High Frequency Quoting, Trading, and the Efficiency of Prices*

Jennifer Conrad
Kenan Flagler Business School
University of North Carolina at Chapel Hill
j_conrad@kenan-flagler.unc.edu

Sunil Wahal
WP Carey School of Business
Arizona State University
Tempe, AZ 85287
Sunil.Wahal@asu.edu

Jin Xiang
Integrated Financial Engineering
51 Monroe Street, Suite 1100
Rockville, MD 20850

April 2014

* We thank Alyssa Kerr and Phillip Howard for research assistance and the Wharton Research Data Service for providing a TAQ-CRSP matching algorithm. We thank Gaelle Le Fol, Mark Seasholes and participants at the 5th Hedge Fund Research Conference (Paris), the Instinet Global Quantitative Equity Conference and the Financial Markets Research Center Conference (Vanderbilt University). We thank the Tokyo Stock Exchange for providing proprietary data around the introduction of Arrowhead. Wahal is a consultant to Dimensional Fund Advisors. DFA and Integrated Financial Engineering provided no funding or data for this research.

High Frequency Quotation, Trading, and the Efficiency of Prices

Abstract

We examine the relation between high frequency quotation and the behavior of stock prices between 2009 and 2011 for the full cross-section of securities in the U.S. On average, higher quotation activity is associated with price series that more closely resemble a random walk, and significantly lower cost of trading. We also explore market resiliency during periods of exceptionally high low-latency trading: large liquidity drawdowns in which, within the same millisecond, trading algorithms systematically sweep large volume across multiple trading venues. Although such large drawdowns incur trading costs, they do not appear to degrade the subsequent price formation process or increase the subsequent cost of trading. In an out-of-sample analysis, we investigate an exogenous technological change to the trading environment on the Tokyo Stock Exchange that dramatically reduces latency and allows co-location of servers. This shock also results in prices more closely resembling a random walk, and a sharp decline in the cost of trading.

1. Introduction

Electronic trading has dramatically changed the way that liquidity is demanded and supplied. Market-making has been re-defined. The NYSE specialist has been replaced with Designated Market Makers (DMMs) and Supplemental Liquidity Providers (SLPs), who do not receive preferential information, and are not subject to negative obligations. Traditional market makers on NASDAQ and other exchanges now use low-latency technology. Other market participants, primarily high frequency trading firms, may also serve as liquidity providers. All participants operate in a fragmented market with multiple exchanges, electronic communications networks (ECNs), alternative trading systems (ATSs) and dark pools, under a regulatory framework provided by Regulation NMS and Regulation ATS.

In this market structure, the returns to liquidity provision are earned through two channels: (a) the liquidity provider takes on inventory, bearing adverse selection risk, and releasing inventory, and/or (b) earning the difference between make/take fees paid by exchanges. The former is reflected in classic market microstructure models such as Ho and Stoll (1981), Glosten and Milgrom (1985) and Kyle (1985). The latter is a more recent market feature, studied by Colliard and Foucault (2012), and Foucault, Kadan and Kandel (2012). A hallmark of the current market structure is that the supply curve(s) offered by liquidity providers are endogenous and can change very quickly. The ability of liquidity suppliers to change the price and quantity of liquidity offered is measured in the nanoseconds at NASDAQ, and in microseconds at NYSE/Arca, BATS and EDGA/X. High frequency changes in the supply curve are frequently lumped under the term “high frequency trading” (HFT) and “algorithmic trading”, although they are more often changes in high frequency *quotations* – while supply curves can move for a variety of reasons, for a trade to occur, demand and supply curves must intersect.

High frequency quotation and trading can have important economic consequences – although the direction of the impact is still being debated. For example, Budish, Cramton and Shim (2013) build a model in which the ability to continuously update order books generates technical arbitrage opportunities and a wasteful arms race in which fundamental investors bear costs through larger spreads and thinner markets. Han, Khapko and Kyle (2014) argue that since fast market makers can cancel faster than slow market makers, this causes a winner’s curse resulting in higher spreads. In contrast, in Aït-Sahalia and Saglam (2013), lower latency generates higher profits and higher liquidity provision. However, in their model, high-frequency liquidity provision declines when market volatility increases, which can lead to episodes of market fragility. And in Baruch and Glosten (2013), frequent order cancellations are a standard part of liquidity provision, and are generated by limit order traders mitigating the risk that their quotes will be undercut (through rapid submissions and cancellations) rather than "a nefarious plan to manipulate the market". While all of these mechanisms are plausible, ultimately, the net effect of high frequency quotation changes on markets is an empirical question.

Different interpretations of the consequences of high-frequency quotation and trading underlie the SEC's (2010) concept release on market structure, which explicitly asks whether high frequency quoting represents "phantom liquidity (which) disappears when most needed by long-term investors and other market participants".[1] These issues are of obvious importance to trading and market microstructure, but are also relevant from a larger economic perspective; that is, measures of value are not unrelated to the architecture of the markets on which assets trade. Duffie (2010) points out that equilibrium asset pricing theories can and perhaps should include the capitalization and willingness of intermediaries to participate in the market. The efficacy of the price formation process is also critical for empirical work in asset pricing: if noise related to market microstructure is too large, the resulting biases overwhelm attempts to understand the dynamics of asset prices (Asparouhova, Bessembinder and Kalcheva (2010)). Finally, non-execution risk related to market microstructure has welfare consequences in that mutually profitable trades may be missed. Bessembinder, Hao and Lemmon (2011), for instance, explicitly study allocative efficiency and price discovery and argue that affirmative obligations can improve social welfare relative to endogenous liquidity provision. Stiglitz (2014) also highlights the welfare debate and expresses skepticism that high frequency quotation/trading is welfare improving. Our purpose is to bring additional data and evidence to bear on these issues

[1] Filings with regulatory bodies, exchanges, trade groups and press accounts, as well as some academic papers, contain numerous suggestions to slow the pace of quotation and trading to what is determined to be a "reasonable" pace. See, for example, the testimony of the Investment Company Institute to the Subcommittee on Capital Markets, US House of Representatives, in which the testifier argues for meaningful fees on cancelled orders as a mechanism to prevent high frequency changes in the supply curve (http://www.ici.org/pdf/12_house_cap_mkts.pdf).

using a comprehensive cross-section of securities, across multiple markets and with a relatively long time-series, while investigating both 'average' and stressed market environments.

We examine the effect of rapid changes in the supply curves of over 3,000 individual securities on the behavior of prices between 2009 and 2011. Our measure of a change in the supply curve is a "quote update", which we define as any change in the best bid or offer (BBO) quote or size across all quote reporting venues. Changes in the supply curve can come from the addition of liquidity to the limit order book at the BBO, the cancellation of existing unexecuted orders at the BBO, or the extraction of liquidity via a trade. We examine the relation between quote updates and variance ratios over short horizons. If high frequency activity merely adds noise to security prices, then we should observe variance ratios substantially smaller than one for securities in which high frequency activity is more prevalent, as high frequency quotation-induced price changes are reversed; Brunnermeier and Pedersen (2009) describe this possibility as "liquidity-based volatility", which might be observed in short-horizon variance ratio tests. The more quickly reversals occur, the quicker variance ratios should converge to one. In contrast, if high frequency quotation is associated with persistent swings away from fundamental values, variance ratios in securities with higher levels of quotations may rise above one.

There is significant variation in update activity across securities. Between 2009 and 2011, in the smallest size quintile of stocks, there is less than one quote update per second. In large capitalization stocks, there are on average over 20 updates to quotes per second. Controlling for firm size and trading activity, average variance ratios (based on 15-second and 5-minute quote midpoint returns) are reliably closer to one for stocks with higher updates. In a robustness check, we examine variance ratios for a subset of securities at higher frequencies (100 milliseconds and 1 second) and the results are unchanged. In addition, the time series average of the cross-sectional standard deviation of variance ratios is lower for stocks with higher updates. The implication is that higher update activity is associated with improved price formation.

Higher updates are also associated with lower costs of trading. Again controlling for firm size and trading activity, average effective spreads (that is, deviations of transaction prices from quote midpoints) are lower for stocks with higher quote updates by 0.5 to 6 basis points. To put this in economic perspective, make or take fees of $0.003 per share for a $60 stock correspond to

0.5 basis points. Make/take fees are important enough to drive algorithmic order routing decisions between venues, indicating that the differences in effective spreads that we observe are at least as economically important.

Effective spreads could narrow because of lower revenue for liquidity providers (lower realized spreads) or smaller losses to informed traders (changes in price impact). Most of the difference in effective spreads in our sample appears to come from a reduction in realized spreads, suggesting that competition between liquidity providers provides incentives to update quotes. Regardless of the source, updates appear to be economically meaningful, rather than merely "quote stuffing" that obfuscates trading intentions. Overall, the data appear to be consistent with the Baruch and Glosten (2013) argument that there is nothing "nefarious" about high frequency quote updates; rather, these updates reflect the way that liquidity is provided in electronic markets.

A common complaint of the current market structure is that it is fragile in the sense that the price of liquidity rises too rapidly, or that liquidity disappears entirely when traders need it most (Stiglitz (2014)). Such episodes, as exemplified by the Flash Crash and individual security "mini-crashes", naturally concern market participants and regulators. We investigate fragility (or rather its mirror image, resilience) by examining price formation and trading costs surrounding large and extremely rapid drawdowns of liquidity. In fragile markets, such liquidity drawdowns should cause price series to deviate from a random walk and future trading costs to rise. We identify liquidity sweeps as multiple trades in a security across different reporting venues with the same *millisecond* timestamp. Such trades are quite common and are simultaneous algorithmic sweeps off the top of each venue's order book, designed to quickly extract liquidity. There are often successive sweeps that, within short periods of time, extract substantial amounts of liquidity. We design a simple algorithm to aggregate successive sweeps into singular liquidity drawdowns and restrict our attention to drawdowns in which at least 10,000 shares are traded. Unsurprisingly, both buyer- and seller-initiated drawdowns incur substantial costs. The average total effective spread paid by liquidity extractors ranges from over 100 basis points in microcap stocks to 17 basis points for securities in the largest size quintile.[2] However, average variance ratios 300 seconds before and after such events are indistinguishable from each other. We similarly see no evidence that effective spreads increase after large buyer- or seller-initiated liquidity drawdowns. On average, the market appears resilient.

Of course, quote updates and prices are endogenous and jointly determined, so that the above set of results do not imply causation. That is, high frequency traders may be more likely to participate, and hence we would be more likely to observe heavy quote updating, in more liquid securities. We perform two tests that help with identification, while not abandoning a large sample approach.

First, we exploit the daily time series variation in quote updates. The daily average number of quote updates closely tracks the VIX. A reduced-form vector autoregression shows that changes in updates are related to lagged innovations in the VIX but not vice-versa, implying that update activity is not merely noise ("servers talking to servers") but related to economic fundamentals. Related, Nagel (2012) uses daily data and finds a predictive connection between return reversals (interpreted as returns to liquidity provision) and the VIX, and argues that the VIX likely proxies for state variables that influence both the demand for liquidity and its supply. To the extent that spreads represent the return to endogenous liquidity supply, quote updates are the tool used by liquidity providers to manage their intraday risk. Given this evidence, it seems unlikely that variance ratios in day *t* are endogenously related to quotation activity in day *t-1*. The implication is that if we use the prior day's updates to sort stocks into low and high update groups, this will mitigate the possibility that an omitted factor is driving both the lagged update measure, and current spreads and variance ratios. Using the previous day's update measures, we continue to find that higher updates are associated with variance ratios closer to one.

Second, we examine an exogenous technological change to trading practices in the Tokyo Stock Exchange. On January 4, 2010 the Tokyo Stock Exchange replaced its existing trading infrastructure with a new system (Arrowhead) that reduced the time from order receipt to posting/execution from one-to-two seconds to less than 10 milliseconds. At that time, the TSE

---

[2] In comparison, Madhavan and Cheng (1997) report average price impact (measured as the price movement from 20 trades prior to a block print) of between 14 and 17 basis points in Dow Jones stocks for 30 days in 1993-1994.

also permitted co-location services and started reporting data in millisecond time stamps (down from minutes). This large change in latency provides us with an exogenous shock that helps identify the impact of high frequency quoting on the price formation process. The fact that it takes place in a non-U.S. market is advantageous in that it serves an out-of-sample purpose. We find that the introduction of Arrowhead resulted in large increases in updates. As with the U.S., spikes in updates correspond to economic fundamentals and uncertainty, such as the earthquake and tsunami that hit Japan in March 2011. Unlike the U.S. our Japanese data allows us to directly observe new order submissions, cancellations and modifications. We find similar increases in all three components of updates after the introduction of Arrowhead, along with similar spikes related to economic shocks. Most importantly, we find a systematic improvement in variance ratios between the three-month period before and after the introduction of Arrowhead in every part of the trading day. There are also beneficial effects on the cost of trading: effective spreads decline by 20 percent on the date of the introduction of the new trading system.

Various studies in this area investigate what high frequency traders do, whether they are profitable, whether they improve or impede price discovery, and other questions of importance (see, e.g., Brogaard, Hendershott and Riordan (2012), Hendershott and Riordan (2012), and Menkveld (2012)). Empirical studies often focus on specific exchanges (e.g., Nasdaq, or Deutsche Borse), a subsample of stocks, or exchange-identified high frequency traders (the so-called Nasdaq dataset). The above papers generally find that, although liquidity-demanding HFT may impose adverse selection costs on other traders, high frequency traders in aggregate either do no harm or occasionally improve liquidity. However, better identification sometimes comes at the cost of generality; we believe there is an advantage to examining a comprehensive cross-section of stocks over a relatively long time-series when drawing conclusions about the effect of high-frequency trading. In addition, our interest is not just in high frequency trading per se but more broadly in supply curves. In that sense, the two empirical papers most similar to ours are Hendershott, Jones, and Menkveld (2011) and Hasbrouck (2013). Hendershott et al. (2011) study 1,082 stocks between December 2002 and July 2003 and, using the start of autoquotes on the NYSE as an exogenous instrument, find that algorithmic trading improves liquidity for large stocks. They observe increases in realized spreads in their sample and speculate that electronic liquidity providers may initially have had market power relative to human traders, although they conclude that this power was temporary (see their Figure 3)). Our data show that if anything, increased quotation activity is associated with decreases in realized spreads, suggesting competition between electronic liquidity providers during our sample period.

Both Hendershott et al. (2011), and our analysis of the Arrowhead introduction, analyze the transition from one microstructure equilibrium to another, where the transition facilitates electronic trading. In comparison, our US analysis is an investigation of the effects of cross-sectional variation in electronic trading across securities. Hasbrouck (2013) is also concerned about the effects of electronic trading on short-term volatility in quotes, since excessive volatility reduces their informational content and can increase execution price risk for marketable orders. Using a sample of 100 stocks in April 2011, he examines variances over time scales as low as 1 millisecond, and finds that these short horizon variances appear to be approximately five times larger than those attributable to fundamental price variance. Like us, Hasbrouck (2013) concludes that it is important to understand the nature of this volatility and its impact on the price formation process.

The remainder of the paper is organized as follows. In Section 2, we describe our sample and basic measurement approach. We discuss the cross-sectional results in Section 3, and present alternative tests, including those based on liquidity sweeps and Japanese data in Section 4. Section 5 concludes.

2. Sample Construction and Measurement

2.1 U.S. Data and Sample

For 2009 we use the standard monthly TAQ data in which quotes and trades are time-stamped to the second. For 2010 and 2011, we use the daily TAQ data (NBBO and CQ files) in which quote and trades contain millisecond timestamps. There are obvious advantages of working with data that have millisecond resolution. For example, we avoid conflation in signing trades, a process necessary for computing effective spreads. In addition, these data are also necessary for identifying liquidity sweeps.

In processing TAQ data, we remove quotes with mode equal to 4, 7, 9, 11, 13, 14, 15, 19, 20, 27, 28 and trades with correction indicators not equal to 0, 1 or 2. We also remove sale condition codes that are O, Z, B, T, L G, W, J and K, quotes or trades before or after trading hours, and locked or crossed quotes. In the millisecond data, we also employ BBO qualifying conditions and symbol suffixes to filter the data. We use an algorithm provided by WRDS (TAQ-CRSP Link Table, Wharton Research Data Service, 2010) that generates a linking table between CRSP Permno and TAQ Tickers. We keep only firms with CRSP share codes 10 or 11 and exchange codes 1, 2, and 3. To ensure that small infrequently traded firms do not unduly influence our results, we remove firms with a market value of equity less than $100 million or a share price less than $1 at the beginning of the month. Although we remove ETFs from the main sample, we retain information on two ETFs (SPY & IWM) for separate tests.

Many of our tests are based on size quintiles because of systematic effects between small and large capitalization firms. We employ the prior month's NYSE size breakpoints from Ken French's website to create quintiles. Using NYSE breakpoints obviously causes the quintiles to have unequal numbers of firms in them, but we end with a better distribution of market capitalization across groups. This method also facilitates comparisons for those interested in the relevance of our results for investment performance and portfolios. On average, we sample more than 3,000 stocks which represent over 95 percent of aggregate U.S. market capitalization.

2.2 Japanese Data and Sample

During our sample period, trading on the TSE is organized into a morning session between 9:00 am and 11:00 am, and an afternoon session from 12:30 to 3:00 pm. Each session opens and closes with a single price auction (known as "*Itayose*"), and continuous trading (known as "*Zaraba*") takes place between the auctions. Under certain conditions (e.g. trading halts), price formation can take place via the *Itayose* method even during continuous trading.

The Tokyo Stock Exchange provided us with two proprietary datasets. The first is for the six months prior to the introduction of Arrowhead (July 1, 2009 to December 31, 2009), and the second is for 15 months after the introduction of Arrowhead (January 4, 2010 to March 31, 2011). The data are organized as a stream of messages that allow us to rebuild the limit order book in trading time. Prior to the introduction of Arrowhead, time stamps are in minutes but updates to the book within each minute are correctly sequenced. After Arrowhead, time stamps are in milliseconds. For each change to the book, we observe the trading mechanism and the status of the book (*Itayose*, *Zaraba*, or trading halts). We also observe the nature of each modification to the book. Specifically, we observe new orders, modifications which do not discard time priority, modifications which result in an order moving to the back of the queue, cancellations, executions and expirations. We also observe special quote conditions and sequential trade quotes. Each dataset is for the largest 300 stocks in First Section of the Tokyo Stock Exchange by beginning-of-month market capitalization. As a result, the sampling of stocks varies slightly over time. Lot sizes for stocks vary cross-sectionally and change over time. The TSE provided us with a separate file that contains lot sizes as well as changes in these sizes, allowing us to compute share-weighted statistics.

2.3 Measuring Changes in the Supply Curve

In a perfect world in which trading occurs in a single market and is entirely transparent, one might observe a complete quote montage that provides a trader with bid and ask prices, and their associated depths. Of course, this would be a static snapshot of a supply curve, but it nonetheless constitutes a complete representation of the extant supply curve. Trading protocols and market structure lead to deviations from this ideal. In the U.S., there are three important considerations. First, securities can trade in multiple venues, rather than just the primary exchange. The implication is that there are multiple supply curves which need to be aggregated to obtain a true supply curve that conforms to price and time priority rules. The Order Protection Rule of Regulation NMS requires that brokers respect the best price across protected venues, giving it importance over any other attribute of best execution, including speed. This effectively aggregates the best price *point* across multiple supply curves, rather than the entire curve.[3] Second, only supply curves from registered exchanges can be aggregated. Dark pools that do not

[3] Such fragmentation also gives rise to latency issues. Exchanges supply data feeds to the Consolidated Quote System/Service (CQS) and the Consolidated Tape System (CTS), which is then used to determine the official NBBO. This NBBO is slower than that generated by a manual consolidation of direct exchange feeds, implying that investors who use the slower official feed can be subject to latency arbitrage.

provide quotations are part of the latent supply curve but do not currently show up in official exchange or consolidated feeds. For example, a stock that shows an aggregated depth (across exchanges) of 500 shares at the ask price might have another 200 shares available in a dark pool pegged to the ask quote, so that the true supply is for 700 shares. Third, hidden orders on exchanges do not show up on the displayed supply curve.

With the above structure in mind, we build our main measure ("quote updates") as the number of changes that occur in the best bid or offer price, or in the quoted sizes at these prices, within a specified time interval across all registered exchanges. There are several advantages to using this method. A venue choice is a decision element common to both liquidity extraction and provision algorithms. Venue choices are often dynamic in nature and can be changed for different child orders generated from the same parent. As a result, inferences from datasets from a particular exchange are subject to potential selection biases.[4] Moreover, under Regulation NMS, flickering quotes, defined as quotes that change more than once per second, are not eligible to set the NBBO.[5] By using the BBO across all exchanges, we include quote changes that are legitimate changes to the supply curve, regardless of whether the change is eligible to set a new NBBO. This is an underestimate of true changes in the supply curve – it does not include dark venues, and also does not include hidden orders. In addition, it does not consider changes to the supply of liquidity outside the best bid or ask prices, the "Level II" of the quote book.

Changes in updates occur as orders are added to each exchange's book, removed from the respective books due to cancellation, or removed due to executions. The first two represent changes to the supply curve through quotation activity. The last is a change in the supply curve caused by intersection with a demand curve through a trade. In deciding to trade, traders likely endogenize the tradeoff between the cost and benefit of monitoring the market (Foucault, Kadan and Kandel (2012)). Trades, by virtue of their capital commitment, have important

[4] For example, suppose one observes trades from a high frequency trader from Exchange X that is known to be cheaper for extracting liquidity for a particular group of stocks. Such a high frequency trader may be providing liquidity in Exchange Y, but a researcher only observing trades on Exchange X would erroneously draw the conclusion that this high frequency trader is a liquidity extractor. There are a variety of reasons why there may be a non-random distribution of trades across trading venues, ranging from concerns about adverse selection to systematic differences in make-take fees. Sampling a larger number of high frequency traders does not necessarily make this potential bias disappear.

[5] Exchanges are free to ignore flickering quotes for trade through protection and many exchanges have rules that explicitly prohibit quote manipulation (e.g. NYSE/Arca Rule 5210).

consequences for the price formation process, impounding information into prices and also demanding/supplying liquidity. We therefore conduct tests controlling for trade frequency.

In Japanese data, we calculate updates in a manner similar to that for the U.S. but without the need to deal with venue fragmentation. There is an added advantage in that, in addition to updates, we separately observe submissions, modifications, cancellations, executions, and expirations. As we show later, these are highly correlated, suggesting that even though updates are the summation of these different behaviors, they are a good proxy for cancellations.

2.4 Measuring Price Efficiency and Execution Quality

2.4.1 Variance Ratio

Following the notation in Lo and MacKinlay (1988), define X_t as the log price process, where $X_t = X_0, X_1, \ldots, X_{nq}$, each interval is equally spaced and there are *nq+1* time series observations. We refer to the price process in generic terms, although in implementation we use NBBO quote midpoints to avoid negative autocorrelation induced by bid-ask bounce. For cross-sectional tests, we measure prices/returns over 15 second and 5 minute intervals. There are 120 15-second returns in a half-hour and we require at least 20 non-zero 15-second returns, to calculate a variance ratio. This ensures that variances, and therefore their ratios, are not degenerate. Our choice of measurement interval is determined by two tradeoffs. The interval needs to be short enough to measure high frequency changes in the supply curve, while preserving time of day effects. The interval also needs to be long enough to reliably measure contemporaneous variance ratios across a large sample of securities. A half hour interval is a reasonable balance between capturing high frequency activity and this econometric necessity.

Given the speed of the quote updating process, it is possible that 15 seconds is too long an interval. In a robustness check, we also measure variance ratios using 100 millisecond and one or two second returns for stocks in the largest size quintile. We do so only for large cap stocks because in other securities, quote midpoints do not change enough in successive 100 millisecond intervals to provide a reliable measure of the variance of returns. In addition, for calculating variance ratios around liquidity sweeps, we use the variability of quote midpoints at 1 and 15 second intervals, in a 300 second period before and after each sweep.

There are T-q+1 (where T = nq) overlapping returns in the data. Comparing midpoint sampling intervals, we generally have q = 10 or 20 in our tests (see Lo and MacKinlay (1989) for a discussion of the choice of q). Given this, the estimate of the mean drift in prices is

$$\hat{\mu} = \frac{1}{nq}\sum_{k=1}^{nq}(X_k - X_{k-1}) = \frac{1}{nq}(X_{nq} - X_0)$$

so that the variance of shorter interval returns (a) is then

$$\hat{\sigma}_a^2(q) = \frac{1}{nq}\sum_{k=1}^{nq}(X_k - X_{k-1} - \hat{\mu})^2$$

To maximize power, we use overlapping q^{th} differences of X_t so that the variance of larger interval (c) returns is

$$\hat{\sigma}_c^2(q) = \frac{1}{nq^2}\sum_{k=q}^{nq}(X_k - X_{k-q} - q\hat{\mu})^2$$

Lo and MacKinlay (1989) recommend estimating variances as follows with a bias correction.

$$\bar{\sigma}_a^2(q) = \frac{1}{nq-1}\sum_{k=1}^{nq}(X_k - X_{k-1} - \hat{\mu})^2$$

$$\bar{\sigma}_c^2(q) = \frac{1}{m}\sum_{k=q}^{nq}(X_k - X_{k-q} - q\hat{\mu})^2$$

where

$$m = q(nq - q + 1)(1 - \frac{q}{nq})$$

The variance ratio test is then the ratio of the two variance estimates, which should be linear in the measurement interval.

$$M_r(q) = \frac{\bar{\sigma}_c^2(q)}{\bar{\sigma}_a^2(q)} - 1$$

Lo and MacKinlay (1988) show that $M_r(q)$ is a linear combination of the first q-1 autocorrelation coefficients with arithmetically declining weights.

2.4.2 Execution Quality

We measure liquidity using effective (percentage) half spreads in standard ways, defined as follows.

$$es_{jt} = q_{jt}(p_{jt} - m_{jt})/m_{jt}$$

where q_{jt} is equal to +1 for buyer-initiated trades and -1 for seller-initiated trades, p_{jt} is the transaction price and m_{jt} is the prevailing quote midpoint. Signing trades in a high frequency quoting and trading environment is extremely noisy if timestamps are in seconds. In calculating effective spreads for trades in 2010 and 2011, there is no mis-measurement because we use data with millisecond timestamps. For some tests that include data from 2009, we use the approach advocated by Holden and Jacobsen (2012) to minimize errors. For tests that require absolute precision in timestamps (such as identifying liquidity sweeps), we only use data from 2010-2011.

We also decompose effective spreads into their components: realized spreads and price impact. Realized spread is computed as follows:

$$rs_{jt} = q_{jt}(p_{jt} - m_{j,t+\tau})/m_{jt}$$

where $m_{j,t+\tau}$ is the quote midpoint τ periods after the trade. The realized spread is a measure of revenue to market makers that nets out losses to better-informed traders. It is conventional in prior studies to set τ to 5 minutes after the trade, under the assumption that liquidity providers close positions after 5 minutes. Given the nature of high frequency quoting and trading in our sample period, a 5 minute interval to close out a position seems exceptionally long for a high-frequency market maker. However, since one can never know the true trading horizons of high-frequency liquidity providers, we estimate realized spreads from one second to 20 seconds after each trade. While computationally challenging, this allows us to examine the full term-structure of realized spreads over a range of intervals. Analogously, we calculate the losses to better-informed traders, or price impact, as follows.

$$pi_{jt} = q_{jt}(m_{j,t+\tau} - m_{j,t})/m_{jt}$$

Realized spreads and price impact represent a decomposition of effective spreads: the identity describing their relationship is exact at particular points in this term structure (values of τ), such that $es_{jt} = rs_{jt} + pi_{jt}$.

3. Cross-Sectional Results

3.1 Quotation and Trading Activity

We calculate the average number of trades and quote updates across securities in a size quintile in a half hour interval, and then average over the entire time series. We also perform the same calculation for SPY and IWM. Panel A of Table 1 shows the average number of trades per second and Panel B shows the average number of quote updates per second. Given our data filters in section 2.1, each quintile is well diversified across a large number of firms. The smallest size quintile has the largest number of firms due to the use of NYSE size breaks, and typically contains micro-cap stocks. Generally, quintiles 4 and 5 contain over 80 percent of the aggregate market capitalization. For readers interested in efficiency outside of small stocks, focusing on these quintiles is adequate to conduct inferences.

The average number of trades and quote updates increases monotonically from small to large firms. The magnitude of the increases is notable. For instance, between 1:00 and 1:30 PM, there are 0.03 trades per second (or 54 trades in the half hour) for the smallest market capitalization securities. In contrast, for stocks in the largest size quintile there is almost one trade per second. The speed of trading in the two ETFs is extremely high. In IWM, there are over four trades per second and in SPY there are almost 14 trades per second. This velocity of trading increases sharply at the beginning and end of the trading day, consistent with an extensive literature in market microstructure. In the last half hour of the trading day when liquidity demands are particularly high, there are over two trades per second in the stocks in the largest size quintile and 39 trades per second in SPY.

Panel B shows the number of quote updates. There are monotonic increases in quote updates across size quintiles. Focusing again on the 1:00 to 1:30 PM window, there are 0.5 quote updates per second in the smallest size quintile and over 12 quote updates per second in the largest size quintile. In IWM and SPY, there are 136 and 91 updates per second. In general, the data show that changes in the supply curve are quite rapid – an order of magnitude more than trades. Their speed underscores the importance of execution risk and latency.

3.2 Quote updates and variance ratios

We sort all stocks within a size quintile into low and high update groups in each half hour based on the median number of updates in the prior half hour. Although we refrain from providing further statistics in a table, we can report that there is ample variation in quote update groups within size quintiles. We calculate the cross-sectional average variance ratio in each half hour and report the time series mean of these cross-sectional averages in Table 2. The standard errors of these means are extremely small because of the smoothing of average variance ratios over large numbers of stocks. To provide a sense of variability, we report time series averages of the cross-sectional standard deviations of variance ratios in parentheses.

Outside of microcap stocks (size quintile 1), average variance ratios are quite close to one. For all intents and purposes, an investor seeking to trade securities in these groups can expect prices to behave, on average, as a random walk over the horizons that we examine. We also calculate but do not report first order autocorrelations of 15 second quote midpoint returns. These autocorrelations are largely indistinguishable from zero. In addition, average variance ratios for SPY and IWM are remarkably close to one.[6]

Our interest is in the difference in variance ratios between high and low update groups. In the vast majority of cases, variance ratios in the high update groups are closer to one than in the low update group. For example, in the two largest size quintiles, which contain the majority of the market capitalization of the U.S. equity markets, average variance ratios are closer to one in high update groups for 17 out of 24 half-hour estimates. In two cases, the average variance ratios are identical and there are 5 cases where high update groups have variance ratios which are further from one. Average cross-sectional standard deviations are also systematically lower for high update groups. In the two largest size quintiles, the cross-sectional standard deviation is lower for high update groups in all 24 cases.

3.3 Separating Quote Updates from Trades

Quote updates can come from additions and cancellations of orders to the order book, or from trades that extract liquidity. One could argue that controlling for trading is unnecessary because all changes to the supply curve are legitimate, regardless of whether they are due to submissions/cancellations or trading. Nonetheless, it is important to understand whether it is

[6] We also calculate but do not report variance ratios and autocorrelations based on transaction prices. As expected, they are influenced by bid-ask bounce. Estimates are available upon request.

differences in the trading frequency across these groups that drive the relationship we observe. Within each size quintile, we sort all stocks in a half hour interval into quintiles based on the number of trades in that interval. Then, within each trade quintile, we further separate stocks into low and high update groups based on the median number of quote updates in the prior half hour. This dependent sort procedure results in 50 groups (5x5x2), and allows us to see the effect of increased quotation activity, holding size and trading activity roughly constant.

Displaying such a large number of estimates is an expositional challenge so we employ two approaches to describe our results. Figure 1 shows average variance ratios for each size and trade quintile, with separate bars for low and high update groups. The three-dimensional graph is tilted to allow the reader to see differences in variance ratios and departures from one. The graph shows that across all size and trade quintiles, average variance ratios are generally closer to one for high update groups. Figure 2 shows a similarly constructed bar graph for the average cross-sectional standard deviation of variance ratios. Cross-sectional standard deviations are systematically lower for high update groups in every size and trade quintile.

Table 3 contains formal statistics for these differences. Since we are interested in departures of variance ratios from one, we calculate the absolute value of the difference in each stock's variance ratio from one ($|VR_i-1|$), and then compute cross-sectional averages for low and high update groups. Panel A shows times series means of the differences in cross-sectional averages (high minus low). The average differences in the distance of variance ratios from one are negative for every size and trade quintile, indicating that high update groups consistently have variance ratios which are closer to one. The magnitudes of the differences are between 0.02 and 0.03, which are sizeable given the average variance ratios in Table 2. Standard errors, which are conservative and based on the time series distribution of differences, are quite small.

It may be that 15 seconds is too long an interval for the purpose at hand. Therefore, we also measure variance ratios by sampling quote midpoints at 100 millisecond and 1 second intervals (q=10), as well as 100 millisecond and 2 second (q=20) intervals. This is only feasible for large stocks because quote midpoints have to change enough over 100 milliseconds to calculate variances. As in our previous tests, we calculate average variance ratios across trade quintiles, and then report differences in the high versus low update groups in Panel B of Table 3.

In trade quintile 1, the differences in variance ratios are statistically indistinguishable from zero. In trade quintiles 2 through 4, the differences are reliably negative, implying that high update groups have variance ratios closer to one even at this higher frequency. In trade quintile 5, the difference in average variance ratios is positive when sampling midpoints at 1 second but indistinguishable from zero when using 2 seconds.

We also calculate average effective spreads across update groups. To compute these, we first calculate share-weighted effective spreads for each stock in a half hour interval, and then average across stocks in a group. These results are presented in Figure 3 and Panel C of Table 3. In every size and trade quintile, the differences are negative; effective spreads are lower for high update quintiles. With the exception of the smallest trade quintile in microcap stocks, all of the reported differences across groups have small standard errors. The magnitude of the differences varies across size and trade quintiles, ranging from 0.25 to 6.10 basis points. In large cap stocks (size quintile 5) and the highest trade quintile, the difference in effective spreads is 1.12 basis points, which we regard as economically large.

Overall, controlling for market capitalization and the level of trading, securities which have increased quotation activity tend to have price processes that more closely resemble random walks and have lower effective spreads.

3.4 Spread Decompositions

Reduction in effective spreads could come from changes in realized spreads to liquidity providers, changes in losses to informed trades (i.e. a change in price impact), or some combination of the two. Changes in price impact could occur either because of a change in the information environment, or because liquidity providers are less likely to be adversely selected. Realized spreads could decline because of competition between liquidity providers, which seems plausible given the investment in infrastructure and trading technology.[7]

The classic Glosten (1987) decomposition of effective spreads into realized spreads and price impact is typically implemented by looking at quote midpoints 5 minutes after each trade.

[7] See, for example, "Raging bulls: How Wall Street got addicted to light speed trading", Wired Magazine, August 2012.

The intuition behind this choice of horizon is that liquidity providers can reasonably offload inventory 5 minutes after acquiring it. While such a horizon is sensible in floor trading dominated by specialists and floor brokers, or in OTC markets operated by traditional market makers, it seems excessively long for high frequency liquidity providers who may take offsetting trades much more quickly. In addition, it is unlikely that horizons over which liquidity-driven return reversals take place are similar across securities with different trading and quoting frequencies. Since we are necessarily agnostic about horizons, we calculate realized spreads over one second intervals ranging from one to 20 seconds after the transaction. An added advantage of doing so is that it allows us to see the full term-structure of realized spreads. With realized spreads in place, price impact over each horizon is simply the difference between effective spreads and realized spreads ($pi_{jt} = es_{jt} - rs_{jt}$).

Implementing this approach in millisecond data is computationally non-trivial because of the enormous volume of within-second quotes. Since full cross-sectional coverage is important to our tests, we sample the time series by randomly selecting two days in each month for 2010-2011. For these 48 days, we calculate share-weighted effective spreads, realized spreads and price impact for each stock, and then average across low and high update groups within size and trade quintiles. As in Table 3, we calculate differences between each of these measures by subtracting the low update group average from its high update group counterpart. Table 4 shows the time series of average differences in basis points. In the interest of space, we only show realized spreads and price impact for one, five, ten and twenty seconds after the transaction.

Consistent with the full sample results in Table 3, average differences in effective half spreads are negative for most size and trade quintiles. In size quintiles 1 through 3, most of the reduction in effective spreads comes from decreases in realized spreads. In size quintiles 4 and 5, the reductions in effective spreads themselves are smaller, but are still due to declines in realized spreads. These results are in contrast to those reported by Hendershott et al. (2011) who report increases in realized spreads between 2001-2005, suggesting that liquidity providers (at least temporarily) earned greater revenues after the advent of autoquoting. Our results suggest that between 2009 and 2011, competition between electronic liquidity providers appears to be sufficient to generate permanent reductions in realized spreads.

3.5 Liquidity Drawdowns and Market Resiliency

In resilient markets, large drawdowns of liquidity should minimally influence prices. The millisecond TAQ data afford the possibility of such a test.[8] We employ a three-step procedure to isolate large liquidity drawdowns. In the first step, we identify multiple trades in a stock with the same millisecond timestamp across more than one reporting venue. We then aggregate these individual sweeps that occur within short durations of each other into "collapsed" sweeps, and focus exclusively on those that extract large amounts of liquidity. Details of the process are described below.

In the first step, we isolate multiple trades with the same millisecond time stamp that originate from different reporting venues.[9] It is critical that trades take place in different venues to ensure that multiple trades with the same millisecond time stamp are not a mechanical artifact of trade reporting and splitting procedures – as would be generated by one large order interacting with multiple small counterparty orders on a single limit order book. Trades across multiple venues in the same millisecond are algorithmic in nature, sweeping the top of various order books in dark and/or lit markets, and therefore represent attempts to rapidly extract liquidity. During our sample period, there are 764 million such liquidity sweeps, comprising $17.6 trillion in volume. Market volume over this period (2010-2011) is approximately $70 trillion, so individual sweeps represent about 25 percent of total volume.

Panel A of Table 6 shows three measures of the size and nature of these sweeps. In each half hour, we calculate the sum of shares traded in all individual sweeps, scaled by total share volume. We then average across stocks in a size quintile and then over time. In small stocks, sweeps represent 13 percent of total volume, rising almost monotonically to 22 percent for large

---

[8] One could look, *ex post*, at cases in which there are dramatic changes in price seemingly caused by innocuously small trades. This is the approach that some market participants take to highlight aspects of market structure to either regulators or the press. A good example is individual stock "flash crashes" systematically documented by Nanex on its website (http://www.nanex.net/FlashCrashEquities/FlashCrashAnalysis_Equities.html). While there are certainly advantages to such small sample detection methods, we prefer to use a larger sample and a more agnostic approach to draw inferences.

[9] Although millisecond timestamps may be the same, this does not necessarily mean that the execution times of such trades are identical. Venues operate on micro and nanosecond time scales so that latency differences between direct feeds from exchanges and the CQS may cause trades to look as though they occur at the same exact time. That is not a problem for our identification, however, because from an economic perspective, the liquidity sweep is still extremely rapid.

stocks. Given fragmentation and the speed at which quotes change, trading algorithms that attempt to extract liquidity quickly across multiple venues must exercise particular care to not violate the trade-through rule of Reg NMS. It is therefore common to employ Intermarket Sweep Orders (ISOs). The second and third columns in Panel A show the number of ISO trades in individual sweeps, scaled by the total number of trades and the total number of ISOs, respectively. Both measures show considerable use of ISOs. ISO usage is higher in large capitalization stocks, often representing over 13 percent of all trades in an interval and over 27 percent of all ISO trades. Figure 4 shows the time series distribution of the size of the sweeps (the relative volume ratio in percent) and the ISO percentage for small and large cap stocks. The graph shows an increase in the size of sweeps and the use of ISOs over time.

We separate our sample of sweeps into buyer- and seller-initiated sweeps, so that the trading we analyze represents rapid drawdowns on either the bid or ask side of the limit order book (and not just fast random trades on both sides of the market).The bottom part of Panel A in Table 6 shows the distribution of the time difference in seconds between successive individual sweeps. At the 25th percentile, the inter-sweep difference in time, across both buys and sells, is less than 1/20th of a second. The median time difference between successive sweeps in small cap stocks is 22 seconds, falling monotonically across size quintiles to only 0.8 seconds in large cap stocks. It is unlikely that sweeps on one side of the market, which occur so closely to one another, are independent. Typical trading algorithms generate waves of child orders that are conditioned on prior executions and desired volume (among other parameters). Therefore, the second step of the procedure we use aggregates closely timed sequential sweeps. To do so, we first calculate the expected time between trades as the median time between trades for each stock-half-hour in the prior month. We then cumulate consecutive buy or sell sweeps together if the time between adjacent sweeps is less than its expected value.

This process summarizes 746 million individual sweeps into 434 million aggregated buyer- and seller-initiated sweeps. By construction, aggregated sweeps are larger and inter-sweep time differences are greater. Since our interest is in resiliency, we further restrict the sample to those sweeps that extract large amounts of liquidity. We use 10,000 shares as a breakpoint to establish large drawdowns of liquidity for two reasons. First, it corresponds to the cutoff for block trades in the upstairs market in the era of pre-electronic trading. Second, it allows for comparisons with the extensive literature on the liquidity and price discovery effects of such block trades.[10] Our final sample consists of about 4.4 million aggregated sweeps each for buyer- and seller-initiated trades.

Panel B shows the distribution of inter-sweep time differences for buyer and seller-initiated aggregated sweeps. If a stock has only one sweep in a day, we cannot calculate an inter-sweep time difference. The first column, labeled "% Unique" show the percentage of such sweeps across size quintiles. In microcaps, over 30 percent of all sweeps are the only sweep in that day. That percentage declines monotonically so that in size quintile 5, less than 4 percent of sweeps are unique to the day. For the remainder of the sample, inter-sweep time differences are large. There is considerable skewness in both distributions, but the average time between sweeps ranges from 2,456 seconds for size quintile 1 to 454 seconds for size quintile 5.

Our interest is in resiliency, the extent to which the market is able to absorb such liquidity shocks without experiencing significant changes in the price formation process. To that end, we estimate variance ratios before and after these large liquidity sweeps. This poses implementation challenges. We need an appropriate time horizon over which to measure returns. To preserve independence, the overall pre- and post- measurement interval needs to be short enough so that there is minimal overlap in successive sweeps in the same stock. Given these considerations, and the distribution of the time differences between consecutive sweeps in Table 6, we define pre- and post-sweep periods as 300 seconds. For measuring variance ratios, this allows us to sample quote midpoints at 1 second and 15 second intervals (q=15). We impose the additional constraint that there be at least 15 non-zero returns in each 300 second interval in order to calculate variance ratios.

Panel A of Table 7 shows average values of $|VR_t-1|$ before and after large liquidity sweeps. Average pre- and post-sweep estimates of variance ratios are quite different from one, and from the unconditional averages in Table 2. The relatively high variance over short horizons

[10] We also identify large sweeps using a stock specific approach. Specifically, we define the relative size of each sweep as the number of shares traded in the collapsed sweep, scaled by average trade size (in shares) in the same stock in the prior month multiplied by the number of trades in the collapsed sweep. We then consider large sweeps as those above the 95th percentile in the distribution of this relative volume ratio. Such a definition includes many liquidity sweeps that are smaller than 10,000 shares. The results from this exercise are available upon request.

that these variance ratios imply is consistent with the "liquidity-based volatility" that Brunnermeier and Pedersen (2010) describe. More importantly, the average post-sweep estimates of $|VR_t\text{-}1|$ are remarkably similar to variance ratios calculated prior to the aggregated sweeps. As a consequence, the values of $\Delta|VR_t\text{-}1|$ are very close to zero. Note that paired *t*-statistics reject the null that $\Delta|VR_t\text{-}1|$ is different from zero; the significant *t*-statistics are not surprising, since the sample sizes are enormous. However, the point estimates consistently show that variance ratios improve slightly after large drawdowns in liquidity.

Panel B shows share-weighted average effective spreads in the 300 seconds before and after these large sweeps. We also report the total effective spread paid by all trades within the sweep itself. Effective spreads before sweeps decline with firm size from roughly 20 basis points to 3 basis points. The average total effective spread, paid by those executing the sweep orders, ranges from around 120 basis points (in size quintile 1) to 17 basis points (for quintile 5). The latter is roughly similar in magnitude to the total price of block trades in Dow Jones stocks reported by Madhavan and Cheng (1997). Post-sweep effective spreads are quite similar to those prior to the liquidity drawdown. Once again, paired *t*-statistics reject the null of equality (albeit with large sample sizes) and, in all cases, the direction of the difference is one in which effective spreads are lower after the liquidity event. It appears that investors are able to extract large amounts of liquidity from the market, with markets replenishing that liquidity quickly and the price discovery process experiencing no significant ill effects.

4. Alternative Tests

The above tests suggest that, in the cross-section, increased quotation activity is associated with variance ratios closer to one and with lower costs of trading. Of course, price efficiency, quotation activity and trading are all endogenous. Although we measure quotation activity (in the half hour) prior to measuring variance ratios, this endogeneity may affect inferences. In this section, we report results from tests that allow for sharper conclusions.

4.1 Exploiting Time Series Variation in Updates

Nagel (2012) reports that the returns to supplying liquidity are strongly related to measures of fundamental volatility, such as VIX. If quote updates are the tool that market makers use to manage their intraday risk exposure (and hence affect the returns that they earn), then updates should also be related to the VIX. We sum the number of updates for each firm-day and then average across firms in a size quintile. Figure 5 plots the daily time series, along with the VIX over the sample period. The graph shows two things. First, the time series variation in updates is large. Although the difference in number of updates across large-cap and small-cap firms causes the variation in updates for smaller size quintiles to be dwarfed when presented in a single figure, separate plots for each size quintile (not shown to conserve space) show similar time series variation in all size quintiles and there is little evidence of a secular trend. Second, the correlation of variation in updates with variation in the VIX is clear. For example, the high volatility period in August 2011 is accompanied by large increases in updates. The local peak in the average number of updates across all firms and VIX on August 8 coincides with a 6.7 percent drop in the S&P 500. This suggests that aggregate uncertainty has a role to play in the intensity of changes in the supply curve.

We investigate these relations more formally using a simple reduced form bivariate vector autoregression of the following form.

$$\Delta Upd_t = \sum_{i=1}^{4} \beta_i \Delta Upd_{t-i} + \sum_{i=1}^{4} \gamma_i \Delta VIX_{t-i} + \varepsilon_{1,t}$$

$$\Delta VIX_t = \sum_{i=1}^{4} \beta\theta_i \Delta Upd_{t-i} + \sum_{i=1}^{4} \varphi_i \Delta VIX_{t-i} + \varepsilon_{2,t}$$

where all changes are in percentages. The system is estimated on daily data from 2009-2011 separately for each size quintile. Simple specification checks (not reported) show that four lags are adequate to capture the dependence structure.

Table 5 reports coefficients with Z-statistics in parentheses. The regressions indicate that changes in updates are positively related to lagged innovations in VIX, suggesting that liquidity suppliers react to changes in the risk environment. Beyond a one-day lag, the influence of VIX wanes quite quickly. Interestingly, the VIX equation shows no connection between changes in the VIX and quote update activity at any lag. This is inconsistent with the belief that high frequency quoting/trading either generates or exacerbates measures of market volatility.

The results of the VAR suggest another test that may help control for endogeneity. Recall that despite the fact that we measure variance ratios (in the half hour) after calculating quote updates, improvements in variance ratios may attract quote updates rather than be caused by them; alternatively, another variable, such as the management of inventory risk, could be driving both variance ratios and quote updates on day t. As another control for this, we perform a robustness check and employ quote updates from the prior day in assigning securities to update groups – it seems unlikely that high-frequency variance ratios in day *t* affect, or are driven by the variables that cause, quotation activity in day *t-1*.[11] Since the results in Table 5 indicate that changes in updates are *negatively* related to their lagged values, this allows for still larger separation between the two. We replicate the tests in Tables 2 and 3 using quote update groups assigned from the prior day.[12] The results are not reported in tables to conserve space but they lead to the same conclusion – higher levels of quote updates are associated with variance ratios closer to one.

4.2 An Out-of-Sample Exogenous Shock: The Introduction of Arrowhead on the TSE

4.2.1 Changes to Trading Protocols

On January 4, 2010, the Tokyo Stock Exchange changed its trading infrastructure (hardware, operating system and software) in a way that facilitates high frequency quoting and trading. Under the prior system, the time between order submission and posting on the book and/or execution ranged between 1 to 2 seconds. In the new system, referred to as Arrowhead, latency dropped to roughly 2 milliseconds. This infrastructure change was accompanied by several other changes, some of which have a bearing on the design of our tests. First, the TSE permitted co-location services so that trading firms were permitted to install servers on the TSE

[11] Inventory positions may affect both quotation activity and variance ratios. If that were the case, however, using the previous day's updates to rank stocks should mitigate this effect – HFT's by definition attempt to end a day's trading with a zero inventory position. Related, evidence presented in Hendershott and Menkveld (2013) suggests that the average inventory half-life for NYSE securities, measured in a sample period *prior* to the widespread use of high-frequency trading, is relatively short (0.92 days).

[12] We check composition of low and high update groups using both the prior half hour and the prior day. On average, in about 25 percent of data, securities fall into different groups based on the different grouping procedures.

Primary Site for Arrowhead. Second, time stamps for data reported to traders and algorithms changed from minutes to milliseconds. Third, the TSE changed its tick size grid. Pre-Arrowhead, stocks were bracketed into 13 price buckets with separate tick sizes. Post-Arrowhead, both the breakpoints and the minimum price variation were changed. For example, pre-Arrowhead, stocks between ¥30M and ¥50M had a tick size of ¥100,000. After Arrowhead, this tick size was reduced to ¥50,000. Importantly, tick size reductions did not take place in all stocks/price grids. Fourth, the TSE instituted a new rule, termed the "sequential trade quote" in which a single order that moves prices beyond a certain price band (referred to as the special quote renewal price interval) triggers a quote/trade condition for one minute. This condition is designed to inform market participants and attract contra-side orders.[13] Fifth, the TSE implemented changes to its procedures designed to slow down trading when there is an order imbalance. Pre-Arrowhead, the TSE employed price limits to trigger special price quote dissemination that would (presumably) attract contra order flow. At the introduction of Arrowhead, these price limits were raised (allowing prices to move more freely), and the allowable range of the next price (the "renewal price interval") was also raised. These changes were not across the board but were based on the prior stock price.

4.2.2 Results

We calculate the average daily number of updates, new order submissions, trades, cancellations, modifications to orders that lose time priority, and modifications to orders that retain time priority. The time series of these cross-sectional averages are displayed in Figure 6. Recall that in the U.S., we do not observe submissions, cancellations and modifications, and can only calculate total updates. In Japan, the correlation between updates and submissions, trades, cancellations and modifications is easily observable, implying that updates are a good instrument for the activity we are interested in. In addition, the spikes in the graph correspond to economic uncertainty. For instance, there is a local spike in updates and their constituents on May 7, 2010, the calendar day after the Flash Crash in the U.S. (recall that the time difference between NY and

[13] These conditions are triggered minimally in our sample period. We reproduce our estimates after removing such quotes and find no difference in results.

Tokyo is 14 hours). We also observe a sharp increase in updates and their components on the Monday (and subsequent days) after the earthquake, tsunami and Fukushima nuclear disaster in March 2011. These easily identifiable spikes reinforce our conclusion from U.S. data that updates respond to real economic uncertainty. Finally, there is a shift in update levels before and after the introduction of Arrowhead. As a formal test, we compute the average number of updates for each security in a three month interval before and after the Arrowhead introduction. The average percentage increase in updates is 18 percent with a paired *t*-statistic of 10.75.

Our primary interest is in whether the shift to systems designed to accommodate high frequency activity is associated with changes in variance ratios and the cost of trading. To be consistent with our U.S. analysis, we continue to calculate variance ratios based on 15 second and 5 minute returns. In the pre-Arrowhead data, time-stamps are at the minute frequency. However, since the sequence of changes to the limit order book is preserved in the data series, we use linear interpolation to calculate prices changes at 15 second intervals (similar to the process used by Holden and Jacobsen (2014) for U.S. markets). Because Figure 6 suggests substantial changes in the information environment over long periods, and because we wish to focus purely on the effects of the Arrowhead introduction, we calculate variance ratios in the three month interval before and after the introduction. We do so only for securities that do not experience a change in tick size so that autocorrelations (and hence variance ratios) are not affected by changes in minimum price variation. For each security, we calculate the deviation in variance ratios from one (|VR-1|) and average over the three month pre- or post-Arrowhead period for each half-hour of trading. We then compute paired differences between the pre and post-period averages for each security. The second column of table 8 shows the cross-sectional average of these paired differences, along with their *t*-statistics. In every half-hour of the trading day, the change in the deviations of variance ratios from one is negative, implying an improvement in variance ratios. The magnitude of the changes in average variance ratios ranges from 0.01 to 0.03, similar to the differences between high update and low update groups in the U.S. (Table 3).

We also calculate share-weighted effective spreads for each stock, again restricting the sample to securities with no changes in tick size. As with variance ratios, we calculate weighted effective spreads in each half hour interval, and then average three months before and after the introduction of Arrowhead. Average paired changes in effective spreads are in the 3rd column of Table 8. In each half hour interval, effective spreads decline by about 2 basis points. An exception is the first half hour in the afternoon trading session, in which the decline is even larger (3.6 basis points). All changes are highly statistically significant. Since pre-Arrowhead effective spreads are approximately 10 basis points, a decline of between two to three basis points is economically meaningful.

The 4th and 5th columns of the table show changes in realized spreads and price impact based on midpoints 10 second after a trade.[14] The results show a substantial decline in realized spread, varying from 5 to almost 9 basis points, and a sharp *increase* in price impacts, varying from 3 to 7 basis points (see Figure 7). Recall that the relation between effective spreads, realized spreads and price impact is mechanical: if effective spreads decline by about 2 basis points and realized spreads also decline, it must be the case that price impact increases.

The increase in price impact associated with an increase in high frequency activity in this market is not consistent with the decline in price impact after autoquote in the U.S. market, observed in Hendershott, Jones and Menkveld (2011). We speculate that this difference is related to different rules on price continuity in the two markets. As described earlier, prior to Arrowhead, the TSE's trading protocols smoothed price paths when an order imbalance occurred – in effect, changes in spread midpoints were suppressed, reducing measured price impacts. Post-Arrowhead, the magnitude of the smoothing is substantially reduced, which would present as an increase in price impact. To verify this, we calculate the frequency of special price quote dissemination before and after the introduction of Arrowhead. We do so for each stock in the three month interval before and after Arrowhead and find that special quote dissemination decreases by an average of 50 percent (*t*-statistic=12.72).

Summarizing, the introduction of Arrowhead trading was designed to reduce the latency of trading on the TSE. Following its introduction, the evidence indicates that variance ratios move closer to 1. The magnitude of this change is very similar to the difference in variance

[14] As with U.S. data, we compute realized spreads and price impact using a variety of horizons but only show one horizon for brevity.

ratios observed in high update and low update groups in the U.S. In addition, effective spreads also decline sharply immediately around the shift to Arrowhead, with the magnitude of this decline roughly similar to the difference in effective spreads in high and low update groups in the U.S. Overall, evidence from this exogenous change in trading platforms in Japan confirms the evidence found in U.S. equity markets, with higher levels of high frequency activity associated with improvements in the price discovery process and lower costs of trading.

5. Conclusion

Market structure in the U.S has undergone a fundamental shift, replacing old-style market makers who had affirmative and negative obligations with intermediaries who provide liquidity endogenously, electronically, and at higher frequency. The U.S. is not unique in experiencing this shift, as markets around the world have similarly transformed themselves. The switch to high frequency quoting and high frequency trading has generated much debate with a long list of researchers, market professionals, and regulators concerned that price discovery and efficiency may be fundamentally harmed.

In this paper, we investigate whether high frequency changes in supply curves have a detrimental effect on the efficiency of prices. On average, the evidence suggests that high frequency quotation activity does not damage market quality. In fact, the presence of high frequency quotes is associated with modest improvements in the efficiency of the price discovery process and reductions in the cost of trading. Even when high frequency trading is associated with a large extraction of liquidity from the market for individual securities, the price discovery process in those securities appears to be quite resilient.

To us, the data broadly show that the electronic trading market place is liquid and, on average, serves investors well. Obviously, some caution in interpretation is warranted. First, although the evidence here suggests that high frequency activity is associated with some improvements in the market's function, these results do not imply that the market *always* functions in this way. An obvious case in point is the Flash Crash. Market practitioners also refer to "mini-crashes" in individual securities in which there are substantial increases or decreases in prices as liquidity disappears and market orders result in dramatic changes in prices.

However, while dislocations are harmful to market integrity, it is important to recognize that they have always occurred in markets (even before the age of electronic trading), just as flickering quotes have existed well before the advent of high frequency quotation (Baruch and Glosten (2013)). If liquidity provision is not mandated by law, liquidity providers can always exit without notice, exposing marketable orders to price risk.[15] From an economic perspective, one key issue is whether markets provide efficient price discovery on average and whether, in expectation, investors can hope to get fair prices. That appears to be the case in our sample. Of course, it is also important to consider whether a particular market structure increases or decreases the propensity for dislocations to occur, as well as how it affects the severity of the dislocation.

We also urge caution in drawing conclusions given potential externalities that we cannot measure. There is a tradeoff between the cost and benefit of monitoring high frequency quotation activity (Foucault, Roëll and Sandas (2003), and Foucault, Kadan and Kandel (2012)). If liquidity suppliers change supply curves in microseconds and liquidity extractors bear the cost of monitoring supply curves, this can have a negative effect on welfare. We provide no evidence on the extent to which high frequency quotation/trading affects welfare. But broadly understanding the influence of rapidly changing supply curves on price formation and the cost of trading is critical because, as Stiglitz (2014) points out, these are 'intermediate' variables that are necessary (but not sufficient) for understanding the role of high frequency quoting/trading on welfare.

[15] The Brady Report (1988) notes that for 31 stocks for which detailed trade data is available, 26% of specialists did not take counterbalancing trades, and their trading reinforced market trends on October 19, 1987. Some specialists in large capitalization securities were short those stocks on that day. On Tuesday, October 20, 1987, 39% of specialists in those 31 securities had trading which reinforced (negative) market trends.

References

Aït-Sahalia, Yacine and Mehmet Saglam, 2013, High frequency traders: Taking advantage of speed, working paper NBER.

Asparouhova, Elena, Hendrik Bessembinder and Ivalina Kalcheva, 2010, Liquidity biases in asset pricing tests, *Journal of Financial Economics*.

Bessembinder, Hendrik, Jia Hao and Michael Lemmon, 2011, Why designate market makers? Affirmative obligations and market quality, working paper, University of Utah.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan, 2011, High frequency trading and price discovery, working paper.

Brunnermeier, Markus and Lasse Pedersen, 2009, Market liquidity and funding liquidity, *Review of Financial Studies* 22, 2201-2238.

Budish, Eric, Peter Cramton and John Shim, 2013, The high-frequency trading arms race: Frequent batch auctions as a market design response, working paper, University of Chicago.

Baruch, Shmuel and Lawrence R. Glosten, 2013, Fleeting orders, working paper, Columbia University.

Colliard, Jean-Edouard and T. Foucault, 2012, Trading fees and efficiency in limit order markets, Review of Financial Studies 2012, 3389-3421.

Duffie, Darrell, 2010, Asset price dynamics with slow moving capital, *Journal of Finance* 65, 1238-1268.

Foucault, T., A. Roëll and P. Sandas, 2003, Market making with costly monitoring: An analysis of the SOES controversy. *Review of Financial Studies* 16, 345-384.

Foucault, T, Ohad Kadan and Eugene Kandel, 2012 Liquidity cycles and make/take fees in electronic markets, *Journal of Finance*, forthcoming.

Glosten, Lawrence, 1987, Components of the bid-ask spread and the statistical properties of transaction prices, *Journal of Finance* 42, 1293-1307.

Han, Jungsuk, Mariana Khapko and Albert Kyle, 2014, Liquidity with high frequency market making, working paper.

Hasbrouck, Joel, 2013, High Frequency Quoting: Short-Term Volatility in Bids and Offers, working paper, NYU.

Hendershott, Terrance and Albert Menkveld, 2013, Price pressures, *Journal of Financial Economics*, forthcoming.

Hendershott, Terrance and Ryan Riordan, 2012, Algorithmic trading and the market for liquidity, forthcoming, *Journal of Financial and Quantitative Analysis*.

Hendershott, Terrance, Charles Jones and Albert Menkveld, 2011, Does algorithmic trading improve liquidity, *Journal of Finance* 66, 1-33.

Ho, Thomas and Hans Stoll, 1981, Optimal dealer pricing under transactions and return uncertainty, *Journal of Financial Economics* 9, 319-381.

Holden, Craig and Stacey E. Jacobsen, (2014), Liquidity measurement problems in fast competitive markets: expensive and cheap solutions, *Journal of Finance*, forthcoming.

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.

Glosten, Lawrence, 1987, Components of the bid-ask spread and statistical properties of transaction prices, *Journal of Finance* 42, 1293-1307.

Kyle, Albert S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315-1336.

Lo, Andrew W., and A. Craig MacKinlay, 1988, Stock prices do not follow random walks: Evidence from a simple specification test, *Review of Financial Studies* 1, 41-66.

Lo, Andrew W., and A. Craig MacKinlay, 1989, The size and power of the variance ratio test in finite samples: A Monte Carlo investigation, *Journal of Econometrics* 40, 203-238.

Madhavan, Ananth and Minder Cheng, 1997, In search of liquidity: Block trades in the upstairs and downstairs markets, *Review of Financial Studies* 10, 175-203.

Menkveld, Albert, 2012, High frequency traders and the new market makers, working paper, VU University Amsterdam.

Nagel, Stefan, 2012, Evaporating liquidity, *Review of Financial Studies* 25, 2005-2039.

Securities and Exchange Commission, 2010, Concept Release on Equity Market Structure.

Stiglitz, Joseph, 2014, Tapping the brakes: Are less active markets safer and better for the economy?, working paper, Columbia University.

Table 1

Average Number of Trades and Quote Updates in Size Quintiles and Selected ETFs

The sample consists of all common stocks (not including ETFs), with a stock price greater than $1 and a market capitalization greater than $100m at the beginning of the month. The sample period is 2009-2011, excluding May 6, 2010. Each firm is placed in a size quintile at the beginning of the month using NYSE breakpoints taken from Ken French's website. A quote update is defined as any change in the prevailing best bid or offer price (BBO), or any change in the displayed size (depth) for the best bid or offer, across all exchanges. For each firm and half hour interval, we calculate the total number of trades and the total number of quote price or quote size changes in that interval. These are averaged across firms in a quintile and then averaged over the time series. Statistics are reported on a per second basis. Separate statistics are shown for two ETFs, SPY (the S&P 500 "spider") and IWM (the iShares Russell 2000 Index ETF).

Size	9:30 10:00	10:00 10:30	10:30 11:00	11:00 11:30	11:30 12:00	12:00 12:30	12:30 1:00	1:00 1:30	1:30 2:00	2:00 2:30	2:30 3:00	3:00 3:30	3:30 4:00
Panel A: Average number of trades per second													
Small	0.05	0.05	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.05	0.10
2	0.15	0.14	0.12	0.11	0.10	0.09	0.08	0.08	0.01	0.10	0.11	0.14	0.28
3	0.33	0.32	0.27	0.24	0.21	0.18	0.17	0.18	0.19	0.22	0.24	0.31	0.60
4	0.65	0.62	0.50	0.44	0.38	0.34	0.32	0.32	0.34	0.40	0.44	0.56	1.06
Large	2.05	1.77	1.41	1.22	1.05	0.92	0.85	0.86	0.90	1.05	1.14	1.42	2.62
SPY	39.28	32.54	24.05	20.43	17.52	14.79	13.36	13.99	14.62	17.92	18.88	22.35	39.01
IWM	13.45	10.88	7.67	6.42	5.43	4.62	4.28	4.40	4.61	5.67	6.07	7.24	13.12
Panel B: Average number of quote updates per second													
Small	1.42	0.98	0.79	0.69	0.62	0.56	0.52	0.52	0.52	0.59	0.61	0.70	1.09
2	3.57	2.85	2.33	2.04	1.79	1.60	1.48	1.49	1.55	1.78	1.82	2.13	3.27
3	6.49	5.82	4.82	4.21	3.64	3.23	3.00	3.04	3.17	3.72	3.81	4.51	6.85
4	10.81	10.79	9.08	7.92	6.70	5.84	5.42	5.49	5.72	6.70	6.99	8.28	12.42
Large	25.52	26.02	21.80	18.80	15.93	13.76	12.69	12.87	13.27	15.45	16.18	19.00	27.55
SPY	376.40	315.86	249.11	211.02	181.34	152.05	135.89	136.48	138.14	158.88	165.07	188.99	288.95
IWM	252.90	219.86	172.82	141.70	120.81	101.24	91.12	91.42	93.30	107.87	113.11	127.48	186.53

Table 2

Average Variance Ratios for Size Quintiles and Selected ETFs

The sample consists of all common stocks, with a stock price greater than $1 and a market capitalization greater than $100m at the beginning of the month from 2009-2011, excluding May 6, 2010. A quote update is defined as any change in the prevailing best bid or offer price (BBO), or any change in the displayed size (depth) for the best bid or offer, across all exchanges. For each firm and half hour interval, we use the median number of quote updates to separate firms into low and high update groups (within each size quintile). For each firm and subsequent half hour interval, we calculate variance ratios based on 15 second and 5 minute quote (NBBO) midpoints. The table shows time series averages of these group variance ratios. The time series average of the cross-sectional standard deviation in parentheses. Separate statistics are shown for two ETFs, SPY and IWM.

Quint./ETF	Upd. Group	10:00-10:30	10:30-11:00	11:00-11:30	11:30-12:00	12:00-12:30	12:30-1:00	1:00-1:30	1:30-2:00	2:00-2:30	2:30-3:00	3:00-3:30	3:30-4:00
Small	Low	0.81	0.78	0.77	0.76	0.74	0.73	0.74	0.74	0.74	0.75	0.76	0.79
		(0.61)	(0.60)	(0.60)	(0.59)	(0.58)	(0.57)	(0.58)	(0.57)	(0.56)	(0.57)	(0.57)	(0.57)
	High	0.87	0.87	0.86	0.84	0.82	0.80	0.80	0.80	0.80	0.81	0.82	0.84
		(0.60)	(0.60)	(0.59)	(0.59)	(0.58)	(0.57)	(0.56)	(0.56)	(0.55)	(0.55)	(0.55)	(0.56)
2	Low	0.99	0.98	0.97	0.97	0.94	0.94	0.94	0.94	0.91	0.94	0.93	0.91
		(0.63)	(0.64)	(0.63)	(0.63)	(0.61)	(0.61)	(0.60)	(0.60)	(0.58)	(0.59)	(0.58)	(0.58)
	High	0.99	1.00	0.98	0.99	0.95	0.94	0.94	0.93	0.90	0.92	0.90	0.88
		(0.60)	(0.61)	(0.61)	(0.62)	(0.60)	(0.59)	(0.58)	(0.58)	(0.54)	(0.55)	(0.54)	(0.54)
3	Low	1.06	1.06	1.06	1.06	1.04	1.02	1.02	1.02	0.97	0.98	0.96	0.92
		(0.65)	(0.66)	(0.66)	(0.66)	(0.65)	(0.64)	(0.63)	(0.63)	(0.58)	(0.59)	(0.57)	(0.56)
	High	1.04	1.05	1.06	1.08	1.04	1.02	1.01	1.01	0.95	0.95	0.92	0.91
		(0.62)	(0.63)	(0.63)	(0.65)	(0.63)	(0.61)	(0.60)	(0.60)	(0.55)	(0.55)	(0.53)	(0.53)
4	Low	1.09	1.10	1.10	1.11	1.08	1.06	1.05	1.04	1.00	0.99	0.97	0.95
		(0.66)	(0.68)	(0.68)	(0.69)	(0.67)	(0.65)	(0.65)	(0.63)	(0.59)	(0.59)	(0.57)	(0.56)
	High	1.05	1.05	1.08	1.10	1.07	1.05	1.04	1.03	0.97	0.98	0.95	0.96
		(0.62)	(0.62)	(0.64)	(0.65)	(0.64)	(0.62)	(0.60)	(0.60)	(0.55)	(0.55)	(0.53)	(0.54)
Large	Low	1.09	1.09	1.11	1.12	1.09	1.07	1.06	1.05	1.00	0.99	0.99	1.01
		(0.65)	(0.65)	(0.67)	(0.68)	(0.65)	(0.64)	(0.62)	(0.62)	(0.57)	(0.57)	(0.55)	(0.56)
	High	1.04	1.03	1.06	1.08	1.06	1.05	1.04	1.03	0.98	0.99	0.99	1.02
		(0.58)	(0.59)	(0.60)	(0.62)	(0.60)	(0.59)	(0.58)	(0.57)	(0.53)	(0.53)	(0.52)	(0.53)
SPY		0.97	0.94	0.99	0.99	0.96	0.97	1.00	0.99	0.96	0.97	1.00	1.03
IWM		1.00	0.99	1.00	1.03	0.97	0.98	1.01	1.00	096	0.96	0.98	0.97

Table 3

Average Differences in Deviations of Variance Ratios from One and Effective Half Spreads between Low and High Update Groups for Size and Trade Quintiles

Stocks are first sorted into size quintile based on prior month NYSE breakpoints, and within size groups into quintiles based on the number of trades in each half hour. Within these groups, stocks are further sorted into low and high update groups, using the median number of updates. We calculate the deviation in variance ratios from one (|VR-1|) for all securities with a group. Similarly, we calculate effective half spreads for all trades in each stock as the scaled difference between the transaction price and prevailing quote mid-point. For each half hour in the day, we then calculate share-weighted average effective half spreads for each stock. We calculate the cross-sectional average effective spread across all stocks in each of these 50 groups. For both variance ratios and effective spreads, we calculate the difference in the cross-sectional average of the low and high update groups (high minus low). The table shows the time series average of these cross-sectional differences. The variance ratios in Panel A are based on sampling midpoints at 15 second and 5 minute intervals. The variance ratios in Panel B are based on sampling midpoints at 100 milliseconds and 1 second, as well as 100 milliseconds and 2 sections. These are for stocks in size quintile 5 only. Differences in effective spreads are in basis points. The sample period is 2009-2011. Standard errors appear in parentheses and are based on the time-series of the cross-sectional averages.

	Quintiles Formed on Number of Trades in Prior Half Hour				
	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Panel A: Differences in Average Variance Ratios (15 seconds, 5 minutes)					
Small	-0.033	-0.028	-0.021	-0.017	-0.018
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
2	-0.018	-0.012	-0.019	-0.020	-0.026
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
3	-0.017	-0.020	-0.022	-0.030	-0.035
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
4	-0.022	-0.028	-0.033	-0.037	-0.034
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Large	-0.027	-0.031	-0.032	-0.031	-0.034
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Panel B: Differences in Average Variance Ratios (Size Quintile 5 only)					
100ms, 1sec.	0.0005	-0.0066	-0.0036	-0.0051	0.0064
	(0.0006)	(0.0005)	(0.0006)	(0.0006)	(0.0006)
100ms, 2sec.	0.0019	-0.0117	-0.0081	-0.0100	0.0007
	(0.0010)	(0.0007)	(0.0007)	(0.0007)	(0.0007)
Panel C: Differences in Effective Spreads					
Small	-6.10	-0.70	-3.68	-4.23	-1.58
	(5.81)	(0.25)	(0.19)	(0.07)	(0.08)
2	-1.27	-0.96	-0.77	-0.66	-0.62
	(0.18)	(0.04)	(0.04)	(0.05)	(0.07)
3	-2.00	-0.12	0.06	-0.31	-0.58
	(0.05)	(0.04)	(0.04)	(0.05)	(0.06)
4	-1.27	0.15	-0.22	-0.91	-1.20
	(0.05)	(0.04)	(0.04)	(0.05)	(0.06)
Large	-0.61	-0.28	-0.46	-1.13	-1.12
	(0.04)	(0.03)	(0.04)	(0.04)	(0.05)

Table 4

Average Differences in Effective Half-Spreads, Realized Spreads and Price Impact between Low and High Update Groups for Size and Trade Quintiles

Effective spreads are calculated as the scaled difference between the transaction price and prevailing quote mid-point. Realized spreads are computed as the price movement from transaction prices to a future quote midpoint, scaled by the midpoint prevailing at the time of the transaction. We use quote midpoints 1, 5, 10 and 20 seconds after the trade. Price impact is the difference between the effective spread of the transaction and its realized spread. For each stock, we calculate share-weighted average effective spreads, realized spreads and price impact. We compute cross-sectional averages for stocks in each size quintile, trade quintile and update group. We then calculate the differences between these cross-sectional averages between low and high update groups (high minus low). The table shows the time series averages of these differences for each size and trade quintile. All estimates are in basis points. The sample consists of all stocks in these groups for two randomly selected trading days in each month for 2010-2011.

Size Quintile	Trade Quintile	Eff. Spreads	Realized Spreads				Price Impact			
			t+1	t+5	t+10	t+20	t+1	t+5	t+10	t+20
Small	1	-36.7	-14.0	-14.6	-14.2	-13.8	-22.7	-22.1	-22.5	-22.9
	2	-11.5	-4.9	-5.5	-5.4	-5.3	-6.6	-6.0	-6.1	-6.2
	3	-7.2	-5.3	-5.4	-5.2	-5.0	-1.9	-1.8	-2.0	-2.2
	4	-7.5	-5.5	-5.3	-5.1	-4.8	-2.0	-2.2	-2.4	-2.7
	5	-4.7	-3.7	-3.7	-3.6	-3.4	-1.0	-1.0	-1.1	-1.3
2	1	-4.3	-2.4	-2.7	-2.6	-2.5	-1.9	-1.6	-1.7	-1.8
	2	-1.0	-0.9	-0.9	-0.9	-0.9	-0.1	-0.1	-0.1	-0.1
	3	-1.1	-0.8	-0.8	-0.8	-0.8	-0.3	-0.3	-0.3	-0.3
	4	-1.0	-0.9	-0.9	-0.9	-0.8	-0.1	-0.1	-0.1	-0.2
	5	-0.7	-0.5	-0.8	-0.7	-0.7	-0.2	0.1	0.0	0.0
3	1	-2.7	-1.6	-1.6	-1.6	-1.5	-1.1	-1.1	-1.1	-1.2
	2	-0.4	-0.5	-0.5	-0.5	-0.5	0.1	0.1	0.1	0.1
	3	-0.4	-0.4	-0.4	-0.4	-0.4	0.0	0.0	0.0	0.0
	4	-0.3	-0.3	-0.4	-0.4	-0.4	0.0	0.1	0.1	0.1
	5	-0.4	-0.4	-0.6	-0.5	-0.5	0.0	0.2	0.1	0.1
4	1	-0.7	-0.6	-0.5	-0.5	-0.4	-0.1	-0.2	-0.2	-0.3
	2	-0.1	0.0	-0.1	-0.1	-0.1	-0.1	0.0	0.0	0.0
	3	0.00	0.0	-0.1	-0.1	-0.1	0.0	0.1	0.1	0.1
	4	-0.1	-0.1	-0.2	-0.2	-0.2	0.0	0.1	0.1	0.1
	5	-0.3	-0.3	-0.5	-0.4	-0.4	0.0	0.2	0.1	0.1
Large	1	-0.1	-0.1	-0.1	-0.1	-0.1	0.0	0.0	0.0	0.0
	2	0.1	0.1	-0.1	-0.1	-0.1	0.0	0.2	0.2	0.2
	3	-0.1	0.0	-0.1	-0.1	-0.1	-0.1	0.0	0.0	0.0
	4	-0.2	-0.2	-0.2	-0.2	-0.1	0.0	0.0	0.0	-0.1
	5	-0.2	-0.1	-0.1	0.0	0.0	-0.1	-0.1	-0.2	-0.2

Table 5

Bivariate Vector Autoregressions of Daily Changes in Quote Updates and VIX

Each day, we sum the number of updates (Upd_t) for each firm and then average across firms in size quintiles. Using daily data, we estimate bivariate reduced form vector autoregressions of the form

$$\Delta Upd_t = \sum_{i=1}^{4} \beta_i \Delta Upd_{t-i} + \sum_{i=1}^{4} \gamma_i \Delta VIX_{t-i} + \varepsilon_{1,t}$$
$$\Delta VIX_t = \sum_{i=1}^{4} \beta\theta_i \Delta Upd_{t-i} + \sum_{i=1}^{4} \varphi_i \Delta VIX_{t-i} + \varepsilon_{2,t}$$

where all changes are in percentages. The equations are estimated separately for each size quintile using daily data from 2009-2011. Z-statistics appear in parentheses.

	Small		Size Quintile 2		Size Quintile 3		Size Quintile 4		Large	
	ΔUpd_t	ΔVIX_t	ΔUpd_t	ΔVIX_t	ΔUpd_t	ΔVIX_t	ΔUpd_t	ΔVIX_t	ΔUpd_t	ΔVIX_t
ΔUpd_{t-1}	-0.500	0.001	-0.493	0.008	-0.488	0.010	-0.487	0.006	-0.474	0.008
	(12.67)	(0.07)	(12.19)	(0.48)	(11.93)	(0.61)	(11.81)	(0.42)	(11.43)	(0.54)
ΔUpd_{t-2}	-0.264	0.017	-0.306	0.012	-0.302	0.016	-0.313	0.010	-0.312	0.009
	(6.11)	(0.86)	(6.94)	(0.63)	(6.81)	(0.90)	(6.96)	(0.58)	(6.93)	(0.56)
ΔUpd_{t-3}	-0.208	0.015	-0.250	-0.011	-0.258	0.001	-0.239	-0.004	-0.235	0.001
	(4.80)	(0.74)	(5.66)	(0.06)	(5.81)	(0.09)	(5.32)	(0.25)	(5.21)	(0.08)
ΔUpd_{t-4}	-0.026	0.031	-0.073	0.023	0.075	0.031	-0.071	0.024	-0.079	0.028
	(0.67)	(1.65)	(1.82)	(1.33)	(1.85)	(1.82)	(1.74)	(1.50)	(1.90)	(1.74)
ΔVIX_{t-1}	0.427	-0.102	0.503	-0.108	0.531	-0.110	0.581	-0.106	0.571	-0.110
	(5.18)	(2.59)	(5.40)	(2.67)	(5.45)	(2.69)	(5.59)	(2.59)	(5.49)	(2.65)
ΔVIX_{t-2}	0.107	-0.075	0.166	-0.073	0.209	-0.080	0.239	-0.073	0.212	-0.072
	(1.27)	(1.86)	(1.73)	(1.76)	(2.09)	(1.91)	(2.23)	(1.72)	(1.98)	(1.71)
ΔVIX_{t-3}	0.029	-0.111	0.144	-0.098	0.201	-0.101	0.196	-0.094	0.183	-0.101
	(0.35)	(2.78)	(1.52)	(2.37)	(2.02)	(2.43)	(1.84)	(2.23)	(1.72)	(2.38)
ΔVIX_{t-4}	-0.028	-0.054	0.027	-0.051	0.066	-0.061	0.065	-0.056	0.078	-0.062
	(0.34)	(1.37)	(0.29)	(1.26)	(0.67)	(1.50)	(0.63)	(1.37)	(0.75)	(1.51)

Table 6

Liquidity Sweeps: Size and Inter-Sweep Time Differences

Individual liquidity sweeps are multiple trades from more than one reporting venue with the same millisecond time stamp. The top part of Panel A shows three statistics. We sum the number of shares traded in all liquidity sweeps in a half hour and divide by the total number of shares traded in the same security. The resulting ratio represents the volume of trading in individual liquidity sweeps scaled by market volume. We similarly sum the number of ISO trades employed in sweeps and scale by the number of market trades (in the second column) or market ISOs (in the last column). The second part of Panel A shows the distribution of the difference in time between successive sweeps. In Panel B, individual liquidity sweeps are cumulated if the time difference between successive individual sweeps in the same stock is less than the median inter-trade time in the same half hour in the prior month. This aggregating procedure is employed separately for buyer- and seller-initiated sweeps. The panel shows the distribution of difference in time between successive collapsed sweeps (separately for buyer- and seller initiated trades) in which more than 10,000 shares are traded.

Panel A: Individual Liquidity Sweeps (trades in the same millisecond)

Ratios of Sweeps to Market Statistics

	Sweep Shrs / Market Shrs	Sweep ISOs / Market Trades	Sweep ISOs / Market ISOs
Small	13.61	7.92	19.72
2	13.45	7.40	18.41
3	26.54	13.48	29.99
4	18.42	10.37	23.03
Large	22.15	12.99	28.01

Time difference in seconds between successive sweeps in the same stock-day

	10^{th} Perc.	25^{th} Perc.	Median	Mean	75^{th} Perc.	90^{th} Perc.
Small	0.01	0.05	22.92	341.29	190.24	751.83
2	0.01	0.02	10.85	133.95	80.03	288.51
3	0.01	0.02	6.85	66.49	46.68	154.04
4	0.00	0.01	3.99	34.39	26.68	84.49
Large	0.00	0.01	0.80	12.46	19.41	31.62

Panel B: Time difference in seconds between successive collapsed liquidity sweeps greater than 10,000 shares

	% Unique	10^{th} Perc.	25^{th} Perc.	Median	Mean	75^{th} Perc.	90^{th} Perc.
Collapsed Buyer-Initiated Sweeps							
Small	30.80	32.32	147.00	748.17	2456.94	2913.31	7498.26
2	18.47	18.52	87.42	432.55	1695.96	1689.12	4868.18
3	11.06	11.67	53.69	271.23	1171.53	1034.32	3074.75
4	8.68	8.21	36.49	173.30	886.61	653.04	2100.75
Large	3.64	3.62	15.20	68.17	454.26	259.13	874.68
Collapsed Seller-Initiated Sweeps							
Small	31.06	34.30	161.83	811.03	2569.59	3125.52	7854.48
2	18.71	18.77	90.37	454.65	1753.06	1774.13	5063.75
3	11.19	11.77	55.33	278.62	1183.89	1055.01	3134.42
4	8.73	8.29	37.06	175.91	892.54	663.33	2117.95
Large	3.64	3.55	15.04	67.67	453.29	258.93	1903.30

Table 7

Average Variance Ratios and Effective Half Spreads Before and After Large Liquidity Drawdowns

The sample consists of all collapsed buyer- and seller-initiated liquidity sweeps in which more than 10,000 shares are traded. We calculate variance ratios and share-weighted effective spreads 300 seconds before the beginning of the aggregated sweep and 300 seconds after the end of the sweep. We also calculate the sum of effective spreads for all trades within the sweep. Panel A reports the absolute value of the distance of the variance ratios from one, as well as the difference between the post- and pre-sweep variance ratios. Panel B reports simple averages of share-weighted pre- and post-sweep effective spreads, the intra-sweep effective spread, and the difference between the post- and pre-sweep effective spreads. All effective spreads are in basis points.

Panel A: |VR-1| Before and After Liquidity Sweeps Greater than 10,000 shares

	Pre-Sweep	Post-Sweep	Δ|VR-1|	t-statistic
Buyer-Initiated				
Small	0.3133	0.3038	-0.0095	-2.88
2	0.3046	0.2948	-0.0097	-5.33
3	0.2991	0.2919	-0.0072	-5.11
4	0.3013	0.2946	-0.0066	-6.37
Large	0.2696	0.2648	-0.0047	-14.90
Seller-Initiated				
Small	0.3114	0.3085	-0.0028	-0.85
2	0.3034	0.2930	-0.0103	-5.86
3	0.2953	0.2899	-0.0054	-3.84
4	0.2950	0.2919	-0.0030	-2.99
Large	0.2681	0.2642	-0.0039	-12.38

Panel C: Average Effective Half-Spreads Around Liquidity Sweep Greater than 10,000 shares

	Pre-Sweep	Intra- Sweep	Post- Sweep	Δ Effec. Spread	t-statistic
Buyer-Initiated					
Small	20.10	121.59	17.80	-2.29	-3.41
2	10.16	65.30	9.30	-0.80	-8.05
3	7.61	49.73	6.93	-0.68	-8.21
4	6.08	34.26	5.24	-0.83	-7.38
Large	3.04	17.18	2.68	-0.36	-9.68
Seller-Initiated					
Small	19.59	115.22	17.77	-1.85	-3.48
2	10.35	65.73	9.41	-0.93	-7.74
3	7.77	48.08	6.98	-0.71	-5.84
4	6.29	33.61	5.27	-1.02	-5.13
Large	3.08	17.37	2.70	-0.38	-14.18

Table 8

Average Paired Differences in Deviations of Variance Ratios from One, Effective Spreads, Realized Spread and Price Impact around the introduction of Arrowhead on the Tokyo Stock Exchange

The sample consist of the 300 largest stocks on the 1st Section of the Tokyo Stock Exchange, minus stocks which experience changes in tick size during the introduction of Arrowhead. We calculate variance ratios over half hour intervals based on 15-second and 5 minute returns, and compute the absolute deviation from one (|VR-1|). For each stock, we average these deviations in each half hour interval in the three months before and after the introduction of Arrowhead, and compute a paired difference. The column labeled $|VR\text{-}1|_{pre}$ - $|VR\text{-}1|_{post}$ shows average paired differences. We compute share-weighted effective, realized spreads (using quote midpoints 10 seconds after the trade), and price impact in each stock-day and half hour interval. We average these in the three months before and after the introduction of Arrowhead, and reported a paired difference (before minus after) in basis points. *T*-statistics appear in parentheses.

	$\lvert VR\text{-}1\rvert_{pre} - \lvert VR\text{-}1\rvert_{post}$	Δ Effective Spread	Δ Realized Spread	Δ Price Impact
9:00 - 9:30	-0.016	2.06	9.15	-7.09
	(1.80)	(16.67)	(54.89)	(40.49)
9:30 - 10:00	-0.037	2.00	8.16	-6.15
	(3.18)	(12.83)	(46.79)	(36.34)
10:00 - 10:30	-0.020	1.94	7.36	-5.43
	(1.97)	(12.18)	(46.75)	(38.29)
10:30 - 11:00	-0.015	1.90	6.85	-4.95
	(1.02)	(11.81)	(42.75)	(37.06)
12:30 - 1:00	-0.019	3.65	7.25	-3.60
	(1.85)	(23.48)	(48.65)	(22.33)
1:00 - 1:30	-0.025	1.93	6.72	-4.79
	(2.25)	(11.73)	(42.52)	(36.20)
1:30 - 2:00	-0.019	1.94	6.45	-4.51
	(1.13)	(11.64)	(44.13)	(34.19)
2:00 - 2:30	-0.027	1.95	6.52	4.48
	(1.97)	(11.62)	(43.97)	(37.93)
2:30 - 3:00	-0.026	1.92	5.76	3.47
	(1.54)	(11.85)	(39.23)	(37.94)

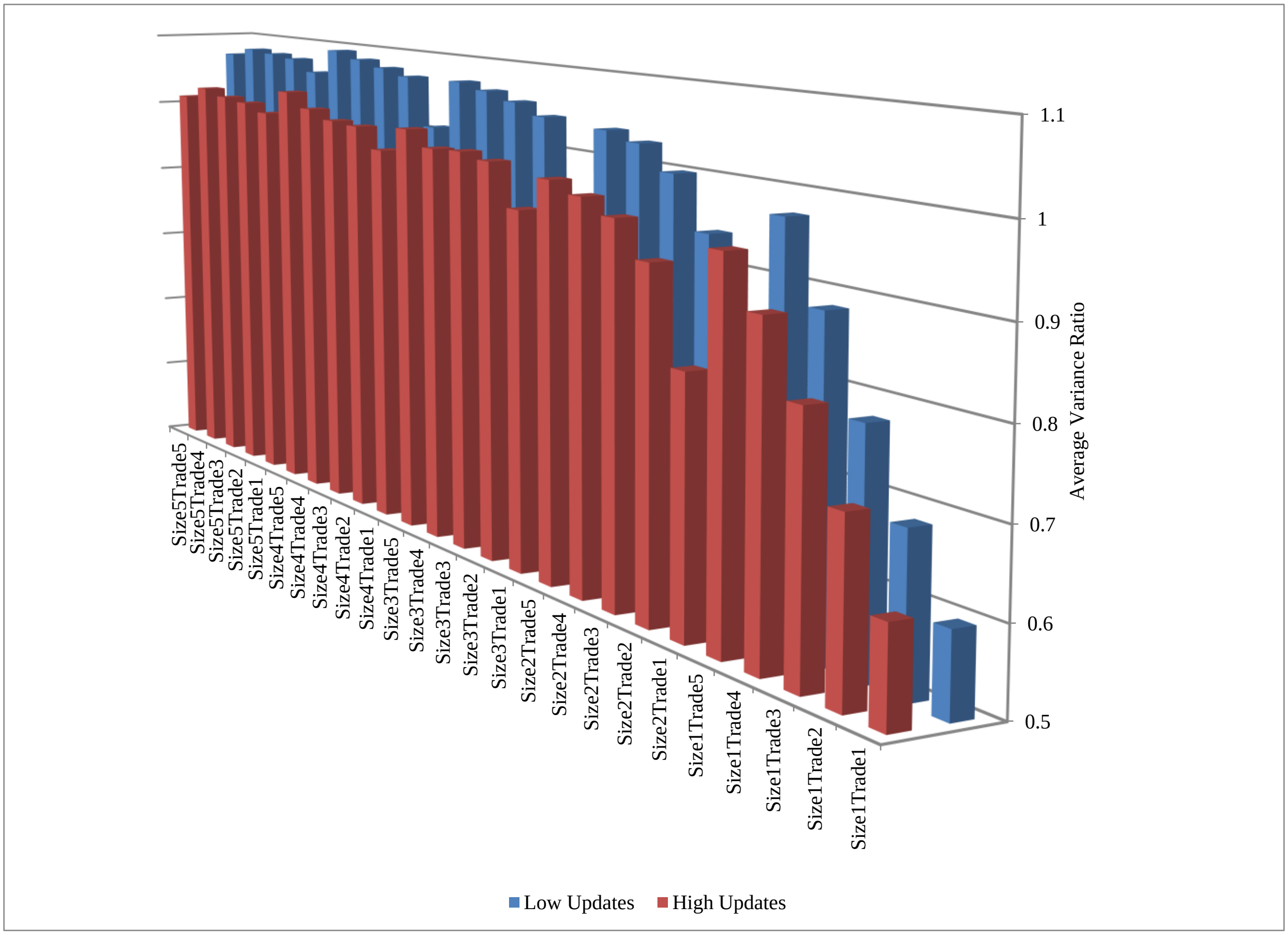

Figure 1: Stocks are sorted into size quintile based on prior month NYSE breakpoints, and within size groups into quintiles based on the number of trades in each half hour. For example, Size1Trade1 contains firms in the smallest size and trade quintile. Within these 25 (5x5) groups, stocks are further sorted into low and high update groups, using the median number of updates. We calculate average variance ratios for all stocks within a group and plot the time series average of these cross-sectional means.

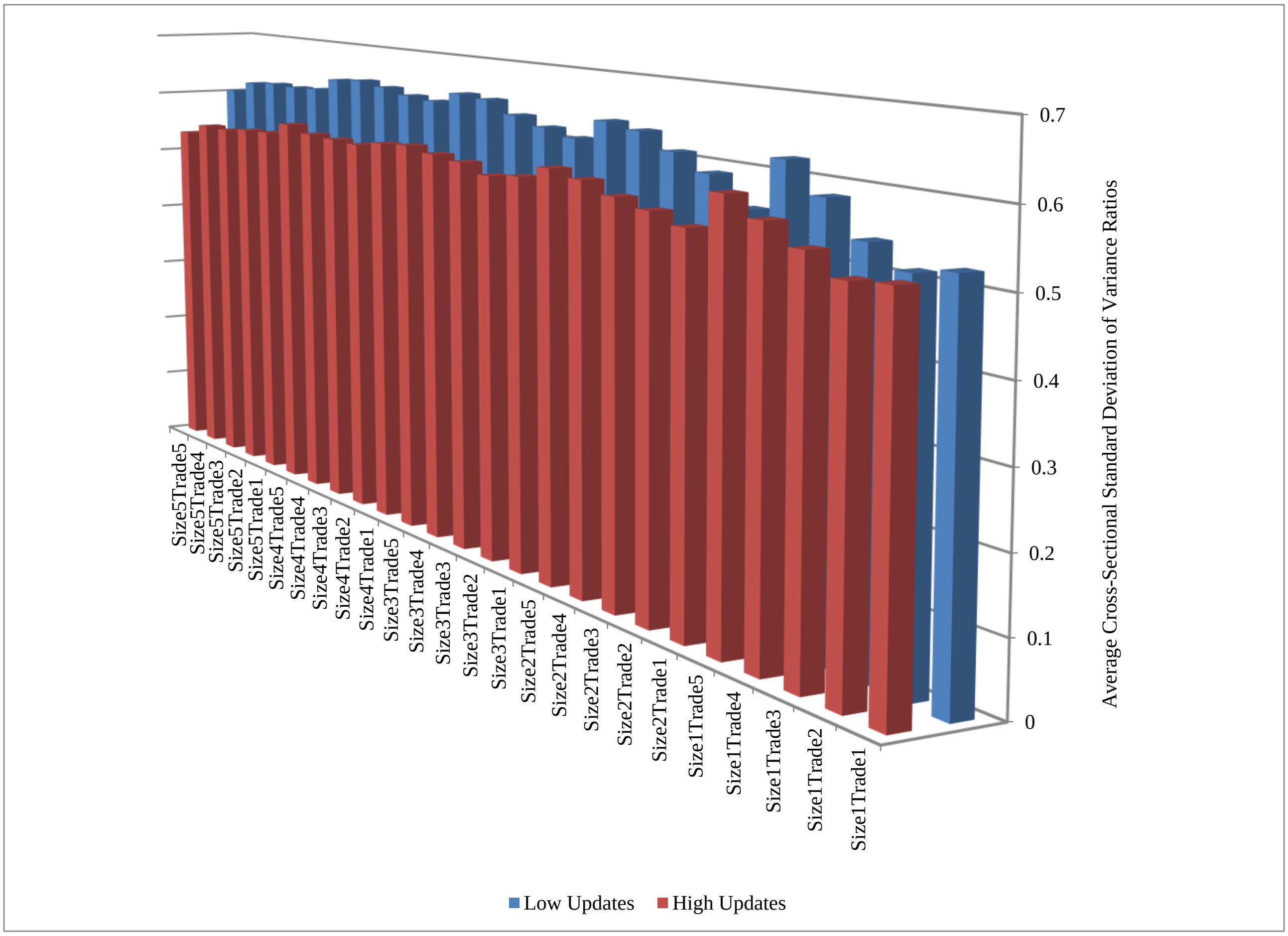

Figure 2: Stocks are sorted into size quintile based on prior month NYSE breakpoints, and within size groups into quintiles based on the number of trades in each half hour. For example, Size1Trade1 contains firms in the smallest size and trade quintile. Within these 25 (5x5) groups, stocks are further sorted into low and high update groups, using the median number of updates. We calculate the cross-sectional standard deviation of variance ratios for all stocks within a group and plot the time series average of these cross-sectional standard deviations.

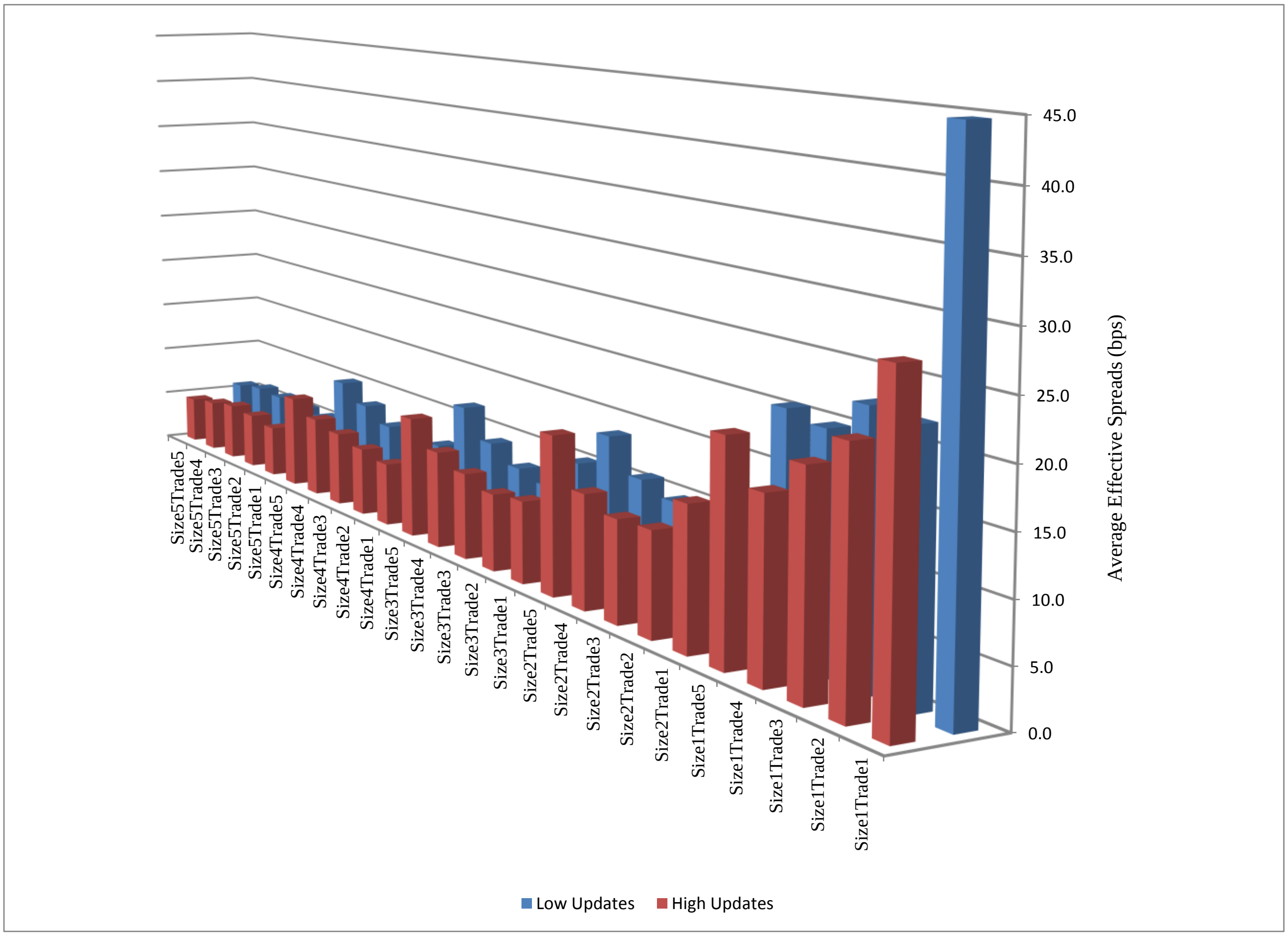

Figure 3: Stocks are sorted into size quintile based on prior month NYSE breakpoints, and within size groups into quintiles based on the number of trades in each half hour. For example, Size1Trade1 contains firms in the smallest size and trade quintile. Within these 25 (5x5) groups, stocks are further sorted into low and high update groups, using the median number of updates. We calculate share-weighted effective half-spreads for all stocks within a group and plot the time series average of these cross-sectional means.

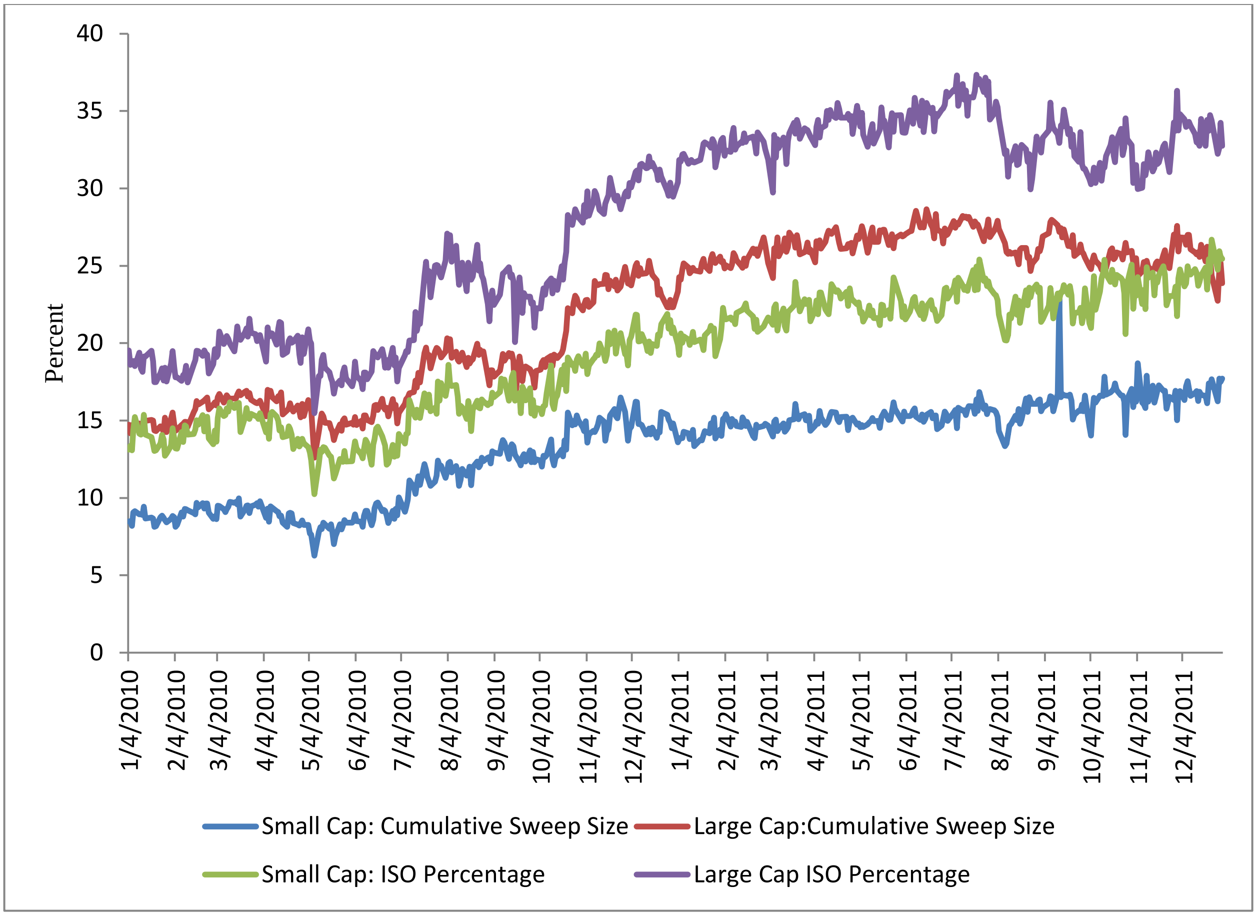

Figure 4: We identify a liquidity sweep as one in which there are multiple trades in the same stock with the same timestamp (to the millisecond) across more than one trade reporting venue. We calculate the cumulative size of sweeps in a stock and half-hour interval as number of shares traded in all sweeps scaled it by the total number of shares traded in that stock-interval. We calculate ISO percentage as the total number of ISO trades in all sweeps in a stock-interval scaled by the total number of ISO trades in that stock-interval. We average across stocks in a size quintile and intervals in a day to obtain daily averages. The figure shows the time series of sweep sizes and ISO percentages for size quintile 1 (small stocks) and size quintile 5 (large stocks).

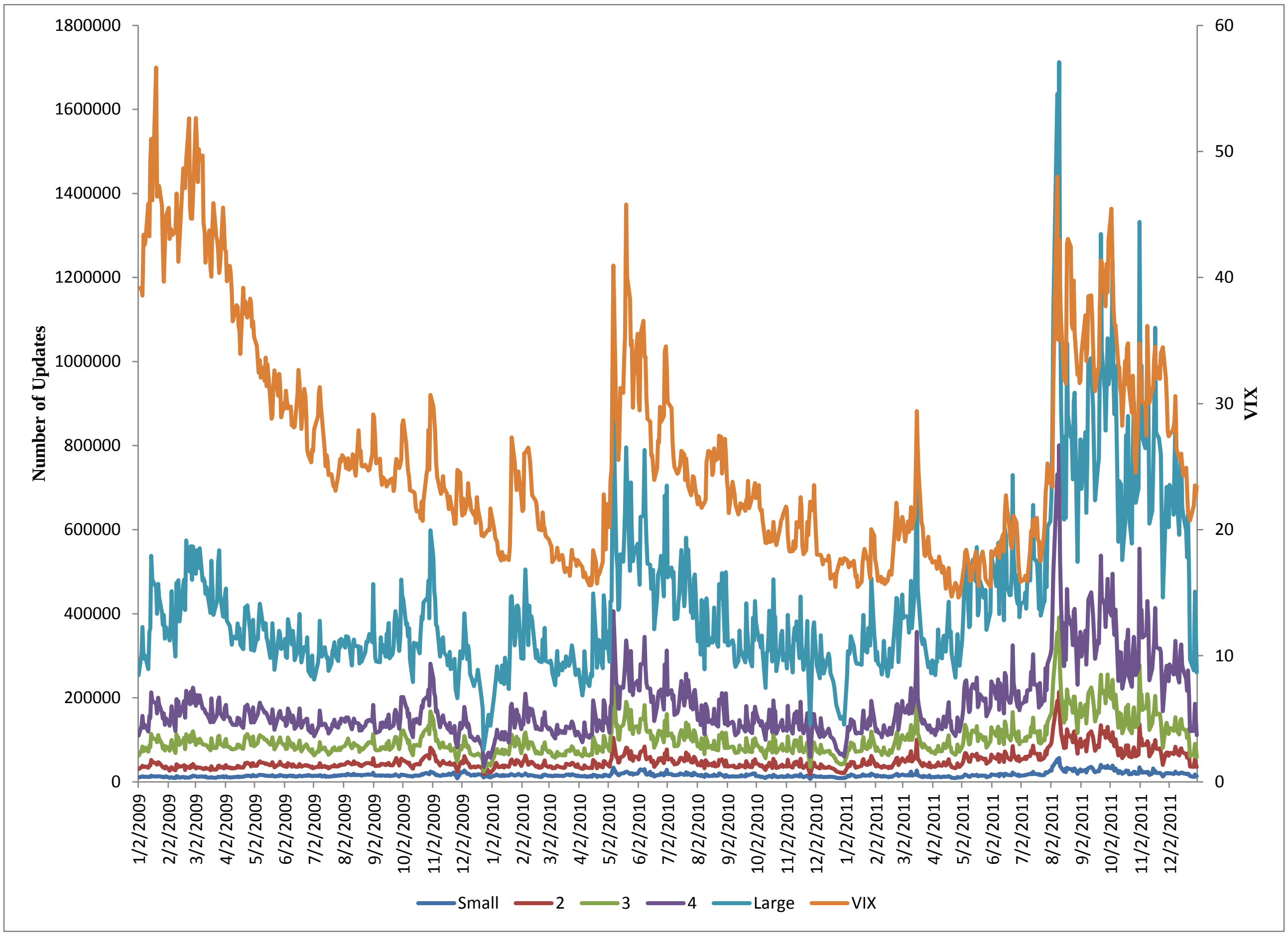

Figure 5: Each day, we sum the number of updates for each firm and then average across firms in size quintiles. Daily closing values of the VIX index are plotted using the right vertical axis.

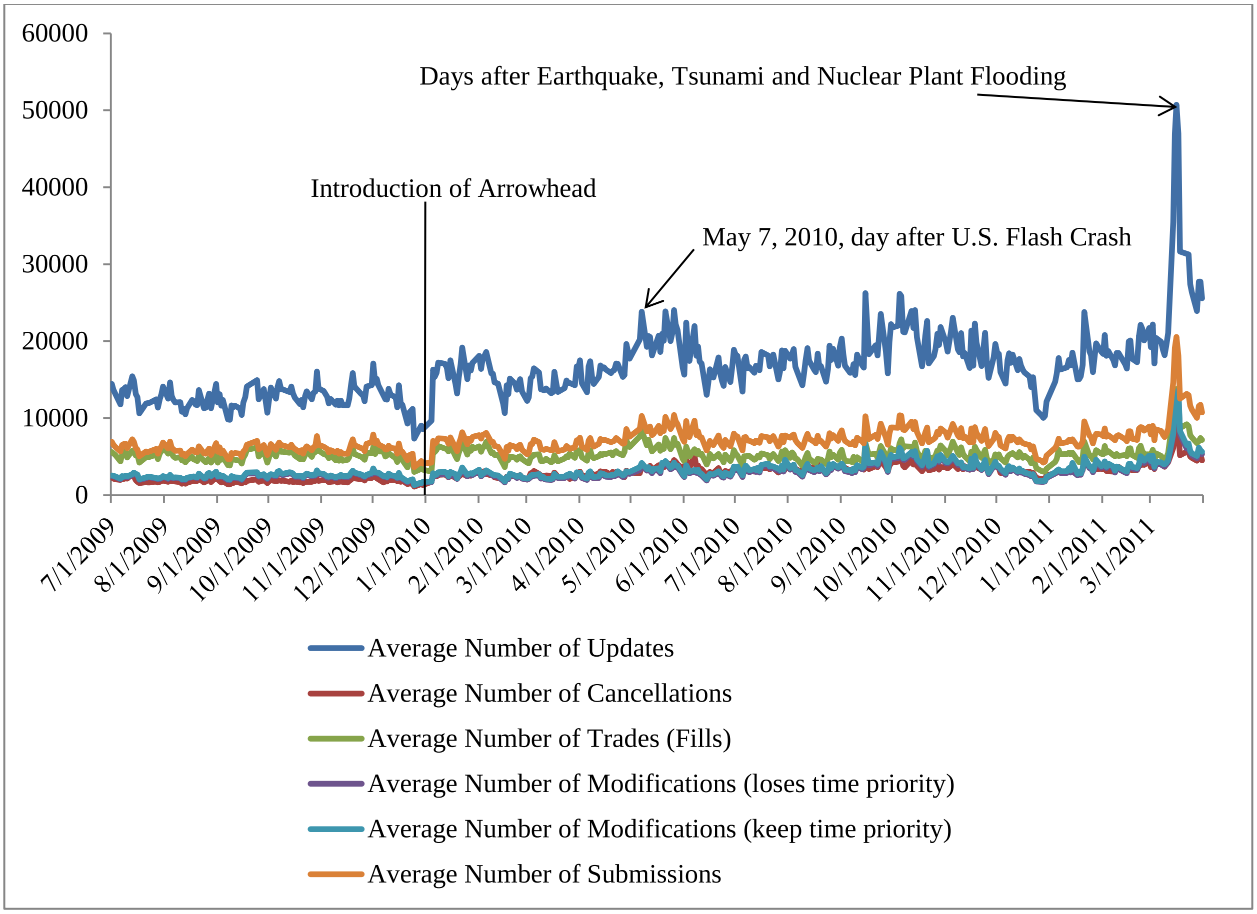

Figure 6: The figure shows the cross-sectional average of the number of updates, cancellations, trades, modifications that lose time priority, modifications that keep time priority and submissions between July 1, 2009 and March 31, 2011. The sample consists of the largest 300 stocks in the first section of the TSE at the beginning of each month.

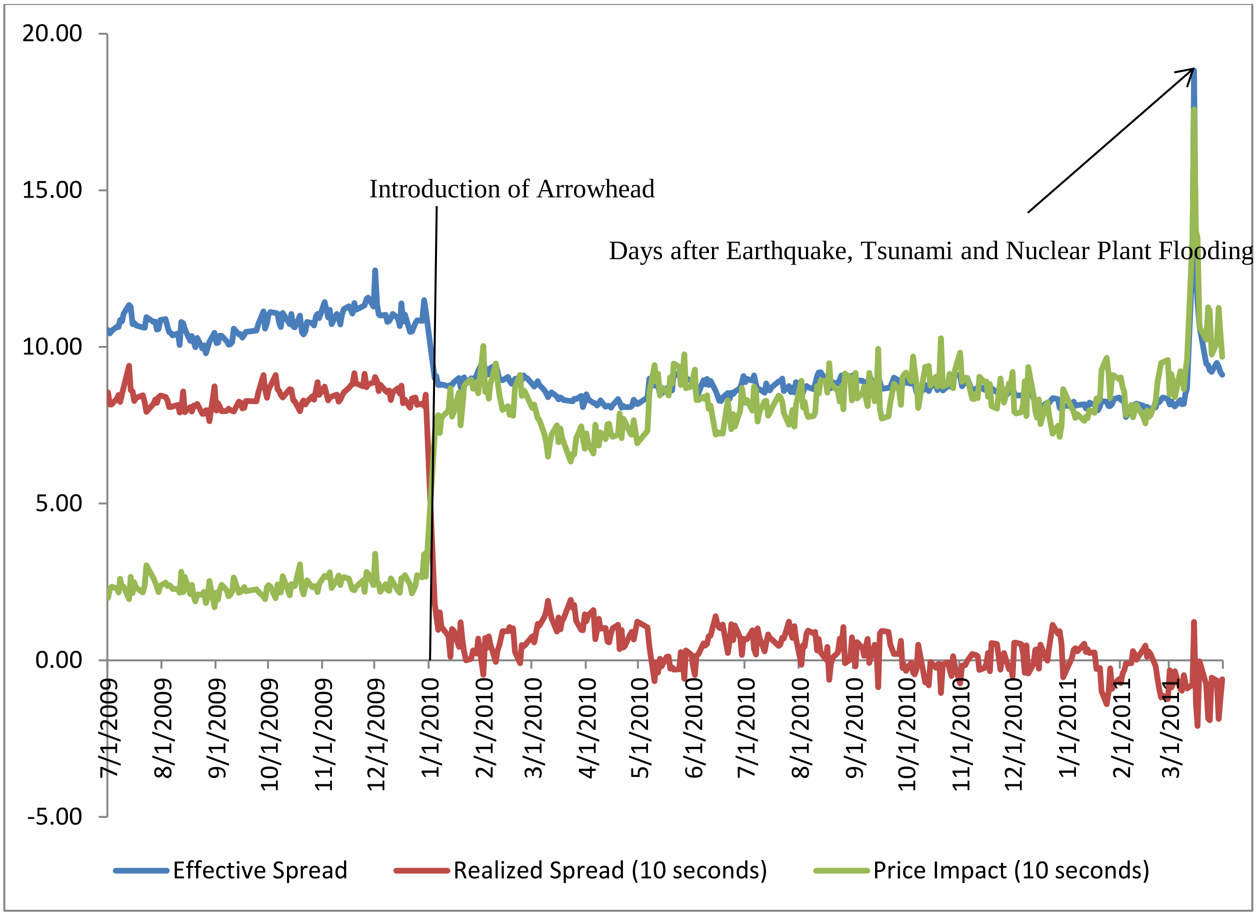

Figure 7: For each stock, we calculate the share-weighted average effective spread, realized spreads and price impact for each day. Realized spreads and price impact are based on quote midpoints 10 seconds after a trade. The figure shows the cross-sectional average on each day from July 1, 2009 to March 31, 2011.