Payment for Order Flow And Asset Choice*

Thomas Ernst† and Chester Spatt‡

January 27, 2023

Abstract

We investigate differences in execution quality and payment-for-order-flow (PFOF) across asset classes. In equities, retail trades receive meaningful price improvement, particularly in tick-constrained stocks, and PFOF is small. In single-name equity options, problematic market structures lead to worse retail price improvement, and PFOF is large. While all option trades execute on-exchange, option exchange rules facilitate internalization. We exploit variation in designated-market-maker (DMM) assignments, minimum tick size, and auction allocation rules, showing that option internalization is imperfectly competitive. Option market structure gives rise to a second potential incentive conflict of brokers: encouraging customers to trade assets offering higher PFOF.

*First Draft: October 2021. For helpful comments, we are thankful to Robert Battalio, Svetlana Bryzgalova, Zhi Da, Brian Hatch, Terrence Hendershott, Steve Heston, Craig Lewis, Saad Ali Khan, Pete Kyle, Dmitriy Muravyev, Bernt Arne Odegaard, Neil Pearson, John Shim, Robert Van Ness, Michael Golding and Edward Monrad at Optiver, several anonymous industry participants, and seminar participants at Carnegie Mellon University, the University of Cincinnati, Hofstra University, the University of Maryland, the Microstructure Exchange, the University of Mississippi, the SFS Cavalcade, the Econometric Society, FMA Europe, FMA Derivatives and Volatility, NBER Big Data meeting, and the University of Washington, and for financial support we gratefully acknowledge the Block Center at Carnegie Mellon.

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

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

I. Introduction

Financial technology has drastically expanded access to the stock market. Today, a retail investor can effortlessly trade stocks on a smartphone, access news on social media, and pay zero fees for both services. While the automation of trading has been a decades-long process, the coronavirus pandemic further accelerated this evolution. Robinhood, LLC (a zero-commission brokerage firm) currently has 22.8 million users, and earns over 80% of its revenue from payment for order flow (PFOF). These are payments from market makers (such as "Citadel") to brokers (such as "Robinhood"), made on the condition that broker privately routes orders to the market maker rather than to the public markets. This practice has raised a series of concerns from both market participants and the U.S. Securities and Exchange Commission (SEC), which has a mandate to maintain fair, orderly, and efficient markets. The first concern is that these trades may not receive the best prices. The second is that the practice of routing customer orders away from markets may lead to wider bid-ask spreads on exchanges. The third is that retail brokerages will have incentives to encourage excess trading.

Within all these concerns, our paper highlights a novel cross-asset difference in PFOF between stocks and options. We investigate the underlying micro-structure of both markets to determine the source of this variation, as well as the respective relationships between PFOF and market quality. In equity markets, PFOF is small, and broker routing to wholesalers who pay PFOF *also benefits retail equity traders* because wholesalers offer smaller bid-ask spreads than the exchanges. Moreover, this occurs even when exchange spreads are at the minimum allowable width, in which case PFOF cannot be leading to artificially wide spreads. In option markets, we find the opposite result: Broker routing to PFOF-paying wholesalers *harms retail option traders.* Thanks to quasi-randomized assignments of Designated Market Makers (DMMs) across option exchanges, we are able to show that PFOF-paying DMMs are causally associated with wider bid-ask spreads. When we look at the PFOF payments made directly to brokers, we show that a similar cross-asset difference occurs: Routing option trades pays brokers substantially more—both in aggregate and per-share—than equity trades. Brokers have more than just an incentive to encourage excess trading of each asset, rather, they have a particular incentive *to encourage excess trading of options.*

This cross-asset difference is magnified by the zero-commission trading environment. From

SEC Rule 606 reports, we estimate the typical PFOF paid to a broker for routing a 100-share option trade is around 40 cents, while the typical payment for routing a 100-share equity trade is around 20 cents. In an era of $5 commissions, an option trade would give a market maker 4% more revenue than an equity trade. With $0 commissions, an option trade would give a market maker 100% more revenue than an equity trade. Differences in prices between stocks and options provide further amplification. The average retail stock trade is in a $25 stock, while the average retail option trade is in a $5 option. A nominal investment of $1,000 in a $25 stock would generate a 40-share equity order worth 8 cents in equity PFOF, while a nominal investment of $1,000 in a $5 option would generate a 200-share option order worth 80 cents in option PFOF. In other words, the same nominal investment in options will generate 10 times as much PFOF as investment in equity. When PFOF is the only source of broker income, this discrepancy in per-asset PFOF creates a very powerful potential incentive misalignment for brokers, with option trades producing substantially more income for the broker than equity trades.

The traditional concern about payment for order flow has not been over asset choice, but instead over execution quality. In equity markets, we carefully document the quality of execution for internalized orders. We identify internalized orders following Boehmer, Jones, Zhang, and Zhang (2021), who highlight sub-penny price improvements. These sub-penny price improvements are offered by market makers to attract retail trades and are distinct from PFOF.[1] While PFOF accrues to the broker, sub-penny price improvement accrues to the customer. We show that the total value of these sub-penny improvements is substantial. In our sample of all U.S. equity trades from January 1, 2019 to October 31, 2021, retail investors receive between $20 and $30 million per month in price improvement. Per trade, sub-penny improvement saves retail investors an average of .5 basis points. Against the concern, raised by SEC Chairman Gensler, that internalization leads to wide bid-ask spreads on exchanges[2], we document that over 50% of sub-penny improvement occurs when bid-ask spreads are at the minimum tick size, and thus the sub-penny improvement

[1]Boehmer et al. (2021) note that sub-penny quoting is not allowed on exchanges, nor is it generally allowed for institutional trades. Sub-penny price improvements therefore offer a method of identifying retail trades which were internalized off-exchange. This method misses retail trades executed on public markets or crossed at the midquote. Our focus is on *internalized* trades; to the extent that trades are executed on-exchange, they are clearly not internalized. Conversely, trades executed at the midquote are internalized, but they receive more than sub-penny improvement, so true execution quality may be even better than we estimate.

[2]In prepared remarks at the 2022 Piper Sandler Global Exchange Conference, SEC Chair Gensler stated "Under the segmentation of the current market, nearly half of trading, along with a significant portion of retail market orders happens away from the lit markets. I believe this may affect the width of the bid-ask spread." Gensler (2021).

often represents a meaningful savings over prevailing spreads.

In option markets, we document that price improvement is very common, but initial option spreads are considerably wider. Due to clearing considerations, option markets do not have off-exchange internalization. Instead, all trading happens on-exchange, with exchange-provided methods for internalizing trades. We contribute an analysis of two key methods of internalizing retail option trades: designated market maker (DMM) assignments, and price improvement mechanism (PIM) auctions. Across both methods, we show that there are limits to competition, which protect market maker internalization revenues, and thus enable market makers to pay more for order flow.

Most U.S. option exchanges have designated market makers (DMMs), who are market makers with both a special obligation to quote securities and special privileges in trading. For the purposes of internalizing trades, the key advantage of a DMM assignment is that a market maker can route an order to their own quote on an exchange, regardless of their time priority or pro-rata share at the NBBO. This self-routing reduces the DMM's incentive to narrow the bid-ask spread, and it prevents any competing market maker from obtaining marketable retail trade flow. Among stocks with just one DMM, we show that DMMs who purchase order flow are associated with wider quoted, effective, and realized spreads, consistent with a direct trade-off between internalizing purchased flow and quoting narrower exchange spreads.

Most stocks do not have a single DMM. While the DMM assignment will be unique at a specific exchange, there are nine exchanges which assign DMMs leading to overlap in coverage. As an example, at the Chicago Board Options Exchange, Global Trading Systems is the DMM in Coca-Cola, while Citadel is the DMM in Walgreens. At the Miami Options Exchange, Global Trading Systems is the DMM in Walgreens, while Citadel is the DMM in Coca-Cola. Comparing option trades in Coca-Cola at CBOE vs. Miami against option trades in Walgreens at CBOE vs. Miami allows us to fit fixed effects for both assets and exchanges, and isolate the role of PFOF. We present evidence that PFOF-paying market-making firms use the specialist system to internalize trades: When a PFOF-paying DMM is present, there are more single-participant trades and these trades earn larger realized spreads, providing direct evidence that DMMs use their self-routing privilege to profitably internalize retail trades.

Market makers can also internalize retail trades in a PIM auction, whereby a market maker brings a retail trade and proposes a price, after which other participants can enter the auction

and propose better prices. These auctions are commonplace, comprising 15% of all trades, and the improvement that the auction generates against the prevailing bid-ask spread ranges from $200 to $600 million per month. These internalization mechanisms appear to generate substantial savings for investors. The bar for improvement in options, however, is much lower than the bar in equity markets. Bid-ask spreads are substantially wider, with minimum ticks on some exchanges ranging from 5 to 10 cents.[3] Option exchanges can also attract wholesalers looking to internalize by adopting rules that favor internalizers.

Our first test of PIM trades examines the winner's curse problem in auctions, where the market maker who initiates an auction has a right to auto-match competing bids. When there are competing bids, auto-matching leads to a tie, and the initiator and competing bidder split the order. The exception is for trades of a single contract, which cannot be split; as a result, any competing bidders receive the order *only if the initiator declines to automatch.* We compare execution quality on 1-contract and 2-contract trades. Trades for just one contract receive less price improvement, and earn higher realized spreads, consistent with a winner's curse preventing competing bids in PIM auctions.

Our second test of PIM trades exploits changes in tick size for a subset of stocks in the Penny Pilot Program, which have a one-penny tick size for options priced below $3.00, and a five-penny tick size above $3.00. We use a regression discontinuity design to examine auctions on options priced just below or just above the $3.00 threshold. Compared to trades just below the cutoff, trades just above receive much larger price improvement, but also have higher realized spreads, consistent with market makers internalizing profitable trades and retaining significant profit for themselves. The increased price improvement around the tick size threshold reflects the increased width of the quotes, and not a reduction in market maker profits in filling orders.

Finally, we compare the performance of retail equity and option portfolios to highlight the dangers of over-trading different assets. From observed retail call option trades of not more than three months maturity, we construct portfolios of 50 assets each. We calculate portfolio returns, assuming the options are held to maturity, and compare against an equivalent portfolio that only trades options. Across over 400,000 samples, the option portfolios have a standard deviation 10

[3]For reference, the minimum tick size in equity markets is 1 cent. While option market making costs are higher (Battalio and Schultz (2011)), the preponderance of price-improving auctions in options reflects that the quotes can be improved for many retail trades.

times higher than that of the stock portfolio—and half of the option portfolios lose more than 90% of their value. Thus the differences across assets for clients are no less stark than the differences for brokers. While a small amount of excess equity trading produces a small PFOF benefit to brokers, the client returns will also generally be similar to the market portfolio return. By contrast, a small amount of excess option trading produces a large PFOF benefit to brokers, and the client will see a return substantially different from the market portfolio return.

While "best execution" can be difficult to define, the goal is straightforward: buying (or selling) at the best available price. In the equity markets, the tight bid-ask spread gives a high bar for price improvement. In the option markets, bid-ask spreads are much wider. While price improvement is common, we show that PFOF-paying DMMs are associated with wider spreads. While the price improvement offered in PIM trades is larger when quoted spreads are wider, the average realized spread of these trades is larger, too. Evaluating broker performance is not just a matter of comparing to one benchmark (like the National Best Bid or Offer), but rather to the best price available.[4] When trades frequently execute at prices better than the best quotes, brokers should deliver such execution quality to their customers.

Compared to equity, option trades have larger discrepancies between the prices obtained and the best quotes reported publicly; this generates potential profit opportunities for market makers, and market makers in options provide higher payments to brokers. We note that this gives rise to a second incentive conflict for brokers, namely a conflict over which assets the customer should trade. While the SEC has scrutinized the "gamification" of trading smartphone apps, the concern has been focused on the overall volume of trading, and not the asset choice. Deviating from a mean-variance optimal portfolio takes a great many equity trades, but just a few option trades. Broker conflicts over asset recommendations are comparatively more difficult to regulate. Rather than being a simple goal of getting the best price, recommending "the best" assets for a client is not a clearly-defined target. Some clients may prefer riskier securities or a more volatile portfolio. Investors with an idiosyncratic risk preference may only be drawn into investing by the allure of potential large positive returns. For these investors, the choice may not be between holding a risky

[4]It also is useful to note that some interpretations of Best Execution standards even suggest that any inducement for order flow (PFOF) should not be taken into account (netted) in assessing the market quality obtained for the customer. Furthermore, it also bears emphasizing that meeting Best Execution responsibilities should reflect the extent of price improvement rather than just simply respecting the NBBO quotes. Indeed, under Best Execution standards brokers are responsible for detecting weaknesses in their own routing at a security and order type level.

portfolio and the market portfolio, but between having a risky investment or not investing at all.[5] It is difficult to say whether a broker does a disservice by catering to these types of investors, and bringing them into investing.

II. Contribution and Literature Review

Payment for order flow has a long history, from regional exchanges attracting broker flow (Chordia and Subrahmanyam (1995)), Bernard Madoff paying for retail brokerage flow in the 1990s, and modeling transaction pricing on option exchanges, analyzed in Battalio, Shkilko, and Van Ness (2016b). Segregation of retail flow can be viewed as either a cream-skimming approach, as in Easley, Kiefer, and O'Hara (1996), or a segmentation and inventory management approach, as in Baldauf, Mollner, and Yueshen (2022). Battalio and Holden (2001) argue segmentation is optimal for retail traders, while Parlour and Rajan (2003) argue it decreases consumer welfare. Many brokers have transitioned to zero-commission trading, with PFOF as their primary revenue source, as documented in Jain, Mishra, O'Donoghue, and Zhao (2020) and Kothari, Johnson, and So (2021). Our paper identifies the importance of cross-asset differences in payments, which is a new incentive conflict arising from the combination of PFOF with zero-commission trading.[6] We show that brokers have a powerful incentive to encourage not just more trading of all assets, but in particular trading of options due to their higher payment for order flow.

Our option analysis utilizes new data on the option market. While the role of option designated market makers has been studied Mayhew (2002), we add novel analysis on DMM ability to internalize retail trades, as well as the differences in outcome depending on whether option DMMs are also wholesalers. This dual role in option markets distinguishes option DMMs from traditional equity DMMs, where their role is studied only as a liquidity provider, as in Bessembinder, Hao, and Zheng (2020), Venkataraman and Waisburd (2007), Clark-Joseph, Ye, and Zi (2017), Foley, Liu, Malinova, Park, and Shkilko (2020), and the related specialist system in Madhavan and Smidt

[5] In questioning by House Committee on Financial Services member Jim Himes (D-CT) suggesting that the proper benchmark should be the market portfolio, Robinhood CEO Vladimir Tenev argued instead that the proper benchmark for these investors is not investing at all. (February 18, 2021: Game Stopped? Who Wins and Loses When Short Sellers, Social Media, and Retail Investors Collide.)

[6] Typical payment for a 100-share trade is 40 cents in options markets and 20 cents in equity markets. On a fixed $5 commission, the option trade earns brokers 4% more than the stock. With a $0 commission, the option pays a broker 100% more than the stock. These differences are per-share, so if the investment is in nominal terms, these price differences can be further amplified by the lower nominal price of options.

(1993). Liquidity providers compete on posting narrow quoted spreads, while internalizing retail spreads is more profitable when spreads are wide. The incentives of a DMM who also internalizes trades are therefore divergent from those of a DMM only providing displayed liquidity, and we show that outcomes do differ, with PFOF-paying DMMs associated with larger bid-ask spreads relative to non-PFOF-paying DMMs.

Options markets typically have higher trading costs than equity markets, in part due to greater difficulty market makers face in hedging, as in Battalio and Schultz (2011). This gamma hedging, in turn, leads to impacts on equity markets, as shown by Ni, Pearson, Poteshman, and White (2021). While option spreads are wide, investors can time their trades (Muravyev and Pearson (2020)) or route to exchanges with particular make-take pricing models (Battalio et al. (2016b)). Battalio, Griffith, and Van Ness (2021) examine the switch of the PHLX options exchange from make-take to PFOF pricing model. While the PFOF exchange pricing model has been studied, our paper instead focuses on the PFOF from wholesalers to brokers. Wholesalers may internalize trades on both PFOF-pricing model exchanges, but also on traditional make-take exchanges or fee-fee exchanges, with retail trades occurring on all platforms.

We also analyze price improvement mechanisms (PIM) in options. PIM trades are auctions which enable potential price improvement for customers, have only recently been separately identified in OPRA data, and are growing substantially in popularity. Three contemporaneous working papers analyze option auctions. Bryzgalova, Pavlova, and Sikorskaya (2022) show that price improvement auctions are correlated with other measures of retail trading and that retail traders fail to optimally exercise calls before dividends. Eaton, Green, Roseman, and Wu (2022) suggest retail traders increase volatility, while Hendershott, Khan, and Riordan (2022) examines price improvement auctions theoretically and empirically, and argues option auctions are consistent with cream-skimming, having lower price impact than limit order book trades. The model also provides tests of competitiveness and their evidence suggests that neither auctions nor the limit order book are perfectly competitive. Our paper documents that auctions are one of two possible routes for internalizing marketable retail trades, with DMM assignments providing an alternative method. Within auctions we highlight the winner's curse problem, showing single-contract trades receive less price improvement and earn higher realized spreads than two-contract trades. We also introduce a novel regression discontinuity analysis utilizing the option Penny Pilot to show that the price

improvement in auctions often can reflect very wide quotes rather than competitive improvement of quotes.

Payment for order flow is related to off-exchange execution, including both broker-affiliated venues and alternative trading systems. Routing to broker-affiliated venues for institutional orders typically produces poor execution quality, as Battalio, Corwin, and Jennings (2016a) and Anand, Samadi, Sokobin, and Venkataraman (2021) show. We show, however, that retail traders receive fairly good execution quality in equity markets. Unlike institutions, retail traders are more likely to execute small individual trades that are smaller than the available quantities at the national best bid or offer. Our work makes extensive use of microsecond timestamps in our analysis of execution quality. Against the standard practice of matching trades and quotes with SIP timestamps (Holden and Jacobsen (2014)), we explore an alternative of using the participant timestamps. In part, we are motivated by Hasbrouck (2018)'s documentation of local exchange liquidity, and Bartlett and McCrary (2019), who document occasional differences between the SIP and proprietary feed of the national best bid and offer. Methodologically, we show that SIP and proprietary feed differences are important for exchange trades, but less important for retail trades. In recent work, Barardehi, Bernhardt, Da, and Warachka (2022) show that sub-penny price improvement occurs not just with retail trades, but in particular when market makers decide internalization is profitable. Consistent with this, we document overwhelmingly positive spreads for internalized trades, but we also note that these trades mostly frequently occur when spreads are at the minimum tick size.

The minimum tick size has come under scrutiny. Li, Wang, and Ye (2021) connect much HFT size to tick-constrained stocks, while Li and Ye (2022) show that some stocks are round-lot constrained, and not tick constrained. Bartlett, McCrary, and O'Hara (2022) highlight how significant information is contained in odd-lot quotes. In our work, we note that when stocks are tick-constrained, internalization is better for investors than crossing the spread, and that banning internalization could not lead to narrower spreads in stocks which are already tick constrained unless tick sizes were also redefined.

Schwarz, Barber, Huang, Jorion, and Odean (2022) place randomized equity trades with a series of brokers, and find that trades frequently receive midquote pricing. We identify equity retail trades by Boehmer et al. (2021) and sign trades according to Barber, Huang, Jorion, Odean, and Schwarz (2022); while this misses equity trades which execute at midquote, our focus is on internalized

trades and their broader effect on market quality, in both equity and option markets. We show that most equity trade occurs in tick-constrained or narrow-spread stocks. For the majority of equity trading, therefore, routing to the exchange and paying the full spread would lead to worse outcomes for retail traders. Hu and Murphy (2022) present evidence that internalization may lead to worse spreads; their focus is not on the average stock, but rather is on the very smallest stocks with wide spreads. Their model of a monopolist internalizer increasing spreads is consistent with our evidence of monopolist option DMMs who pay for order flow being associated with wider spreads.

The negative relationship between trading and returns has been studied over time in various contexts ranging from the institutional setting in Jensen (1968) to the individual-investor setting of Barber and Odean (2000). Robinhood investors, in particular, have earned lower profits with their trading, as documented in Barber, Huang, Odean, and Schwarz (2021), and among stocks, Greenwood, Laarits, and Wurgler (2022) show that retail traders like low-priced, high-volatility stocks. Our paper highlights the significant asset-choice side dimension to over-trading. We document that while a portfolio drawn from observed retail equity trades delivers close to the market return, a portfolio drawn from observed retail option trades delivers a substantially lower return and higher variance.

III. Internalization in U.S. Equities

Brokers have a best execution requirement on behalf of their clients. The Financial Industry Regulatory Authority (FINRA) explicitly defines this best-execution requirement via Rule 5310, which requires that members "shall use reasonable diligence to ascertain the best market for the subject security, and buy or sell in such market so that the resultant price to the customer is as favorable as possible under prevailing market conditions."

In modern U.S. equity markets, determining the best market for a security is a sophisticated process. Trading takes place both across several exchanges and in non-exchange platforms, with data centers located hundreds of miles apart. This physical distance complicates the market search; while brokers can view the current best quotes at each exchange, sending an order to any exchange will take time. Even at the speed of light, during the transit time, quotes may change. Thus the

routing decision of a broker must take into account not just the market now, but also latency in the current pricing as well as what the market may be milliseconds into the future.

Against this challenge finding the best exchange quote, brokers may route to an off-exchange trading venue (or internalize the order). This carries the potential advantage of receiving price improvement compared to the prevailing quotes, either from a mid-quote matching facility like a dark pool, or wholesale liquidity provider offering reduced spreads to retail clients.

We measure the price improvement and execution quality provided by brokers. We evaluate aggregate broker performance with a careful technical analysis of the markets. First, we measure prices against both the exchange timestamps, and the slower Securities Information Processor (SIP) timestamps. We find that differences are common for exchange trades, but less common for off-exchange trades. Differences are uncommon for likely retail trades with de minimis price improvement. These likely retail trades are also far less likely to see changes in the price both before and after the trade, suggesting no incentive to artificially delay execution by a few microseconds. We also find that even the de-minimis price improvement of 20 or 30 hundreths of a cent add up to substantial savings for investors, and most stocks are quoted close to the minimum 1 cent bid-ask spread.

A. Data

We use NYSE TAQ (Trade and Quote) data from January 1, 2019 to October 31, 2021. We examine all securities which have a closing price of at least $1 and trade on at least three-fourths of the days of our sample; this yields a sample of 6,009 individual securities. All trades and quotes are cleaned according to the techniques described in Holden and Jacobsen (2014).

B. Price Improvement

Brokers have flexibility in routing their client orders, and an obligation to obtain the best price possible for their clients. Wholesalers can internalize trades off-exchange, provided brokers route to them. To induce brokers, wholesalers can offer payment for order flow and sub-penny price improvement. We examine the payments for order flow in Section V, and analyze the sub-penny price improvements here. Boehmer et al. (2021) document these sub-penny price improvements, and use them to identify retail trades. While their focus is on the predictive power of retail

trades, our focus will be on the execution quality of retail traders as well as the level of sub-penny improvements.

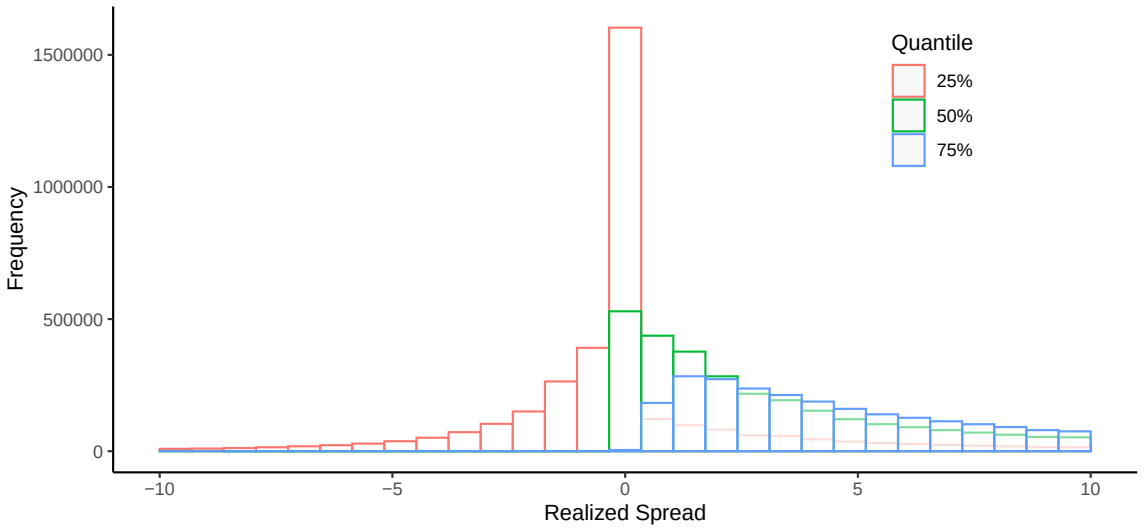
Sub-penny trades are defined as trades in which the price has a sub-penny component, but is not a mid-quote trade. As an example, a trade price of $10.2515 has a sub-penny component of 15 hundredths of a cent. Following Boehmer et al. (2021), we define a trade as a sub-penny improvement if the sub-penny component falls between (0, 40) and (60, 100). These trades are predominantly retail trades, as institutions typically trade either on-exchange, at the mid-quote, or at even penny increments.

Market makers offer sub-penny price improvement to attract retail flow. Under a Glosten and Milgrom (1985) model, bid-ask spreads depend on the ratio of informed to uninformed traders. If retail traders are less informed than the general population of traders, market makers could charge a smaller bid-ask spread to retail traders. We directly confirm this with observed realized spreads, comparing the price at the time of trade to the mid-quote one second after the trade. These realized spreads reflect the potential profits of a market maker; positive values indicate market makers make money, while negative values indicate market makers loose money. We measure realized spreads for all stock-day observations in our sample. Figure 1 plots the distribution of realized spreads, both for all trades and for the subset of trades which receive sub-penny price improvement. Realized spreads are overwhelmingly positive for retail trades, with positive values at the 25%, 50%, and 75% quantiles of realized spreads. In contrast, across all trades, the 25% quantile is frequently negative (most, but not all, stock-days have a negative observation for the 25% quantile), suggesting market makers loose money on a portion of all trades. As a result, internalized retail trades appear to be far more profitable for market makers in the short run, and providing a sub-penny price improvement is one way for market makers to induce brokers to route more of these profitable retail trades.

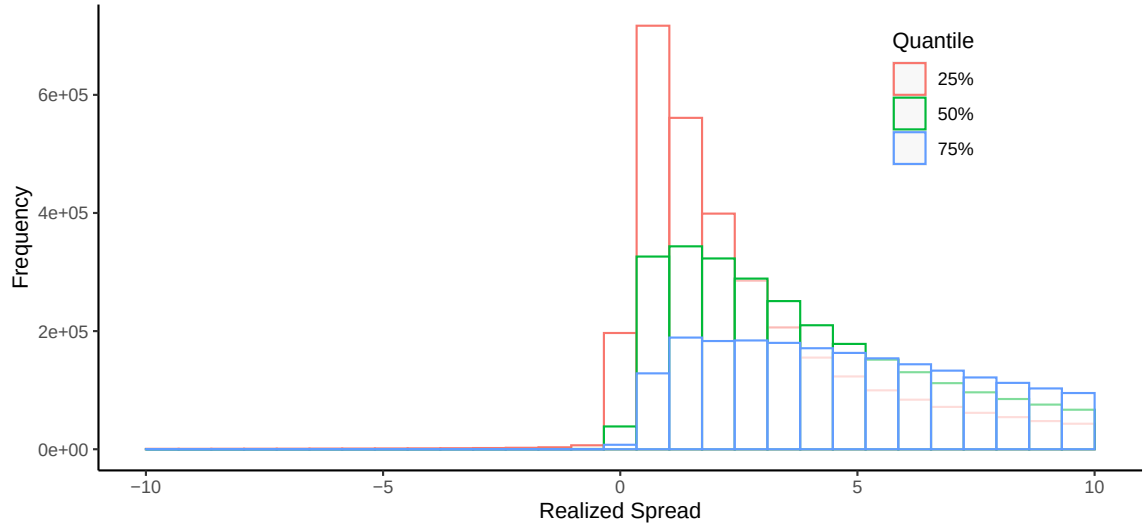
We calculate the combined value of all sub-penny improvements, and find that the sub-penny improvements have substantial total value. Figure 2 plots the monthly total improvements, with most months having between $20 and $30 million in sub-penny price improvement. By total improvement, we calculate the total value of only the sub-penny portion of improvement. For example, if a trade executes at $10.2515, the total improvement is $0.0015, or fifteen hundredths of a cent. For a trade which executes at $10.2595, the total improvement is $0.0005, or five hundredths of a cent. It is these fractional cents which we add up to arrive at the monthly totals.

Figure 1. Comparison of Realized Spreads. For each stock-day observation, we measure realized spreads at the 25%, 50%, and 75% quantiles. Realized Spreads are measured as the signed price difference between the order price and the mid-quote one second after the trade, and reflect the short-term profit available to market-making. Panel A plots the distribution of realized spreads for all trades, while Panel B plots the distribution of realized spreads for trades which receive sub-penny price improvement.

Panel A: Realized Spreads for All Trades

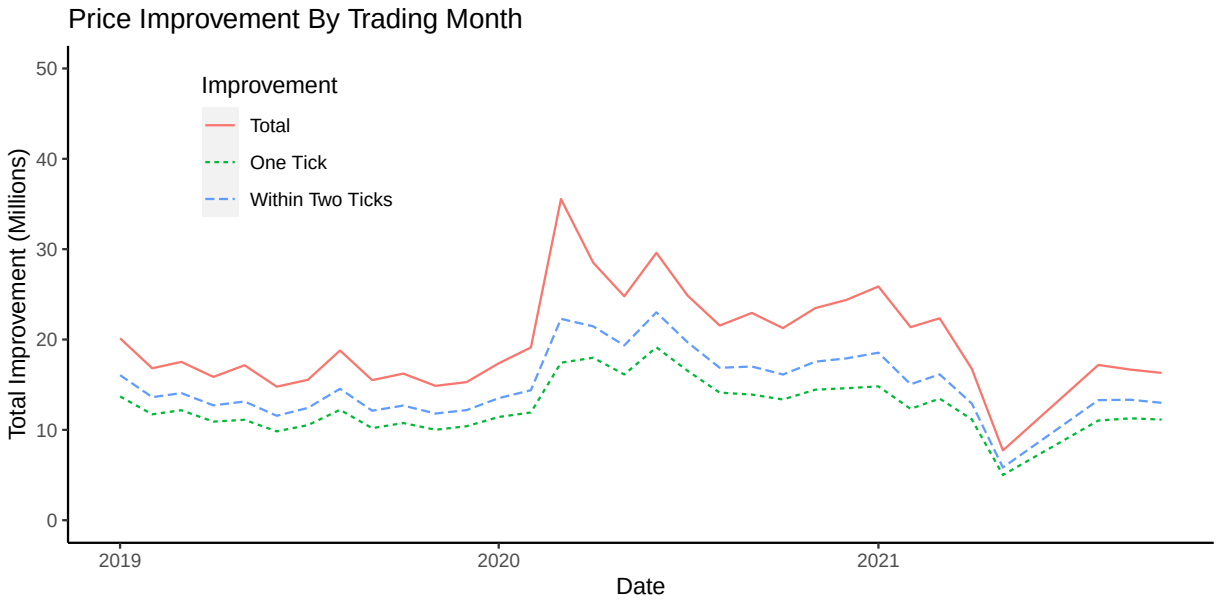


Panel B: Realized Spreads for Sub-penny Trades



While these volumes are substantial in part due to the tremendous volume of U.S. equities trades, they also represent a meaningful portion of transaction costs. With monthly transaction

Figure 2. Price Improvement By Month. Total price improvement for sub-penny trades, by month. Following Boehmer et al. (2021), sub-penny trades are defined as trades in which the price has a sub-penny component between (0, 40) and (60, 100). For example, a trade price of $10.2515 has a sub-penny price improvement of 15 hundreths of a cent. We calculate the total dollar value of all such sub-penny price improvements, and plot the monthly total with the red solid line. We separately calculate the total value of sub-penny improvements which occur when the quoted spread is at one tick (green dashed line) or at two ticks (blue dashed line).



volumes of sub-penny trades ranging between $300 billion and $1 trillion, the average sub-penny price improvement is around half a basis point. This is a substantial amount of price improvement, considering the liquidity of U.S. equity markets. As Figure 2 shows, around half of the total sub-penny price improvement occurs when stocks are at the minimum one-tick spread, i.e., a single penny bid-ask spread.

In the Internet Appendix B, we estimate improvement by comparing the execution price to the best bid or offer at the time of trade. This risks errors in timing. Consider, for example, a trade to buy stock in a rising market. If the timestamp on the trade is delayed (relative to timestamps on quotes), the buy order will appear to have executed at a very low price and received generous improvement. If we do measure improvements against the prevailing spread, we get substantially larger estimates of improvement.

To further capture the value of these price improvements, we consider how market-making profits would be impacted if there were to be no sub-penny price improvements. We use realized

spreads as a measure of market-making profits, as the realized spread compares the signed trade price against the mid-quote some time interval after the trade.

For each sub-penny price-improved trade, we also consider the national best bid or offer, $NBBO$, at the time of the trade. We compare two different measures of realized spreads: one measured against the actual trade price, and one measured using the trade price adjusted for the sub-penny improvement. That is, for a trade price of \$10.2585, the realized spread without improvement would be calculated using a price of \$10.26. This measure of realized spreads without sub-penny price improvement could be thought of as the total potential revenue available to a market maker, while the measure of realized spreads using the actual trade price is the total profit. If sub-penny price improvement is viewed as an expense, the ratio of the two realized spread measures captures the share of total revenue devoted to sub-penny price improvement.

Formally, for a trade price P_T, trade sign Y, sub-penny improvement Sub_t, and midquote m which occurs X seconds after the trade, we define the two possible definitions of a realized spread:

$$\text{Realized\_With\_Improvement}_t = Y(P_t - m_{t+X}) \tag{1}$$

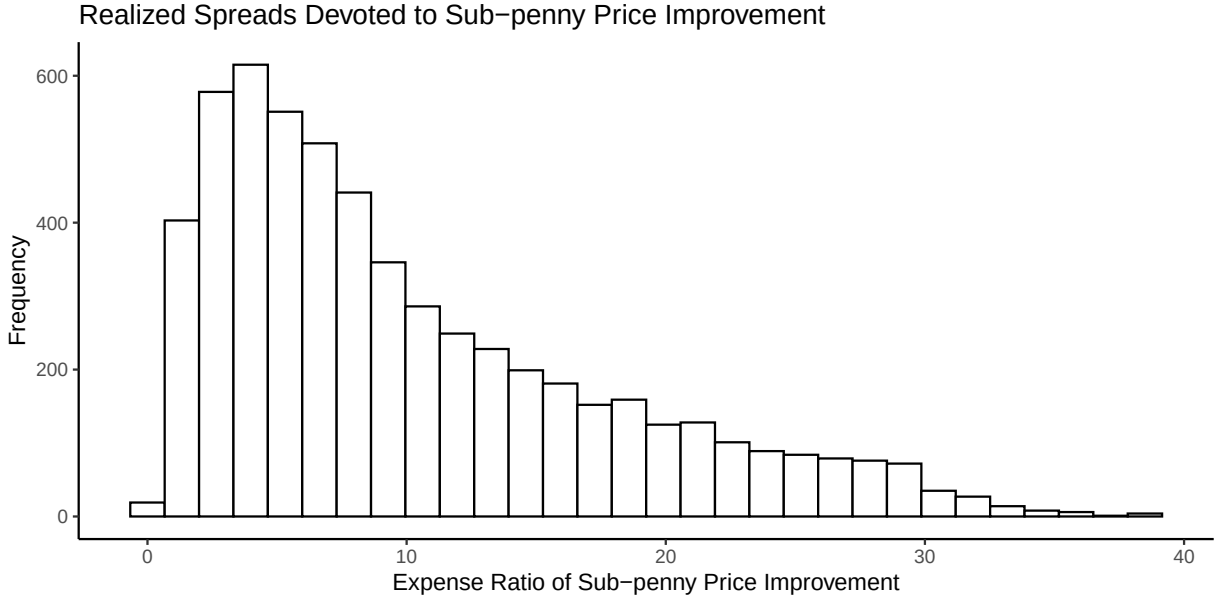
$$\text{Realized\_No\_Improvement}_t = Y(P_t - m_{t+X}) + Sub_t \tag{2}$$

$$\text{Expense\_Ratio} = 1 - \frac{\sum \text{Realized\_With\_Improvement}}{\sum \text{Realized\_No\_Improvement}} \tag{3}$$

For each stock, we calculate the value of this expense ratio, and plot the distribution of this ratio across stocks in Figure 3. The average stock has approximately a 10% difference between the two measures. If we consider the total revenue from retail trades to be the realized spread on these trades calculated using the price without sub-penny improvement, around 10% of this total revenue goes to offering sub-penny price improvements.

These calculations of market maker profits do not take into account how the NBBO itself depends on market maker behavior. Ernst, Sokobin, and Spatt (2021) examine how the off-exchange trades can influence on-exchange trading. This element of the value of off-exchange trades will not be captured in the analysis of this paper, but we do examine the relationship between sub-penny price improvement and quoted spreads. As Figure 2 shows, much improvement already occurs when spreads are at the minimum allowed spread. For these trades when spreads are already at the minimum, no alternative routing or internalization could lead to narrower quoted spreads.

Figure 3. Price Improvement As Fraction of Realized Spreads. Realized spreads for sub-penny price-improved trades are as much as 40% lower than a realized spread measured against the contemporaneous national best bid or offer. For all sub-penny trades, we calculate the realized spreads using both the trade price and the NBBO (Equation 3). The realized spread using the trade price reflects market maker profits, while the realized spread using the NBBO reflects market maker revenue. Total realized spreads in sub-penny trades are typically between 5% and 20% lower than the total realized spreads on those same trades using the NBBO rather than the trade price. This suggests roughly 5 to 20% of the total revenue market makers make from retail trades is allocated to sub-penny price improvement.



C. Regression Analysis of Equity Price Improvement

To formally measure the relationship between retail improvement and market conditions, we estimate three regressions. Regression 1 estimates the relationship between the share of trades receiving price improvement and market conditions. Regression 2 estimates the relationship between the share of realized spreads devoted to sub-penny price improvement and market conditions. Regression 3 estimates the relationship between the share of trades which have quote changes around the time of the trade and market conditions.

REGRESSION 1: *For each stock i on date t, we estimate:*

$$\begin{aligned} ImprovementShare_{it} = \quad & \alpha_0 Closing\_Price_{it} + \alpha_1 Absolute\_Intraday\_Return_{it} \\ & + \alpha_2 Mean\_Quoted\_Spread_{it} + \alpha_3 Mean\_Realized\_Spread_{it} \\ & + \alpha_4 Off\_Exchange\_Non\text{-}Subpenny\_Share_{it} + X + \epsilon_{it} \end{aligned}$$

Results are presented in Table 1. Improvement Share is the dollar volume share of trades receiving sub-penny price improvement. Closing Price is measured in dollars, Absolute Intraday Return is the percentage intraday return from open to close, mean quoted spread is the trade-weighted mean, realized spread is measured at the one second level, off-exchange share is the share of dollar volume executed off-exchange, and X is a fixed effect estimated for the each stock (Table 1, Column 1) or date (Table 1, Column 2).

Large absolute returns are strongly associated with lower shares of trades receiving sub-penny price improvement. A 1% larger absolute return is associated with a 34% reduction in the volume share of trades receiving sub-penny price improvement, consistent with sub-penny trading increasing less rapidly than exchange trading in times of volatility or with market makers giving less sub-penny improvement in times of market volatility. Consistent with the former hypothesis, the Off-Exchange Non-Subpenny Share is positively correlated with sub-penny improvement: that is, in the time series, the same conditions at which there is substantial off-exchange trading also have substantial sub-penny trading. In the cross-section, more off-exchange trading is associated with fewer orders receiving sub-penny price improvement, perhaps consistent with potential substitution. While sub-penny price improvement from the spread is substantial and appears to save investors substantial transaction costs, if these investors could receive execution at the midquote, this would be an even larger savings. As midquote facilities are by their very nature "dark" in the sense that they have no pre-trade transparency, we cannot directly test whether there is any potential midquote volume available at the time trades are given sub-penny price improvement.

For quoted spreads, larger quoted spreads and realized spreads are both associated with a larger share of orders receiving sub-penny price improvement. As the minimum quoted spread is one penny, larger quoted spreads reduce the value to investors of sub-penny price improvement, especially when compared to the half-spread reduction crossing at the midquote would provide.

Table I: Price Improvement Share. This table estimates Regression 1. ImprovementShare measures the volume share of prices which receive sub-penny price improvement. Closing Price is measured in dollars, while Absolute Intraday Return and Off-Exchange Non-Subpenny Share are measured in percentages. The level of observations is the stock-day level. Column (1) has a fixed effect for each date, while column (2) has a fixed effect for each stock. Standard errors are clustered at the stock and day level.

	Dependent variable:	
	ImprovementShare	
	(1)	(2)
Closing Price	-0.012^{***}	
	(0.002)	
Absolute Intraday Return	-34.210^{***}	0.735
	(4.743)	(0.713)
Mean Quoted Spread (BPS)	0.004^{***}	-0.002^{***}
	(0.001)	(0.0004)
Mean Realized Spread (BPS)	0.069^{***}	0.034^{***}
	(0.004)	(0.002)
Off-Exchange Non-Subpenny Share	0.043^{***}	-0.231^{***}
	(0.005)	(0.004)
Observations	3,689,300	3,689,300
R^2	0.059	0.524
Adjusted R^2	0.058	0.523
Residual Std. Error	11.995 (df = 3688636)	8.538 (df = 3683531)
Note:		*p<0.1; **p<0.05; ***p<0.01

Larger realized spreads are strongly associated with a larger share of orders receiving improvement, with a 10 basis point increase in realized spreads being associated with a 0.7% increase in the share of trades receiving improvement. With larger realized spreads, trading at the prevailing quotes becomes more profitable for market makers, and purchasing trade volume becomes more attractive. As Figure 3 shows, while the median sub-penny price improvement costs market makers around 10% of the total realized spread, there is considerable variation across stocks.

To test how the profitability of internalization changes with market conditions, we test Regression 2. As defined in Equation 3, the realized ratio is the ratio between two different measures of realized spreads: one measured against the actual trade price, and one measured against the national best bid or offer. The measure of realized spreads using the NBBO could be thought of as the total potential revenue available to a market maker, thus the difference between the two realized spreads represents the cost to the market maker of offering the sub-penny price improvement. This ratio is defined only for sub-penny improved trades, and to reduce the variance of this ratio we exclude any stock-day observations with fewer than 50 sub-penny trades.

REGRESSION 2: *For each stock i on date t with at least 50 sub-penny trades, we estimate:*

$$\begin{aligned} Realized\_Ratio_{it} = \quad & \alpha_0 Closing\_Price_{it} + \alpha_1 Absolute\_Intraday\_Return_{it} \\ & + \alpha_2 Mean\_Quoted\_Spread_{it} + \alpha_3 Mean\_Realized\_Spread_{it} \\ & + \alpha_4 Off\_Exchange\_Non\text{-}Subpenny\_Share_{it} + X + \epsilon_{it} \end{aligned}$$

Results of this estimation are presented in Table II. The mean realized spread is measured in basis points and at the one-second time interval. Larger realized spreads are associated with a smaller share of the realized spread being devoted to sub-penny price improvement. Mean quoted spreads are also associated a smaller share of the spread being devoted to improvement, suggesting that the sub-penny price improvement is more lucrative for wholesalers when spreads are wider, though by assumption, we are only counting the sub-penny share of price improvement. In the Internet Appendix B we make the more permissive definition of improvement as all improvement relative to the NBBO on sub-penny trades, and find that larger quoted spreads are instead associated with *more* improvement, suggesting market makers may have fixed costs to cover which

prevent additional improvement for narrow-spread stocks.

At a one-tick bid-ask spread, the potential realized spreads are small, so the sub-penny improvement represents a larger share of market maker revenues. With stock fixed effects, the association between intraday returns and realized ratio is negative, suggesting that on days when an individual stock makes a large move, market makers devote a smaller share of realized spreads to sub-penny improvement. With date fixed effects, the association is positive, suggesting that stocks which are more volatile than the average stock also see a larger share of realized spreads devoted to sub-penny improvement.

To test whether market conditions have a similar relationship with retail trade timing as they do with non-retail trades, we estimate Regression 3.[7] Quote_Difference_Share measures the percentage of trades for which the quotes matched at one time differ from the quotes matched at another time. We evaluate three time comparisons: quotes 3 milliseconds prior to execution against quotes at execution, quotes at the SIP time of a trade against quotes at the participant timestamp, and quotes 3 milliseconds after execution against quotes at execution. We plot the distribution of differences for each of these timestamps in Figure 12 in Appendix A.

REGRESSION 3: *For each stock i on date t, we estimate:*

$$\begin{aligned} Quote\_Difference\_Share_{it} = \quad & \alpha_0 Closing\_Price_{it} + \alpha_1 Absolute\_Intraday\_Return_{it} \\ & + \alpha_2 Mean\_Quoted\_Spread_{it} + \alpha_3 Mean\_Realized\_Spread_{it} \\ & + \alpha_4 Off\_Exchange\_Non\text{-}Subpenny\_Share_{it} + X + \epsilon_{it} \end{aligned}$$

Results of Regression 3 are presented in Table III. Larger returns are strongly associated with a larger quote difference share at all time horizons. An intraday return is a prerequisite for price changes. The coefficient estimate is much larger for non-retail trades than it is for retail trades, but the mean level of quote differences, as well as the variance, are higher for non-retail trades.

Higher closing prices are also associated with a larger quote difference share, as stocks with a larger price typically have more frequent price changes, given the minimum one-penny bid-ask spread. Stocks with higher realized spreads have lower quote difference shares; one potential explanation is that quote difference shares are often reversals, which reduce realized spreads. The

[7] We provide a detailed examination of the timestamps and trade-matching in Appendix A.

Table II: Realized Spread Share. This table estimates Regression 2. Realized Ratio, defined in Equation 3, is the ratio on realized spreads for sub-penny improved orders, and compares the realized spread calculated with the trade price against the realized spread calculated with with no sub-penny price improvement. A larger ratio indicates a larger share of market-maker revenue goes to offering sub-penny price improvement. Closing Price is measured in dollars, while Absolute Intraday Return and Off-Exchange Share are measured in percentages. The level of observations is the stock-day level; to reduce noise in the realized spread ratio, we exclude stock-day observations with less than 50 sub-penny trades. Odd columns have a fixed effect for each date, while even columns have a fixed effect for each stock. Standard errors are clustered at the stock and day level.

	Dependent variable:					
	Realized Ratio 3ms		Realized Ratio 1s		Realized Ratio 30s	
	(1)	(2)	(3)	(4)	(5)	(6)
Closing Price	-0.034^{***}		-0.036^{***}		-0.036^{***}	
	(0.006)		(0.006)		(0.006)	
Absolute Intraday Return	-16.579^{***}	-20.905^{***}	-13.848^{***}	-20.102^{***}	-13.848^{***}	-20.102^{***}
	(2.898)	(3.003)	(2.678)	(2.912)	(2.678)	(2.912)
Mean Quoted Spread (BPS)	-0.018^{***}	-0.007^{***}	-0.018^{***}	-0.007^{***}	-0.018^{***}	-0.007^{***}
	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
Mean Realized Spread (BPS)	-0.063^{***}	-0.030^{***}	-0.071^{***}	-0.035^{***}	-0.071^{***}	-0.035^{***}
	(0.005)	(0.003)	(0.006)	(0.003)	(0.006)	(0.003)
Off Exchange Non-Subpenny Share	0.059^{***}	-0.026^{***}	0.054^{***}	-0.027^{***}	0.054^{***}	-0.027^{***}
	(0.006)	(0.002)	(0.006)	(0.002)	(0.006)	(0.002)
Date Fixed Effect	X		X		X	
Stock Fixed Effect		X		X		X
Observations	3,052,180	3,052,180	3,045,926	3,045,926	3,045,926	3,045,926
R^2	0.186	0.652	0.165	0.579	0.165	0.579
Adjusted R^2	0.185	0.652	0.164	0.579	0.164	0.579
Residual Std. Error	9.953	6.508	11.202	7.954	11.202	7.954
Degrees of Freedom	3051516	3046411	3045262	3040157	3045262	3040157
Note:						*p<0.1; **p<0.05; ***p<0.01

effect of off-exchange trades differs between the retail and non-retail trades.

Across all specifications, market conditions explain a large portion of the variance in quote difference shares for all trades, and a much smaller portion of the variance in quote difference shares for retail trades. The R^2 for the non-retail trades is between 20 to 30%, while the R^2 for the retail trades is below 2%. As retail traders do not time their trades at the millisecond level, as a result, their trades are unlikely to benefit from executing a few milliseconds faster or slower.

Table III: Quote Difference Share. This table estimates Regression 3. Quote Difference Share measures the percentage of trades which have a quote change from one timestamp to another. Columns (1) and (2) measure how often the quote from 3 milliseconds before a trade differs from the quote at the time of trade. Columns (3) and (4) measure how often the quote matched to a trade's SIP timestamp differs from the quote matched to the trade's participant timestamp. Columns (5) and (6) measure how often the quote from 3 milliseconds after a trade differs from the quote at the time of trade. Closing Price is measured in dollars, while Absolute Intraday Return and Off-Exchange Share are measured in percentages. Odd columns estimate the effect for all trades, while even columns estimate the effect for retail trades only. Standard Errors are clustered at the stock and day level.

	Dependent variable: Quote Difference Share					
	3 Milliseconds Prior		SIP vs. Participant		3 Milliseconds Posterior	
	All Trades	Sub-penny	All Trades	Sub-penny	All Trades	Sub-penny
	(1)	(2)	(3)	(4)	(5)	(6)
Closing Price	0.060***	0.003***	0.031***	−0.001***	0.045***	0.002***
	(0.011)	(0.0005)	(0.006)	(0.0001)	(0.008)	(0.0002)
Absolute Intraday Return	114.835***	2.712***	62.596***	0.338***	104.029***	2.347***
	(15.701)	(0.495)	(8.671)	(0.129)	(14.096)	(0.324)
Median Quoted Spread (BPS)	0.004*	0.002***	−0.0004	0.001***	0.004**	0.001***
	(0.002)	(0.0002)	(0.001)	(0.0001)	(0.002)	(0.0001)
Realized Spread (BPS)	−0.221***	−0.001	−0.119***	−0.001**	−0.181***	−0.001***
	(0.012)	(0.001)	(0.007)	(0.0003)	(0.011)	(0.0003)
Off-Exchange Non-Sub-penny Share	−0.471***	0.011***	−0.282***	0.004***	−0.421***	0.0001
	(0.007)	(0.001)	(0.005)	(0.0002)	(0.006)	(0.0002)
Observations	3,689,300	3,689,300	3,689,300	3,689,300	3,689,300	3,689,300
R^2	0.286	0.029	0.223	0.013	0.275	0.054
Adjusted R^2	0.286	0.029	0.223	0.013	0.275	0.054
Residual Std. Error	17.829	2.308	12.049	1.338	15.294	0.890
Note:						(df = 3688636) *p<0.1; **p<0.05; ***p<0.01

IV. Internalization in Options Markets

Options trade in a significantly different regulatory framework than U.S. equities. Unlike equities, options carry risks of counter-party default. The Options Clearing Corporation oversees the clearing of all options trades in U.S. single name equities, and requires that these options be traded on an exchange. There is no off-exchange internalization or dark trading; unlike U.S. equities, all option trades occur on lit exchanges.

Trades can, however, still be internalized, and the exchanges offer trading procedures which facilitate this process. We highlight two particular trading methods which facilitate internalization. The first method, Designated Market Maker (DMM) assignments, allow wholesalers to directly internalize trades on an exchange, and are described in detail in Sections IV B and C. The second, Price Improvement Mechanism (PIM) trades, which create auctions with the opportunity to provide price improvement (which accrues to the customer, and not the broker), are described in detail in Section IV D and E.

There are 16 competing options trading venues, and wholesalers looking to internalize a trade have a choice of exchange. This leads to potential competition among some exchanges to offer terms that are particularly favorable to a wholesaler looking to internalize. In recent remarks, the SEC chair called for regulators to "to draw upon lessons from the options market, focusing on assuring full competition among all market participants to provide the best prices for retail investors" Gensler (2022). Our analysis highlights the extent to which competition in the options market is not exemplary, but is in fact limited, with the auctions and DMM allocations both favoring the internalizer and limiting competition.

A. Data

We obtain all U.S. equity options trades reported by the Options Price Reporting Authority (OPRA) from November 4, 2019 to December 31, 2021 through SpiderRock.[8] From SpiderRock, we also obtain matching quotes for the option as well as the underlying asset, and the option Greek values. For matching quotes, we obtain both the best bid and best ask at the exchange where a trade occurs, and the national best bid and ask across all exchanges.

[8]SpiderRock Holdings, LLC is a Chicago-based market data vendor and offers SpiderRock EXS, an agency broker-dealer. Due to a limitation of the data provider, there is some missing data during January, 2021.

We sign trades according to Lee and Ready (1991), using the local exchange best bid or offer for "auto-electronic trades" (OPRA Code 73) and using the NBBO for "Single Leg Auction Non ISO" (OPRA Code 97). For calculating realized spreads, we use the option NBBO at 1 minute and 10 minutes after the trade.

B. Designated Market Maker (DMM) Internalization

Designated Market Maker (DMM) assignments provide a powerful method for internalizing trades. Exchanges appoint DMMs, also called specialists, at the stock-specific level. Each exchange independently assigns DMMs to stocks, creating stock-exchange level variation in DMM assignments. These DMM assignments grant special rights and obligations to the holder. One particularly strong advantage for internalizing trades is that if a DMM has a quote at the NBBO, the DMM can route any order not exceeding five hundred shares to execute entirely against its own quote.[9]

As an example, suppose two market makers have quotes at the best ask at the CBOE exchange: one from Citadel, for 500 shares, and one from Belvedere, for 2,000 shares. CBOE uses a pro-rata model, so an incoming order would normally be split according to how much depth each market maker displays; for example, an incoming market order for 500 shares would be allocated with 100 shares to Citadel and 400 to Beveldere. Citadel is, however, the DMM in Walgreens options at the CBOE exchange. Consequently, if Citadel routes a customer order for 500 shares to the exchange, Citadel has a right to claim an allocation of all 500 shares thanks to its DMM status.

The allocation right of the DMM allows the DMM to internalize orders they route to the exchange regardless of their position in the price-time or pro-rata queue at the NBBO. This is a powerful advantage for internalization: if market makers wish to internalize order flow, they can use a DMM seat to guarantee their ability to do so, and do not need to offer a superior quote over competitors to internalize the trade. Competing market makers, in turn, will not be able to trade against this internalized flow, regardless of their pro-rata depth or time priority at the NBBO.

From SEC 606 reports, we identify five key PFOF DMMs in the options space: Citadel, Wolver-

[9]Option contracts trade only in round lots. By 500 shares, we refer to an option trade of 5 contracts, with each contract defined on 100 shares of the underlying security. For orders above 500 shares, the DMM receives a 500-share guaranteed allocation, and further allocation depends on price-time or pro-rata rules. The allocation may also give the DMM an out-sized allocation relative to its share of displayed liquidity, but not a 100% allocation as for trades below 500 shares.

ine Execution Services, Susquehanna International Group, Morgan Stanley, and Dash IMC. Together, these DMM firms account for approximately 98% of total PFOF in options trades, as noted in Table IX. We compare these firms to the following DMM firms which pay nothing or very little for order flow: Belvedere, Two Sigma, GTS, XR Securities, Cutler, Simplex, Hudson River Trading, and Optiver.

We use PFOF as an indicator which takes the value 1 if the exchange's specialist assignment for the option is a DMM who is a major provider of PFOF (Citadel, Wolverine, Susquehanna, Morgan Stanley, and Dash IMC). This indicator takes the value zero for other DMMs. For this section, we restrict our analysis to trades on exchanges which use a DMM, and we only analyze trades which OPRA reports as trade type 73, with the trade executing as "(Auto) Electronic". These are regular electronic trades on the exchange, for which the DMM receives the potential advantage.

To test whether market-makers use their DMM seats to internalize orders, we examine the share of orders which involve multiple participants. OPRA data utilizes "printing on the passive side": that is, if an incoming market order executes against multiple passive limit orders, there is a separate print for each passive limit order, with the same price and timestamp, but potentially different quantities. We define the variable $MultipleParticipant$ as taking the value 1 if a trade involves multiple participants on the liquidity-supplying side, and 0 otherwise. We then estimate Regression 4 as a logit model. $PFOF$ takes the value 1 if the DMM at exchange j for asset k pays PFOF. Controls X include the option Greeks (vega, gamma, theta, and the absolute value of delta), the option price, the option size, a fixed effect for each exchange, and a fixed effect for each symbol.

REGRESSION 4: *For each option trade i in security j on exchange k on date t:*

$$Pr(MultipleParticipant_{ijkt}) = \quad \alpha_0 + \alpha_1 PFOF_{jk} + X_{ijkt} + \epsilon_{ijkt}$$

Results of Regression 4 are presented in Table IV. For orders of 500 shares or less, when the DMM at an exchange pays payment for order flow, trades are approximately 20% less likely to have multiple participants compared to when the DMM does not pay for order flow. This is consistent with PFOF-paying DMMs using their 500-share advantage to internalize entire trades for themselves.

We then consider what effect this internalization may have on market participants. We exploit the 500 share cutoff for DMM internalization to estimate a regression discontinuity. Trades at or below 500 shares can be entirely internalized by the DMM, while shares above 500 shares will not be fully internalized by the DMM; to capture the full order, the DMM would have to use a PIM trade, described in Sections IV D and E. In Section C, we also consider the special case where only one DMM is assigned to a stock across all exchanges. While this has identification advantages, it restricts analysis to a set of stocks with only one DMM, which tend to be small stocks.

To analyze the benefit of the DMM-internalization decision, we estimate Regression 5, which examines changes in outcomes around the 500-share cutoff. We define $BelowCutoff$ as 1 when an order is for 400 or 500 shares, and 0 when an order is for 600 or 700 shares. Of interest is the interaction term α_3 between PFOF and $BelowCutoff$. Shares of 500 or less can be fully internalized by the specialist, while shares of 600 or more cannot be fully internalized.

REGRESSION 5: *For each option trade i in security j on exchange k on date t:*

$$\begin{aligned} RealizedSpreadBPS_{ijkt} = \quad & \alpha_0 + \alpha_1 PFOF_{jk} + \alpha_2 BelowCutoff_{ijkt} \\ & + \alpha_3 PFOF_{jk} * BelowCuttoff_{ijkt} + X_{ijkt} + \epsilon_{ijkt} \end{aligned}$$

Results of Regression 5 are presented in Table V. Realized spreads for the 400 to 500 share orders, which can be fully internalized via the DMM allocation, are 15 basis points higher when the DMM at an exchange pays PFOF compared to when the DMM does not pay PFOF. This is consistent with these DMMs utilizing the 500-share DMM allocation to internalize trades, and earning extra profit from these trades. Our measure exploits the DMM assignments, and allows us to fit both stock and exchange fixed effects. We do not see whether any individual trade involves the DMM; as a result, our 15 basis point estimate is a potentially large underestimate, as this 15 basis points is an average across all trades for 400 to 500 shares in stocks which have a PFOF-paying DMM, and not just those for which the DMM participated.

Table IV: Probability of Multiple Participants. This table estimates Regression 4 with a logit model for auto electronic trades occurring in the 30 trading days starting on July 1, 2020. The dependent variable, MultipleParticipant, takes the value 1 if a trade involves multiple individuals on the quoting side, and zero otherwise. Our primary coefficient of interest, α_1 uses PFOF: an indicator for whether a DMM at a specific exchange is a major PFOF firm. We restrict to trades between 200 and 500 shares, and estimate a logit model with a fixed effect for each stock and exchange.

	Dependent Variable: Multiple Participants
PFOF	-0.21^{***}
	(0.01)
Absolute Delta	0.08^{***}
	(0.01)
Vega	0.13^{***}
	(0.01)
Gamma	0.08^{*}
	(0.03)
Theta	-0.01^{***}
	(0.00)
Price	-0.00^{*}
	(0.00)
Quoted Spread	0.00^{***}
	(0.00)
Deviance	1685442.41
Num. obs.	10321868
Num. groups: prtExch	9
Num. groups: SYM_ROOT	1860

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table V: Spreads and DMM Allocations. This table estimates Regression 4 for auto electronic trades occurring in the 30 trading days starting on July 1, 2020. PFOF takes the value 1 if the stock-exchange DMM pays PFOF, and 0 otherwise. Belowcutoff takes the value 1 for orders of 400 to 500 shares, and 0 for orders of 600 to 700 shares. We include the option Greek values, a fixed effect for each stock, and a fixed effect for each exchange.

	Realized Spread (BPS at 1m)	Realized Spread (BPS at 10m)
	(1)	(2)
PFOF	3.495	4.205
	(2.199)	(3.494)
$BelowCutoff$	-8.783^{***}	-5.946^{*}
	(2.078)	(3.301)
Absolute Delta	-354.162^{***}	-278.093^{***}
	(2.084)	(3.310)
Vega	-47.742^{***}	-33.586^{***}
	(2.039)	(3.240)
Gamma	-29.395^{***}	-47.053^{***}
	(4.549)	(7.227)
Theta	-0.642^{*}	-3.288^{***}
	(0.339)	(0.539)
Price	0.914^{***}	0.653^{***}
	(0.030)	(0.047)
QuotedSpread	17.657^{***}	20.220^{***}
	(0.361)	(0.573)
PFOF:$BelowCutoff$	13.366^{***}	15.343^{***}
	(2.351)	(3.735)
Observations	4,360,399	4,360,399
R^2	0.061	0.027
Adjusted R^2	0.061	0.026
Residual Std. Error (df = 4356683)	834.686	1,326.098
Note:		*p<0.1; **p<0.05; ***p<0.01

C. *Single DMM Stock Analysis*

To analyze how PFOF affects execution quality, we consider a simple question: does routing appear worse when it occurs at an exchange where a DMM pays PFOF? As an alternative test of the effect of PFOF on option transaction costs, we conduct a test using cases where, across all exchanges, there is a single firm assigned as a DMM. We then look at how quoted, effective, and realized spreads vary with the characteristic of this sole DMM, for all trades across all exchanges.

While we do not observe the individual brokers routing the trades, we observe which DMMs pay PFOF from SEC 605 Reports. We rely on the fact that DMM assignments at each exchange are quasi-random[10], which gives us an exogenous way to measure how execution quality varies with whether the DMM does or does not provide PFOF.

REGRESSION 6: *For each option trade i in security j on exchange k on date t:*

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

In contrast with Regression 4, we now use PFOF_j to refer to whether trades in stock j have a DMM who is a primary PFOF provider. In addition to the controls previously used (vega, gamma, theta, the absolute value of delta, price, and size), we add the total dollar value traded each option during our sample period. Due to the small volume in some of the symbols, we also use a weighted-OLS, with the total dollar value traded in each option as the weight in the regression. We no longer have a fixed effect for each stock (to avoid co-linearity with PFOF, since there is a single DMM per stock), but continue to fit a fixed effect for each exchange and cluster standard errors by exchange.

Estimates of Regression 6 are presented in Table VI. When the sole DMM is a PFOF-paying firm, we find that across all auto-electronic option trades, effective spreads are a statistically significant 2.5% higher, quoted spreads a statistically significant 10% higher, and realized spreads are 0.5% higher, though the effect on realized spreads is not statistically significant. Among PIM-only trades, we find that realized spreads are a statistically significant 0.5% larger. These increases in effective and quoted spreads suggest that overall option trading costs are higher in stocks where the sole

[10]To the extent to which DMM assignments are not perfectly randomized, this failure to perfectly randomize would suggest a different restriction to competition, whereby certain firms would have an advantage over competing market makers in obtaining and permanently retaining DMM seats in specific stocks. We do not explore this potential additional limitation to competition.

DMM pays PFOF compared to stocks where the sole DMM does not make PFOF payments, and that PIM trades are more profitable.

In the words of SEC Chairman Gensler: "Under the segmentation of the current market, nearly half of trading, along with a significant portion of retail market orders happens away from the lit markets. I believe this may affect the width of the bid-ask spread." Gensler (2021). While the reference to off-exchange trading implicitly ties this quote to the market for U.S. equities, the market for U.S. option trades, instead, has both many retail trades segmented in PIM trades and bid-ask spreads, which are frequently not constrained by the minimum tick-size.

Table VI: Option Spreads With a Single DMM. This table estimates Regression 6. Dependent variables are the effective, quoted, or realized spread, all measured in basis points. We restrict analysis to the subset of options which have a single DMM across all exchanges; columns (1), (2) and (3) are for all auto-electronic trades, while column (4) is restricted to PIM trades. PFOF is an indicator for whether the single DMM in that stock is one of the major PFOF-paying firms. Estimates are weighted-least-squares regression, using the total traded value of each security as the weight. There is a fixed effect for each exchange, and standard errors are clustered by exchange.

	All Trades			*PIM Trades*
	Effective Spread (BPS)	Quoted Spread (BPS)	Realized Spread (BPS)	Realized Spread (BPS)
	(1)	(2)	(3)	(4)
PFOF	2,557.517***	5,043.822***	423.570	589.124**
	(614.264)	(1,488.428)	(262.706)	(235.531)
Delta	−1,990.167***	−4,304.953***	−745.640***	−964.422***
	(311.946)	(553.353)	(106.073)	(51.573)
Vega	−7,867.649***	−14,863.660***	−3,426.131***	−2,730.568
	(2,199.695)	(4,293.287)	(887.138)	(1,909.834)
Gamma	−3,879.496**	−9,053.006***	−3,598.972***	−4,743.129***
	(1,477.710)	(2,668.710)	(801.746)	(497.264)
Theta	1,199.447	446.404	549.329	822.368
	(701.838)	(1,569.901)	(474.279)	(2,089.562)
Price	−104.257***	−239.499***	−34.513***	−91.236***
	(25.396)	(64.178)	(10.589)	(15.186)
Size	−0.0001	0.001	0.00003	−0.044*
	(0.0005)	(0.002)	(0.0003)	(0.020)
Symbol Volume	0.00000	0.00001*	0.00000*	0.00001
	(0.00000)	(0.00000)	(0.00000)	(0.00001)
Days To Maturity	6.212**	10.745**	3.500***	0.235
	(2.446)	(4.506)	(1.176)	(2.026)
Observations	637,634	637,634	637,634	82,445
R^2	0.026	0.029	0.010	0.020
Adjusted R^2	0.026	0.029	0.010	0.019
Residual Std. Error	55,021,976.000	104,692,039.000	32,036,385.000	26,606,592.000
Note:				*p<0.1; **p<0.05; ***p<0.01

D. Price Improvement Mechanism (PIM) Trades

Price improvement mechanims (PIM) trades provide a second method for internalizing a trade. In such a mechanism, an order is advertised, and market makers have 100 milliseconds in which to place bids. In the OPRA datafeed, these trades are formally defined as"Single Leg Auction Non ISO," and coded as order type 97. These trades are predominantly retail trades. In these single-leg auction trades, regular trade execution stops, and the proposed trade goes through a two-sided auction mechanism with an exposure period. Exchanges may refer to these trades as "Price Improvement Mechanism", "Customer Best Execution (CUBE)," or an "Automated Improvement Mechanism." The improved trade price is a benefit accruing to customers, and not the broker acting on their behalf, just as sub-penny price improvement in equities accrues to customers and not brokers.

PIM trades are overwhelmingly retail trades. Like sub-penny equity trades, however, the decision to internalize a trade is endogenous, and many retail option trades do not receive improvement. While DMM assignments may influence the decision of a market maker to use or not use a PIM Auction, they do not play a direct role in the PIM trading rules.

Any market maker can initiate a PIM auction, and any market maker can ostensibly participate in the auction by submitting a competing bid, but the exchanges often adopt rules that discourage anyone other than the market maker initiating the auction from participating in the trade. As an example, the NASDAQ PHLX exchange allows a market maker who proposes internalizing a trade in a Price Improvement Mechanism to selectively auto-match any competing bids in the auction. This gives the market maker a powerful first-mover advantage, and a winner's curse problem: a competing market maker will only ever win the auction outright if the initiator (who is better informed about the source of the order) decides not to match at that price. When there are multiple bids at a price level, market makers also receive a guaranteed fill of at least one contract or 40% of the original size of the order, whichever is larger. These rules both decrease the potential competition for PIM auctions.

Suppose a retail customer wishes to buy a call option for 500 underlying shares in Amazon, and the current price for this national best ask for this option is $2.50. If Morgan Stanley receives this order from the retail customer's broker, Morgan Stanley will be unable to use a DMM allocation,

as they do not have any DMM seats in Amazon. An alternative method of internalizing the order would be to use a price improvement mechanism trade. Morgan Stanley would go to an exchange offering this order type, and start an auction at $2.49. Competing market makers could place bids to fill the order at a price even lower that $2.49; if they do so, Morgan Stanley can elect to automatically match competing bids. This auto-match right is given to Morgan Stanley as the initiator of the auction. If there are no competing bids, Morgan Stanley will directly trade with the customer for $2.49. If Morgan Stanley does not elect to auto-match bids and is outbid, the competitor will win the auction and trade against the retail customer. If there is a competing bid and Morgan Stanley elects to auto-match bids, allocation will be pro-rata but with Morgan Stanley, as the initiator, receiving an out-sized share as a bonus for initiating the auction.

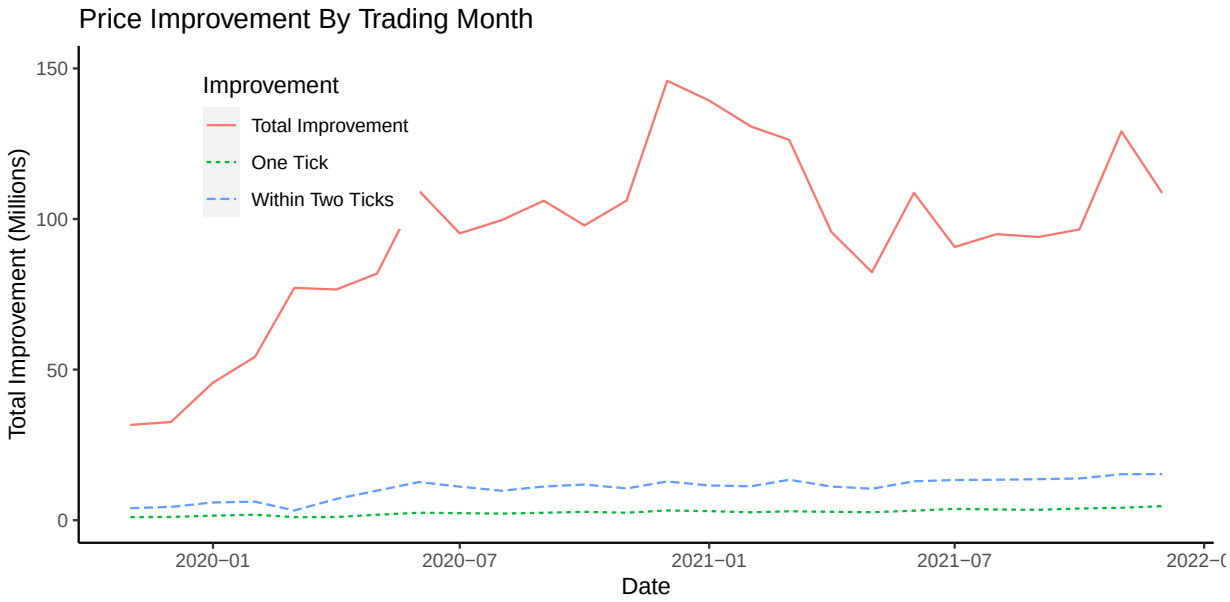
The Price Improvement Mechanisms must offer improvement against the bid-ask spread.[11] Just as in equities, the NBBO provides an ultimate lower bound on the execution quality of retail traders, and price improvement offers potentially better prices to customers. The extent to which this is meaningful improvement, however, partially hinges on the competitiveness of the quotes.

Figure 4 plots the total improvement by trading month, with the total aggregate monthly improvement varying between $50 and $150 million per month. Around 5 to 10% of trades in our sample go through a price improvement mechanism. Across our entire sample period, PIM trades receive over $2.4 billion in price improvement compared to the NBBO best bid or best ask. Unlike in equities, however, the bid-ask spread in options is rarely constrained by tick size. Figure 4 breaks down price improvement by the NBBO spread at the time of the auction. Tick size in options markets can vary with the option symbol: some have a one-penny bid-ask spread, some have a five-penny bid-ask spread, and some have a ten-penny bid-ask spread. Across these definitions, however, most of the aggregate price improvement comes from trades when the bid-ask spread is wider than two ticks. This price improvement can reflect savings for retail traders over the NBBO, but to some extent it also depends on the competitiveness of the bid-ask spread. If the bid-ask spread is very wide in a particular symbol, most trades may go through PIM auctions, and the wide spreads may not be an accurate reflection of the true savings. We examine the competitiveness of

[11] There is one notable exception: size improvement. Some auctions offer an auction option to provide size improvement. These auctions are allowed to match, rather than improve, the NBBO price provided the order size meets a certain threshold (often 5,000 shares). Flash orders, which are coded as regular electronic trades in OPRA data feeds, also typically offer size improvement.

PIM auctions in the next section.

Figure 4. Option Price Improvement By Month. We plot price improvement across all option trades. Approximately 5 to 10% of option trades are traded in Price Improvement Mechanism (PIM) trades. We plot the total daily improvement with the solid red line. We plot total improvement when spreads are approximately one tick (5 cents) with the green dashed line, and improvement when spreads are two ticks (10 cents) with the blue dashed line. Total improvement is $12.8 billion over our sample period.



E. Competition in PIM Trades

OPPRA flags PIM trades as type 97, "single-leg non-ISO auction", but there is no information in either the consolidated OPRA feed, or the individual exchange feeds, about the bidding activity in the auctions. Only the outcome price and quantities are observed, but not the bidders nor their identities. To explore the competitiveness of these option auctions, we explore two identification strategies. The first is a size-based cutoff, and the second is to use a regression discontinuity design from the option tick-size pilot program.

Our size-based cutoff is guided by the initiator allocation advantage in PIM trades. For single-contract trades, a market maker who initiates and auto-matches competing bids always wins the entire allocation. In contrast, for two or more contracts, a market maker who initiates and faces competing bids will receive less than the full allocation. As a result, there is a powerful winner's curse problem for single-contract trades: competing bidders will only ever win an allocation if the

initiating market maker, who is best informed about the source of the trade, declines to auto-match their bid.

Motivated by this Winner's Curse problem, we define Winner's Curse to take the value of 1 if a trade is for exactly 100 shares, and 0 otherwise. We then estimate Regression 7 using PIM-mechanism trades of exactly 1 or 2 contracts (i.e. options defined on 100 or 200 underlying shares).

REGRESSION 7: *For each option trade i in security j on exchange k on date t:*

$$RealizedSpreadBPS_{ijkt} = \quad \alpha_0 + \alpha_1 WinnersCurse_{ijkt} + X_{ijkt} + \epsilon_{ijkt}$$

Results of Regression 7 are presented in Table VII. For trades of exactly one contract, we find that they receive slightly less price improvement, measured as a percentage of the NBBO spread.[12] Realized spreads are 15 to 20 basis points higher, suggesting market makers make considerably larger profits internalizing trades for a single contract, for which they can guarantee to receive the entire contract by electing to auto-match competing bids.

As an alternative way to evaluate execution quality, we consider stocks on the Penny Pilot Program. We focus on the set of stocks which have a sharp-cutoff in tick-size around \$3. Below \$3, these stocks trade with a one-penny bid-ask spread, while above \$3, these stocks trade with a five-penny bid-ask spread.[13] We examine how price-improvement mechanism trades vary in execution quality around this sharp cutoff with Regression 8. We use a regression discontinuity design and restrict our analysis to trades priced between \$2.75 and \$3.25. We then define WideTick as taking the value 1 for stocks on the penny list priced above \$3, and 0 when they are priced below \$3.

REGRESSION 8: *For each option trade i in security j on exchange k on date t:*

$$PriceImprovement_{ijkt} = \quad \alpha_0 + \alpha_1 \, WideTick_{jk} + X_{ijkt} + \epsilon_{ijkt}$$

Results of Regression 8 are presented in Table VIII. Trades with a wide tick size receive slightly more price improvement, as a percentage of the NBBO bid-ask spread. Despite this larger im-

[12] We use the NBBO spread for PIM trades as the exchange holding a price-improvement auction does not have to have a quoted spread at the NBBO, and frequently has a bid-ask spread worse than the NBBO.

[13] In other words, limit orders may be priced in penny increments, like \$2.98, \$2.99, \$3.00, but then jump to \$3.05, \$3.10, and \$3.15.

Table VII: Price-Improvement and Share Cutoff. This table estimates Regression 7 for PIM trades occurring in the 30 trading days starting on July 1, 2020. Winner's Curse takes the value 1 for orders of 100 shares, and 0 for orders of 200 shares. We include the option Greek values, a fixed effect for each stock, and a fixed effect for each exchange.

	Dependent variable:		
	Price Improvement (Percentage of Spread)	Realized Spread (BPS at 1 minute)	Realized Spread (BPS at 10 minutes)
	(1)	(2)	(3)
Winner's Curse	-0.207^{***}	16.842^{***}	19.979^{***}
	(0.019)	(0.487)	(0.863)
Absolute Delta	4.349^{***}	-512.125^{***}	-575.033^{***}
	(0.040)	(1.062)	(1.883)
Vega	1.247^{***}	-99.195^{***}	-103.070^{***}
	(0.032)	(0.840)	(1.489)
Gamma	-8.888^{***}	-68.517^{***}	-65.861^{***}
	(0.095)	(2.504)	(4.442)
Theta	0.161^{***}	0.975^{***}	0.493^{**}
	(0.005)	(0.121)	(0.214)
Price	0.013^{***}	-0.309^{***}	0.006
	(0.001)	(0.014)	(0.025)
Quoted Spread	-4.344^{***}	97.965^{***}	88.452^{***}
	(0.017)	(0.454)	(0.804)
Observations	13,695,206	13,695,206	13,695,206
R^2	0.068	0.070	0.028
Adjusted R^2	0.068	0.070	0.028
Residual Std. Error	27.957	735.118	1,303.914
Note:			*p<0.1; **p<0.05; ***p<0.01

provement, however, trades with a wide tick are associated with larger realized spreads, suggesting the larger tick size leads to higher profits to market makers, suggesting the combination of features of PIM trades which inherently favor the initiator of the auction (the auto-match bid option, the out-sized allocation, and the reduction in trading fees) functionally limit competition in the auctions.

Table VIII: Price-Improvement and Tick Size. This table estimates Regression 8 for PIM trades in securities participating in the Penny Pilot Program, priced between $2.75 and $3.25 occurring in the 30 trading days starting on July 1, 2020. WideTick takes the value 1 for orders priced above $3, and 0 for orders priced below $3. We include the option Greek values, a fixed effect for each stock, and a fixed effect for each exchange.

	Dependent variable:		
	Price Improvement (Percent of Spread)	Realized Spread (BPS at 1m)	Realized Spread (BPS at 10m)
	(1)	(2)	(3)
WideTick	0.707***	4.185**	12.319**
	(0.135)	(2.125)	(5.932)
absDe	1.907***	−15.232	−16.507
	(0.679)	(10.652)	(29.739)
prtVe	6.972***	−12.625	134.228**
	(1.323)	(20.770)	(57.988)
prtGa	−33.010***	43.452	648.404***
	(5.093)	(79.957)	(223.235)
prtTh	−0.535***	0.554	−5.728
	(0.107)	(1.678)	(4.684)
prtPrice	1.317***	−25.452***	−37.138*
	(0.489)	(7.674)	(21.425)
QuotedSpread	−6.803***	262.468***	184.718***
	(0.209)	(3.273)	(9.139)
Observations	558,389	558,389	558,389
R^2	0.033	0.016	0.002
Adjusted R^2	0.032	0.015	0.001
Residual Std. Error	25.545	401.034	1,119.663
Note:			*p<0.1; **p<0.05; ***p<0.01

V. Transaction Revenue and Payments

We obtain SEC 606 data on payment for order flow from five brokerages: TD Ameritrade, Robinhood, E∗Trade, Charles Schwab, and Vanguard. The first four brokerages were the four largest recipients of payment for order flow in 2020, while Vanguard is a large brokerage which does not take payment for order flow for equity, and stopped accepting PFOF in options in July 2021. Fidelty, an additional large retail brokerage, does not take any PFOF, so all 606 reports have zeros for payments received. Total payments by each broker are plotted in Figure 5. Across our sample period, we document over $3 billion in total payment for order flow.

Payments vary considerably by asset class. Figure 6 plots the payment by each asset type. Options are by far the largest share of PFOF, with around 65% of all PFOF. Non-S&P 500 stocks account for 30% of PFOF (note that ETFs, even those focused on S&P stocks, will be categorized as non-S&P), and individual S&P 500 stocks account for just 5% of all PFOF. These percentages are based on the total value of the payments, but the payment rate per order is also unequal. Averaged across the main brokerages, options pay around 40 cents per 100 shares, while stocks pay around 20 cents per 100 shares. These per-share differences understate the nominal value difference. Among options trades receiving price improvement, the median price is $5. Thus a $1,000 investment in options generates a far larger share volume than an investment in equities; consequently, the option investment also generates a far larger per-trade payment than an investment in equities.

The total value of PFOF to brokers is smaller than the value of price improvement given to customer orders. In equity markets, we can accurately identify sub-penny retail improvement. As Figure 2 notes, this improvement can range between $50 and $150 million per month, which is more than the monthly equity PFOF of between $40 and $100 million (Figure 5).

Routing is concentrated among a small number of wholesalers. Table IX documents the routing behavior across asset classes. In each asset class, the top two firms receive 70% of all broker order routing. The top 4 firms receive over 90% of all broker routing in each asset class.

Figure 5. PFOF By Broker. Payment for order flow has increased over time. For our 606 data from January 1, 2020 to July 2021, we document $3.2 billion in payment for order flow across the five brokerages. Note that Vanguard does not take equity PFOF, ceased taking option PFOF in July 2021, and its total PFOF is small enough that it is not visible relative to the other firms. Payment for order flow is relatively stable by brokerage, with the exception of Robinhood. Robinhood grew from $20 million in January 2020 to $67 million in July 2021, with a peak of $120 million in February 2021.

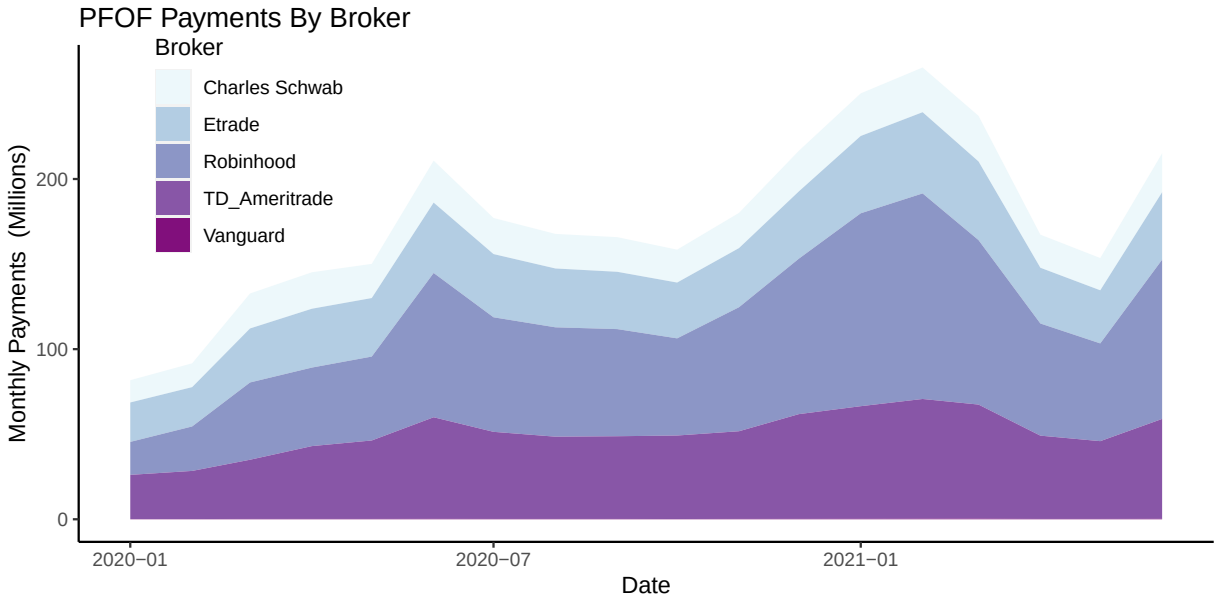
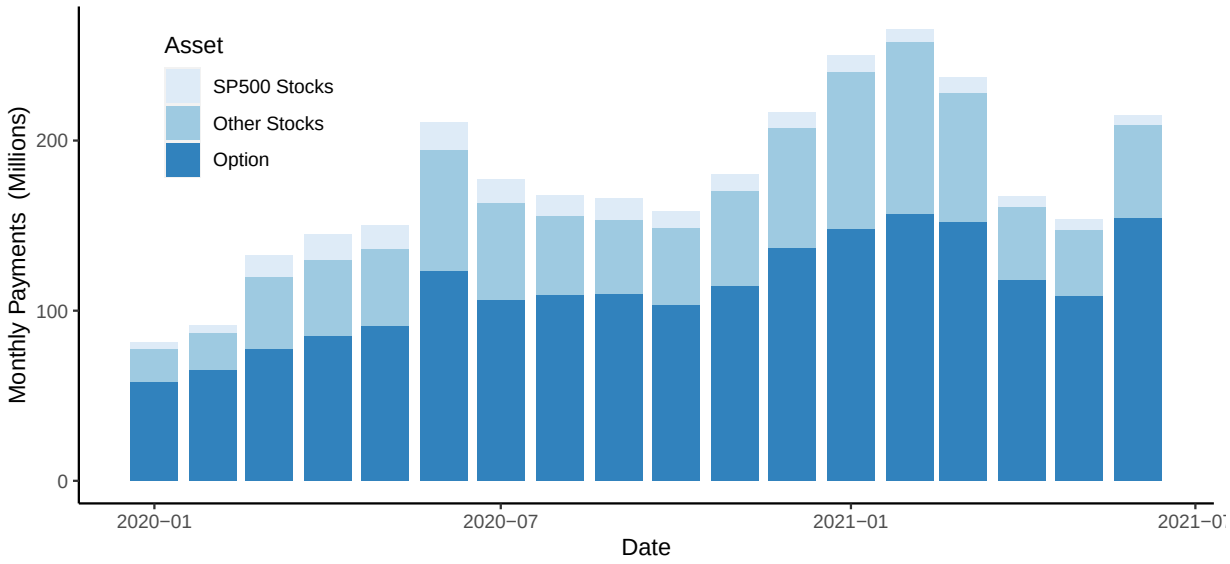


Figure 6. PFOF By Asset Class. There are large differences in payments across assets, both in total value and the payment per share. Panel A presents total payments. Most payment for order flow (65%) comes from options markets. The remainder comes from non-S&P 500 Stocks (30%) or S&P 500 stocks (5%). Panel B presents the payment rate per 100 shares traded. Options are consistently the highest in payments, averaging 40 cents per 100 shares traded.

Panel A: Total Payments



Panel B: Payment Per 100 Shares

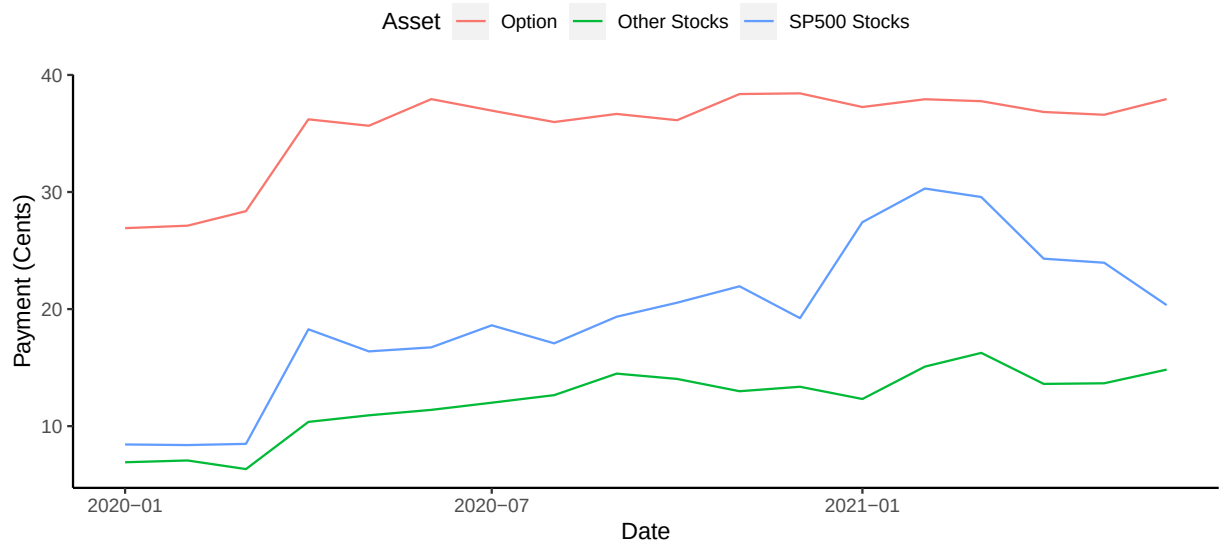


Table IX: Routing Destinations. A very small number of firms receive a very large share of order routing. Citadel and Virtu receive almost all equity routing, with G1X and Two Sigma receiving almost all of the remainder. Virtu is not active in options payment for order flow. In options, Citadel, Global Execution Brokers, Wolverine, Morgan Stanley, and Dash IMC receive almost all the order flow.

Options

Firm	Total Orders (Millions)	Order Share
Citadel	852.18	42.20
Global Ex. Brokers (SIG)	572.56	28.40
Wolverine	288.80	14.30
Morgan Stanley	156.88	7.80
Dash IMC	121.15	6.00
G1X	18.33	0.90
Two Sigma	5.13	0.30
Citigroup	3.58	0.20

Other Stocks

Firm	Total Orders (Millions)	Order Share
Citadel	395.42	40.70
VIRTU	292.01	30.10
G1X	125.61	12.90
Two Sigma	88.67	9.10
UBS	27.51	2.80
Wolverine	27.43	2.80
CBOE	8.78	0.90
Nasdaq	4.38	0.50
Jane Street Capital	0.62	0.10
Etrade	0.00	0.00
Susquehanna	0.00	0.00

SP500 Stocks

Firm	Total Orders (Millions)	Order Share
Citadel	73.11	41.40
VIRTU	55.05	31.20
G1X	24.92	14.10
Two Sigma	11.11	6.30
UBS	6.67	3.80
CBOE	1.95	1.10
Nasdaq	1.91	1.10
Wolverine	1.84	1.00
Jane Street Capital	0.09	0.00
Susquehanna	0.00	0.00

VI. Asset Returns and Broker Conflict of Interest

Options have more volatile payoffs than stocks. Within options, there is extreme variation between the volatility of at-the-money options and far out-of-the money options. To provide a basic estimate of the effect of option trades on retail trades, we conduct a bootstrapped estimation from our dataset of all retail trades.

Our goal is to develop a objective assessment of how volatile the portfolios of retail traders are in practice. To do this from anonymous trade data, we make a set of assumptions, each with its own advantages and disadvantages:

1. Samples of 50 trades. For each date, we draw 1,000 samples of 50 trades each. The selection of 50 trades per portfolio creates a reasonably diverse portfolio.

2. Samples are drawn from PIM trades, as these trades are overwhelmingly retail trades. We note two limitations to our data: first, not all PIM trades are retail and second, not all retail trades go through PIM. We believe the latter issue is the more serious: some retail trades will not go through PIM, and the decision of a wholesaler not to internalize these trades may reflect that they may be less volatile, different maturities, or have lower spreads.

3. Samples of only call options. Retail traders will certainly buy both calls and puts, but buying a put leaves investors short the equity premium. By restricting to only calls, the option portfolios are long the equity premium, and we compare them against equivalent equity portfolios.

4. Samples with a three-month horizon. On any given day, observed retail option trades will have a variety of maturities. We restrict our analysis to only draw from observed retail trades with a maturity between 1 day and 3 months. To calculate 3-month returns, we calculate the option return at maturity, and then calculate the stock return between the maturity date and the end of the 3-month window. The geometric combination of these two returns gives an overall return to the sample window, and reflects the returns of an investor who held the option to maturity, and then held the stock to the ending period. This understates the volatility difference between an all-option and all-stock portfolio, while replacing in-the-money options with another draw from observed option trades would overstate the volatility

difference.

5. Samples are benchmarked against the same all-stock portfolio. One key concern with retail traders is that they may select stocks which deviate from the market portfolio. To account for this, we compare the option portfolio return against the return of a portfolio comprised of the same underlying stock names.

6. Samples are assumed equally weighted. Out-of-the-money options often have very low prices. When we draw a sample of 50 retail option trades, we combine trades across the portfolio in an equal-weighted average, rather than a value-weighted average.

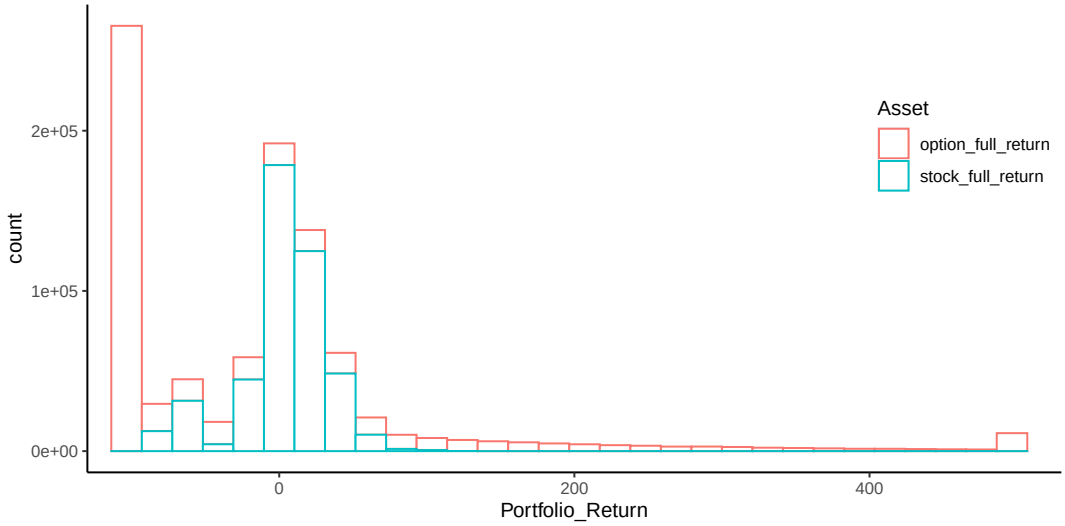
As an example, on January 2, 2020, we construct 10,000 portfolios, each comprised of 50 option trades drawn from that day's PIM trades in call options. As an example option trade, a call option on APPL with a maturity date of February 2, 2021 and a strike price of 280 sold for \$22.42. On February 2, 2021, the closing price of Apple was \$313.03, giving a total return on the option of $\frac{(33.03-22.42)}{22.42} - 1 = 47.3\%$, while the closing price 3 months after January 2, 2020 was \$244.93, an 21.8% drop from the maturity date. The overall return for buying this option on January 2, 2020, and then holding the asset until April 2, 2020 was $(1 + 47\%)(1 - 21.8\%) - 1 = 15\%$. This 15% return is then equal-weighted with the other 49 returns on option trades from the portfolio.

We plot distributions of returns, combined across all dates in Figure 7. The distribution of option returns has very high skew. Over 50% of all option portfolios lose 90% or more of their value, while 2.3% of option portfolios have a return greater than 500%. There is considerable time-series variation, however.

We plot the daily median returns of these portfolios in Figure 8. Across our sample, the median stock portfolio closely matches the return on the S&P 500 Index, as measured by the return of the S&P 500 ETF, SPY. The median option portfolio return varies considerably from the return of the S&P 500 Index, with a considerably higher return in some months, and considerably lower return in others.

For brokers, the difference in returns is, in a way, no less stark. Options trades pay roughly double what equity trades pay. Thus rather than a distribution of possible returns, as investors may obtain in Figure 8, the broker has a binary set of two possible payoffs. If investors are investing a fixed nominal amount, the low nominal price of options can further amplify this difference in

Figure 7. Comparison of Distribution of Realized Returns. We simulate returns from the empirical distribution of PIM option trades. For each day, we draw 1,000 portfolios of 50 option trades. We plot the distribution of returns assuming each option is held to maturity. Over 75% of option trades lose money; over half of the option portfolios lose more than 90% or more of their value. Portfolios have a very positive skew, with 2.3% of option portfolios having a return exceeding 500% (which we plot as a single mass point at 500%).



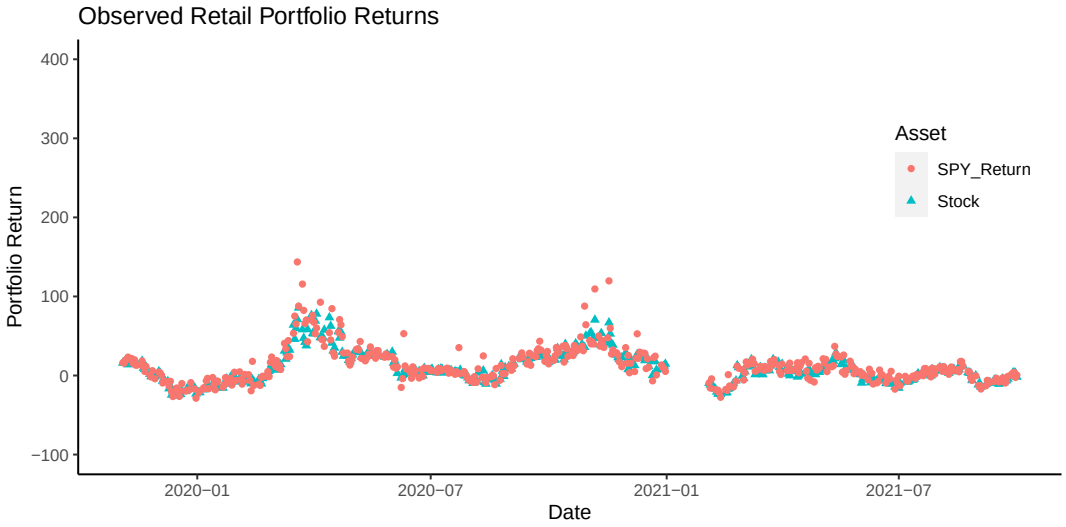
payments. For example, an equally-priced $1,000 investment in a stock with a share price of $100 would give 20 cents worth of payment for order flow. In contrast, an equally-priced $1,000 investment in an equity option with a price of $10 would give approximately $4.00 worth of payment for order flow.

This incentive conflict between the broker and client is far larger in magnitude than the conflict of interest over routing choice. As we note in Appendix A, sub-penny improved trades are rarely sensitive to timing, as retail trades typically do not arrive at times when market prices are changing. Executing a millisecond faster or slower is important for trades on exchanges, but far less important for internalized trades. Moreover, these trades frequently occur at minimum one-tick spreads, meaning prevailing quotes are already at the minimum feasible. In options markets, spreads are wider and there is more potential room for price improvement. But even in options markets, spreads are measured in basis points, i.e. hundredths of one percent. Differences in portfolio returns between stocks and options, however, are measured in percentage points, and often exceed double-digit differences.

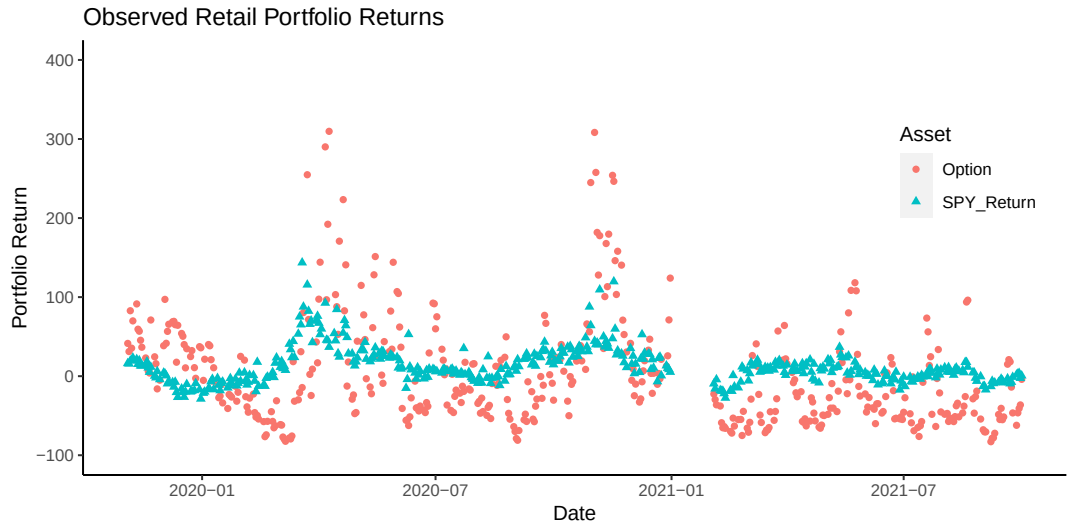
Under the SEC's Regulation Best Interest, brokers have a mandate to "Act in the best interest

Figure 8. Comparison of Realized Returns Through Time. We simulate returns from an empirical distribution of option trades. We draw 1,000 portfolios of 50 option trades, and calculate returns assuming each option is held to maturity. Note that there are some missing dates in January 2021 due to missing data from SpiderRock, our data provider. Panel B plots the distribution of the daily median of portfolio returns, while Panel A plots the distribution of the daily median of portfolio returns for a portfolio which invests in stocks rather than options. The option portfolios have considerably higher volatility.

Panel A: Median Realized Returns for Retail Stock Trades



Panel B: Median Realized Returns for Retail Option Trades



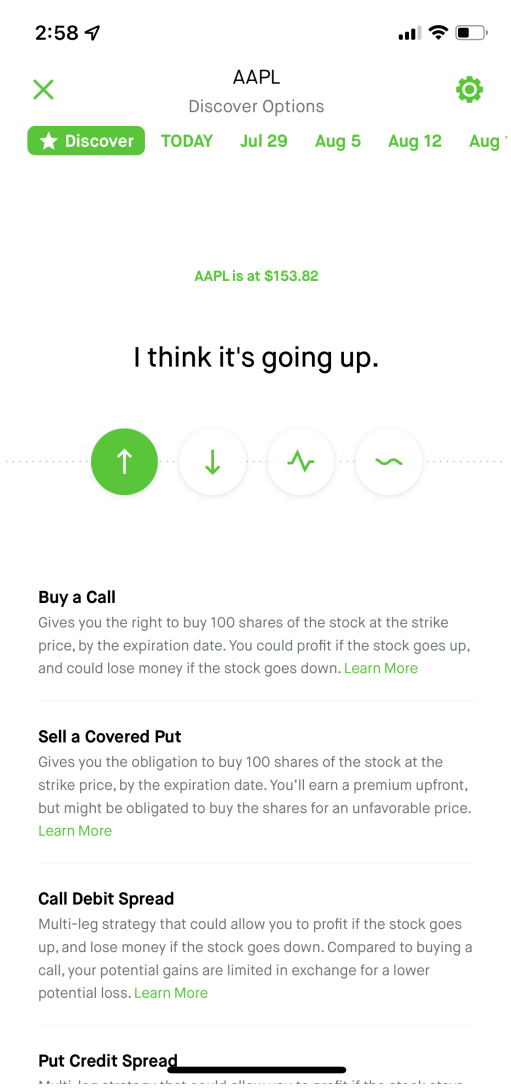
of the retail customer at the time the recommendation is made, without placing the financial or other interest of the broker-dealer ahead of the interests of the retail customer." The SEC recognizes that what constitutes a recommendation is "not susceptible to a bright line definition,"

but depends upon whether the communication "reasonably would influence an investor to trade a particular security or group of securities[14]". To illustrate potential differences in recommendations, we highlight two application screenshots from two large brokerages in Figure 9. The first screenshot is from Robinhood LLC, which describes its company goal as attempting to "democratize finance for all, regardless of a customer's background, income, or wealth." Consistent with a potential democratizing goal, the app reminds users of which options profit when the stock price goes up, though this also raises questions about the suitability of these assets for customers for whom this reminder is warranted. No maturity date is displayed by default, but users must select a maturity date, starting with daily options, then weekly options, and the ability to scroll to monthly options. In contrast, we also show a screenshot from Interactive Brokers LLC, which describes its mission as competing "on price, speed, size, diversity of global products and advanced trading tools." Selecting option trading immediately brings up the option chain with one month expiration date, and the ability to scroll to weekly or longer-dated options. The app interface is more crowded, and does not remind users that a call option profits when stock prices increase.

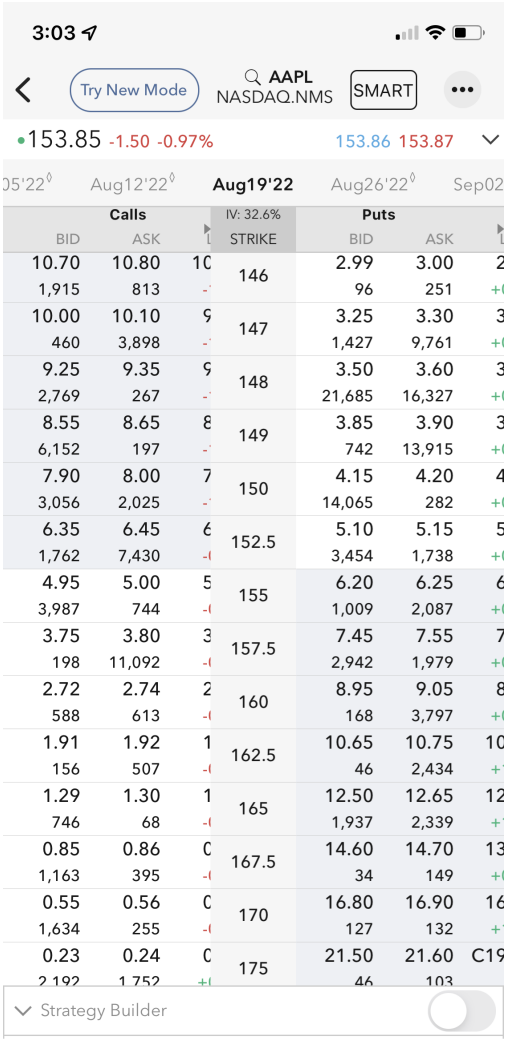
Users have agency in picking assets, in following or rejecting broker recommendations, and in choosing a brokerage firm. While 605 reports show that customers of some brokers trade options more frequently than customers of others, the customer choice of broker is endogenous. In Figure 10, we plot Google search interest in trading searches for Robinhood and Interactive Brokers. Google search users interest enter as many or more searches for "options trading Robinhood" as they do for "stock trading Robinhood." In contrast, Google search users search for "options trading Interactive Brokers" less often than "stock trading Interactive Brokers." These differences in search activity reflect a potential innate preference difference between the users of respective platforms, suggesting that customers with certain return or asset preferences do indeed gravitate toward certain brokers.

[14]See "Regulation Best Interest: A Small Entity Compliance Guide." Securities and Exchange Commission (2019)

Figure 9. Differences in User Interface. Brokers make a series of user interface design choices, each of which shapes the user experience. Panel A presents a screenshot from the Robinhood app, which prominently features a reminder of which options profit when the stock price rises. Users must select an expiration date, with daily and weekly options displayed first. Panel B presents a screenshot from the Interactive Brokers app, which begins with the option chain for one-month options.



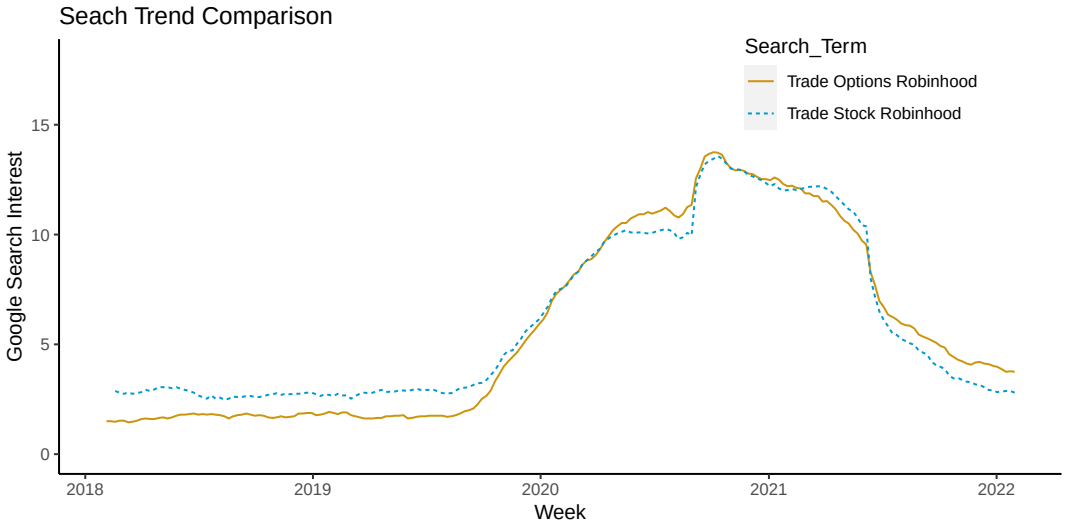
Panel A: Robinhood App



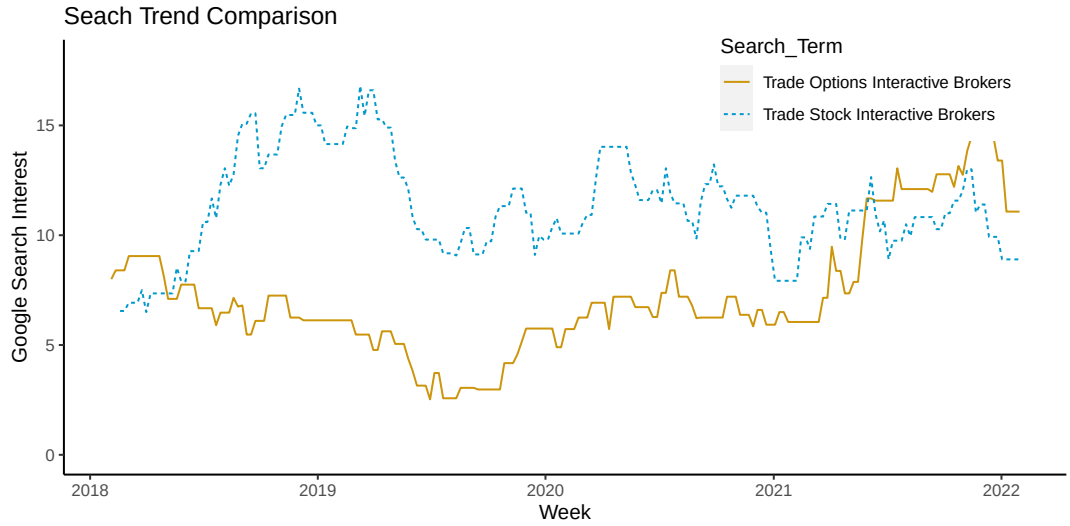
Panel B: Interactive Brokers App

Figure 10. Comparison of Google Search Interest. Google search history offers potential insight into the interest of brokerage customers. We plot the relative search interest of "trade options *broker*" where broker takes the value of either "Interactive Brokers" or "Robinhood". For Interactive Brokers, "trade stock" generally has more search interest than "trade options". For Robinhood, "trade stock" has more roughly equal search interest to "trade options."

Panel A: Google Search Interest For Robinhood



Panel B: Google Search Interest for Interactive Brokers



VII. Conclusion

We explore the underlying differences in execution quality in equity and option markets, as well as the associated PFOF in these respective markets. Relative to equity markets, option markets have much wider bid-ask spreads, more price improvement relative to spreads, and larger payments for order flow. We examine the underlying competitiveness of the rules around internalizing within the respective markets, and how some option trading rules protect profits from internalization. In turn, these profits from internalizing motivate high payments to brokers to secure order flow, with options orders consistently obtaining higher PFOF than equity orders. The resulting cross-asset variation in payments gives rise to substantially different financial rewards to brokers based on the types of assets their clients trade.

The traditional concern around payment for order flow focuses on best-execution of an individual trade. While the SEC can easily confirm whether trade prices are at least as good as posted quotes, market makers may be willing to improve on quotes. This unpublished potential improvement is difficult to measure, and the segregation of retail order flow has the potential to make on-exchange posted quotes worse. As a result, measuring execution quality is not just about comparing prices with the best displayed quotes, but about the possible price obtainable off-exchange, and regulators must further consider the equilibrium impact of segregating order flow.

In equity markets, we show that around half of all trades occur when bid-ask spreads are constrained at a one-penny bid-ask spread. When stocks are tick constrained, routing internalized trades to the exchange would do nothing to the width of the quoted bid-ask spread unless the minimum tick size on exchanges were also changed. In comparison to equity markets, option markets are far less likely to be tick-constrained, particularly in the contracts traded by retail investors. While all option trades must be done on-exchange, we document two limits to competition: designated market-maker assignments and PIM auctions. With DMM assignments, market makers at the NBBO can internalize trades for 5 contracts or less regardless of their price-time or pro-rata position. Within the subset of stocks with a single DMM, we show that PFOF-paying DMMs are associated with wider bid-ask spreads. Within PIM trades, an initiating market maker has the ability to auto-match competing bids and receives a larger allocation in the event of a tie. We exploit variation in the tick size pilot to show that while larger minimum tick sizes lead to more

price improvement, the realized spreads are also larger. In other words, the large discount given in the PIM auction says more about the wideness of the quotes than the competitiveness of the auction.

The impact of asset choice on broker PFOF revenue is, to the best of our knowledge, a previously unstudied aspect of payment for order flow. While PFOF has ushered in zero-commission trading, the PFOF-only business model pays brokers far more when their clients trade options than when their clients trade stocks. Differences in execution quality are measured in basis points, but the difference between an equity or option investment return is typically measured in double-digit percentages. Brokers do have suitability standards around the assets they can offer clients. Distinguishing between a broker who pushes high-variance securities to gain higher PFOF, and a broker who merely satisfies client demands for high-variance securities, however, is a difficult task. In comparison, measuring execution quality against posted spreads is far easier. Developing more competitive price improvement mechanisms in the auction market, with different participants on an even footing in the auction, has the potential to reduce the profitability—and the associated PFOF—surrounding internalizing option trades, and potentially turn broker incentives to a more equal level between equity and options.

REFERENCES

Anand, Amber, Mehrdad Samadi, Jonathan Sokobin, and Kumar Venkataraman, 2021, Institutional Order Handling and Broker-Affiliated Trading Venues, *Review of Financial Studies* 34, 3364–3402.

Baldauf, Markus, Joshua Mollner, and Bart Z. Yueshen, 2022, Siphoned Apart: A Portfolio Perspective on Order Flow Fragmentation, *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 .

Barber, Brad M, Xing Huang, Philippe Jorion, Terrance Odean, and Christopher Schwarz, 2022, A (Sub) penny For Your Thoughts: Tracking Retail Investor Activity in TAQ, *Available at SSRN: 4202874* .

Barber, Brad M, Xing Huang, Terrance Odean, and Christopher Schwarz, 2021, Attention Induced Trading and Returns: Evidence from Robinhood Users, *Journal of Finance, Forthcoming* .

Barber, Brad M, and Terrance Odean, 2000, Trading is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors, *Journal of Finance* 55, 773–806.

Bartlett, Robert P, and Justin McCrary, 2019, How Rigged are Stock Markets? Evidence from Microsecond Timestamps, *Journal of Financial Markets* 45, 37–60.

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

Battalio, Robert, Shane A. Corwin, and Robert Jennings, 2016a, Can Brokers Have it All? On the Relation Between Make-take Fees and Limit Order Execution Quality, *Journal of Finance* 71, 2193–2238.

Battalio, Robert, Todd Griffith, and Robert Van Ness, 2021, Do (Should) Brokers Route Limit Orders to Options Exchanges that Purchase Order Flow?, *Journal of Financial and Quantitative Analysis* 56, 183–211.

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 Paul Schultz, 2011, Regulatory Uncertainty and Market Liquidity: The 2008 Short Sale Ban's Impact on Equity Option Markets, *Journal of Finance* 66, 2013–2053.

Battalio, Robert, Andriy Shkilko, and Robert Van Ness, 2016b, To Pay or be Paid? The Impact of Taker Fees and Order Flow Inducements on Trading Costs in US Options Markets, *Journal of Financial and Quantitative Analysis* 51, 1637–1662.

Bessembinder, Hendrik, Jia Hao, and Kuncheng Zheng, 2020, Liquidity Provision Contracts and Market Quality: Evidence from the New York Stock Exchange, *Review of Financial Studies* 33, 44–74.

Boehmer, Ekkehart, Charles M Jones, Xiaoyan Zhang, and Xinran Zhang, 2021, Tracking Retail Investor Activity, *Journal of Finance* 76, 2249–2305.

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

Chordia, Tarun, and Avanidhar Subrahmanyam, 1995, Market Making, the Tick Size, and Payment-for-order Flow: Theory and Evidence, *Journal of Business* 543–575.

Clark-Joseph, Adam D, Mao Ye, and Chao Zi, 2017, Designated Market Makers Still Matter: Evidence from Two Natural Experiments, *Journal of Financial Economics* 126, 652–667.

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 Roseman, and Yanbin Wu, 2022, Retail Option Traders and the Implied Volatility Surface, *Available at SSRN: 4104788* .

Ernst, Thomas, Jonathan Sokobin, and Chester Spatt, 2021, The Value of Off-Exchange Data .

Foley, Sean, Anqi Liu, Katya Malinova, Andreas Park, and Andriy Shkilko, 2020, Cross-Subsidizing Liquidity, Technical report, Working Paper, Macquarie University.

Gensler, Gary, 2021, "Prepared Remarks at the Global Exchange and FinTech Conference", Speech: Prepared Remarks at the Global Exchange and FinTech Conference [Accessed: 2022 08 29].

Gensler, Gary, 2022, "Market Structure and the Retail Investor:", Remarks Before the Piper Sandler Global Exchange Conference, Washington, D.C. [Accessed: 2022 08 01].

Glosten, Lawrence R, and Paul R Milgrom, 1985, Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders, *Journal of Financial Economics* 14, 71–100.

Greenwood, Robin, Toomas Laarits, and Jeffrey Wurgler, 2022, Stock Market Stimulus, Technical report, National Bureau of Economic Research.

Hasbrouck, Joel, 2018, Price Discovery in High Resolution, *Journal of Financial Econometrics* .

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

Holden, Craig W., and Stacey Jacobsen, 2014, Liquidity Measurement Problems in Fast, Competitive Markets: Expensive and Cheap Solutions, *Journal of Finance* 69, 1747–1785.

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

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* .

Jensen, Michael C, 1968, The Performance of Mutual Funds in the Period 1945-1964, *Journal of Finance* 23, 389–416.

Kothari, SP, Travis L Johnson, and Eric C So, 2021, Commission Savings and Execution Quality for Retail Trades, *Available at SSRN* .

Lee, Charles MC, and Mark J Ready, 1991, Inferring Trade Direction From Intraday Data, *Journal of Finance* 46, 733–746.

Li, Sida, Xin Wang, and Mao Ye, 2021, Who Provides Liquidity, and When?, *Journal of Financial Economics* 141, 968–980.

Li, Sida, and Mao Ye, 2022, The Optimal Price of a Stock: A Tale of Two Discretenesses, *Available at SSRN: 3763516* .

Madhavan, Ananth, and Seymour Smidt, 1993, An Analysis of Changes in Specialist Inventories and Quotations, *Journal of Finance* 48, 1595–1628.

Mayhew, Stewart, 2002, Competition, Market Structure, and Bid-ask Spreads in Stock Option Markets, *Journal of Finance* 57, 931–958.

Muravyev, Dmitriy, and Neil D Pearson, 2020, Options Trading Costs Are Lower Than You Think, *Review of Financial Studies* 33, 4973–5014.

Ni, Sophie X, Neil D Pearson, Allen M Poteshman, and Joshua White, 2021, Does Option Trading Have a Pervasive Impact on Underlying Stock Prices?, *Review of Financial Studies* 34, 1952–1986.

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* .

Securities and Exchange Commission, 2019, "Regulation Best Interest: A Small Entity Compliance Guide", [Accessed: 2022 09 05].

Venkataraman, Kumar, and Andrew C. Waisburd, 2007, The Value of the Designated Market Maker, *Journal of Financial and Quantitative Analysis* 42, 735–758.

Appendix A. Time Sensitivity of Trades

Evaluating whether a price is a good price depends on the other prices available in the market. At very short time horizons, the geography of the market plays a critical role. Prices at a specific point in time are only available from a specific location. In this subsection, we analyze the timing of trades, doing a technical analysis of the timestamps available in TAQ. We find that the differences in timestamps are crucial for exchange trades, with sizable price changes occurring over small time horizons. We find, however, that the retail trades with sub-penny price improvement occur almost exclusively at times when prices are not changing. For these retail trades, executing slightly earlier or later would make a difference only for a small share of trades.

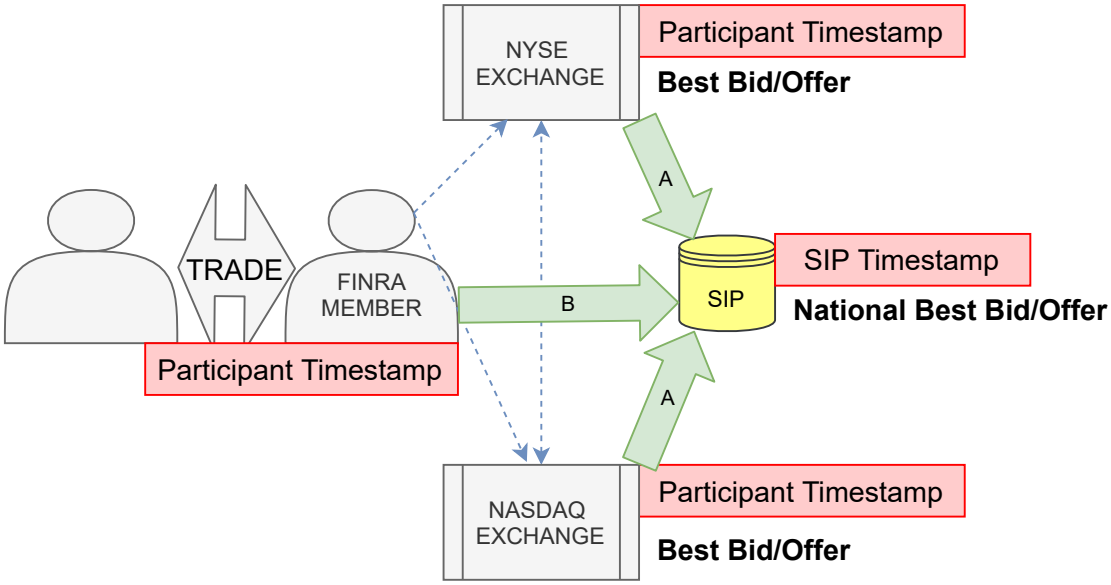
The Securities Information Processor (SIP) calculates, disseminates, and records the official National Best Bid or Offer (NBBO). TAQ data, which this study will use, comes from SIP records. All trades and quotes in TAQ are timestamped twice, with a participant timestamp and a SIP timestamp.[15] Trades and quotes are first stamped with a participant timestamp at the facility at which they occur. For example, a trade occurring at the NYSE would be stamped by NYSE according to the time displayed on the NYSE's clock. All trades and quotes are reported to the the SIP, where the SIP timestamp is assigned according to the SIP's clock.

There are two SIP facilities: the Consolidated Tape Association (CTA) and the Unlisted Trading Privileges (UTP). The CTA SIP, historically associated with the NYSE and co-located in the current NYSE data center, processes data for any securities listed at the NYSE's family of exchanges, including Tape A and Tape B securities. The UTP SIP, historically associated with NASDAQ and co-located in the current NASDAQ data center, processes data for any securities listed at NASDAQ's family of exchanges, including Tape C securities.

The multiplicity of timestamps and SIPs, depicted in Figure 11, presents several possibilities for matching trades and quotes. Holden and Jacobsen (2014) match trades and quotes according to their SIP timestamps. This works well conditional on trades and quotes being reported to the SIP with equal latencies. When the trade and quote updates occur at different exchanges, however, the latency to report either to the SIP will be unequal. As an example, consider a NASDAQ-listed stock. If there is a trade at NYSE, and a few microseconds later a quote update at NASDAQ,

[15] There is a third timestamp field in the data, the TRF Timestamp. For a detailed discussion and analysis of the TRF timestamp, see Ernst et al. (2021).

Figure 11. Timestamps and Data Map. All trades and quotes in TAQ are first timestamped by the participant, and second by the SIP. Exchanges report both trades and quotes to the SIP (Green Arrows A). FINRA members report off-exchange trades to the SIP (Green Arrow B). While the SIP disseminates an official National Best Bid or Offer, each exchange or broker can monitor market conditions via direct feeds (blue dotted arrows).



the update will reach the UTP SIP long before the trade. Matching this NASDAQ quote with the NYSE trade according to the SIP timestamps will lead to matching a trade with a quote that happened after the trade.

We explore an alternative, that of matching trades and quotes according to their participant timestamps. This has the advantage of eliminating discrepancies in the exchange-to-SIP latency, as the participant timestamps record when a trade or quote took place, not when it was recorded by the SIP. We feel this presents a plausibly more accurate comparison of a trade price against prevailing market conditions.

No approach, however, is perfect. For example, matching participant timestamps fails to take into account both the time and geographic position at which the market participant send a trade message. In addition to matching trades to the nearest quote under either timestamp, we explore matching trades to quotes slightly before or after the trade.

The SIP and participant-timestamp matched quotes differ around 35% of the time (see Figure 5). This means effective spreads, quoted spreads, and price impacts will differ depending on the method used, and the execution quality a customer receives may be sensitive to the microsecond-

level routing decision made by a broker. For sub-penny trades, however, we find that differences between the SIP and participant-timestamp matched quotes are far less likely, averaging below 10% per day. Differences in quotes before the trade are more common than differences in quotes after the trade, with the midquote 1 to 3 milliseconds before a trade differing from the participant-matched quote around 60% of the time for non-sub-penny trades, and 35% of the time for sub-penny trades. Differences in quotes after a trade are less likely, with the midquote 1 to 3 milliseconds after the SIP-matched quote averaging 35% of the time for non-sub-penny trades and below 10% for sub-penny trades.

Across stocks, differences in the matched quotes are more common in higher-priced stocks. Figure 13 plots the percentage of trades for which the SIP-matched quote differs from the participant-matched quote across stock-day observations. Higher-priced stocks differ on a larger percentage of days than low-priced stocks, reflecting the greater likelihood of many small price changes in the higher-priced stocks.

As a further test of timing differences, we examine differences in the SIP and participant timestamp matched quotes for each of Tape A, Tape B, and Tape C securities. Tapes A and B are both processed by the CTA SIP, while Tape C securities are processed by the UTP SIP. During out sample period, the UTP SIP disseminated trades for broadcast less than 20 microseconds after receiving them, while the CTA SIP disseminated trades for broadcast between 100 and 200 microseconds in 2019 and most of 2020, and less than 30 microseconds for 2021. There is also a geographic difference, with the UTP located in central New Jersey, next to the NASDAQ exchange, and the CTA SIP located in northern New Jersey, next to the NYSE exchange. Despite these differences, however, the percentage of trades with a difference between the SIP-matched quote and participant-timestamp matched quote is similar across Tapes A and C (Figure 14). Tape B has substantially fewer differences between the SIP-timestamp-matched quote and participant-timestamp-matched quote. Tape B is far more likely to have ETF listings than either Tape A or Tape C, and the lower (portfolio) volatility of ETF shares would explain the substantially decreased likelihood of a price difference between the SIP-matched and participant-matched quotes.

Figure 12. Midquote Differences by Venue Type. Trades can be matched against quotes according to one of two timestamps (described in Figure 11). On a typical day, the midquote matched via participant timestamps differs from the midquote matched via SIP timestamps around 35% of the time for non-sub-penny trades, but less than 10% of the time for sub-penny trades. We also measure how often the quotes differ 1 or 3 milliseconds prior, and 1 or 3 milliseconds after the trade. Across all time horizons, differences in quotes are far less likely for sub-penny trades.

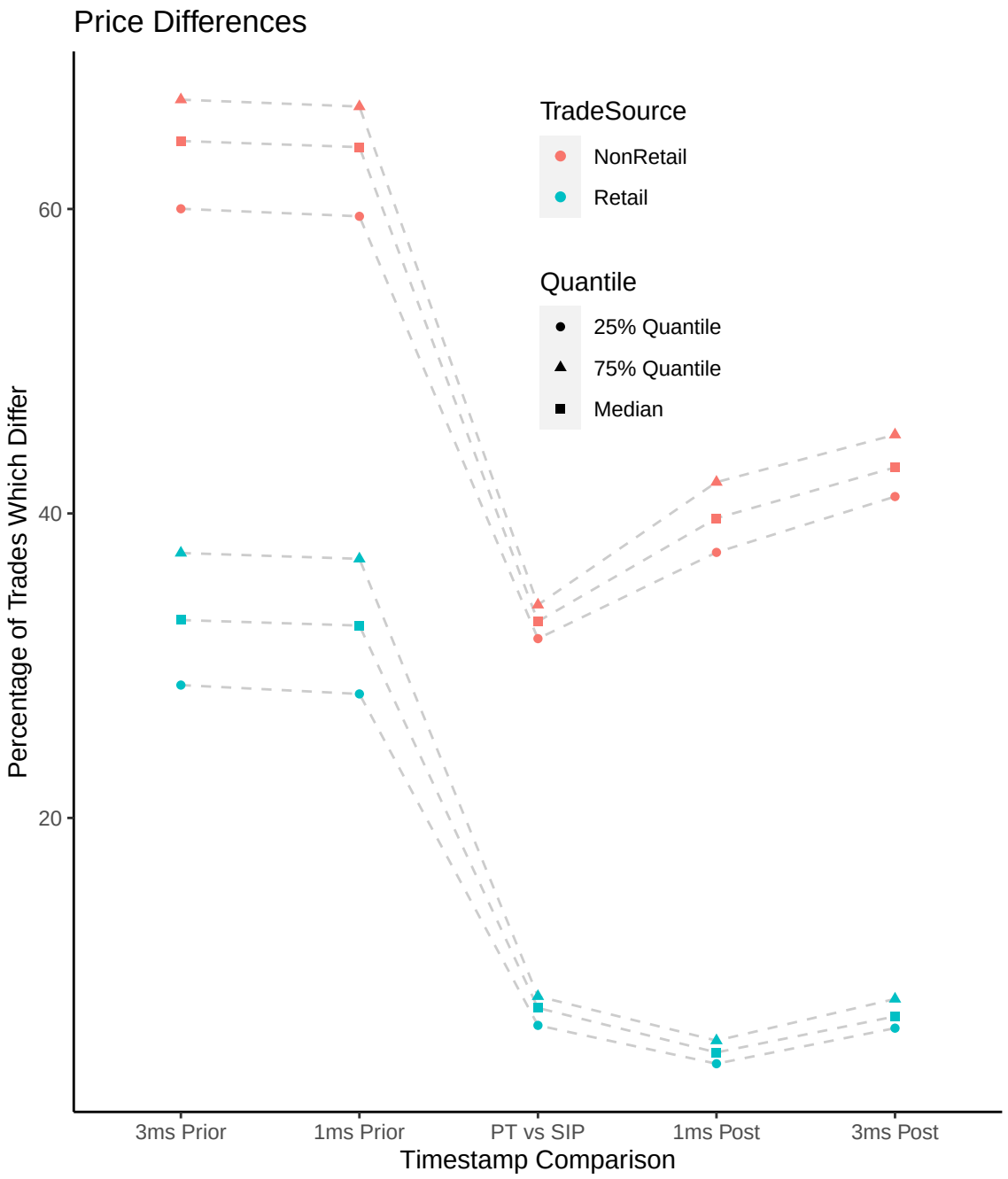


Figure 13. Midquote Differences By Stock Characteristics.. We plot the percentage of trades for which the SIP-timestamp-matched quotes differ from the participant-timestamp-matched quotes across stock-day observations. Stocks with a share price greater than $100 (dashed purple) have far more days with a high number of differences between the timestamps compared to stocks with a share price less than $15 (solid red line), reflecting the fact that small price changes are more frequent in higher-priced stocks.

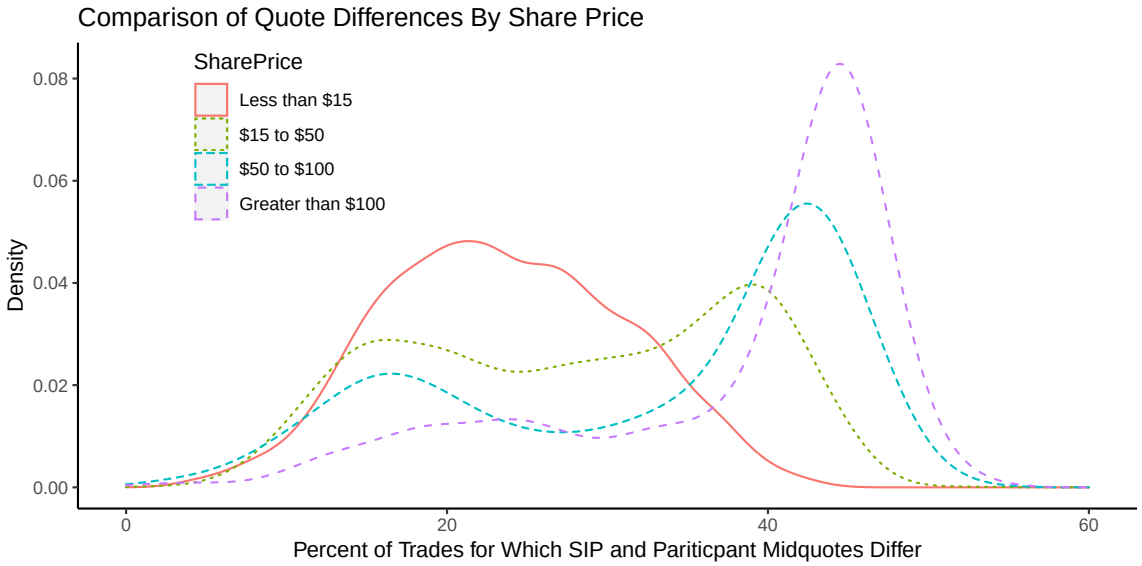
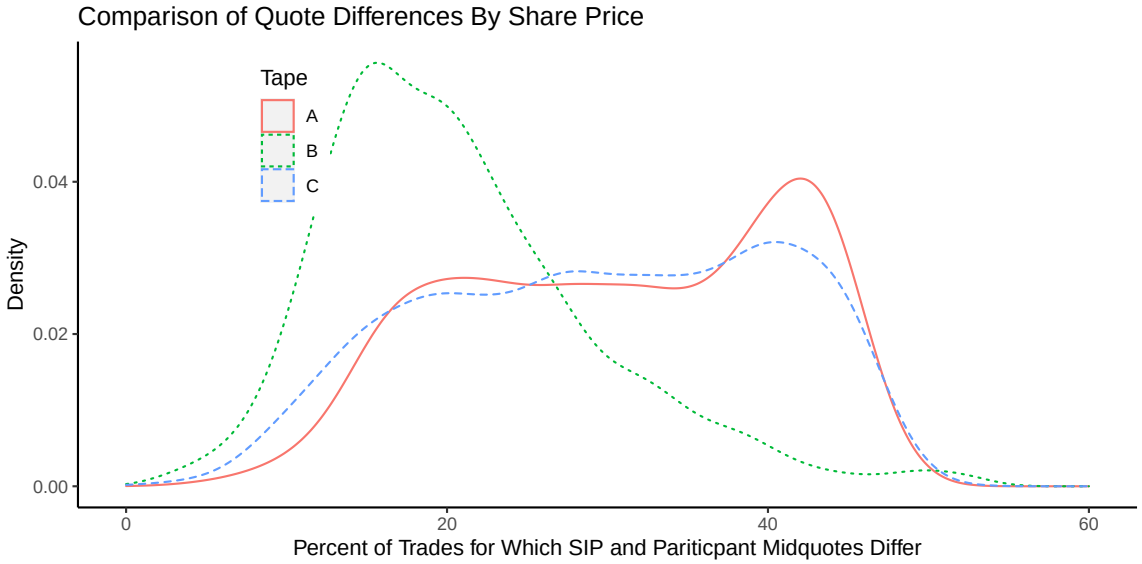


Figure 14. Midquote Differences By Tape. We plot the percentage of trades for which the SIP-timestamp-matched quotes differ from the participant-timestamp-matched quotes across stock-day observations. Stocks listed on Tapes A and C (solid red and dashed blue lines) are far more likely to have differences between the SIP-matched quotes and Participant-matched quotes compared to Tape B securities. Tapes A and B are both processed by the CTA SIP, while Tape C trades are processed by the UTP SIP.



Internet Appendix B. Total Improvement Based on NBBO

We measure price improvement for sub-penny trades as only the sub-penny portion of the order. We could alternatively measure sub-penny improvement against the prevailing NBBO, matched via either the SIP or participant timestamp.

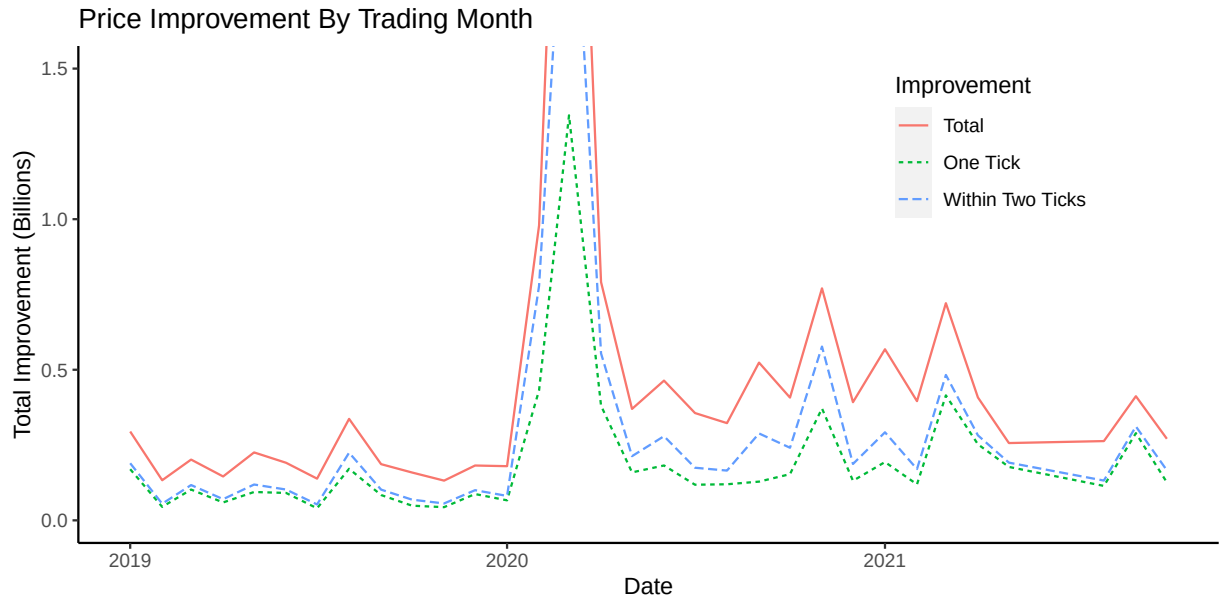
This alternative measure of improvement yields significantly larger estimates, but is open to errors in timestamps of trades. For example, participant timestamps for off-exchange trades are only accurate to the nearest millisecond. Matching quotes, therefore, may be off by as much as a millisecond, meaning trades with large improvement are also possibly trades in a volatile market. Note that during the highest volatility of the COVID-19 pandemic, estimated price improvement under the NBBO-based measure substantially increases (Figure 15).

On the other hand, if wholesalers do provide more than a sub-penny improvement, this alternative measure captures such improvement. When spreads are wider than one penny, providing more than one penny improvement, but less than a full half-spread of improvement, is possible. We redo our analysis on the sub-penny improvement using this alternative NBBO-based measure in this Appendix.

While these volumes are substantial in part due to the tremendous volume of U.S. equities trades, they also represent a meaningful portion of transaction costs. With monthly transaction volumes of sub-penny trades ranging between $300 billion and $1 trillion, the sub-penny price improvements amount to an average of a 5 basis point improvement. This is a substantial amount of price improvement. As Figure 2 shows, around half of the total sub-penny price improvement occurs when stocks are at the minimum one-tick spread, i.e. a single penny bid-ask spread.

To further capture the value of these price improvements, we consider how market-making profits would be impacted were there to be no sub-penny price improvements. We use realized spreads as a measure of market-making profits, as a realized spread compares the signed trade price against the midquote some time interval after the trade. For each sub-penny price-improved trade, we also consider the national best bid or offer, $NBBO$, at the time of the trade. We compare two different measures of realized spreads: one measured against the actual trade price, and one measured against the national best bid or offer. The measure of realized spreads using the NBBO could be thought of as the total potential revenue available to a market maker, while the measure

Figure 15. Price Improvement By Month. Total price improvement for sub-penny trades, by month. Following Boehmer et al. (2021), sub-penny trades are defined as trades in which the price has a sub-penny component between (0, 40) and (60, 100). For these trades, we calculate improvement as the difference between the price and the prevailing NBBO. For example, for a buy order at \$10.2575 with a best offer of \$10.27, the improvement would be \$0.125. We separately calculate the total value of price improvements which occur when the quoted spread is at one tick (red solid line) or at two ticks (blue dashed line). Note that March 2020, with \$3.3 billion in total improvement, is a substantial outlier and is not captured by the axis of this chart.



of realized spreads using the actual trade price is the total profit. If sub-penny price improvement is viewed as an expense, the ratio of the two realized spread measures captures the share of total revenue devoted to sub-penny price improvement.

Formally, for a trade price P_T, national best bid or offer NBBO, trade sign Y, and midquote m which occurs X seconds after the trade, we define the two possible definitions of a realized spread:

$$\text{Realized\_With\_Improvement}_t = Y(P_t - m_{t+X}) \tag{B1}$$

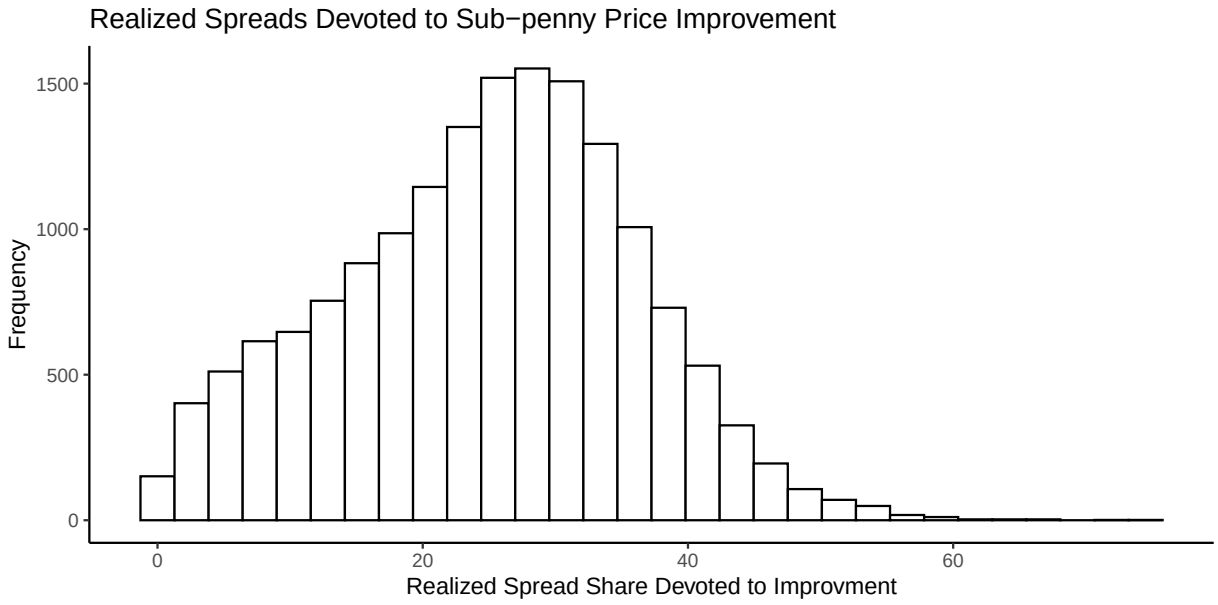
$$\text{Realized\_No\_Improvement}_t = Y(NBBO_t - m_{t+X}) \tag{B2}$$

$$\text{Alt\_Expense\_Ratio} = 1 - \frac{\sum \text{Realized\_With\_Improvement}}{\sum \text{Realized\_No\_Improvement}} \tag{B3}$$

For each stock, we calculate the value of this expense ratio, and plot the distribution of this ratio across stocks in Figure 3. The average stock has a 25% difference between the two measures. If we consider the total revenue from retail trades to be the realized spread on these trades calculated

using the NBBO, around 25% of this total revenue goes to offering sub-penny price improvements.

Figure 16. Price Improvement As Fraction of Realized Spreads. Sub-penny price-improvement reduces realized spreads by between 20 to 40%, measured against the contemporaneous national best bid or offer. For all sub-penny trades, we calculate the realized spreads using both the trade price and the NBBO (Equation B3). The realized spread using the trade price reflects market maker profits, while the realized spread using the NBBO reflects market maker revenue. In the average stock, total realized spreads in sub-penny stocks are around 25% lower than the total realized spreads on those same trades using the NBBO rather than the trade price. This suggests around 25% of the total revenue market makers make from retail trades is allocated to sub-penny price improvement.



To test how the profitability of internalization changes with market conditions, we test Regression 2. As defined in Equation 3, realized ratio is the ratio between two different measures of realized spreads: one measured against the actual trade price, and one measured against the national best bid or offer. The measure of realized spreads using the NBBO could be thought of as the total potential revenue available to a market maker, thus the difference between the two realized spreads represents the cost to the market maker of offering the sub-penny price improvement. This ratio is defined only for sub-penny improved trades, and to reduce the variance of this ratio we exclude any stock-day observations with fewer than 50 sub-penny trades.

REGRESSION 9: *For each stock i on date t with at least 50 sub-penny trades, we estimate:*

$$\begin{aligned} Realized\_Ratio_{it} = \quad & \alpha_0 Closing\_Price_{it} + \alpha_1 Absolute\_Intraday\_Return_{it} \\ & + \alpha_2 Mean\_Quoted\_Spread_{it} + \alpha_3 Mean\_Realized\_Spread_{it} \\ & + \alpha_4 Off\_Exchange\_Share_{it} + X + \epsilon_{it} \end{aligned}$$

Results of this estimation are presented in Table X. Results are generally consistent with Table X, with the exception that the closing price here is not significant, and here the relationship between Realized Ratios and intraday returns is unambiguously negative.

Table X: Realized Spread Share. This table estimates Regression 9. Realized Ratio, defined in Equation B3, is the ratio on realized spreads for sub-penny improved orders, and compares the realized spread calculated with the trade price against the realized spread calculated with the prevailing best bid or offer. A larger ratio indicates a larger share of market-maker revenue goes to offering sub-penny price improvement. Closing Price is measured in dollars, while Absolute Intraday Return and Off-Exchange Share are measured in percentages. The level of observations is the stock-day level; to reduce noise in the realized spread ratio, we exclude stock-day observations with less than 50 sub-penny trades. Odd columns have a fixed effect for each date, while even columns have a fixed effect for each stock. Standard errors are clustered at the stock and day level.

	Dependent variable:					
	Realized Ratio 3ms		Realized Ratio 1s		Realized Ratio 30s	
	(1)	(2)	(3)	(4)	(5)	(6)
Closing Price	−0.0003		0.002		0.004	
	(0.003)		(0.003)		(0.003)	
Absolute Intraday	−14.923***	−11.264***	−8.109***	−6.785***	11.151***	2.433
Return	(2.103)	(1.859)	(1.430)	(1.503)	(2.244)	(1.636)
Mean Quoted	0.024***	0.008***	0.026***	0.010***	0.026***	0.009***
Spread (BPS)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
Mean Realized	−0.068***	−0.071***	−0.083***	−0.083***	−0.103***	−0.093***
Spread (BPS)	(0.007)	(0.004)	(0.007)	(0.004)	(0.007)	(0.004)
Off-Exchange	0.087***	−0.031***	0.081***	−0.028***	0.071***	−0.012***
Date Fixed Effect	X		X		X	
Stock Fixed Effect		X		X		X
Observations	3,044,374	3,044,374	3,029,709	3,029,709	2,935,098	2,935,098
R^2	0.056	0.240	0.050	0.206	0.038	0.148
Adjusted R^2	0.056	0.238	0.050	0.205	0.038	0.146
Residual Std. Error	15.177	13.633	16.107	14.732	18.565	17.483
Degrees of Freedom	3043710	3038605	3029045	3023940	2934434	2929329
Note:						*p<0.1; **p<0.05; ***p<0.01