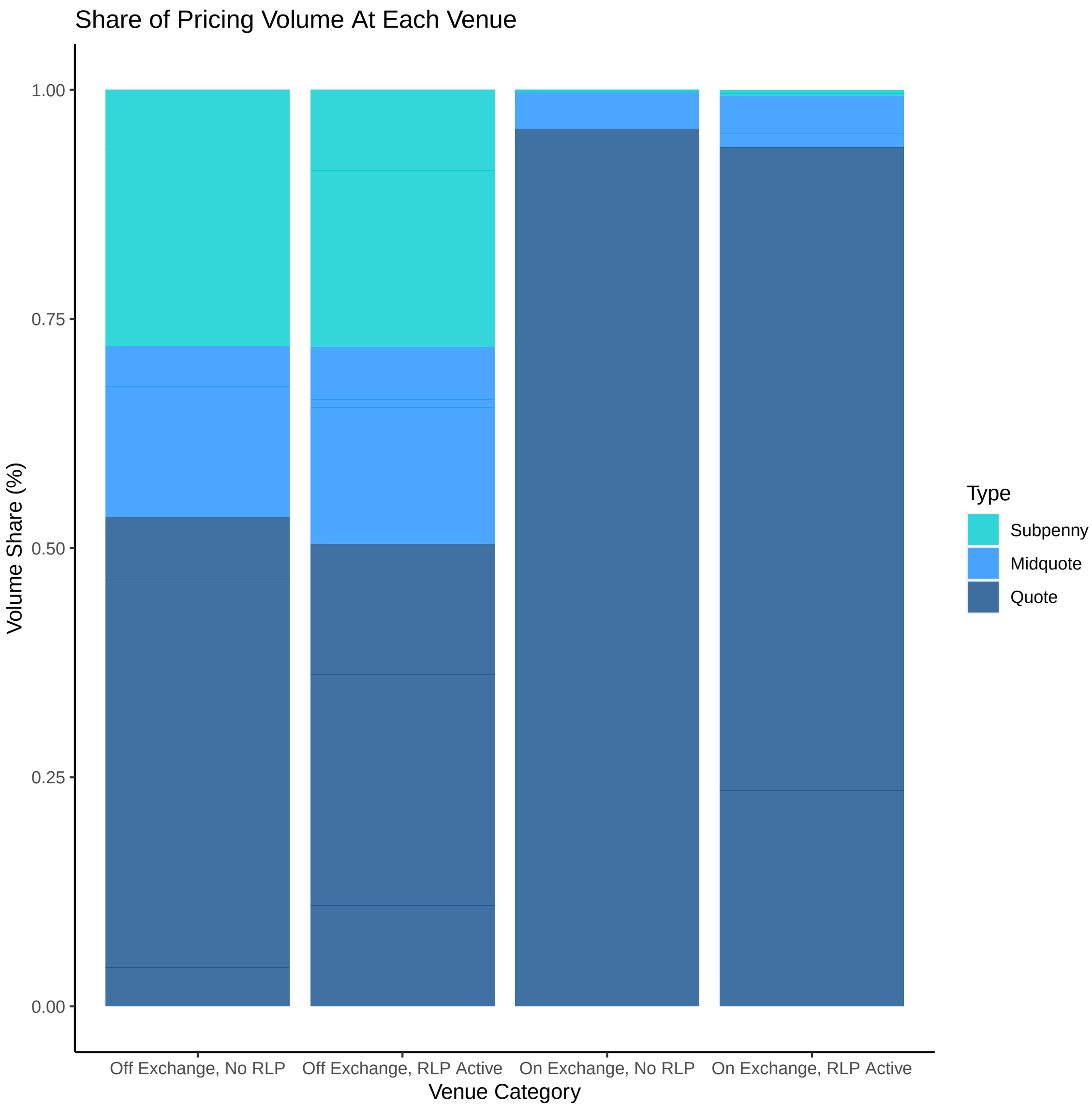
**Panel A: Exchange Trades**. We define a flag as 'active' if there is contemporaneous RLP interest at the exchange where trade occurs.

| Asset | RLP | Volume | | | Percent | | |
|---|---|---|---|---|---|---|---|
| | Flag | mid-quote | At Quote | Subpenny | Mid-quote | At Quote | Subpenny |
| ETF | Active | 2259 | 35868 | 414 | 3.2 | 51 | 0.6 |
| ETF | None | 317 | 9160 | 55 | 0.5 | 13 | 0.1 |
| Russell_1000 | Active | 3162 | 60890 | 235 | 1.7 | 32 | 0.1 |
| Russell_1000 | None | 1933 | 50111 | 99 | 1.0 | 27 | 0.1 |
| Russell_2000 | Active | 283 | 5376 | 13 | 1.5 | 28 | 0.1 |
| Russell_2000 | None | 306 | 6012 | 2.6 | 1.6 | 31 | 0.01 |
| TickConstr. | Active | 2767 | 40510 | 400 | 3.3 | 48 | 0.5 |
| TickConstr. | None | 676 | 12773 | 67 | 0.8 | 15 | 0.1 |

**Panel B: Off-Exchange Trades**. We define a flag as 'active' if there is contemporaneous RLP interest disseminated by any exchange.

| Asset | RLP | Volume | | | Percent | | |
|---|---|---|---|---|---|---|---|
| | Flag | mid-quote | At Quote | Subpenny | Mid-quote | At Quote | Subpenny |
| ETF | Active | 3626 | 9070 | 5751 | 5.2 | 13 | 8.2 |
| ETF | None | 537 | 1909 | 1134 | 0.8 | 2.7 | 1.6 |
| Russell_1000 | Active | 8638 | 20795 | 10127 | 4.6 | 11 | 5.4 |
| Russell_1000 | None | 5895 | 16953 | 8756 | 3.1 | 9.0 | 4.7 |
| Russell_2000 | Active | 743 | 2161 | 822 | 3.8 | 11 | 4.2 |
| Russell_2000 | None | 723 | 2107 | 833 | 3.7 | 11 | 4.3 |
| TickConstr. | Active | 4740 | 9643 | 6456 | 5.6 | 12 | 7.7 |
| TickConstr. | None | 1257 | 3079 | 1869 | 1.5 | 3.7 | 2.2 |

**Figure 3. Volume Share of Venues**. We plot the percentage of volume which executes either at the quote, at the mid-quote, or at a sub-penny price for both on-exchange and off-exchange venues. On both types of venues, a higher percentage of volume occurs at the quote when there is no RPI Flag active, and a higher share of volume executes at sub-penny and mid-quote prices when the RPI Flag is active.




total U.S. equity volume (Adinarayan (2021)), implying that up to 1 in 10 retail trades in liquid stocks or ETFs executes through an RLP, falling to fewer than 1 in 20 for less liquid stocks.[18]

[18]These numbers represent an upper bound estimate, as midquote executions may include non-retail executions on exchange. Using only on-exchange sub-penny volume, which is only retail under current rules, establishes a lower bound estimate of 1 in 20 to 1 in 50 for ETFs and tick-constrained stocks (using the RLP sub-penny share of 0.2 to 0.5%).

# IV. Theoretical Framework

This section develops a simple model of the wholesaler's choice between internalization and routing to an RLP. The model incorporates three institutional features: RLP orders are non-displayed, so the wholesaler generally cannot observe the execution price available in the RLP before routing (the IEX RLP, whose midpoint-only pricing makes its price observable, is the exception we examine in Section V.C); RLP orders are not protected quotations; and the wholesaler has full discretion over whether to internalize or route each order. The model generates three testable predictions, each examined in the following sections: internalized orders deliver greater average price improvement than RLP orders, RLP execution share increases in high-cost market conditions, and wholesalers internalize some orders even when the RLP offers the best available execution price.

All prices are in effective-spread units; trade direction is suppressed. There are two periods, $t \in \{1,2\}$. The NBBO half-spread $\bar{s} > 0$ is constant across periods. An execution at effective spread $s_t \in [0,\bar{s}]$ delivers price improvement $\bar{s} - s_t$ to the retail trader.

**Internalization.** At $t = 1$, the wholesaler receives a retail order and privately observes an execution cost $c \sim F$ with support $[0,\bar{c}]$, where $0 < \bar{c} < \bar{s}$. The cost $c$ captures adverse selection, inventory, and hedging costs. If the wholesaler internalizes, it chooses an execution spread $s \in [0,\bar{s}]$ and earns current-period profit $s - c$.

Execution quality determines future order flow: a spread of $s$ today generates continuation order flow $\mu(s)$ at $t = 2$, with $\mu'(s) < 0$ (better execution attracts more flow). This captures the broker-evaluation channel through which execution quality disciplines wholesalers (Ernst et al.

(2026)). The execution cost is persistent; for simplicity, the period-2 cost is also $c$.[19] Since $t = 2$ is terminal, the wholesaler always chooses $\bar{s}$ at time $t = 2$ and earns $\bar{s} - c$ per order. The total internalization payoff is therefore

$$U(s,c) = \mu(s)(\bar{s} - c) + s - c. \tag{1}$$

The term $s - c$ is the current-period profit. The term $\mu(s)(\bar{s} - c)$ is the continuation value: $\mu(s)$ future orders, each at margin $\bar{s} - c$.

**RLP routing.** The RLP execution price $s_R$ is drawn from a distribution $G$ on $[0, \bar{s}]$ and is independent of $c$. Because RLP liquidity is not displayed, the wholesaler cannot observe the realization of $s_R$ before the routing decision. Routing to RLP forgoes current-period profit but retains the continuation value generated by the resulting execution quality. Since $s_R$ is unobserved at the time of routing, the expected continuation flow is $\mathbb{E}[\mu(s_R)]$, and the routing payoff is

$$W(c) = \mathbb{E}[\mu(s_R)](\bar{s} - c). \tag{2}$$

**Routing decision.** At $t = 2$, the wholesaler always internalizes. At $t = 1$, the wholesaler internalizes if and only if

$$V(c) \equiv \max_{s \in [0, \bar{s}]} U(s,c) \geq W(c). \tag{3}$$

**Assumption 1** (Interior solution)**.** The function $\mu$ is twice continuously differentiable, nonnegative,

[19] Perfect persistence is adopted for tractability. An AR(1) cost process yields the same qualitative results.

strictly decreasing, and strictly concave on $[0,\bar{s}]$:

$$\mu(s) \geq 0, \qquad \mu'(s) < 0, \qquad \mu''(s) < 0.$$

In addition, the parameters satisfy the boundary restrictions

$$1+\mu'(0)\bar{s} > 0, \qquad 1+\mu'(\bar{s})(\bar{s}-\bar{c}) < 0. \tag{4}$$

The concavity assumption is a tractability condition. It says that the loss of future order flow from worsening execution quality becomes larger as the execution approaches the NBBO benchmark. The boundary restrictions guarantee an interior solution for the internalization price for every cost realization $c \in [0,\bar{c}]$.

**Lemma 1.** *Under Assumption 1, for each $c \in [0,\bar{c}]$ there is a unique interior optimal internalization price $s^*(c)$ satisfying*

$$1+\mu'(s^*(c))(\bar{s}-c) = 0. \tag{5}$$

*Moreover, $s^*(c)$ is strictly increasing in $c$:*

$$\frac{ds^*(c)}{dc} = \frac{\mu'(s^*(c))}{\mu''(s^*(c))(\bar{s}-c)} > 0. \tag{6}$$

*Equivalently, the price improvement offered by internalization, $\bar{s}-s^*(c)$, is strictly decreasing in $c$.*

**Assumption 2.** $V(\bar{c}) < W(\bar{c})$.

Assumption 2 requires that at the highest cost realization, the wholesaler strictly prefers routing to the RLP. This ensures that routing to RLP is not strictly dominated.

**Proposition 1** (Cutoff routing)**.** Under Assumption 1 and Assumption 2, there exists a unique cutoff $c^\dagger \in (0, \bar{c})$ such that

$$V(c^\dagger) = W(c^\dagger). \tag{7}$$

The wholesaler internalizes if and only if $c \leq c^\dagger$ and routes the order to the RLP if and only if $c > c^\dagger$.

**Corollary 1** (Selection into the RLP)**.** The probability that an order is routed to the RLP is

$$\Pr(\text{RLP}) = 1 - F(c^\dagger). \tag{8}$$

If a market state, such as volatility, shifts the distribution of $c$ upward in the sense of first-order stochastic dominance, then the RLP share increases.

We test this prediction in Section V.B, where RLP execution share rises with spreads, volatility, and other proxies for market-making cost.

The cutoff also generates a price-improvement ranking once the unobserved RLP price is linked to the marginal internalization price. Define internalized price improvement as

$$PI^I(c) = \bar{s} - s^*(c), \tag{9}$$

and RLP price improvement as

$$PI^R = \bar{s} - s_R. \tag{10}$$

**Proposition 2.** If

$$-\mu'\big(\mathbb{E}[s_R]\big)\,\big(\bar{s}-\bar{c}\big) > 1, \tag{11}$$

then

$$\mathbb{E}[PI^I \mid \text{internalized}] > \mathbb{E}[PI^R \mid \text{RLP}]. \tag{12}$$

Proposition 2 shows that when the broker's performance-based order-flow allocation is sufficiently sensitive to execution quality, internalized orders deliver greater average price improvement than RLP orders. The intuition is straightforward: condition (11) ensures that at $\mathbb{E}[s_R]$, the marginal loss of future order flow deters even the highest-cost internalizer from quoting that spread. Every internalized order therefore receives a spread below $\mathbb{E}[s_R]$, and because RLP execution quality is independent of $c$, the ranking follows. Note that the condition is sufficient, not necessary. Section V.A tests this prediction: internalized orders deliver approximately twice the price improvement of RLP orders.

**Corollary 2** (Internalization despite zero-spread RLP execution)**.** Suppose $s_R = 0$ almost surely, so that

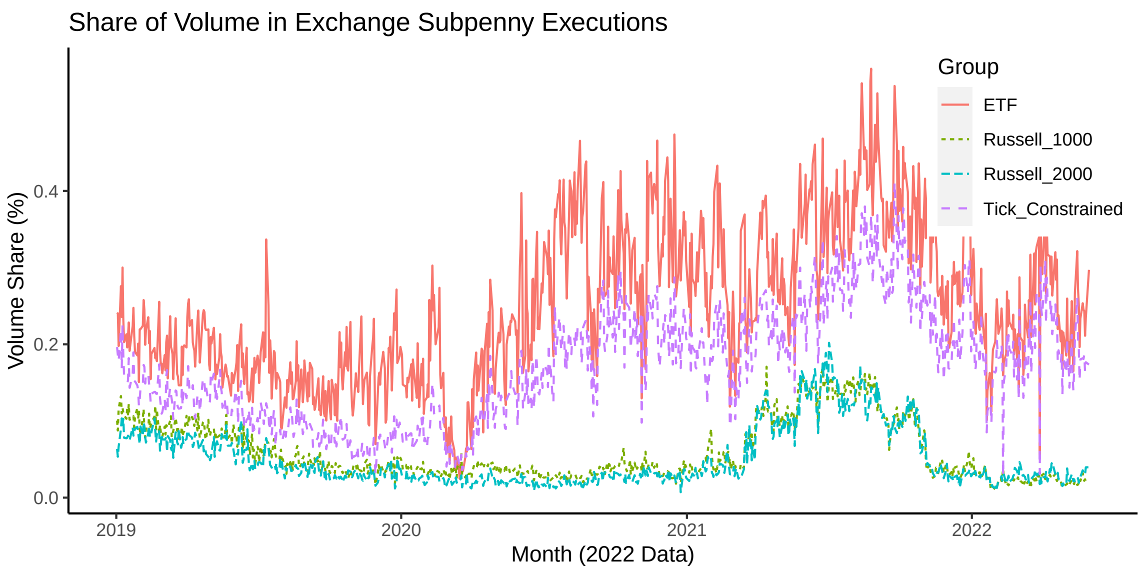
$$W(c) = \mu(0)(\bar{s}-c).$$

There still exists a cutoff $c^\dagger \in (0,\bar{c})$ such that the wholesaler internalizes if and only if $c \leq c^\dagger$. The wholesaler therefore internalizes some orders even when the RLP always provides the best possible execution price.

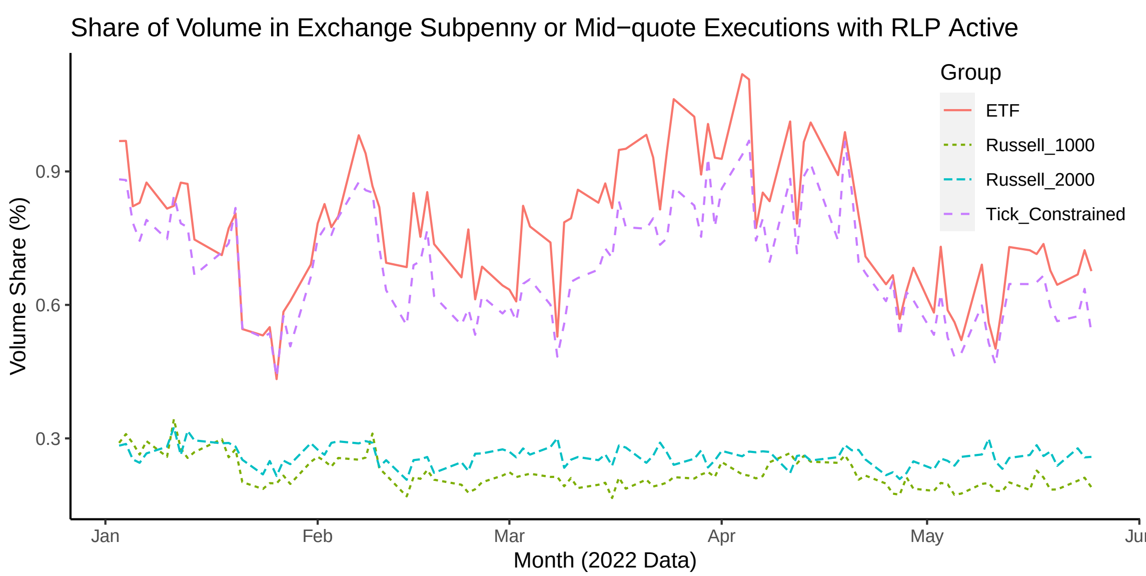
Internalization allows the wholesaler to earn current-period profit by quoting a positive spread. For low-cost orders, the optimal spread $s^*(c)$ is also very small, so the loss of future order flow relative to the RLP's perfect execution is small. The boundary restriction $1+\mu'(0)\bar{s} > 0$ guarantees that at $s = 0$, the marginal spread revenue exceeds the marginal continuation-value loss, making internalization strictly preferred to routing. Section V.C tests this prediction directly: wholesalers internalize orders even when the IEX RLP quotes a price that would deliver superior execution.

**Figure 4. Volume Share of Sub-penny Retail Liquidity Programs**. For each day, we plot the share of volume which executes in a retail liquidity program, compared to total volume. Our sample can be divided into three groups: stocks in the Russell 1000 index, stocks in the Russell 2000 index, and ETFs from our sample. We provide a fourth (overlapping) group, "Tick Constrained", comprised of any stock or ETF which meets the criteria of having a quoted bid-ask spread of one penny at least 50% of the day for at least one-third of the days of our sample. Panel A presents the volume share of only exchange sub-penny executions while a RLP indicator is active, while Panel B presents the volume share of all exchange sub-penny or mid-quote executions while a RLP indicator is active.

**Panel A:** Volume Share for Exchange Sub-penny (Non-Mid-quote) Executions with an RPI Flag Active




**Panel B:** Volume Share for Exchange Sub-penny or Mid-quote Executions with an RPI Flag Active



# V. RLP Performance

Despite having quoted liquidity available for a large portion of the trading day, the comparatively small trading volume of RLPs has raised questions about their effectiveness. In its now-withdrawn Order Competition Rule, the SEC had proposed scrapping all non-IEX RLPs, noting that "exchanges would not be allowed to operate a mechanism limited, in whole or in part, to segmented orders" Securities and Exchange Commission (2022).[20] The SEC's economic analysis suggested that RLPs failed because wholesalers "lack the incentives to route the individual investor order flow with lower adverse selection risk" Securities and Exchange Commission (2022). We investigate this claim and find a more nuanced picture: multiple distinct forces shape RLP performance, and the patterns documented below are jointly rationalized by the model of wholesaler routing developed in Section IV.

Section V.A documents that RLPs deliver less price improvement on average than off-exchange wholesalers. This gap is consistent with selective routing: wholesalers internalize their most profitable orders and route higher-cost orders to RLPs, so the price improvement ranking reflects order composition rather than inherent venue inferiority. The model formalizes this through a cost-based cutoff in which competitive discipline from brokers ensures that internalized orders receive a spread below the expected RLP spread. Non-displayed prices and limited visible size further constrain the attractiveness of RLPs as a routing destination.

Section V.B examines when RLP activity increases. RLP volume share rises with volatility and wider spreads while off-exchange share declines, consistent with wholesalers routing more orders past the cutoff as execution costs increase. On the supply side, RLP liquidity providers concentrate

[20]See page 139, Question 34 Securities and Exchange Commission (2022).

their quotes in calm periods with balanced order flow.

Section V.C focuses on the IEX RLP, whose mid-quote-only design makes the RLP price fully transparent. Using proprietary data from a major retail brokerage, we find that wholesalers sometimes internalize orders at prices worse than mid-quote even when the IEX RLP offers mid-quote execution. Because RLPs are not protected quotations, wholesalers can trade through the IEX RLP without matching its price. This pattern is consistent with the model: internalization generates current-period profit that makes it worthwhile even when the RLP offers superior execution.

Finally, Section V.D uses the SEC Tick Size Pilot as a natural experiment. When trading increments are restricted, RLP activity drops to near zero, confirming that the RLP mechanism depends critically on sub-penny pricing flexibility. When only quoting increments are widened, RLP execution quality improves significantly while off-exchange effective spreads are unchanged, consistent with the wholesaler market being sufficiently competitive that spreads reflect execution costs rather than the quoted spread benchmark (Ernst et al. (2026)). The differential response further indicates that wholesalers adjust the composition of orders they route to RLPs in response to changes in market structure.

## A. *Price and Size in RLP Trades*

All U.S. brokers have a best-execution duty to obtain the best prices possible for their clients. In practice, brokers evaluate the venues they use and route more future order flow to venues delivering better prices on average.[21] Retail marketable orders are routed almost exclusively to off-exchange wholesalers (Battalio and Jennings (2022)). Wholesalers then control subsequent routing: they

[21] Ernst et al. (2026), Dyhrberg et al. (2025), and Huang, Lee, Jorion, and Schwarz (2023) provide further information and analysis of how brokers obtain best execution for customers.

may internalize a given order—filling it out of their own inventory—or externalize it by routing to an RLP, alternative trading system, or exchange. Wholesalers are evaluated on the performance of the entire bundle of orders they receive, regardless of how individual orders are handled. Orders that arrive at RLPs are therefore typically orders that wholesalers have chosen to re-route, rather than orders sent directly by brokers.

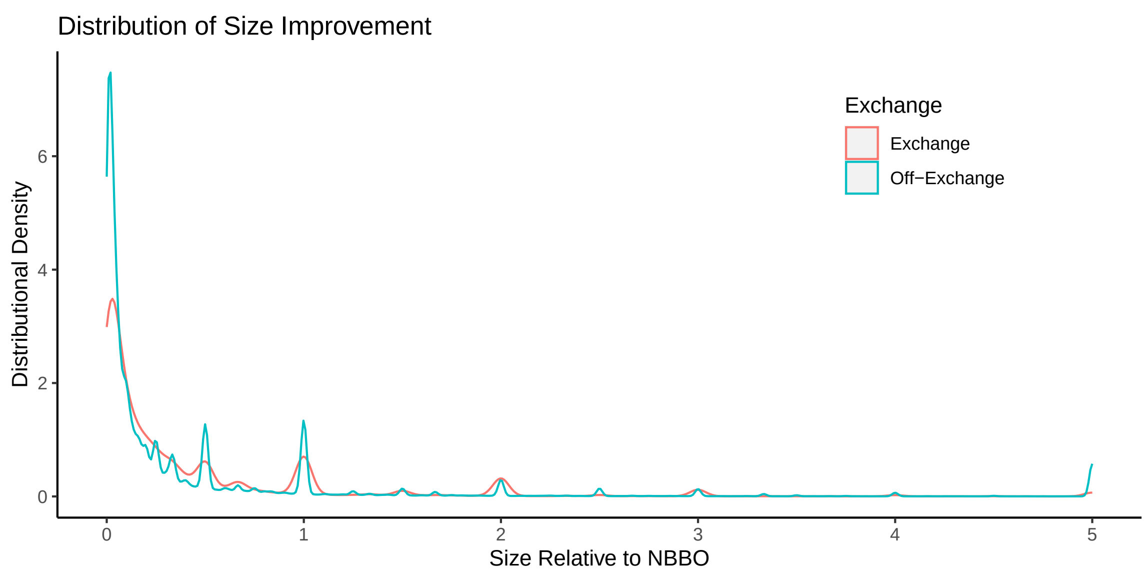
RLPs deliver less price improvement on average than off-exchange wholesalers. Across our sample, the average sub-penny price improvement in RLP trades is 11 mils, compared to 19 mils for off-exchange sub-penny trades, though some off-exchange trades may not be retail. This gap helps explain why brokers route almost exclusively to wholesalers rather than directly to exchange RLPs.[22] Using SEC Rule 605 reports, Dyhrberg et al. (2025) estimate that for retail trades in S&P 500 stocks, wholesalers offer on average 27% of the quoted half-spread as price improvement, while Schwarz et al. (2022) estimate retail investors receive an average of 60% of the bid-ask spread, though their sample is tilted toward smaller orders. This price improvement gap is consistent with two interpretations. First, wholesalers may selectively route their higher-cost orders to RLPs while internalizing the more profitable ones—a cream-skimming channel in which the gap reflects order composition rather than venue quality. Second, RLPs may simply deliver worse execution due to their structural limitations, including non-displayed prices and limited available size. The evidence in this subsection cannot disentangle the two channels; we return to this question in Section V.C, where the IEX RLP's observable midpoint price allows us to rule out a pure venue-quality explanation.

The non-displayed nature of RLPs also limits visibility into available size. Because resting

[22] Within RLP trades, we estimate sub-penny improvement at 11.0 mils across all trades and 10.8 mils among S&P 500 trades. Given that around 80% of RLP volume occurs in S&P 500 stocks, the 27% of the quoted half-spread benchmark from Dyhrberg et al. (2025) is the more appropriate comparison.

quantities are hidden, we can only infer available liquidity from realized trade sizes: a sub-penny exchange trade confirms that at least that many shares were available. The majority of observed RLP trades are for odd lots, with very few for 200 shares or more. Figure 5 depicts the distribution of order sizes for on-exchange and off-exchange sub-penny trades, expressed as a fraction of the size available at the NBBO. While odd-lot trades are common at both venues, a far larger share of off-exchange sub-penny trades exceed the available size at the best bid or offer: over 2.1% of off-exchange sub-penny trades are for more than five times the NBBO size, compared to only 0.7% of on-exchange sub-penny trades.

**Figure 5. Size Distribution**. We plot the distribution of order sizes, as a percentage of the NBBO, for all sub-penny trades occurring in the stocks of our sample on January 3, 2022. We truncate the distribution of orders at 5 times the NBBO. Of all sub-penny trades, 2.1% of all off-exchange sub-penny trades are larger than five times the NBBO, while 0.7% of on-exchange sub-penny trades are larger than five times the NBBO.




Taken together, these facts place RLPs between off-exchange wholesalers and regular exchange trading in the price improvement hierarchy: RLPs deliver less price improvement than wholesalers but more than standard exchange execution. This ordering is consistent with a pecking order of venues (Menkveld et al. (2017)), with RLPs serving as high-cost, high-immediacy venues that wholesalers turn to when internalization is unattractive. Non-display compounds the difficulty: a wholesaler routing to an RLP can expect only about 11 mils of price improvement and limited size, even when better-priced, deeper liquidity is in fact available—but the wholesaler cannot observe it. The next subsection examines *when* RLP volume increases, providing further evidence on the selection channel.

## *B. RLPs and Market Conditions*

In filing for SEC approval, the NYSE highlights that segmentation in the Retail Liquidity Program "is intended to allow retail order flow to receive potential price improvement as a result of their order flow being deemed more desirable by liquidity providers" SEC (2023). Whether this premise holds depends on the market conditions under which RLP liquidity is available and the composition of orders that reach it.

Table II presents summary statistics on the price impact of orders. Exchange RLP orders carry similar price impact to off-exchange orders: when trades are signed with the Lee-Ready algorithm, exchange RLP orders have a mean price impact of 0.17 basis points at 30 seconds, compared to 0.22 basis points for off-exchange trades. When trades are signed using sub-penny prices, as suggested by Schwarz et al. (2022), the mean price impact of exchange RLP trades is 0.16 basis points, comparable to the 0.18 basis point mean of off-exchange sub-penny trades. The similarity between RLP and off-exchange price impacts indicates that the retail orders reaching RLPs carry comparable adverse selection to any off-exchange trades executed in sub-pennies. These observed price impacts, however, are endogenous to the composition of orders that reach RLPs and the market conditions under which RLP liquidity is available. We next examine the determinants of

RLP activity, both across assets and over time.

**Table II: Price Impact of Trades**. We present summary statistics on price impacts, which measure the signed difference between mid-quote following the trade and the mid-quote at the time of the trade, for all trades on January 3, 2022. We divide trades based on whether they occur on or off-exchange, and whether they occur at the mid-quote, sub-penny increment, or an even-penny increment. Price impacts are measured at the 1-second and 30-second intervals.

Panel A: Price Impact at 1 Second (in Basis Points) signed with Lee-Ready

| | | Standard | Quantiles | | | | |
|---|---|---|---|---|---|---|---|
| Group | Mean | Dev | 5% | 25% | 50% | 75% | 95% |
| Exchange RLP | 0.06 | 0.68 | -0.63 | 0.00 | 0.00 | 0.10 | 1.00 |
| Exchange | 0.48 | 57.81 | -1.25 | -0.00 | -0.00 | 0.53 | 4.13 |
| Off-Exchange Mid-Quote | -0.01 | 1.15 | -0.75 | 0.00 | 0.00 | 0.00 | 0.69 |
| Off-Exchange Sub-penny | 0.04 | 1.17 | -0.87 | -0.00 | -0.00 | -0.00 | 1.08 |
| Off-Exchange Even Penny | 0.04 | 1.17 | -1.05 | 0.00 | 0.00 | -0.00 | 1.50 |

Panel B: Price Impact at 30 Seconds (in Basis Points) signed with Lee-Ready

| | | Standard | Quantiles | | | | |
|---|---|---|---|---|---|---|---|
| Group | Mean | Dev | 5% | 25% | 50% | 75% | 95% |
| Exchange RLP | 0.17 | 5.22 | -5.88 | -1.38 | -0.00 | 1.64 | 6.74 |
| Exchange | 0.80 | 8.48 | -10.20 | -1.77 | 0.00 | 3.17 | 12.97 |
| Off-Exchange Mid-Quote | -0.10 | 8.15 | -11.21 | -2.06 | 0.00 | 1.85 | 11.08 |
| Off-Exchange Sub-penny | 0.22 | 8.67 | -11.03 | -1.92 | -0.00 | 2.35 | 12.21 |
| Off-Exchange Even Penny | 0.22 | 8.67 | -11.80 | -2.06 | -0.00 | 2.50 | 12.74 |

Panel C: Price Impact at 30 Seconds (in Basis Points) signed with sub-penny prices

| | | Standard | Quantiles | | | | |
|---|---|---|---|---|---|---|---|
| Group | Mean | Dev | 5% | 25% | 50% | 75% | 95% |
| Exchange RLP | 0.16 | 5.20 | -5.88 | -1.38 | -0.00 | 1.61 | 6.63 |
| Off-Exchange Sub Penny | 0.18 | 8.36 | -10.55 | -1.68 | 0.00 | 1.90 | 11.33 |

There are substantial differences in RLP activity across assets, as shown in Figure 4. To characterize these differences, we estimate Regression 1 for the RPI volume share of each stock $i$ in our sample.

**REGRESSION 1.** Using observations at the asset $i$ level, we estimate:

$$\begin{aligned}\text{RPI\_Volume\_Share}_i = \quad & \alpha_0 + \alpha_1 Percent\_Time\_At\_Minimum\_Spread_i + \alpha_2 Market\_Cap_i \\ & + \alpha_3 Average\_Volume_i + \varepsilon_i\end{aligned}$$

Results of Regression 1 are presented in Table III. We estimate volume as a percentage of total volume and as a percentage of total sub-penny volume. Exchange RLP volume is considerably larger when assets spend a larger percentage of the day at the minimum bid-ask spread, consistent with RLPs serving as a venue for sub-penny price improvement in tick-constrained stocks. Proportional RLP volume is also larger for larger market-cap stocks and for stocks with higher average trading volume. The low usage of RLPs in small, less-liquid stocks is notable given that retail interest in trading these names is substantial (Dyhrberg et al. (2025)). The gap likely reflects limited willingness among liquidity providers to post RLP quotes for illiquid names: as Dyhrberg et al. (2025) argue, wider quoted spreads may not translate into higher market-making profits due to long inventory holding times and low trading volumes.

Liquidity in RLPs is at-will: market makers can provide liquidity at any time, but they are under no obligation to do so.[23] This contrasts with broker-wholesaler relationships, where wholesalers are expected to fulfill any orders that brokers send. Retail investors placing market orders may arrive at any time in the day, including during periods of stress. Even under the generous assumption that their orders are uncorrelated with aggregate institutional order flow, half their order flow would

[23] We note that the NYSE RLP does have a requirement that retail liquidity providers offer price-improving RLP limit orders for at least 5% of the trading day on a certain fraction of trading days to qualify for superior trading fee / rebate pricing. As Figure 2 makes clear, this threshold is low compared to the percentage of time that RLP orders are active.

**Table III: Cross-Sectional Variation in Volume Shares**. This table estimates Regression 1 with sub-penny volume, measured as a percentage of all volume, and as a percentage of sub-penny priced volume. For stock $i$, $\textit{Percent Time At Minimum Spread}_{it}$ measures the average, for each stock, percentage of the trading day with a quoted bid-ask spread of one penny. *Market Cap* measures the market capitalization of the stock in billions, and Average Volume measures the average trading volume in billions. Observations are at the stock (or ETF) level for the sample of securities described in Section III.B.

| | *Dependent variable:* | |
|---|---|---|
| | Percentage of All Volume (1) | Percentage of Only Sub-penny Volume (2) |
| Market Cap | $0.120^{***}$ | $0.926^{**}$ |
| | (0.035) | (0.417) |
| Percent Time at Minimum Spread | $0.001^{***}$ | $0.013^{***}$ |
| | (0.0001) | (0.001) |
| Average Volume | $0.041^{***}$ | $0.380^{***}$ |
| | (0.004) | (0.042) |
| Constant | $0.058^{***}$ | $0.931^{***}$ |
| | (0.002) | (0.030) |
| Observations | 2,590 | 2,590 |
| $R^2$ | 0.159 | 0.108 |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

be in the same direction as aggregate institutional order flow.[24] To investigate the relationship between RLP volume and market conditions, we estimate Regression 2, which relates volume shares and price impacts to volatility, spread tightness, and absolute intraday returns, with fixed effects for each stock and date:

[24] A second issue is whether there is any institutional interest at all. Dyhrberg et al. (2025) and Battalio, Jennings, Saglam, and Wu (2022) highlight that retail interest vastly exceeds institutional interest in less liquid securities.

**REGRESSION 2.** We estimate, using observations of asset $i$ on date $t$:

$$\begin{aligned}\text{VolumeShare}_{it} = \quad & \alpha_0 + \alpha_1 Percent\_Time\_At\_Minimum\_Spread_{it} + \alpha_2 Volatility_{it} \\ & + \alpha_3 Absolute\_Intraday\_Return_{it} + X_{it} + \varepsilon_{it}\end{aligned}$$

Results of Regression 2 are presented in Table IV. We focus on off-exchange sub-penny trades as a close analogue of on-exchange RLP trading.[25] Higher volatility or larger absolute returns both lead to a larger share of sub-penny trading occurring on exchange through RLPs, and a lower share occurring off-exchange. When volatility is higher, on-exchange RLP trades have much higher price impacts, while off-exchange sub-penny trades have no statistically significant change in price impacts, both when exchange RLP liquidity is available and when it is not. These results are consistent with wholesalers internalizing a smaller share of retail order flow when market conditions deteriorate, routing more orders to exchange RLPs. In the pecking order of Menkveld et al. (2017), RLPs behave as an intermediate venue: both the volume share and the price impact of RLP trades increase with volatility, consistent with RLPs receiving orders that wholesalers are unwilling to internalize.

Since RLP liquidity provision is entirely discretionary, the market conditions under which liquidity is available also matter. Figure 6 plots the ratio of quoted spreads at various time offsets around a trade to the spread at the time of the trade, separately for periods when the RLP flag is active and when it is not. We divide trading volume into on-exchange and off-exchange trades, and further into sub-penny, mid-quote, and at-quote bins. For each individual stock, we observe

[25] They are not a perfect analogue. Barardehi, Bernhardt, Da, and Warachka (2022) document that sub-penny trading may be driven not by the activity of retail investors, but the extent to which better improvement opportunities (such as mid-quote trading) are unavailable.

**Table IV: Panel Variation in Volume Share and Price Impact.** This table estimates Regression 2 with sub-penny volume, expressed as a percentage of total trading volume, and price impact, measured in basis points 30 seconds after the trade. Observations are at the stock-day level. Volatility measures the standard deviation of 15-minute price changes. Percent time at Minimum Spread measures the percentage of time the stock spread is a single tick, while absolute intra-day return measures the absolute value of the intra-day return. We include a fixed effect for each stock and date, and cluster standard errors by stock and by date. Note that Price Impact cannot be calculated when there is zero volume, thus Columns 4, 5, and 6 differ in the number of stock-days with zero volume in each category.

| Dependent Variable: | *Volume Share* | | | *Price Impact* | | |
|---|---|---|---|---|---|---|
| Venue: | Exchange | Off | Off | Exchange | Off | Off |
| RLP Active: | TRUE | TRUE | FALSE | TRUE | TRUE | FALSE |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Percent Time At | 0.001 | $-0.028^{***}$ | $0.027^{***}$ | 0.021 | 0.050 | $-0.072^{*}$ |
| Minimum Spread | (0.010) | (0.005) | (0.009) | (0.021) | (0.038) | (0.037) |
| Volatility | $6.592^{***}$ | $-2.464^{***}$ | $-4.128^{***}$ | $9.520^{**}$ | $2.281^{***}$ | 1.221 |
| | (0.510) | (0.261) | (0.471) | (4.582) | (0.381) | (0.745) |
| Absolute Intraday | $1.324^{***}$ | $-0.819^{***}$ | $-0.505^{***}$ | $-0.650$ | 0.251 | $-0.162$ |
| Return | (0.041) | (0.034) | (0.032) | (0.662) | (0.306) | (0.127) |
| Observations | 1,965,888 | 1,965,888 | 1,965,888 | 682,727 | 1,771,969 | 1,885,905 |
| $R^2$ | 0.417 | 0.248 | 0.380 | 0.013 | 0.002 | 0.003 |
| *Note:* | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

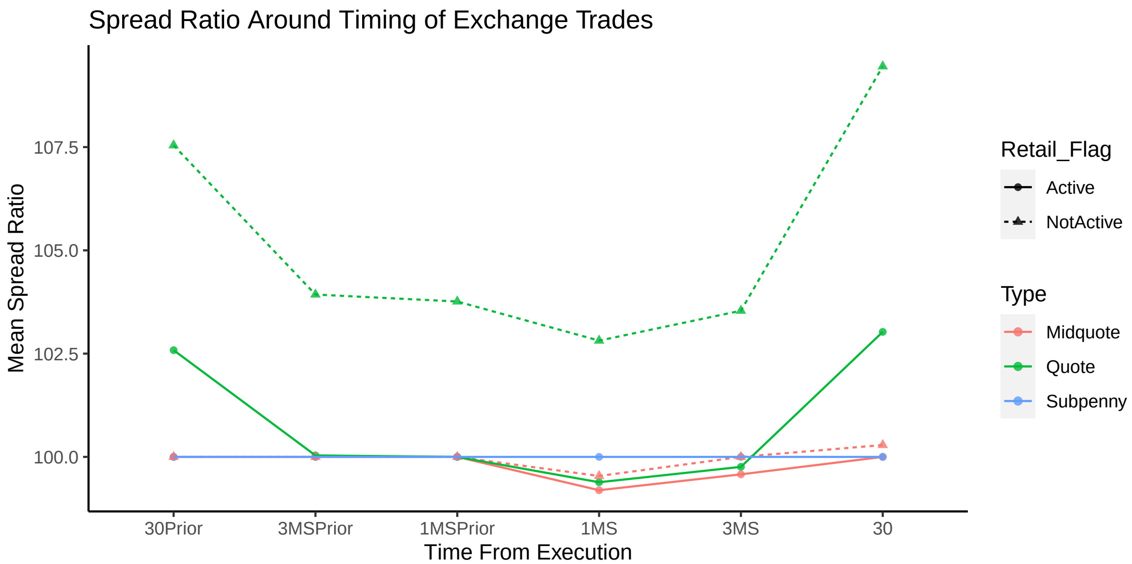
the quoted bid-ask spread $q_{t+i}$ at offsets of $\pm 30$ seconds, $\pm 3$ milliseconds, and $\pm 1$ milliseconds, calculate the average spread $\bar{q}_{t+i}$ separately for when the RPI Flag is or is not active, and plot $100 \times \frac{\bar{q}_{t+i}}{\bar{q}_t}$, so that a value of 100 represents no change in spreads. When the retail liquidity flag is active, spread ratios are stable, with quoted spreads only 0 to 2% wider before and after the trade. When the flag is not active, quoted spreads are 4 to 7% wider before and after the trade, particularly for on-exchange trades at the quote. This pattern suggests that RLP liquidity providers concentrate their quotes in periods when market conditions are stable.

Order imbalances tell a consistent story. Figure 7 presents the distribution of order imbalances for periods when the RLP flag is active and when it is not. When the flag is active, order imbalances are tightly clustered around zero, with roughly equal numbers of buys and sells. When the flag is not active, the distribution has much larger variance, with a greater likelihood of large directional imbalances. This is consistent with the discretionary nature of RLP liquidity provision: market makers post quotes when they expect balanced order flow and withdraw when they anticipate directional pressure.

Taken together, these results suggest that RLP activity is shaped by both demand and supply forces. On the demand side, wholesalers route a larger share of retail order flow to exchange RLPs when market conditions deteriorate, increasing both the volume share and the price impact of RLP trades. On the supply side, RLP liquidity providers concentrate their quotes in calm periods with stable spreads and balanced order flow.

**Figure 6. Spread Ratios Around Trades**. We plot the change in spreads around the time of a trade, separately for trades occurring when the RPI Flag is active, and trades occurring when it is not active. For trades occurring at time $t$, we calculate the quoted spread $q_t$ as well as the quoted spread $q_{t+i}$ occurring at a time-offset of $i$. We then calculate the mean quoted spreads $\bar{q}_t$ and $\bar{q}_{t+i}$, and plot $100 \times \frac{\bar{q}_{t+i}}{\bar{q}_t}$, so that a value of 100 indicates no change in spreads, and plot the median across stock-days. We also calculate this value for time offsets of 1 millisecond, 3 milliseconds, and 30 seconds both before and after trade.

**Panel A:** Spread Ratio For Exchange Trades




**Panel B:** Spread Ratio For Off-Exchange Trades

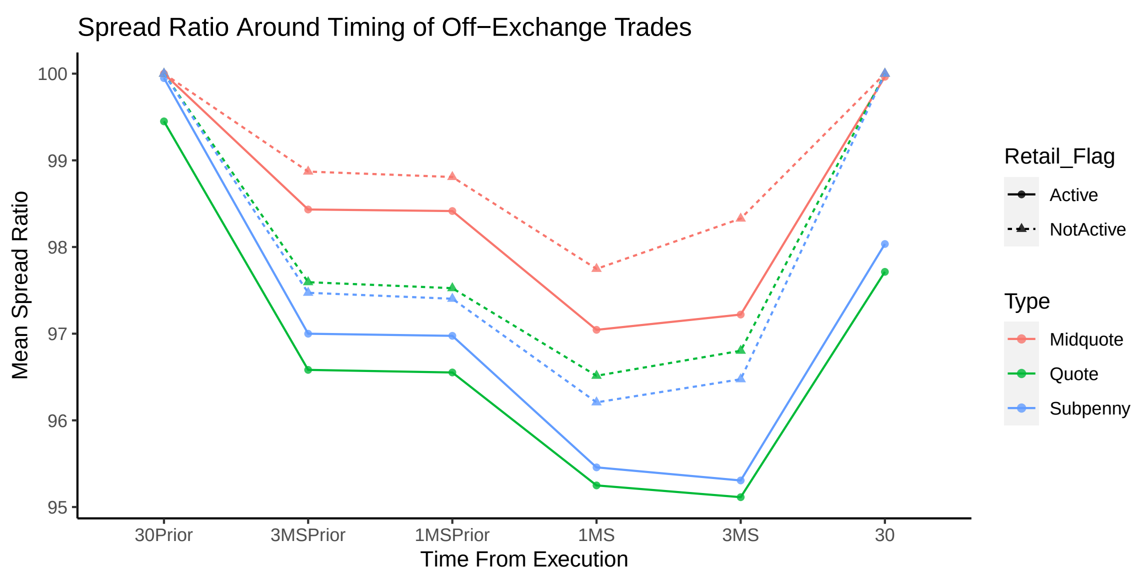
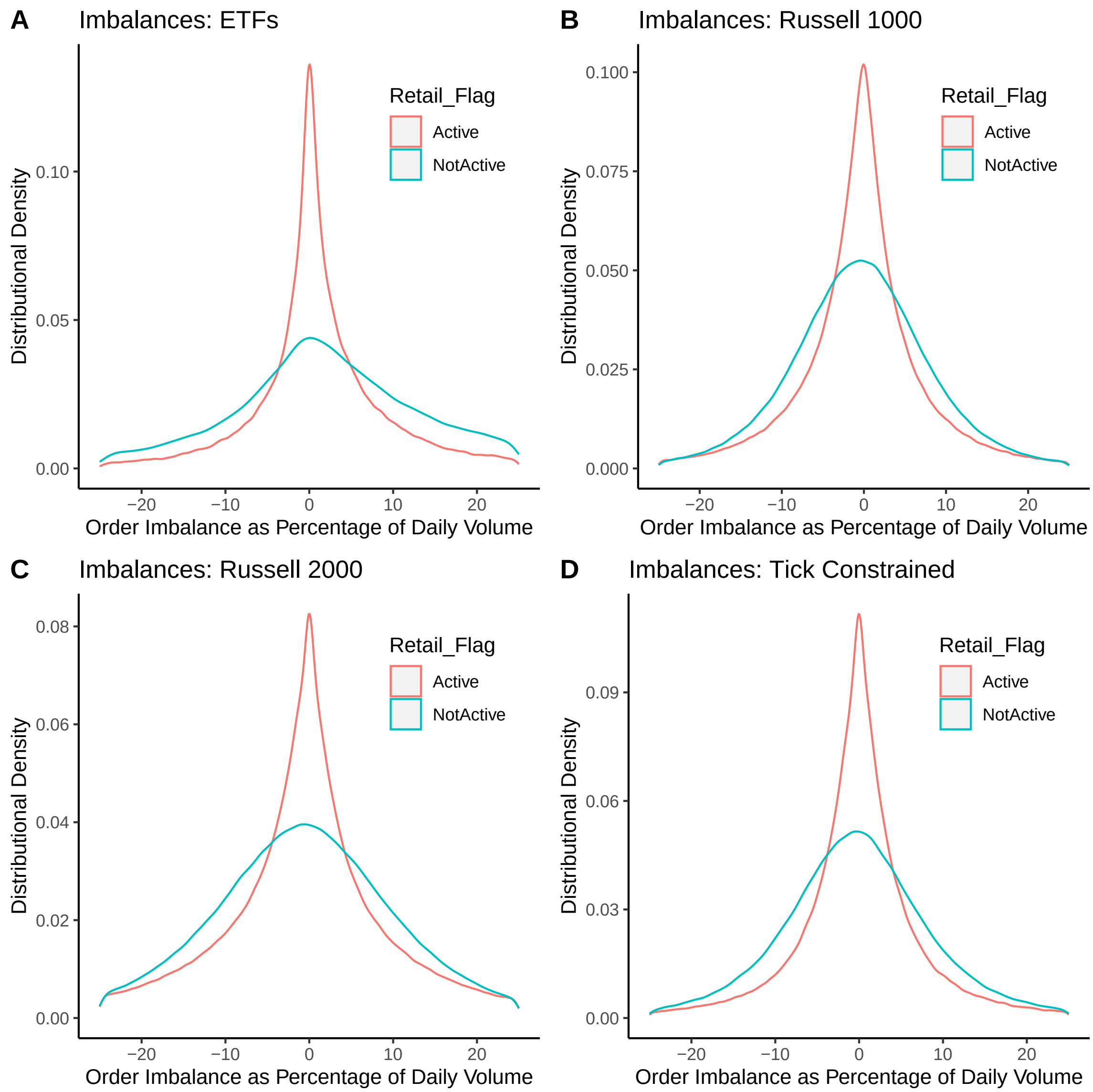


**Figure 7. Distribution of Order Imbalances**. For each stock-day observation, we calculate the total order imbalance among trades occurring when the RLP Flag is active, and the total order imbalance among trades occurring when the RLP Flag is not active. For stock $i$ on date $t$ with flag $j$, imbalance is calculated as $Imbalance_{ijt} = \frac{\sum Buy_{ijt} - \sum Sell_{ijt}}{\sum Buy_{ijt} + \sum Sell_{ijt}}$. We plot the distribution of imbalances, with the tails truncated to an imbalance of $\pm 50\%$. Panel A presents the imbalance distribution for ETFs, Panel B for stocks in the Russell 1000 Index, Panel C for stocks of the Russell 2000 Index, and Panel D for stocks and ETFs which are tick-constrained, defined as having at a one-penny bid-ask spread at least 50% of the trading day for at least one-third of the trading days in our sample.



## C. IEX RLP and Price Visibility

The IEX RLP only allows traders to post limit orders at the mid-quote. This design distinguishes the IEX RLP from other programs, where liquidity providers may price orders at any sub-penny increment that improves on the NBBO. Because the IEX RLP accepts only mid-quote orders, the RLP price is fully transparent: when the IEX RLP flag is active, market participants know that limit orders in the RLP are priced at the midpoint, though the available quantity remains unspecified beyond whether at least 100 shares are available.

The IEX RLP provides a revealed-preference measure of institutional interest in trading with retail investors at mid-quote. Using non-public CAT data, the SEC estimates that a substantial fraction of retail orders currently executed at prices worse than the midpoint could hypothetically trade against existing non-displayed mid-quote liquidity on exchanges and ATSs.[26] This estimate, however, assumes that retail orders could access this hidden liquidity. The IEX RLP provides a direct test: any market participant interested in trading with retail at mid-quote can post a limit order in the IEX RLP through their broker. Figure 2 shows that the IEX RLP has interest for less than 15% of the trading day on average, and two-sided interest for less than 2%. Despite the apparent abundance of hidden mid-quote liquidity elsewhere, few institutional traders choose to direct that liquidity toward retail order flow through the IEX RLP.

A natural explanation for this low participation is concern over information leakage. Figure 8 plots mid-quote trading volume at IEX over time. Total hidden mid-quote orders at IEX, including both RLP and traditional hidden orders, comprise around 1% to 1.5% of total U.S. equity trading

[26] The SEC reports that "On average, 51% of the shares of individual investor marketable orders internalized by wholesalers are executed at prices less favorable than the NBBO midpoint ('Wholesaler Pct Exec Shares Worse Than Midpoint'). Out of these individual investors shares that were executed at prices less favorable than the midpoint, on average, 75% of these shares could have hypothetically executed at a better price against the non-displayed liquidity resting at the NBBO midpoint on exchanges and NMS Stock ATSs." Securities and Exchange Commission (2022)

volume, with no obvious change around the creation of the IEX RLP on October 1, 2019. The IEX RLP began distributing an indicator of RLP interest on October 13, 2021. Total mid-quote volume does not appreciably change following the introduction of this indicator, suggesting that information leakage from the RLP flag alone is not the primary determinant of posting behavior. However, the RLP flag can reveal information beyond its mere existence. Standard hidden mid-quote orders offer no indication of trade direction: a mid-quote execution does not reveal which party was the buyer. RLP flags, by contrast, can reveal direction if the flag is one-sided, or if a retail market order consumes all RLP liquidity and triggers a flag update.[27] As Figure 8 shows, total mid-quote volume at IEX is over 10 times higher than RLP mid-quote volume, which may, in part, reflect concerns from institutional investors over this form of post-trade transparency.[28]

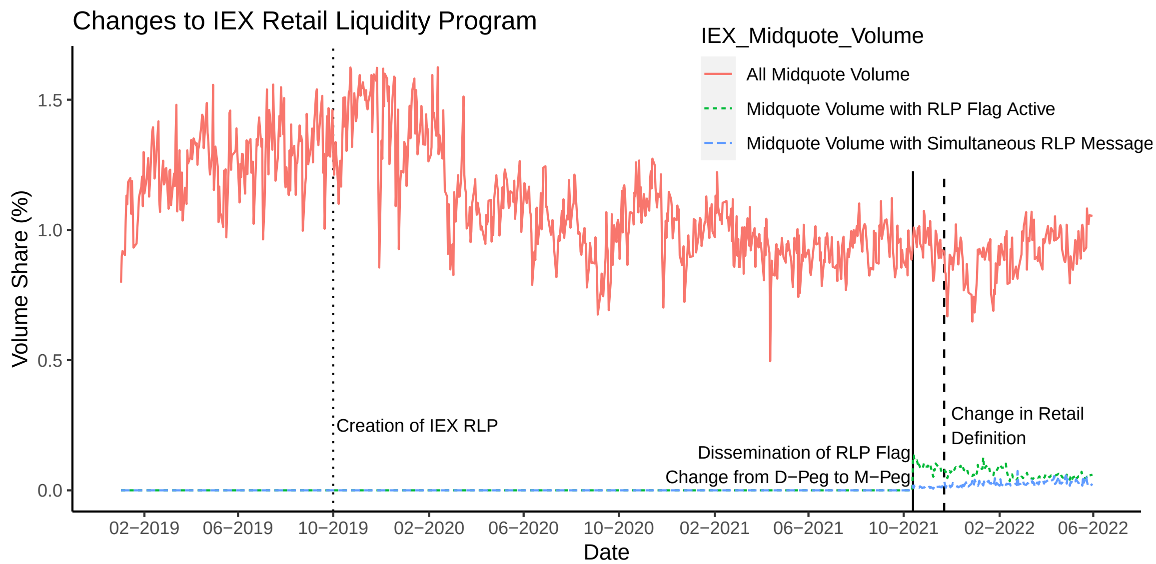
An alternative explanation is that institutional traders posting hidden mid-quote liquidity are not seeking to trade specifically with retail investors. Posting in the IEX RLP restricts the order to retail counterparties only, limiting execution opportunities, particularly as individual retail orders are often small.

We investigate the distribution of trade prices as a function of the IEX RLP status using public data. Figure 9 presents the distribution of sub-penny trade prices separately for periods when the IEX RLP has no interest, one-sided interest, and two-sided interest. When the IEX RLP has two-sided liquidity (that is, interest in both buying from and selling to retail at mid-quote), both

[27]For example, if there is two-sided RLP liquidity, but there is a message of a mid-quote trade and a simultaneous message that the RLP changes from two-sided to bid-side interest only, market participants could infer that there was a retail order to buy which took out the ask. A standard (i.e., not RLP) hidden mid-quote order would avoid this information leakage.

[28]From the TAQ data, it is impossible to determine the exact portion of orders that are retail orders in the IEX RLP program, but we can estimate an upper and lower bound. For the upper bound, we count all mid-quote orders which occur when the IEX retail flag is active, though some of this volume may include non-retail mid-quote orders interacting with hidden mid-quote liquidity which is posted outside the RLP. For the lower bound, we measure mid-quote volume which has a simultaneous message update for the RLP program; this measures only retail orders which consume the available RLP liquidity (necessitating an updated RLP message), but will miss retail orders which do not consume all available RLP liquidity and therefore send no update message.

**Figure 8. IEX Mid-quote Volume and Key RLP Rule Changes**. The IEX Retail Liquidity Program was introduced on October 1, 2019, with only hidden discretionary midpoint-peg orders. On October 13, 2021, the Retail Liquidity Program changed the RLP order type to a midpoint-peg order and began dissemination of an indicator of whether there was RLP interest. On November 22, 2021, the requirement that retail investors submit no more than 390 orders per day was lifted. We plot total mid-quote volume on IEX (as a percentage of total equities trading volume) with the solid (red) line. We plot mid-quote volume which occurs during the time that the IEX RLP is active with the dotted (green) line. We plot the total mid-quote volume which occurs simultaneously with an RLP message with the dashed (blue) line.




on-exchange and off-exchange sub-penny trades are more concentrated at mid-quote, with mid-quote trading comprising over 50% of all trading volume. There are also some off-exchange trades at mid-quote when the IEX RLP has two-sided liquidity, though Battalio et al. (2022) document that many sub-penny trades are non-retail.

Public data does not reveal whether participants are retail investors. To further investigate how retail orders interact with IEX RLP liquidity, we obtain a sample of trades from a large retail brokerage. Our sample covers all retail orders for 100 shares or less on three days (March 13, 14, and 16, 2023) in a set of liquid symbols.[29] Figure 10 plots the cumulative distribution function

[29]The symbols are: AAPL, AMZN, CRDO, DIA, GOOG, IVV, IWM, JPM, MU, PPG, QQQ, SGH, SPY, V, VICR, and VOO.

of price improvement, measured as a percentage of the quoted bid-ask spread. Price improvement of 50% corresponds to execution at the NBBO mid-quote. When the IEX RLP is active, retail investors receive more price improvement: 75% of orders receive mid-quote or better when the flag is active, compared to 60% when it is not.

However, the share of retail orders receiving mid-quote when the IEX RLP is active is less than 100%. While market makers are disciplined by brokers on the aggregate execution quality they deliver, for individual trades they may prefer to internalize even when the IEX RLP is available, sometimes at prices worse than mid-quote. Market makers are able to do so because retail liquidity programs are not protected quotations. For other RLPs, the lack of trade-through protection is less consequential because quotes are hidden and the available price is unknown. For the IEX RLP, however, the mid-quote price is known, making the absence of protection particularly salient: wholesalers can observe that mid-quote liquidity is available yet "trade through" the IEX RLP without matching that price. This behavior is a direct implication of Corollary 2: for low-cost orders, the wholesaler's current-period spread revenue outweighs the marginal loss of future order flow from providing slightly less price improvement than the RLP's mid-quote, making internalization strictly preferred even when the RLP offers a known, superior price.

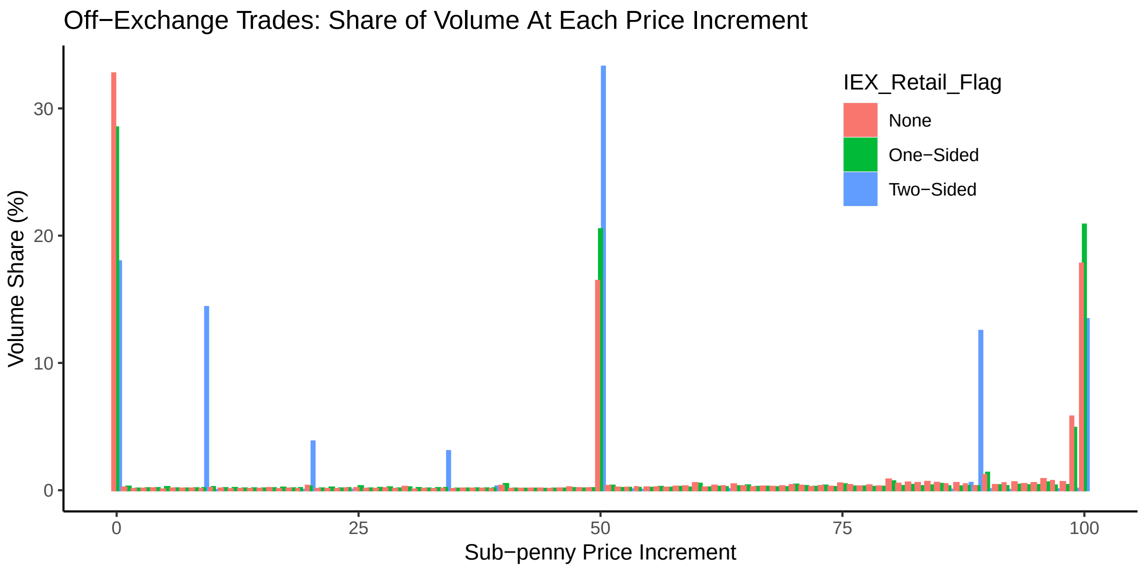
Execution frictions also contribute. We do not observe the market view of the wholesaler at the time the routing decision is made. In addition to the latency introduced by the IEX speed bump, there is latency between the wholesaler and potential venues, and trying multiple venues may take milliseconds rather than microseconds. Not all wholesalers may connect to, or route to, the IEX RLP, and the flag indicates available liquidity for only a brief portion of the trading day. We also note that while some retail orders receive prices worse than mid-quote when the IEX RLP is active, others receive prices *better* than mid-quote even when the IEX RLP is not active.

When the IEX RLP has only one-sided interest in the same direction as the retail order (i.e., opposite to the direction needed for execution), retail investors receive less price improvement.[30] Retail investors obtain 5% less price improvement when the IEX RLP has same-direction interest than when the IEX RLP has no interest at all. Although routing to the IEX RLP is voluntary, wholesalers appear to provide less price improvement when the RLP signal reveals same-direction interest. For example, retail buy orders obtain worse price improvement when there is IEX RLP interest on the bid side only compared to when there is no IEX RLP interest at all.

[30]By one-sided same direction interest, we refer to one-sided ask side interest when there is a retail sell order, or one-sided bid side interest when there is a retail buy order. Note that 'bid' and 'ask' refer to the limit order interest in buying or selling, and the price in the IEX RLP is always the NBBO mid-quote.

**Figure 9. Volume Share of Off-Exchange Sub-Penny Prices**. For each possible sub-penny price increment, we plot the volume occurring at this price increment, as a percentage share of all volume occurring on that venue. We use data from trades occurring on January 3, 2022, and separate trade volume into three categories: trades occurring when the IEX RLP has no interest, trades occurring when the IEX RLP has one-sided interest, and trades occurring when the IEX RLP has two-sided interest. Trades are restricted to be 100 shares or less, as the RLP only indicates availability for at least 100 shares. Panel A presents the distribution of trades for off-exchange trades. Panel B presents the distribution of trades for on-exchange trades. Note that on-exchange sub-penny trades can only occur in increments of tenths of a cent.

**Panel A:** Price Improvement Distribution for Off-Exchange Trades




**Panel B:** Price Improvement Distribution for On-Exchange Trades

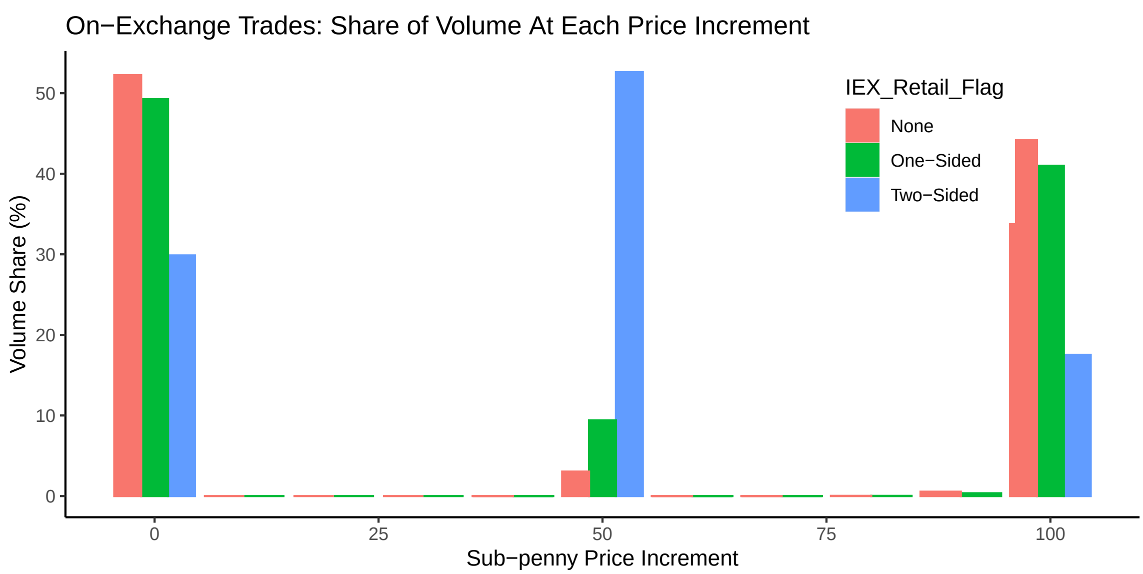
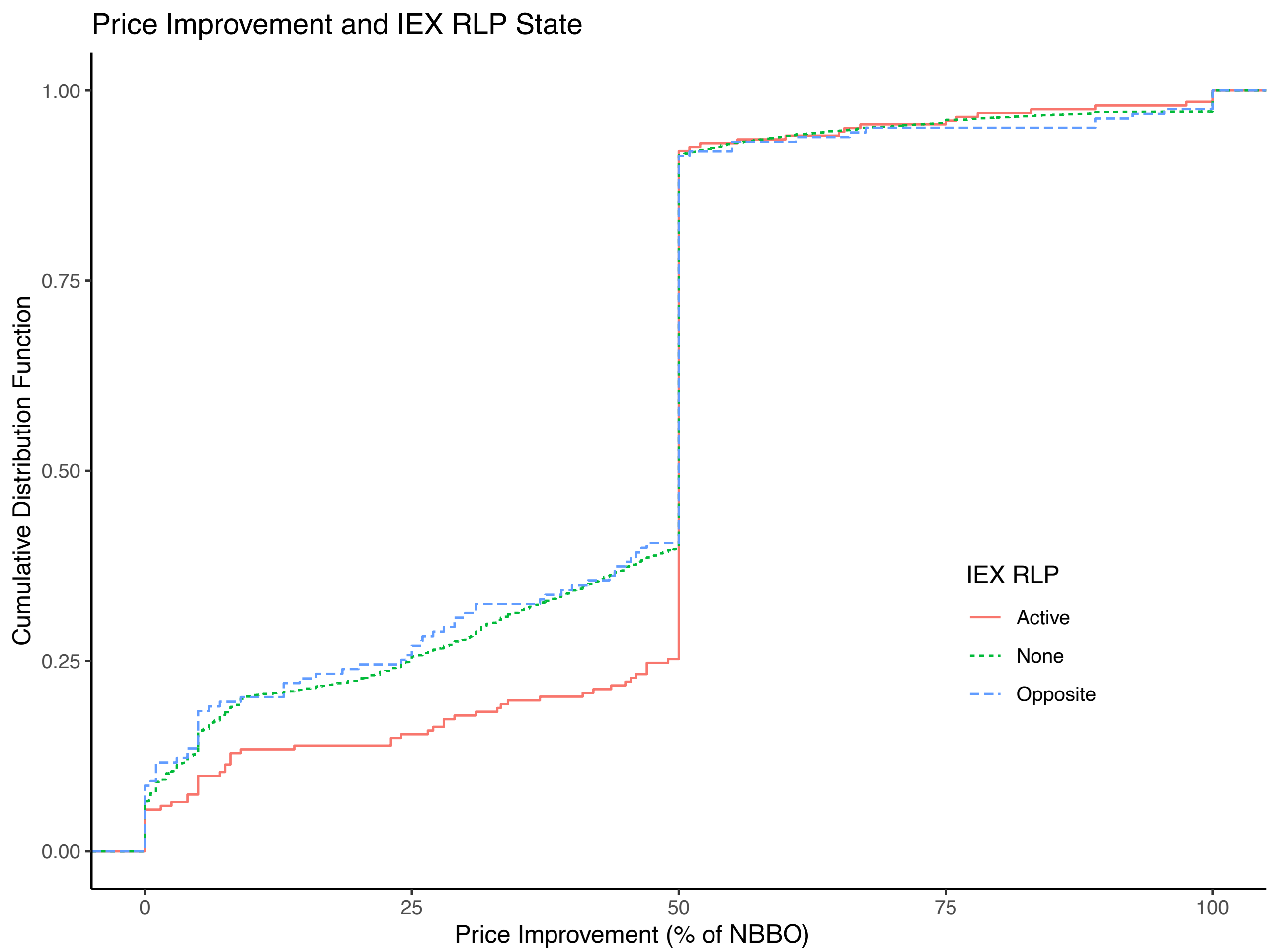


**Figure 10. Retail Trades and the IEX RLP**. We obtain a sample trades of retail investors from a large retail brokerage. Our sample covers all retail orders for 100 shares or less on three days (March 13, 14, and 16, 2023) in a set of liquid symbols. We plot the empirical cumulative distribution function of price improvement (as a percentage of the NBBO quoted spread), received by retail orders. Active refers to either two-sided IEX RLP interest, bid interest at the time of a retail sell order, or ask interest at the time of a retail buy order. None refers to no active RLP flag. Opposite refers to *only* RLP bid interest at the time of a retail sell order, or *only* RLP ask interest at the time of a retail buy order.



## D. *Retail Liquidity Programs and the Tick Size Pilot*

While brokers can route directly to exchanges, including Retail Liquidity Programs, the majority of retail orders flow to off-exchange wholesalers (Battalio and Jennings (2022)), who then decide whether to internalize the order or route it to an RLP. The SEC Tick Size Pilot provides a natural experiment to study how changes in quoting and trading increments affect RLP usage and the relative execution quality of RLPs and off-exchange venues. We focus on the lifting of the pilot on October 1, 2018, when 995 stocks transitioned from treatment to control groups. Figure 11 plots volume executed in tick size pilot stocks around this transition, showing a sharp increase in sub-penny volume and a sharp decrease in mid-quote volume.

The three treatment groups impose progressively stricter restrictions. Group 1 securities must be quoted in 5-cent increments, though RLP orders can still be entered at sub-penny prices. Group 2 securities must be both quoted and traded in 5-cent increments, with an exception allowing retail orders to trade in 0.5-cent increments. Group 3 imposes the same restrictions as Group 2, plus a trade-at rule preventing off-exchange price-matching of displayed quotes, while retaining the same retail exception. To estimate the impact of these restrictions on RLP activity, we estimate the following Regression:

**REGRESSION 3.** Using asset $i$ on date $t$, we estimate:

$$\text{RLP\_Volume\_Share}_i = \alpha_0 + \alpha_1 TreatmentGroup_i + \alpha_2 TSP_t + \alpha_3 TreatmentGroup_i * Lift_t + X_{it} + \varepsilon_{it}$$

Table V presents estimates of Regression 3; the coefficient of interest is $\alpha_3$, the interaction term between the treatment group and the $TSP$, an indicator that takes the value one while the tick size pilot is active.

For Group 1, where only the quoting increment is widened, exchange RLP sub-penny volume is indistinguishable from the control group during the pilot period. For Groups 2 and 3, where trading increments are also restricted, exchange RLP sub-penny volume drops to near zero (Figure 11, Panel A). The contrast between these groups and Group 1 indicates that the competitive advantage of RLPs depends critically on sub-penny pricing flexibility: once trade-level increments are constrained to round pennies only, the RLP mechanism effectively shuts down. Off-exchange sub-penny volume declines across all treatment groups, though the decline for Groups 2 and 3 is substantially larger than for Group 1, consistent with the direct prohibition on most sub-penny trading in Groups 2 and 3 versus a behavioral response to wider quoted spreads in Group 1.

Mid-quote volume increases across all groups, which raises the possibility that midquote RLP volume increases for Groups 2 and 3; public data cannot distinguish between non-retail and retail midquote volume. To address this concern, we estimate Regression 3 using Group 1 as the omitted group where sub-penny RLP trading is allowed, and *SubBanned* as an indicator for Groups 2 and 3, where sub-penny RLP trading is banned. We focus on midquote volume as the outcome variable, to compare whether midquote volume in Groups 2 and 3 differs from midquote volume in Group 1. Results of this estimation are presented in Table VI. All estimates of $\alpha_3$, the interaction between the TSP and the ban on subpenny pricing, are not statistically significant, and point estimates are negative for three of the four columns. This evidence that midquote volume did not increase differentially in groups 2 and 3 relative to group 1 (consistent with Figure 11, Panel B) confirms that when RLP sub-penny volume is prohibited, traders do not switch to trading in the RLP at midquote.

To examine how execution quality responds to wider quoting increments, we re-estimate Regression

3 using only control and Group 1 stocks, isolating the effect of a 5-cent minimum quote increment while leaving trading increments unrestricted. Relative to control stocks, the quoted spread of Group 1 stocks widens substantially and highly significantly during the pilot period. We then examine how effective spreads, price improvement, and price impact change for subpenny trades, both on-exchange and off-exchange. Results of this estimation are presented in Table VII.

First, effective spreads and price impact for off-exchange subpenny trades do not change ($\alpha_3$ is statistically indistinguishable from zero). The insensitivity of off-exchange effective spreads to wider quoting increments is consistent with wholesaler pricing being determined mainly by their own cost structure, including execution and inventory costs, rather than by the quoted spread as an external benchmark. As documented in Dyhrberg et al. (2025), the wholesaler market is sufficiently competitive that spreads largely reflect the cost of providing execution.

By contrast, RLPs deliver meaningfully better execution quality when quoted spreads widen, with Group 1 effective spreads 27 basis points smaller than effective spreads in the control while the pilot is active. There is a concurrent decline in price impact of 26 basis points, suggesting that the order flow sent to the RLP is comparatively less adversely selected by the same magnitude. These results are consistent with the RLP providing a valuable intermediate venue between off-exchange trading and on-exchange trading at a wide spread. In the pecking order theory of Menkveld et al. (2017), RLPs offer intermediate adverse selection and effective spread, with off-exchange trades having the lowest adverse selection and spread, and on-exchange trades having the highest adverse selection and spread.

**Table V: Volume Around Lifting of Tick Size Pilot** We estimate Regression 3 with data from the Tick Size Pilot from August 1 to December 31, 2018. TSP is an indicator which takes the value 1 before October 1, 2018 (when treatment of stocks was active for the stocks in our sample). We track how sub-penny and mid-quote volume changes, as a percentage of total volume, split by whether the volume occurs when a retail liquidity program flag is or is not active for that particular symbol. Controls include the market cap, time at one tick, and average volume.

| | *Dependent variable: Volume* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Exchange | | | Off-Exchange | | | |
| Pricing Increment: | Sub | Mid | Mid | Sub | Sub | Mid | Mid |
| RPI Flag: | T | T | F | T | F | T | F |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| TG1 | $-0.030^{***}$ | $0.065^{***}$ | $0.085^{***}$ | $-0.131$ | $0.011$ | $-0.050$ | $0.086^{***}$ |
| | (0.009) | (0.022) | (0.008) | (0.083) | (0.029) | (0.050) | (0.020) |
| TG2 | $0.002$ | $0.059^{***}$ | $0.023^{***}$ | $-0.332^{***}$ | $-0.005$ | $-0.032$ | $-0.036^{*}$ |
| | (0.009) | (0.022) | (0.008) | (0.083) | (0.029) | (0.051) | (0.020) |
| TG3 | $-0.017^{*}$ | $0.053^{**}$ | $0.088^{***}$ | $-0.317^{***}$ | $0.095^{***}$ | $0.004$ | $0.105^{***}$ |
| | (0.009) | (0.022) | (0.008) | (0.083) | (0.029) | (0.051) | (0.020) |
| TSP | $0.016^{**}$ | $-0.049^{***}$ | $-0.128^{***}$ | $1.265^{***}$ | $0.004$ | $0.220^{***}$ | $-0.133^{***}$ |
| | (0.007) | (0.017) | (0.006) | (0.064) | (0.023) | (0.039) | (0.015) |
| TG1∗TSP | $0.020$ | $1.889^{***}$ | $-0.085^{***}$ | $-0.673^{***}$ | $-0.613^{***}$ | $1.864^{***}$ | $-0.199^{***}$ |
| | (0.014) | (0.034) | (0.013) | (0.129) | (0.046) | (0.079) | (0.031) |
| TG2∗TSP | $-0.284^{***}$ | $2.204^{***}$ | $-0.018$ | $-4.077^{***}$ | $-0.935^{***}$ | $3.540^{***}$ | $0.047$ |
| | (0.014) | (0.035) | (0.013) | (0.131) | (0.046) | (0.080) | (0.031) |
| TG3∗TSP | $-0.270^{***}$ | $2.657^{***}$ | $0.006$ | $-4.060^{***}$ | $-1.102^{***}$ | $4.819^{***}$ | $0.259^{***}$ |
| | (0.014) | (0.035) | (0.013) | (0.130) | (0.046) | (0.080) | (0.031) |
| Observations | 147,477 | 147,477 | 147,477 | 147,477 | 147,477 | 147,477 | 147,477 |
| $R^2$ | 0.019 | 0.146 | 0.081 | 0.074 | 0.031 | 0.078 | 0.049 |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01

**Table VI: Midquote Volume Around Lifting TSP**. We estimate Regression 3 with data from the Tick Size Pilot from August 1 to December 31, 2018. TSP is an indicator which takes the value 1 before October 1, 2018 (when treatment of stocks was active for the stocks in our sample). *SubBanned* is an indicator which takes the value 1 for groups 2 and 3, where RLPs are prohibited from executing sub-penny trades; the reference group is Group 1, where quoting changes are the same but RLPs can still execute sub-penny trades. We track how sub-penny and mid-quote volume changes, as a percentage of total volume, split by whether the volume occurs when a retail liquidity program flag is or is not active for that particular symbol. Controls include the market cap, time at one tick, and average volume.

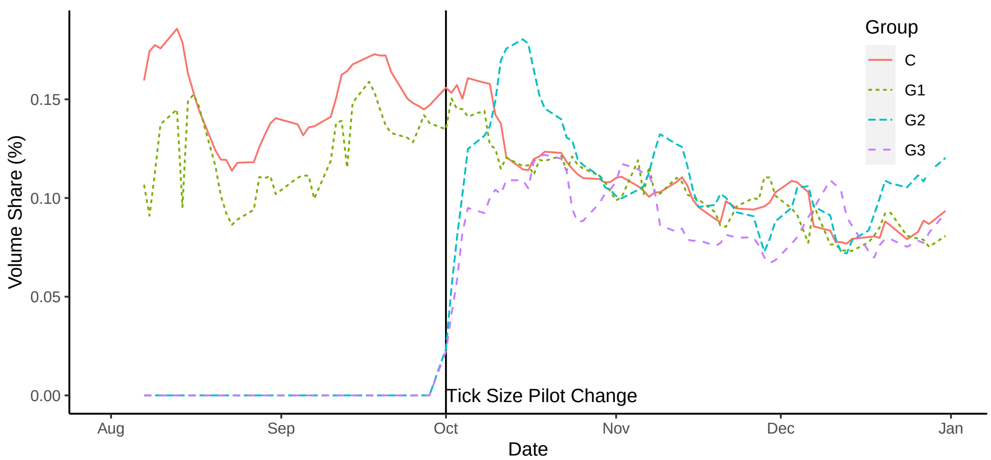
| | *Dependent variable: Volume* | | | |
|---|---|---|---|---|
| | Exchange | | Off-Exchange | |
| Pricing Increment: | Mid | Mid | Mid | Mid |
| RPI Flag: | T | F | T | F |
| | (1) | (2) | (3) | (4) |
| AvgVolume | 5.743*** | 1.731** | 26.298*** | 6.195*** |
| | (1.624) | (0.778) | (3.451) | (1.651) |
| SubBanned | −0.100** | −0.124*** | 0.090 | −0.197*** |
| | (0.044) | (0.021) | (0.094) | (0.045) |
| TSP | 2.136*** | −0.284*** | 3.075*** | −0.405*** |
| | (0.057) | (0.027) | (0.122) | (0.058) |
| SubBanned∗TSP | −2.910 | −0.296 | −3.093 | 0.134 |
| | (1.816) | (0.869) | (3.857) | (1.845) |
| Observations | 8,354 | 8,354 | 8,354 | 8,354 |
| $R^2$ | 0.193 | 0.127 | 0.093 | 0.069 |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 | | |

**Table VII: Trade Measures Around Lifting of Tick Size Pilot** We estimate Regression 3 with data from the Tick Size Pilot from August 1 to December 31, 2018. TSP is an indicator which takes the value 1 before October 1, 2018 (when treatment of stocks was active for the stocks in our sample). We focus on control stocks against treatment group 1, which saw a minimum quote requirement increased to 5 cents. We examine Exchange sub-penny trades (which are RLP trades) in columns 1 through 3, and for reference, estimate off-exchange sub-penny trades in columns 4-6. Effective Spreads, Price Improvement, and Price Impact are all measured in basis points. Observations are at the individual trade level, and we cluster standard errors by symbol and day.

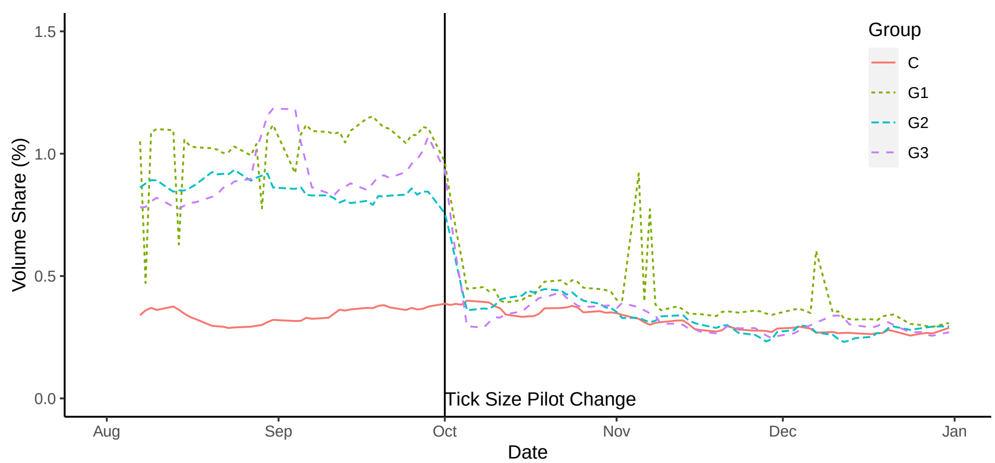
| | *Exchange* | | | *Off-Exchange* | | |
|---|---|---|---|---|---|---|
| | Effective Spread | Price Improvement | Price Impact | Effective Spread | Price Improvement | Price Impact |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| TimeAtOneTick | $-0.184^{***}$ | $-0.210^{***}$ | $-0.084^{***}$ | $-0.329^{***}$ | $-0.395^{***}$ | $-0.082^{***}$ |
| | (0.029) | (0.033) | (0.014) | (0.043) | (0.055) | (0.012) |
| Volume | 0.107 | $0.117^{*}$ | 0.109 | 0.379 | 0.414 | $0.257^{**}$ |
| | (0.159) | (0.061) | (0.101) | (0.233) | (0.260) | (0.114) |
| $\lvert AbsIntradayReturn \rvert$ | $253.181^{***}$ | $-248.108^{***}$ | $174.716^{***}$ | $-63.393$ | $-212.131^{***}$ | 21.424 |
| | (63.830) | (46.236) | (49.327) | (52.168) | (58.446) | (32.900) |
| TG1 | $-20.428^{**}$ | $-31.530^{**}$ | $-9.353^{*}$ | $-39.956^{***}$ | $-42.978^{**}$ | $-10.913^{**}$ |
| | (9.055) | (14.189) | (5.066) | (15.186) | (19.344) | (4.548) |
| TSP | $10.788^{**}$ | $-35.752^{***}$ | $15.338^{***}$ | $-20.116^{***}$ | $-31.677^{***}$ | 0.567 |
| | (5.060) | (6.028) | (3.676) | (6.119) | (7.566) | (2.236) |
| TG1∗TSP | $-27.776^{***}$ | $177.423^{***}$ | $-26.438^{***}$ | 9.986 | $111.088^{***}$ | $-3.537$ |
| | (7.073) | (15.136) | (5.808) | (10.581) | (13.338) | (3.847) |
| Constant | $78.966^{***}$ | $157.077^{***}$ | $25.852^{***}$ | $175.311^{***}$ | $222.349^{***}$ | $36.040^{***}$ |
| | (7.432) | (11.564) | (3.969) | (11.793) | (15.618) | (3.239) |
| Observations | 1,406,107 | 1,406,112 | 1,406,075 | 5,686,680 | 5,687,034 | 5,686,245 |
| $R^2$ | 0.032 | 0.050 | 0.004 | 0.013 | 0.017 | 0.001 |
| *Note:* | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

**Figure 11. Volume Share of RLPs In Tick Size Pilot**. For each day, we plot the share of volume which executes in a retail liquidity program, compared to total volume. Panel A presents this figure for Sub-penny executions, while Panel B presents this figure for sub-penny and midquote volume while the RLP Flag is active (note that some mid-quote executions may be non-retail activity). Groups refers to the SEC Tick Size Pilot Groups: Control, Group 1 (5 cent quote), Group 2 (5 cent trade and quote), Group 3 (5 cent trade and quote and trade-at rule). Retail activity of at least 0.5 cents is exempt from the 5 cent trade and trade-at rules.

**Panel A:** Volume Share for Exchange Sub-penny (Non-Mid-quote) Executions with an RPI Flag Active




**Panel B:** Volume Share for Exchange Sub-penny or Mid-quote Executions with an RPI Flag Active



# VI. Conclusion

This paper studies Retail Liquidity Programs to understand when on-exchange mechanisms can compete for retail order flow and explore why existing programs have failed to gain traction. We identify three key facts: internalized orders deliver greater price improvement than RLP orders, RLP execution share increases with market-making costs, and wholesalers internalize some orders even when the RLP offers superior execution. These facts provide a basis for a stylized model of a wholesaler's routing decision in which a cost-based cutoff determines whether orders are internalized or routed to the RLP.

We trace the limited competitive impact of RLPs to three distinct frictions: non-displayed liquidity, absence of order protection under Regulation NMS, and wholesaler routing discretion. These frictions interact but none is intrinsic to the RLP mechanism itself. Using the SEC's Tick Size Pilot, we show that when quoting increments widen, RLP execution quality improves significantly while off-exchange effective spreads are unchanged. This differential response indicates that wholesaler pricing reflects underlying execution costs rather than the quoted spread benchmark, consistent with a competitive wholesaler market (Ernst et al. (2026)). The competitive interpretation implies that the binding constraint on RLP effectiveness is not the pricing behavior of wholesalers, but the institutional frictions that limit the observability, enforceability, and routing of retail-only liquidity.

Our findings have implications for market design proposals aimed at increasing competition for retail order flow. Simply creating a venue is insufficient when prices are hidden, quotes are unprotected, and routing is discretionary. However, the Tick Size Pilot evidence demonstrates that under appropriate conditions, RLPs do provide an intermediate price between off-exchange prices and the full spread. Reforms that address the three frictions we identify—improving price transparency, extending order protection to retail-only liquidity, and strengthening best-execution standards for routing decisions—could make on-exchange competition viable. The SEC's recent proposal to rescind the Order Protection Rule (Rule 611) leaves routing and order protection (for RLPs) unchanged, as RLP orders currently do not have order protection, but would also allow more flexibility in order types. Under 611, order protection also comes with specific guidelines on what constitutes a protected order; without 611, retail liquidity programs may be able to operate with increased price transparency. More broadly, the model's routing framework applies to any setting in which intermediaries choose between internalizing and routing to a competing venue, providing a lens through which to evaluate proposed reforms to retail market structure.

# REFERENCES


Adinarayan, Thyagaraju, 2021, Retail Traders Account for 10% of U.S. Stock Trading Volume - Morgan Stanley, Reuters.

Aramian, Fatemeh, and Carole Comerton-Forde, 2023, Retail Trading in European Equity Markets, Working Paper.

Baldauf, Markus, Joshua Mollner, and Bart Z. Yueshen, 2022, Siphoned Apart: A Portfolio Perspective on Order Flow Segmentation, Available at SSRN 4173362.

Barardehi, Yashar, Dan Bernhardt, Zhi Da, and Mitch Warachka, 2022, Institutional Liquidity Demand and the Internalization of Retail Order Flow: The Tail Does Not Wag the Dog.

Bartlett, Robert P, Justin McCrary, and Maureen O'Hara, 2022a, A Fractional Solution to a Stock Market Mystery, Available at SSRN.

Bartlett, Robert P, Justin McCrary, and Maureen O'Hara, 2022b, The Market Inside the Market: Odd-Lot Quotes, Available at SSRN 4027099.

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?, Technical report, Working Paper.

Battalio, Robert, and Robert Jennings, 2023, On the Potential Cost of Mandating Qualified Auctions for Marketable Retail Orders, Available at SSRN: 4403047.

Battalio, Robert, Robert Jennings, Mehmet Saglam, and Jun Wu, 2022, Identifying Market Maker Trades as "Retail" From TAQ: No Shortage of False Negatives and False Positives.

Bishop, Allison, 2022, "The SEC Isn't Mad at PFOF. They're Just Disappointed", Proof Trading, [Accessed: 2022 02 01].

Chung, Kee H, Dominik Rösch, and Yuhan Zhang, 2023, Execution Quality of Retail Trades on Exchanges: Evidence from the NYSE Retail Liquidity Program .

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, 2025, The retail execution quality landscape, *Journal of Financial Economics* 168, 104051.

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.

Eaton, Gregory W, T Clifton Green, Brian S Roseman, and Yanbin Wu, 2022, Retail Trader Sophistication and Stock Market Quality: Evidence from Brokerage Outages, *Journal of Financial Economics* 146, 502–528.

Ernst, Thomas, Andrey Malenko, Chester S Spatt, and Jian Sun, 2026, What does best execution look like?, *Available at SSRN 6635238* .

Ernst, Thomas, Chester Spatt, and Jian Sun, 2025, Would Order-By-Order Auctions Be Competitive?, *Journal of Finance* 80, 1879–1927.

Foley, Sean, and Tālis J Putniņš, 2016, Should We be Afraid of the Dark? Dark Trading and Market Quality, *Journal of Financial Economics* 122, 456–481.

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

Huang, Xing, Mina Lee, Philippe Jorion, 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, *Quarterly Review of Economics and Finance* 80, 367–373.

Jain, Pankaj K, Suchi Mishra, Shawn O'Donoghue, and Le Zhao, 2020, Trading Volume Shares and Market Quality in a Zero Commission World, Available at SSRN 3741470.

Kwan, Amy, Ronald Masulis, and Thomas H McInish, 2015, Trading Rules, Competition for Order Flow and Market Fragmentation, *Journal of Financial Economics* 115, 330–348.

Menkveld, Albert J, Bart Zhou Yueshen, and Haoxiang Zhu, 2017, Shades of Darkness: A Pecking Order of Trading Venues, *Journal of Financial Economics* 124, 503–534.

O'Hara, Maureen, 2015, High Frequency Market Microstructure, *Journal of Financial Economics* 116, 257–270.

Parlour, Christine A., and Uday Rajan, 2003, Payment for Order Flow, *Journal of Financial Economics* 68, 379–411.

Schwarz, Christopher, Brad M Barber, Xing Huang, Philippe Jorion, and Terrance Odean, 2022, The 'Actual Retail Price' of Equity Trades, Available at SSRN 4189239.

Schwenk-Nebbe, Sander, 2021, The Participant Timestamp: Get The Most Out Of TAQ Data, Available at SSRN 3984827.

SEC, 2013, Self-Regulatory Organizations; The NASDAQ Stock Market LLC; Order Granting Approval to Proposed Rule Change, as Modified by Amendment No. 1, to Establish the Retail Price Improvement Program on a Pilot Basis until 12 Months from the Date of Implementation.

SEC, 2023, Self-Regulatory Organizations; NYSE Arca, Inc.; Notice of Filing and Immediate Effectiveness of Proposed Rule Change to Amend the NYSE Arca Equities Fees and Charges.

Securities and Exchange Commission, 2022, Order Competition Rule, Proposal for Rule 615. [Accessed: 2022 12 16].

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

# Appendix A. Proofs

*Proof of Lemma 1.* Fix $c \in [0, \bar{c}]$. The first and second derivatives of $U$ with respect to $s$ are

$$U_s(s,c) = 1 + \mu'(s)(\bar{s} - c), \qquad U_{ss}(s,c) = \mu''(s)(\bar{s} - c).$$

Since $c \leq \bar{c} < \bar{s}$ and $\mu''(s) < 0$, we have $U_{ss}(s,c) < 0$ for all $s \in [0, \bar{s}]$. Hence $s \mapsto U(s,c)$ is strictly concave.

The boundary restrictions imply that the maximizer is not at either endpoint. At the lower endpoint, using $\mu'(0) < 0$ and $\bar{s} - c \leq \bar{s}$,

$$U_s(0,c) = 1 + \mu'(0)(\bar{s} - c) \geq 1 + \mu'(0)\bar{s} > 0.$$

At the upper endpoint, using $\mu'(\bar{s}) < 0$ and $\bar{s} - c \geq \bar{s} - \bar{c}$,

$$U_s(\bar{s},c) = 1 + \mu'(\bar{s})(\bar{s} - c) \leq 1 + \mu'(\bar{s})(\bar{s} - \bar{c}) < 0.$$

Strict concavity and the two boundary inequalities therefore imply that there is a unique interior maximizer $s^*(c) \in (0, \bar{s})$, and it is characterized by the first-order condition

$$1 + \mu'(s^*(c))(\bar{s} - c) = 0.$$

This is equation (5).

Because $U_{ss}(s^*(c), c) < 0$, the implicit function theorem implies that $s^*(c)$ is continuously differentiable in $c$. Differentiating the first-order condition with respect to $c$ gives

$$\mu''(s^*(c))(\bar{s}-c)\frac{ds^*(c)}{dc} - \mu'(s^*(c)) = 0.$$

Thus

$$\frac{ds^*(c)}{dc} = \frac{\mu'(s^*(c))}{\mu''(s^*(c))(\bar{s}-c)}.$$

The numerator is negative, while the denominator is also negative because $\mu''(s^*(c)) < 0$ and $\bar{s} - c > 0$. Hence $ds^*(c)/dc > 0$. Since price improvement under internalization is $\bar{s} - s^*(c)$, internalized price improvement is strictly decreasing in $c$. □

*Proof of Proposition 1.* Let

$$m_R \equiv \mathbb{E}[\mu(s_R)] \qquad \text{and} \qquad \Delta(c) \equiv V(c) - W(c).$$

By Lemma 1, the maximizer $s^*(c)$ is unique, interior, and continuously differentiable. The envelope theorem therefore gives

$$V'(c) = \frac{\partial U(s^*(c), c)}{\partial c} = -\mu(s^*(c)) - 1. \tag{A1}$$

Since $W(c) = m_R(\bar{s} - c)$, $W'(c) = -m_R$. Hence

$$\Delta'(c) = m_R - \mu(s^*(c)) - 1. \tag{A2}$$

The function $s^*(c)$ is strictly increasing by Lemma 1, and $\mu$ is strictly decreasing. Therefore $c \mapsto \mu(s^*(c))$ is strictly decreasing, so $c \mapsto \Delta'(c)$ is strictly increasing. Thus $\Delta$ is strictly convex on

$[0,\bar{c}]$.

Next we show that low-cost orders are strictly internalized. At $c = 0$,

$$W(0) = m_R \bar{s}.$$

For every realization $x \in [0,\bar{s}]$ of $s_R$,

$$V(0) = \max_{s\in[0,\bar{s}]} U(s,0) \geq U(x,0) = \mu(x)\bar{s} + x.$$

Taking expectations over $x = s_R$ gives

$$V(0) \geq m_R \bar{s} + \mathbb{E}[s_R].$$

If $\mathbb{E}[s_R] > 0$, then $V(0) > W(0)$. If $\mathbb{E}[s_R] = 0$, then, because $s_R \geq 0$, $s_R = 0$ almost surely. In that case $W(0) = \mu(0)\bar{s} = U(0,0)$. The lower-boundary restriction in Assumption 1 gives

$$U_s(0,0) = 1 + \mu'(0)\bar{s} > 0,$$

so there exists $\varepsilon > 0$ such that $U(\varepsilon,0) > U(0,0)$. Therefore $V(0) > U(0,0) = W(0)$. In all cases,

$$\Delta(0) > 0. \tag{A3}$$

Assumption 2 states that

$$\Delta(\bar{c}) = V(\bar{c}) - W(\bar{c}) < 0. \tag{A4}$$

By continuity of $\Delta$, equations (A3) and (A4) imply that there exists at least one $c^\dagger \in (0,\bar{c})$ such that $\Delta(c^\dagger) = 0$, or equivalently $V(c^\dagger) = W(c^\dagger)$.

It remains to prove uniqueness. Suppose, toward a contradiction, that there are two roots $c_1 < c_2$ in $(0,\bar{c})$. Since $c_2$ lies strictly between $c_1$ and $\bar{c}$, strict convexity implies

$$\Delta(c_2) < \frac{\bar{c}-c_2}{\bar{c}-c_1}\Delta(c_1) + \frac{c_2-c_1}{\bar{c}-c_1}\Delta(\bar{c}).$$

The right-hand side is strictly negative because $\Delta(c_1) = 0$ and $\Delta(\bar{c}) < 0$. This contradicts $\Delta(c_2) = 0$. Hence the root $c^\dagger$ is unique.

Finally, the routing decision is to internalize if and only if $\Delta(c) \geq 0$. Since $\Delta(0) > 0$, $\Delta(\bar{c}) < 0$, and $\Delta$ has a unique zero, continuity implies that $\Delta(c) > 0$ for $c < c^\dagger$ and $\Delta(c) < 0$ for $c > c^\dagger$. At $c = c^\dagger$ the wholesaler is indifferent, and the decision rule in (3) assigns the tie to internalization. Therefore the wholesaler internalizes if and only if $c \leq c^\dagger$ and routes to the RLP if and only if $c > c^\dagger$. $\square$

*Proof of Corollary 1.* By Proposition 1, the RLP event is $\{c > c^\dagger\}$. With the convention $F(x) = \Pr(c \leq x)$,

$$\Pr(\text{RLP}) = \Pr(c > c^\dagger) = 1 - F(c^\dagger),$$

which proves (8).

For the comparative-static statement, consider two cost distributions $F_0$ and $F_1$ over the same cost support, holding the payoff primitives fixed. If $F_1$ shifts costs upward relative to $F_0$ in the sense of first-order stochastic dominance, then $F_1(x) \leq F_0(x)$ for every $x$. Since the cutoff $c^\dagger$ is determined by the payoff primitives and does not depend on the distribution of $c$,

$$\Pr_{F_1}(\text{RLP}) - \Pr_{F_0}(\text{RLP}) = \left[1 - F_1(c^\dagger)\right] - \left[1 - F_0(c^\dagger)\right] = F_0(c^\dagger) - F_1(c^\dagger) \geq 0.$$

Thus the RLP share weakly increases. The increase is strict whenever $F_1(c^\dagger) < F_0(c^\dagger)$. □

*Proof of Proposition 2.* Let

$$x \equiv \mathbb{E}[s_R].$$

Since $s_R \in [0, \bar{s}]$, we have $x \in [0, \bar{s}]$. Condition (11) is equivalent to

$$1 + \mu'(x)(\bar{s} - \bar{c}) < 0.$$

For any $c \in [0, \bar{c}]$, $\bar{s} - c \geq \bar{s} - \bar{c}$. Since $\mu'(x) < 0$,

$$1 + \mu'(x)(\bar{s} - c) \leq 1 + \mu'(x)(\bar{s} - \bar{c}) < 0.$$

Therefore

$$U_s(x, c) = 1 + \mu'(x)(\bar{s} - c) < 0 \qquad \text{for every } c \in [0, \bar{c}]. \tag{A5}$$

For fixed $c$, the map $s \mapsto U_s(s, c)$ is strictly decreasing because $U_{ss}(s, c) < 0$. Since $U_s(s^*(c), c) = 0$, equation (A5) implies

$$s^*(c) < x = \mathbb{E}[s_R] \qquad \text{for every } c \in [0, \bar{c}]. \tag{A6}$$

In particular, (A6) holds on the internalization event $\{c \le c^\dagger\}$. Hence

$$\mathbb{E}[s^*(c) \mid \text{internalized}] < \mathbb{E}[s_R]. \tag{A7}$$

The routing event depends only on $c$, while $s_R$ is independent of $c$ by assumption. Therefore

$$\mathbb{E}[s_R \mid \text{RLP}] = \mathbb{E}[s_R].$$

Combining this equality with (A7) gives

$$\mathbb{E}[s^*(c) \mid \text{internalized}] < \mathbb{E}[s_R \mid \text{RLP}].$$

Subtracting both sides from $\bar{s}$ yields

$$\mathbb{E}[PI^I \mid \text{internalized}] = \bar{s} - \mathbb{E}[s^*(c) \mid \text{internalized}] > \bar{s} - \mathbb{E}[s_R \mid \text{RLP}] = \mathbb{E}[PI^R \mid \text{RLP}],$$

which proves (12). □

*Proof of Corollary 2.* Under the maintained assumptions of Proposition 1, suppose $s_R = 0$ almost surely. Then

$$W(c) = \mu(0)(\bar{s} - c).$$

At $c = 0$,

$$W(0) = \mu(0)\bar{s} = U(0,0).$$

By the lower-boundary restriction in Assumption 1,

$$U_s(0,0) = 1 + \mu'(0)\bar{s} > 0.$$

Thus there exists $\varepsilon > 0$ such that $U(\varepsilon, 0) > U(0,0)$, and therefore

$$V(0) > U(0,0) = W(0).$$

Assumption 2 gives $V(\bar{c}) < W(\bar{c})$. Proposition 1 then implies that there is a unique cutoff $c^{\dagger} \in (0, \bar{c})$ such that the wholesaler internalizes if and only if $c \leq c^{\dagger}$ and routes to the RLP if and only if $c > c^{\dagger}$.

Because $c^{\dagger} > 0$ and the cost distribution has support $[0, \bar{c}]$, the set $\{c \leq c^{\dagger}\}$ has positive probability. Hence some orders are internalized even though $s_R = 0$ almost surely, which is the best feasible execution spread in the model. $\square$