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
Discover Financial Services
2402 W. Beardsley Road
Phoenix, AZ 85027

Forthcoming, *Journal of Financial Economics*

November 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, Terry Hendershott, Mark Seasholes and participants at the 5th Hedge Fund Research Conference (Paris), the Instinet Global Quantitative Equity Conference, the Financial Markets Research Center Conference (Vanderbilt University), the FTSE World Investment Forum, and the Stern Microstructure Conference. 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 Discover Financial Services 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 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

The effect of high frequency trading on market quality is important, and has generated strong interest among academics, practitioners and regulators. Models of the effect of high frequency trading on markets generate different predictions, depending on their assumptions and their focus. 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. Similarly, Han, Khapko and Kyle (2014) argue that since fast market makers can cancel quotes faster than slow traders, 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. In their model, however, high-frequency liquidity provision declines when market volatility increases, which can lead to episodes of market fragility. 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).

There are two strands to the complementary empirical literature. The first can be broadly characterized as examining the behavior of high-frequency traders and estimating their effect on markets: researchers use datasets that explicitly identify high frequency traders (by some definition), explore the trading strategies they use, test whether those strategies are profitable, and whether they impede or improve price discovery. For example, Brogaard, Hendershott and Riordan (2014), Carrion (2013), and Hirschey (2013) all use Nasdaq-identified high frequency traders (the so-called "Nasdaq dataset") and collectively find that HFT are modestly profitable in aggregate, that they both demand and supply liquidity, and that they appear to impose some adverse selection costs on other traders. Similarly, Hagströmer and Norden (2013) use data on 30 stocks from NASDAQ-OMX Stockholm and find evidence which indicates that market-making HFTs reduce short-term volatility. The advantage of this first group of studies is that identification of high frequency traders is relatively clear-cut. The disadvantage is that only activity on that identifying exchange can be precisely measured. In a fragmented market like the U.S., with substantial variation in access (make-take) fees across exchanges and dark pools, it is

entirely possible that HFT behavior in one market (e.g. Nasdaq) may not reflect aggregate market behavior, or inform overall market prices. The latter is our primary concern.[1]

The second strand of the empirical literature looks at outcomes in a conditional setting: identifying changes in market structure that facilitate high frequency activity and examining the consequences. The most important of these papers is Hendershott, Jones and Menkveld (2011), which examines 1,082 stocks between December 2002 and July 2003. Using the start of autoquotes on the NYSE as an exogenous instrument, they find that algorithmic trading improves liquidity. More recent studies which examine changes in the trading environment in smaller markets draw similar conclusions. For example, Menkveld (2012) examines the effect of the introduction of an electronic exchange (including a large HFT) on trading in a sample of 32 Dutch stocks, and Riordan and Storkenmaier (2012) examine the effect of an upgrade of trading systems in 98 stocks on the Deutsche Borse.

These papers provide evidence on the effects of high frequency *trading*. There is very little evidence on a critical aspect of current market structure: that of high frequency *quoting* – essentially, the speed of the market environment. The relative scarcity of this evidence is surprising, since many of the theoretical papers in this area describe the speed of changes to the supply curve – which is more closely related to quotations than to trades. And, regulators certainly care about high frequency quotations. Although the SEC's Concept Release on Equity Market Structure (2010) highlights HFT's as "professional traders acting in a proprietary capacity that engage in strategies that generate a large number of trades on a daily basis", the concept release also recognizes the importance of high frequency quoting in that it might represent "phantom liquidity (which) disappears when most needed by long-term investors".[2] That is, high frequency quoting generates execution risk, which has welfare consequences and is

[1] 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.

[2] 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).

an important characteristic of market structure. For example, Hasbrouck (2013) notes that the execution risk caused by high-speed changes in quotes may not be diversifiable, with slower traders always losing to faster traders.[3] Biais, Foucault and Moinas (2014) investigate policy approaches, including a Pigovian tax, which may mitigate externalities due to differences in traders' ability to process the amount of information generated by the market, such as the volume of high-frequency quotations. Stiglitz (2014) expresses skepticism that high frequency quotation/trading is welfare improving and makes a case for slower markets.

Our purpose is to provide large sample evidence on the influence of high frequency quoting on market quality. We do not look at the trading strategies of identified high-frequency traders; instead, we examine market outcomes. We conduct two types of tests: (a) unconditional tests designed to provide evidence for a comprehensive cross-section of securities over a long and relatively recent time series, and (b) conditional tests which measure the effect of high-frequency traders during different types of market conditions and over changes in market structure. The latter examine both 'average' and stressed market environments, and separately, a change in trading protocols. Each of the tests described below either provides new evidence on the influence of high-frequency quotations on markets, and/or fills a gap in our understanding of high-speed markets.

Our sample comes from the two largest equity markets in the world: the full cross-section of securities in the U.S., and the largest 300 stocks on the Tokyo Stock Exchange. The sample period is 2009-2011 for the U.S. and 2010-2011 for Japan. The breadth of the cross-section allows for general conclusions and recent data are important since there have been significant changes in market structure in the past few decades. It is also critical that the time-series be long enough to generate statistical power, particularly since cross-sectional independence is likely to be low. Finally, the Japanese data allow for an out-of-sample test in which we can estimate how an exogenous change in the speed of the market changes price discovery and the average cost of trading.

[3] Using a sample of 100 stocks in April 2011, Hasbrouck (2013) 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.

Our measure of high frequency quoting, which we refer to as quote updates, is any change in the best bid or offer (BBO) quote or size across all quote reporting venues. Each such change can be triggered by 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. Our first test examines the relation between quote updates and variance ratios over short horizons. A benchmark variance ratio of one is consistent with a random walk in prices, which is typically associated with weak form market efficiency.[4] 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 quotations are associated with persistent swings away from fundamental values, or slow adjustments to shocks, variance ratios in securities with higher levels of quotations may rise above one.

Between 2009 and 2011, in the smallest size quintile of stocks, there is less than one quote update per second. In large capitalization stocks, on average, changes to the top of the limit order book occur every 50 milliseconds. 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. We also examine variance ratios for a subset of securities at higher frequencies (100 milliseconds compared to 1- and 2-second returns) and find largely similar results. The time series average of the cross-sectional standard deviation of variance ratios is also lower for stocks with higher updates, implying that higher update activity is associated with lower variability in deviations from a random walk.

Higher updates are associated with lower costs of trading. Again controlling for firm size and trading activity, average effective spreads 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

[4] See, e.g., Campbell, Lo and MacKinlay (1997) for an excellent discussion of the random walk hypothesis and variance ratio tests.

for a $60 stock correspond to 0.5 basis points. Make/take fees are important enough to drive differences in (algorithmic) order routing between exchanges, implying 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 comes from a reduction in realized spreads, suggesting that increased competition between liquidity providers provides incentives to update quotes. Regardless of the source, the magnitude of the effect of higher updates on market metrics appears to be economically meaningful; increases in the number of updates represent more than simply the addition of noisy data which must be processed and filtered out to assess market conditions. In general, these results are consistent with the Baruch and Glosten (2013) argument that there is nothing nefarious about high frequency quote updates: the speed of updating may have increased but high-speed updates still represent the provision of liquidity and, on average, allow for information to be reflected in prices.

Our second test concerns the fragility of the market. A common complaint (e.g. Stiglitz (2014)) of the current market structure is that it is fragile in that the price of liquidity rises too rapidly, or that liquidity disappears entirely, when traders need it most. 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 could cause price series to deviate from a random walk and future trading costs to rise. On the other hand, if markets are resilient, then high-frequency liquidity providers should continue to supply liquidity following a significant positive shock to liquidity demand, and market quality measures in a high-frequency environment should not deteriorate after such an event.

We begin by identifying 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. Indeed, these algorithmic sweeps are often part of successive sweeps that,

within short periods of time, extract even larger amounts of liquidity. We design a simple algorithm to aggregate successive sweeps into singular liquidity drawdowns and examine 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.[5] Drawdowns in securities with higher updates incur lower costs, consistent with the idea that updates are correlated with liquidity provision. In addition, average variance ratios estimated in the 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 cross-sectional tests 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 additional tests that help with identification, while not abandoning a large sample approach.

First, we exploit the daily time series variation in quote updates. If high-frequency traders are drawn to trading in securities which are more liquid and more efficiently priced, then lagged reductions in effective spreads and lagged differences in the deviations of variance ratios from one should be important determinants of high frequency updating. However, a reduced-form vector autoregression shows the opposite result: particularly in large capitalization securities, prior day *increases* in effective spreads, and prior day *increases* in the deviation of variance ratios from one, are associated with higher updates. The implication is that a 'habitat' effect is not driving our cross-sectional results.

Related, we also find that the daily average number of quote updates closely tracks the VIX. The VAR shows that daily changes in updates are related to lagged innovations in the VIX but not vice-versa, implying that update activity is not merely noise but related to economic fundamentals. If quote updates are the tool used by liquidity providers to manage their intraday

[5] 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.

risk, it seems unlikely that variance ratios in day *t* are driving 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 also permitted co-location services and started reporting data in 100 millisecond increments (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. Unsurprisingly, 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 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 roughly 10% on the date of the introduction of the new trading system. Overall, the data suggest that facilitation of high frequency quotation has, on net, beneficial effects in the second largest equity market in the world.

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.

Many of our tests are based on size quintiles because of significant differences 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% 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 reported to us 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: 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. The data also identify 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 quote updates

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 for all registered exchanges. We construct this manually for each exchange, rather than relying on the official NBBO. 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. Moreover, under Regulation NMS, flickering quotes, defined as quotes that change more than once per second, are not eligible to set the NBBO.[6] By using the BBO across all exchanges, we include quote changes that are legitimate changes to the tip of each exchange's liquidity supply curve, regardless of whether the change is eligible to set a new NBBO. This is an underestimate because it does not include dark venues, and also does not include hidden orders. In addition, it does not consider changes to the totality of the supply curve, that is liquidity outside the best bid or ask prices ("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 through quotation activity. The latter is a change in the tip of the liquidity supply curve caused by a prior intersection with a demand curve (i.e. a trade). Trades, by virtue of their capital commitment, have important 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 during our sample period. 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 $X_t = X_0, X_1, \ldots, X_T$. 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

[6] 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).

intervals, and estimate the ratio of the variance of these returns over measurement intervals of a half-hour. 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.[7]

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 milliseconds, 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.

Each return interval is equally spaced so that there are T = nq returns in the measurement interval, where n and q are integers greater than one.[8] There are T-q+1 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

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

[7] An alternative to traditional variance ratios is to measure pricing errors using a Hasbrouck (1993) VAR model. However, such an approach is more appropriate for pricing errors associated with trades, whereas our interest is in standing quotes. In addition, the computational burdens of the Hasbrouck method, particularly in the millisecond data environment and for a large cross-section of securities, are considerable.

[8] Note that n is the number of non-overlapping long-horizon returns in the measurement interval and q represents the number of non-overlapping short-horizon intervals that are included in the long-horizon return.

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

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

A random walk in the underlying returns implies that variances are linear in the measurement interval. Given the definitions above, this implies that the ratio of $\bar{\sigma}_c^2(q)$ to $\bar{\sigma}_a^2(q)$, or the ratio of scaled large interval returns' variance to short interval returns' variance, should be equal to 1. Therefore, a test of the random walk hypothesis is

$$M_r(q) \equiv \frac{\bar{\sigma}_c^2(q)}{\bar{\sigma}_a^2(q)} - 1 = 0$$

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 estimate 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 (2014). 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. The horizon can be thought of as long enough to incorporate the permanent impact of the trade so that quotes are subsequently stabilized, and temporary effects dissipated. One can also think of the horizon as one in which liquidity providers can close their positions. Under either view, 5 minutes may be excessively long given the speed of quoting and trading in our sample period. We therefore 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.

We also 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 Aativity

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. 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% 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. This velocity of trading increases sharply at the beginning and end of the trading day. 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.

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 0.52 quote updates per second in the smallest size quintile and over 12 quote updates per second in the largest size quintile. In general, the data show that changes to the top of the book are an order of magnitude faster than trades – quoting occurs at a much higher frequency than trading. The speed of these changes 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. 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 averaging of 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.[9]

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 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 graph shows that across all size and trade quintiles, average variance ratios are generally closer to one for high update

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

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 our interest is 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. We then calculate the difference in cross-sectional averages (high minus low) and report time series means of these differences in Panel A. 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 where the test has more power. Thus, even at horizons measured in fractions of a second, high frequency quoting seems to be positively associated with price series that are closer to a random walk.

3.4 Effective spreads

To calculate average effective spreads across update groups, 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 0.06 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. We can assess economic significance either by considering the relative magnitude of quoted spreads or comparing the differences in effective spreads to access fees. Quoted spreads for large cap stocks vary from 2 to 10 basis points over our sample period. Access fees of $0.003 for an average stock price of $60 amount to 0.5 basis points. By either measure, the differences in effective spreads that we observe across low- and high-update groups appear to be economically large.

We also estimate Fama-MacBeth regressions for each half hour, where variance ratios and effective spreads are dependent variables, and market capitalization, price levels and various measures of trading activity are independent variables. Time series plots of regression slopes (not reported) show largely the same result as the results in Table 3: securities with higher levels of quotation activity tend to have price processes that more closely resemble random walks and lower effective spreads.

3.4.1 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. We calculate effective and realized spreads over one second intervals ranging from one to 20 seconds--effectively a full

term structure of the spread decomposition over this horizon. 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. Because 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. To conserve space, we only show realized spreads and price impact for one, five, ten and twenty seconds.

Consistent with the full sample results in Table 3, average differences in effective half spreads are negative for most size and trade quintiles.[10] 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. The decline in realized spreads that we find also differs from the results in Riordan and Storkenmaier (2011), who find a sharp and persistent increase in realized spreads following a system upgrade on Deutsche Borse that resulted in a decline in latency. Our results suggest that between 2009 and 2011, competition between electronic liquidity providers appears to be sufficient to generate reductions in realized spreads for those securities with a higher-speed trading environment.

3.5 Liquidity drawdowns and market resiliency

[10] The effective spreads calculated from millisecond data are less variable than the effective spreads in Table 3, which includes both the second data (2009) and the millisecond data (2010-2011). We conjecture that the higher variability is a result of noise in signing trades when non-millisecond data are used. To test this, we repeated our tests in Table 3 using only 2010-2011 millisecond data and find that standard errors are much smaller. The general pattern of differences in effective spreads between high and low update groups remains similar.

In resilient markets, large drawdowns of liquidity should minimally influence the future supply of liquidity. The millisecond TAQ data afford the possibility of such a test.[11] We employ a three-step procedure to isolate large liquidity drawdowns. We first identify multiple trades in a stock with the same millisecond timestamp across more than one reporting venue ("sweeps"). We then aggregate 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. 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 represent attempts to rapidly extract liquidity. During our sample period, there are 764 million such liquidity sweeps, comprising $17.6 trillion in volume.

In small stocks, sweeps represent 13% of total volume, rising almost monotonically to 22% for large 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 Regulation NMS. We observe considerable use of Intermarket Sweep Orders (ISOs) in algorithmic sweeps, and although not reported, the data show an increase in the size of sweeps and the use of ISOs over time.

We separate the 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 median time difference between successive sweeps in small cap stocks is 22 seconds, falling monotonically across size

[11] 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). We believe that our method provides results that are more representative of the conditions faced by market participants in a large cross-section of securities. In particular, our method does not require cherry-picking the data on price changes or measures of market quality, allowing for general inferences.

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. A graphical illustration of this process is provided in Figure 4.

By construction, aggregated sweeps are larger and inter-sweep time differences are greater. To focus on large liquidity demands, we further restrict the sample to aggregated sweeps that cumulatively extract 10,000 shares. There are two reasons to impose this restriction. 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. The final sample consists of about 4.4 million aggregated sweeps. The average time between successive aggregated sweeps ranges from 2,456 seconds for size quintile 1 to 454 seconds for size quintile 5.[12]

Our interest is in 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 pre- and post- measurement interval needs to be short enough so that there is minimal overlap in successive sweeps in the same stock. Given this consideration, and the distribution of the time differences between consecutive sweeps, we define pre- and post-sweep periods as 300 seconds. Pre-sweep periods include all observations up to the beginning of the sweep and post-sweep observations commence immediately after the end of the sweep. For measuring variance ratios, this allows us to sample quote midpoints at 1 second and 15 second

[12] We also identify large sweeps using a stock specific approach. We compute a relative volume ratio as the number of shares traded in the collapsed sweep, scaled by the product of average trade size (in shares) in the prior month multiplied and the number of trades. 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.

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 5 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 implied by these variance ratios is consistent with the "liquidity-based volatility" that Brunnermeier and Pedersen (2010) describe. However, it is also related to a more subtle aspect of our short-horizon variance ratio measure. Recall that variances can only be computed if quote midpoints move enough (i.e. that the variance measure in returns is non-degenerate), and that we require at least 15 non-zero returns to calculate variance ratios. In our sample, quote midpoints exhibit very little movement at the 1- and 15-second horizon in the 300 second interval before/after an aggregated sweep. For example, even in large cap stocks, we are only able to reliably calculate variance ratios for 31% of the sample. That is, the short-horizon variance measures used in our calculation of variance ratios around sweeps are biased towards larger price movements – and the relative magnitude of the bias increases for shorter horizons. This implies that our variance ratio measures are biased downwards.

With that in mind, our interest is primarily in *changes* in variance ratios around the liquidity drawdown. Panel A shows that the values of $\Delta|VR_t-1|$ are very close to zero. Unsurprisingly given the large sample sizes, paired *t*-statistics reject the null that $\Delta|VR_t-1|$ is different from zero. However, the direction of the change in point estimates consistently show that variance ratios improve slightly (rather than worsen) 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. Again, paired *t*-statistics reject the null of equality 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.[13]

Changes in the price process and/or trading costs before and after sweeps may be fundamentally different for stocks with low versus high quote updates. Therefore, we also calculate changes in variance ratios and effective spreads before, during, and after sweeps after placing each sweep into the triple sorted size-trade-update groupings used in Table 3. The changes in variance ratios and effective spreads are small and negative, similar to those in Table 5 (and therefore not reported in a separate table). There is, however, one key difference. Intra-sweep effective spreads (i.e. the cost of the liquidity drawdown) are significantly *higher* in groups with fewer updates in the prior half hour, with differences ranging in value from 7 to 37 basis points. The implication, consistent with the cross-sectional results in Table 3, is that update activity is associated with lower trading costs.

4. Alternative tests

The cross-sectional tests suggest that 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 prior to measuring variance ratios, this endogeneity may affect inferences. In this section, we report results from tests that allow for causal interpretations.

4.1 Exploiting time series variation in updates

Nagel (2012) reports that 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 (as well as 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

[13] One might be concerned that requiring data before and after a sweep might introduce a selection bias if sweeps cause a significant decline in liquidity. In that situation, market data would be available before a sweep but not after. We verify that this does not occur in our sample.

across firms in a size quintile. Figure 5 plots the daily time series, along with the VIX over the sample period. The graph shows that the time series variation in updates is large, and the correlation of updates with 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% drop in the S&P 500. This suggests that aggregate uncertainty has a role to play in the intensity of changes in the liquidity supply curve.

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

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

$$\Delta VIX_t = \sum_{i=1}^{4} \acute{\beta}_i \Delta Upd_{t-i} + \sum_{i=1}^{4} \acute{\gamma}_i \Delta VIX_{t-i} + \sum_{i=1}^{4} \acute{\kappa} \Delta MQ_{t-i} + \varepsilon_{2,t}$$

$$\Delta MQ_t = \sum_{i=1}^{4} \grave{\beta}_i \Delta Upd_{t-i} + \sum_{i=1}^{4} \grave{\gamma}_i \Delta VIX_{t-i} + \sum_{i=1}^{4} \grave{\kappa} \Delta MQ_{t-i} + \varepsilon_{3,t}$$

where all changes are in percentages and *MQ* refers to a measure of market quality. 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 6 reports coefficients with Z-statistics in parentheses. In Panel A, market quality (*MQ*) is measured using the deviations in variance ratios from one (Δ|VR-1|), while in Panel B, market quality is measured using effective spreads. We draw two inferences from these regressions. First, we examine the time-series relation between measures of market quality and changes in updates. Consider the possibility that, rather than increases in updates driving increases in liquidity and improvements in price discovery, the causation is reversed: high-frequency traders increase their activity in securities because those securities are relatively more liquid and their prices are relatively more efficient. Such a relation may occur because it is easier for HFT to manage risk in such an environment. In that case, one would expect changes in updates in period *t* to be negatively related to lagged changes in measures of market quality. In quintiles 1-4, most of the coefficients on lagged values of ΔMQ are statistically indistinguishable

from zero in both Panels. In quintile 5, the coefficients on ΔMQ are positive and, particularly for variance ratios measures, statistically significant. That is, if lagged effective spreads are higher, and lagged variance ratios are further from one, high frequency activity *increases*. This result is more consistent with high-frequency traders stepping in to provide liquidity and facilitate price discovery, rather than being attracted to improvements in liquidity and efficient pricing.

The second inference that we draw from these results is that changes in updates are positively related to lagged innovations in VIX. The econometric interpretation is that changes in VIX Granger-cause changes in updates. The economic interpretation is that liquidity suppliers react to changes in the risk environment by changing the frequency of updates. Beyond a one-day lag, the influence of VIX wanes quite quickly. Interestingly, the VIX equation shows that changes in quote update activity do not Granger-cause changes in the VIX 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. Despite the fact that we measure variance ratios after calculating quote updates, and the positive sign on market quality measures in the update equation above, it still may be the case that improvements in variance ratios 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 a control for these potential omitted variables, we 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*.[14] Since the results in Table 6 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.[15] The results are not reported in tables to conserve space but they lead to the

[14] 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).

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

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 for the cash equities market so that trading firms were permitted to install servers on the TSE Primary Site for arrowhead. Second, time stamps for data reported in the public data feed changed from minutes to 100 milliseconds increments. In July 2012, this reporting was further reduced to millisecond increments. The data provided to us, however, contain true millisecond timestamps for the entire post-arrowhead period. Third, the TSE changed its tick size grid. Pre-arrowhead, stocks were bracketed into 9 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 (twice 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.[16] Fifth, the TSE implemented changes to procedures which had been designed to slow down trading in the event of an order imbalance. Pre-arrowhead, the TSE employed price limits to trigger special price quote dissemination that would (presumably) attract contra order flow.

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

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 stock price levels.

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, suggesting that in the U.S. as well, updates are a good instrument for the rapid supply curve changes in which we are interested. 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 Tokyo is 14 hours). There is a sharp increase in updates and their components on the Monday after the earthquake, tsunami and Fukushima nuclear disaster in March 2011. 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% with a paired *t*-statistic of 10.75.

Our 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 isolate 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. To do so, we calculate the minimum and maximum price level for each security in the before and after window. We then identify securities that *ex post* remained in the same tick size price grid. For such paired comparisons, the sample consists of 229 securities.

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 7 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. Average paired changes in effective spreads are in the 3rd column of Table 7. In each half hour interval, effective spreads decline by approximately 1 and 2 basis points. The largest decline is in the first half hour in the afternoon trading session, at 2.69 basis points. All changes are highly statistically significant. Since pre-arrowhead effective spreads are approximately 10 basis points, a decline of between one to two basis points is economically large.

The 4th and 5th columns of the table show changes in realized spreads and price impact based on midpoints 10 second after a trade.[17] The results show a substantial decline in the realized spread, varying from 5 to almost 9 basis points, and an *increase* in price impact, varying from 4 to approximately 8 basis points. Figure 7 shows that the changes occur sharply at the introduction of arrowhead. Recall that the relation between effective spreads, realized spreads and price impact is mechanical: if effective spreads decline by about 1-2 basis points and realized spreads decline by a larger amount, it must be the case that price impact increases.

[17] As with U.S. data, we compute realized spreads and price impact using a variety of horizons but only show one horizon (10 seconds) for brevity.

The increase in price impact associated 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 itself as an increase in price impact. To test this conjecture, 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% (*t*-statistic=12.72) following arrowhead introduction. To further test whether differences in special quote dissemination influence the measures of price impact, we separate all stocks (without tick size changes) into two groups based on whether the percentage increase in special quote dissemination is above or below the median. In each half hour interval, the increase in price impact following the introduction of arrowhead is significantly higher (by between 1-2 basis points) for stocks which experienced a larger decline in special quote dissemination following arrowhead.

Summarizing, the introduction of arrowhead trading was designed to reduce the latency of trading on the TSE. Following its introduction, variance ratios move closer to 1. The magnitude of this change is very similar to the difference in variance ratios observed in high update and low update groups in the U.S. 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, the exogenous change in trading platforms in Japan confirms the evidence from U.S. equity markets, that higher levels of high frequency activity are 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 many researchers, market professionals, and regulators concerned that execution costs or execution risks may increase and that price discovery and efficiency has been, or may be, harmed.

The evidence suggests that, on average, high frequency quotation activity does not damage market quality. In fact, the presence of high frequency quotes is associated with improvements in the efficiency of the price discovery process and reductions in the cost of trading. Even when high frequency trading is associated with large extractions of liquidity in individual securities, the price 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. Some caution is warranted. Although the evidence suggests that high frequency activity is associated with some improvements in the market's function, the results do not imply that the market *always* functions in this way. An obvious case in point is the Flash Crash. 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 price changes. While dislocations are harmful to market integrity, it is important to recognize that some discontinuities 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. If liquidity provision is not mandated by law, liquidity providers can always exit without notice, exposing marketable orders to price risk. Designated Market Maker contracts are one way to mitigate this, as in Stockholm, Paris and other exchanges around the world (including the NYSE). These contracts generally do not focus on preventing dislocations but can still be potentially beneficial. Indeed, Bessembinder, Hao and Zheng (2013) argue that affirmative obligations of market makers can improve social welfare relative to endogenous

liquidity provision.[18] 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. The effects of recent regulatory changes to the market structure on market volatility, and the propensity of dislocations to occur, is the subject of our current research.

Caution is also warranted 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. However, understanding the influence of rapidly changing supply curves on price formation and the cost of trading is nevertheless important because as Stiglitz (2014) points out, these are 'intermediate' variables that are necessary (although not sufficient) for understanding the role of high frequency quoting/trading on welfare.

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

References

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

Asparouhova, E., Bessembinder, H., and Kalcheva, I., 2010, Liquidity biases in asset pricing tests, *Journal of Financial Economics*.

Biais, B., Foucault, T. and Moinas. S, 2014, Equilibrium fast trading, working paper, Toulouse School of Economics.

Bessembinder, H., Hao, J. and Zheng, K., 2013, Market making obligations and firm value, working paper, University of Utah.

Brogaard, J., Hendershott, T., and Riordan, R., 2011, High frequency trading and price discovery, working paper.

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

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

Baruch, S. and Glosten, L., 2013, Fleeting orders, working paper, Columbia University.

Campbell, J., Lo, A. and MacKinlay, A., 1997, The Econometrics of Financial Markets, Princeton University Press.

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

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

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

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

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

Han, J., Khapko, M. and Kyle, A., 2014, Liquidity with high frequency market making, working paper.

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

Hendershott, T. and Menkveld, A., 2013, Price pressures, *Journal of Financial Economics*, forthcoming.

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

Hendershott, T., Jones, C. and Menkveld, A., 2011, Does algorithmic trading improve liquidity, *Journal of Finance* 66, 1-33.

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

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

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

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

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

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

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

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

Riordan, R., and Storkenmaier, A., 2011, Latency, liquidity and price discovery, working paper Karlsruhe Institute of Technology.

Stiglitz, J., 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

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.

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
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

Table 2

Average variance ratios for size quintiles

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 is in parentheses.

Quintile	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)

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 share-weighted effective half spreads for all trades in each stock and half hour. 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 midpoints at 15 second and 5 minute intervals. The variance ratios in Panel B are for stocks in size quintile 5 only and are based on midpoints at 100 milliseconds and 1 second, as well as 100 milliseconds and 2 sections. 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 (basis points)					
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	0.0	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

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 B: Average effective half-spreads around liquidity sweeps 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 6

Vector autoregression of daily changes in quote updates, Vix and market quality measures

We estimate reduced form vector autoregressions with daily data of the form:

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

$$\Delta VIX_t = \sum_{i=1}^{4} \acute{\beta}_i \Delta Upd_{t-i} + \sum_{i=1}^{4} \acute{\gamma}_i \Delta VIX_{t-i} + \sum_{i=1}^{4} \acute{\kappa} \Delta MQ_{t-i} + \varepsilon_{2,t}$$

$$\Delta MQ_t = \sum_{i=1}^{4} \grave{\beta}_i \Delta Upd_{t-i} + \sum_{i=1}^{4} \grave{\gamma}_i \Delta VIX_{t-i} + \sum_{i=1}^{4} \grave{\kappa} \Delta MQ_{t-i} + \varepsilon_{3,t}$$

The number of updates and two measures of market quality (effective spreads and (Δ|VR-1|) are averaged across firms in a quintile. Z-statistics appear in parentheses.

Panel A: Market quality (MQ) measured as Δ|VR-1|

	Small			Size Quintile 2			Size Quintile 3			Size Quintile 4			Size Quintile 5		
	ΔUpd_t	ΔMQ_t	ΔVIX_t	ΔUpd_t	ΔMQ_t	ΔVIX_t	ΔUpd_t	ΔMQ_t	ΔVIX_t	ΔUpd_t	ΔMQ_t	ΔVIX_t	ΔUpd_t	ΔMQ_t	ΔVIX_t
ΔUpd_{t-1}	-0.500	-0.001	0.001	-0.493	-0.017	0.008	-0.483	-0.023	0.012	-0.470	-0.017	0.011	-0.454	-0.032	0.013
	(12.64)	(0.33)	(0.09)	(12.14)	(1.89)	(0.49)	(11.76)	(1.88)	(0.70)	(11.22)	(1.25)	(0.66)	(10.75)	(2.11)	(0.80)
ΔUpd_{t-2}	-0.263	-0.012	0.018	-0.303	-0.019	0.010	-0.296	-0.022	0.015	-0.300	-0.016	0.008	-0.292	-0.026	0.008
	(6.07)	(2.11)	(0.89)	(6.83)	(1.90)	(0.57)	(6.62)	(1.61)	(0.84)	(6.57)	(1.05)	(0.44)	(6.37)	(1.56)	(0.47)
ΔUpd_{t-3}	-0.206	-0.006	0.014	-0.241	-0.09	-0.002	-0.248	-0.017	0.001	-0.222	-0.006	-0.005	-0.206	-0.017	0.001
	(4.74)	(1.06)	(0.72)	(5.41)	(0.88)	(0.14)	(5.55)	(1.29)	(0.07)	(4.85)	(0.44)	(0.27)	(4.48)	(1.05)	(0.07)
ΔUpd_{t-4}	-0.025	-0.007	0.031	-0.065	-0.016	0.021	-0.068	-0.032	0.028	-0.067	-0.013	0.019	-0.072	-0.013	0.022
	(0.63)	(1.31)	(1.66)	(1.60)	(1.76)	(1.22)	(1.66)	(2.57)	(1.68)	(1.61)	(0.95)	(1.17)	(1.71)	(0.91)	(1.34)
ΔVIX_{t-1}	0.428	-0.002	-0.102	0.515	-0.001	-0.110	0.548	0.007	-0.109	0.603	-0.019	-0.109	0.593	-0.014	-0.112
	(5.14)	(0.26)	(2.58)	(5.52)	(0.09)	(2.71)	(5.62)	(0.26)	(2.68)	(5.81)	(0.56)	(2.62)	(5.73)	(0.39)	(2.72)
ΔVIX_{t-2}	0.113	0.011	-0.074	0.175	0.025	-0.076	0.214	0.017	-0.08	0.236	-0.009	-0.082	0.206	0.011	-0.082
	(1.34)	(0.99)	(1.83)	(1.82)	(1.15)	(1.82)	(2.13)	(0.57)	(2.05)	(2.20)	(0.26)	(1.94)	(1.93)	(0.30)	(1.92)
ΔVIX_{t-3}	0.031	-0.018	-0.115	0.151	-0.002	-0.098	0.203	-0.003	-0.099	0.198	-0.299	-0.093	0.177	-0.022	-0.101
	(0.37)	(1.60)	(2.85)	(1.58)	(0.11)	(2.36)	(2.04)	(0.12)	(2.39)	(1.87)	(0.83)	(2.21)	(1.67)	(0.57)	(2.39)
ΔVIX_{t-4}	-0.032	0.023	-0.055	0.029	0.026	-0.051	0.069	0.046	-0.063	0.066	0.011	-0.059	0.071	0.016	-0.064
	(0.39)	(2.05)	(1.39)	(0.32)	(1.21)	(1.26)	(0.71)	(1.53)	(1.54)	(0.64)	(0.32)	(1.44)	(0.69)	(0.45)	(1.57)
ΔMQ_{t-1}	0.036	-0.650	0.001	0.131	-0.675	-0.017	0.197	-0.622	0.007	0.285	-0.630	0.010	0.289	-0.688	0.004
	(0.13)	(17.90)	(0.01)	(0.83)	(18.39)	(0.26)	(1.66)	(16.95)	(0.16)	(2.58)	(16.88)	(0.24)	(2.82)	(18.40)	(0.12)
ΔMQ_{t-2}	0.183	-0.521	0.021	0.175	-0.449	-0.054	0.182	-0.388	-0.055	0.197	-0.410	-0.067	0.240	-0.486	-0.053
	(0.59)	(12.48)	(0.15)	(0.93)	(10.34)	(0.67)	(1.32)	(9.11)	(0.96)	(1.55)	(9.50)	(1.33)	(1.97)	(10.96)	(1.09)
ΔMQ_{t-3}	0.175	-0.336	-0.071	0.302	-0.262	-0.046	0.219	-0.206	-0.021	0.210	-0.236	-0.036	0.275	-0.251	-0.035
	(0.57)	(8.08)	(0.49)	(1.61)	(6.04)	(0.58)	(1.58)	(4.83)	(0.37)	(1.64)	(5.44)	(0.72)	(2.26)	(5.66)	(0.72)
ΔMQ_{t-4}	0.014	-0.179	-0.048	0.241	-0.128	-0.04	0.130	-0.120	-0.052	0.033	-0.119	-0.073	0.006	-0.111	-0.081
	(0.05)	(4.98)	(0.38)	(1.52)	(3.50)	(0.68)	(1.08)	(3.27)	(1.05)	(0.30)	(3.17)	(1.66)	(0.06)	(2.96)	(1.98)

Panel B: Market quality (MQ) measured as change in effective spreads															
ΔUpd_{t-1}	-0.500	-0.032	0.004	-0.491	-0.050	0.013	-0.490	-0.083	0.017	-0.498	-0.075	0.011	-0.489	-0.054	0.010
	(12.65)	(0.75)	(0.26)	(12.12)	(0.72)	(0.75)	(11.87)	(1.21)	(1.03)	(11.95)	(1.17)	(0.70)	(11.66)	(1.03)	(0.65)
ΔUpd_{t-2}	-0.264	0.034	0.018	-0.307	0.055	0.012	-0.306	0.074	0.017	-0.317	0.023	0.011	-0.321	-0.028	0.009
	(6.11)	(0.71)	(0.91)	(6.96)	(0.72)	(0.67)	(6.83)	(0.99)	(0.93)	(6.94)	(0.34)	(0.63)	(7.02)	(0.49)	(0.53)
ΔUpd_{t-3}	-0.204	-0.020	0.019	-0.245	-0.084	0.003	-0.252	-0.087	0.009	-0.238	-0.091	0.001	-0.248	-0.062	0.001
	(4.72)	(0.42)	(0.96)	(5.53)	(1.10)	(0.20)	(5.61)	(1.16)	(0.49)	(5.22)	(1.29)	(0.05)	(5.41)	(1.08)	(0.03)
ΔUpd_{t-4}	-0.026	-0.068	0.032	-0.074	-0.082	0.022	-0.078	-0.101	0.030	-0.075	-0.073	0.023	-0.085	-0.061	0.026
	(0.67)	(1.56)	(1.71)	(1.82)	(1.17)	(1.30)	(1.91)	(1.47)	(1.77)	(1.82)	(1.14)	(1.44)	(2.06)	(1.16)	(1.60)
ΔVIX_{t-1}	0.448	0.259	-0.049	0.534	0.334	-0.058	0.545	0.412	-0.062	0.559	0.421	-0.061	0.521	0.361	-0.064
	(5.00)	(2.62)	(1.16)	(5.36)	(1.94)	(1.36)	(5.28)	(2.38)	(1.45)	(5.09)	(2.49)	(1.42)	(4.74)	(2.60)	(1.49)
ΔVIX_{t-2}	0.074	0.159	-0.073	0.131	0.280	-0.075	0.159	0.346	-0.085	0.170	0.384	-0.079	0.169	0.436	-0.072
	(0.81)	(1.58)	(1.46)	(1.29)	(1.59)	(1.72)	(1.51)	(1.96)	(1.94)	(1.52)	(2.22)	(1.80)	(1.52)	(3.08)	(1.63)
ΔVIX_{t-3}	0.056	-0.023	-0.080	0.178	0.024	-0.069	0.232	0.061	-0.074	0.208	0.100	-0.068	0.159	0.123	-0.072
	(0.61)	(0.23)	(1.84)	(1.75)	(0.14)	(1.58)	(2.21)	(0.36)	(1.70)	(1.86)	(0.58)	(1.55)	(1.43)	(0.87)	(1.64)
ΔVIX_{t-4}	-0.033	0.266	-0.047	0.048	0.427	-0.050	0.075	0.451	-0.060	0.058	0.383	-0.054	0.018	0.255	-0.068
	(0.04)	(2.67)	(1.11)	(0.49)	(2.49)	(1.18)	(0.73)	(2.62)	(1.64)	(0.53)	(2.28)	(1.22)	(0.17)	(1.86)	(1.59)
ΔMQ_{t-1}	-0.016	-0.257	-0.054	-0.015	-0.178	-0.032	0.001	-0.198	-0.032	0.034	-0.233	-0.031	0.070	-0.349	-0.037
	(0.45)	(6.43)	(3.18)	(0.68)	(4.48)	(3.21)	(0.05)	(4.98)	(3.24)	(1.36)	(5.58)	(3.06)	(2.27)	(8.92)	(3.08)
ΔMQ_{t-2}	0.025	-0.076	-0.005	0.016	-0.055	-0.000	0.026	-0.081	0.001	0.041	-0.083	0.002	0.050	-0.146	-0.002
	(0.69)	(1.87)	(0.33)	(0.71)	(1.37)	(0.03)	(1.10)	(2.00)	(0.14)	(1.55)	(2.05)	(0.19)	(1.53)	(3.51)	(0.23)
ΔMQ_{t-3}	-0.035	-0.036	-0.028	-0.026	-0.011	-0.014	-0.028	-0.009	0.014	-0.020	-0.014	-0.014	0.028	-0.004	-0.010
	(0.96)	(0.89)	(1.62)	(1.13)	(0.29)	(1.14)	(1.16)	(0.24)	(1.45)	(0.77)	(0.36)	(1.42)	(0.86)	(0.12)	(0.83)
ΔMQ_{t-4}	-0.027	-0.004	-0.007	-0.014	0.021	0.001	-0.006	-0.018	0.001	0.001	-0.021	-0.001	0.043	-0.049	0.010
	(0.75)	(0.11)	(0.44)	(0.63)	(0.54)	(0.09)	(0.27)	(0.47)	(0.11)	(0.02)	(0.55)	(0.10)	(1.38)	(1.25)	(0.85)

Table 7

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 229 largest stocks on the 1st Section of the Tokyo Stock Exchange that did not experience changes in tick size in the three month period before and after 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 $|VR-1|_{post} - |VR-1|_{pre}$ 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 (after minus before) in basis points. *T*-statistics appear in parentheses.

	$\|VR-1\|_{post} - \|VR-1\|_{pre}$	Δ Effective Spread	Δ Realized Spread	Δ Price Impact
9:00 - 9:30	-0.016	-1.16	-8.89	7.73
	(1.80)	(11.12)	(45.93)	(39.32)
9:30 - 10:00	-0.037	-1.00	-7.81	6.80
	(3.18)	(8.87)	(39.20)	(35.13)
10:00 - 10:30	-0.020	-1.01	-6.90	5.92
	(1.97)	(7.95)	(40.36)	(37.58)
10:30 - 11:00	-0.015	-0.90	-6.40	5.50
	(1.02)	(7.52)	(36.35)	(36.52)
12:30 - 1:00	-0.019	-2.69	-6.88	4.19
	(1.85)	(22.68)	(42.16)	(22.92)
1:00 - 1:30	-0.025	-0.90	-6.18	5.29
	(2.25)	(7.38)	(37.26)	(36.04)
1:30 - 2:00	-0.019	-0.89	-5.94	5.04
	(1.13)	(7.25)	(38.86)	(34.29)
2:00 - 2:30	-0.027	-0.88	-5.95	5.06
	(1.97)	(7.18)	(39.41)	(37.32)
2:30 - 3:00	-0.026	-1.00	-5.07	4.14
	(1.54)	(7.56)	(37.06)	(39.33)

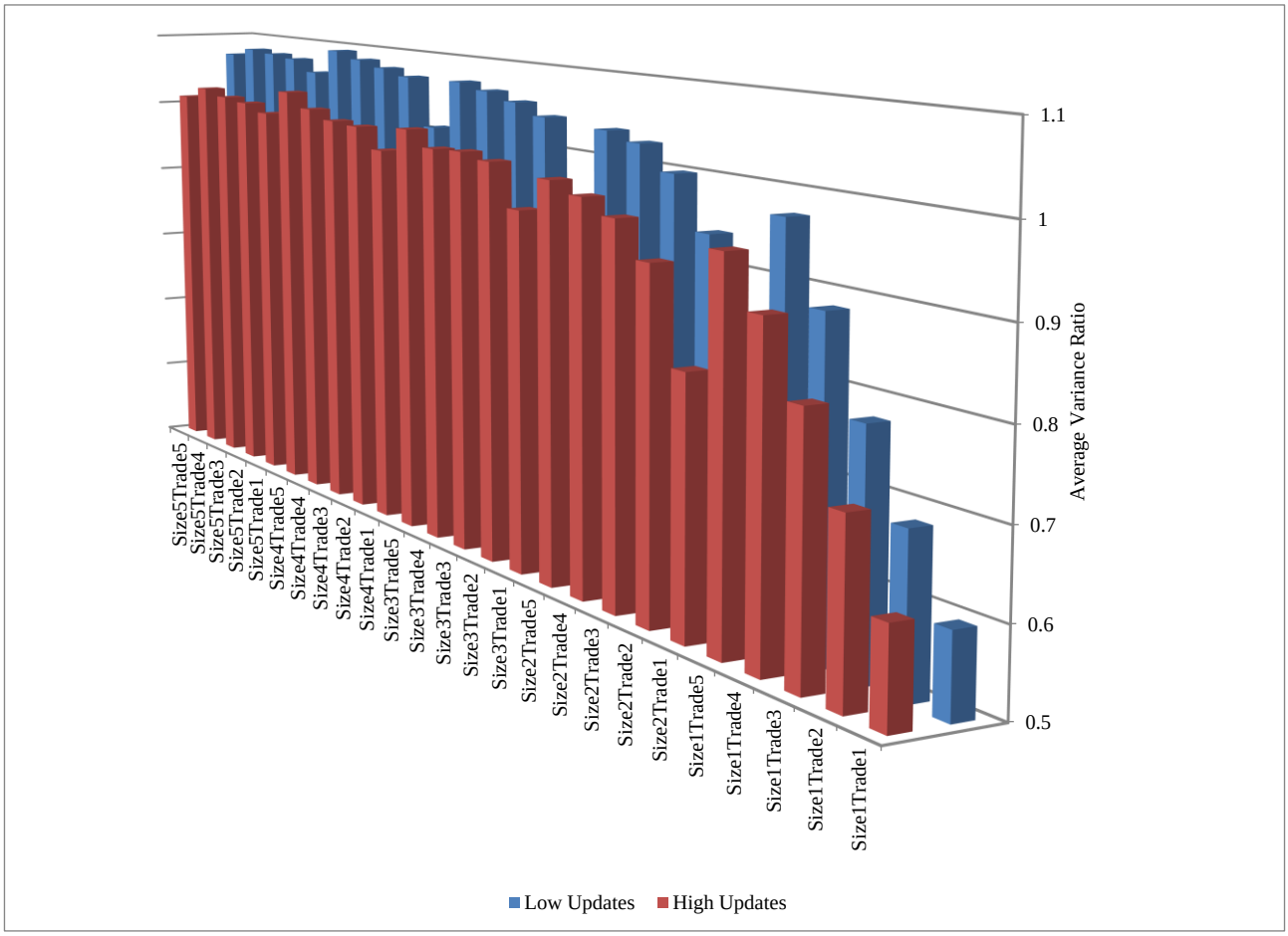

Figure 1: Stocks are sorted into size quintile based on prior month NYSE breakpoints, and within size groups into quintiles based on the number of trades in each half hour. For example, Size1Trade1 contains firms in the smallest size and trade quintile. Within these 25 (5x5) groups, stocks are further 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.

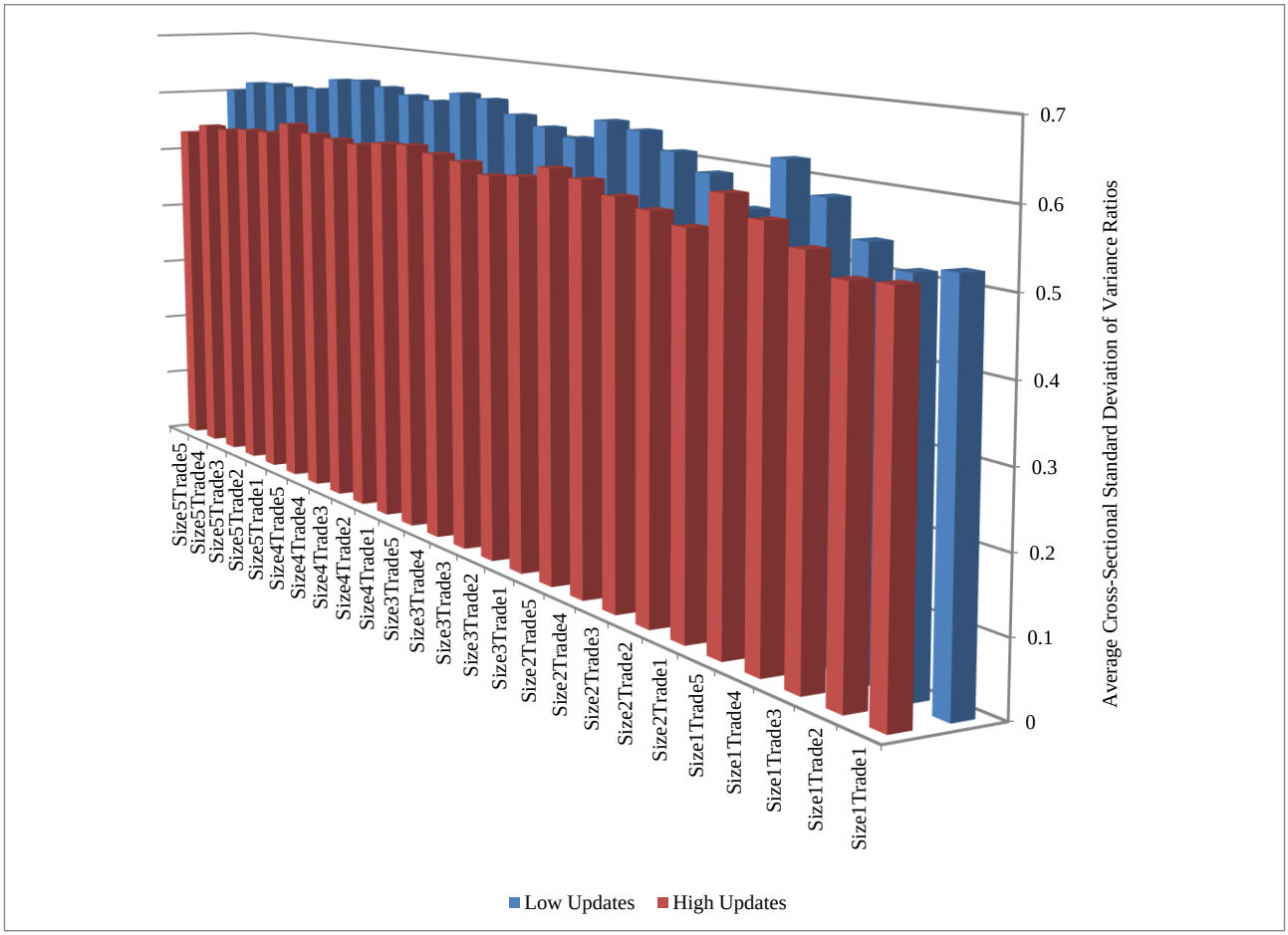

Figure 2: Stocks are sorted into size quintile based on prior month NYSE breakpoints, and within size groups into quintiles based on the number of trades in each half hour. For example, Size1Trade1 contains firms in the smallest size and trade quintile. Within these 25 (5x5) groups, stocks are further 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.

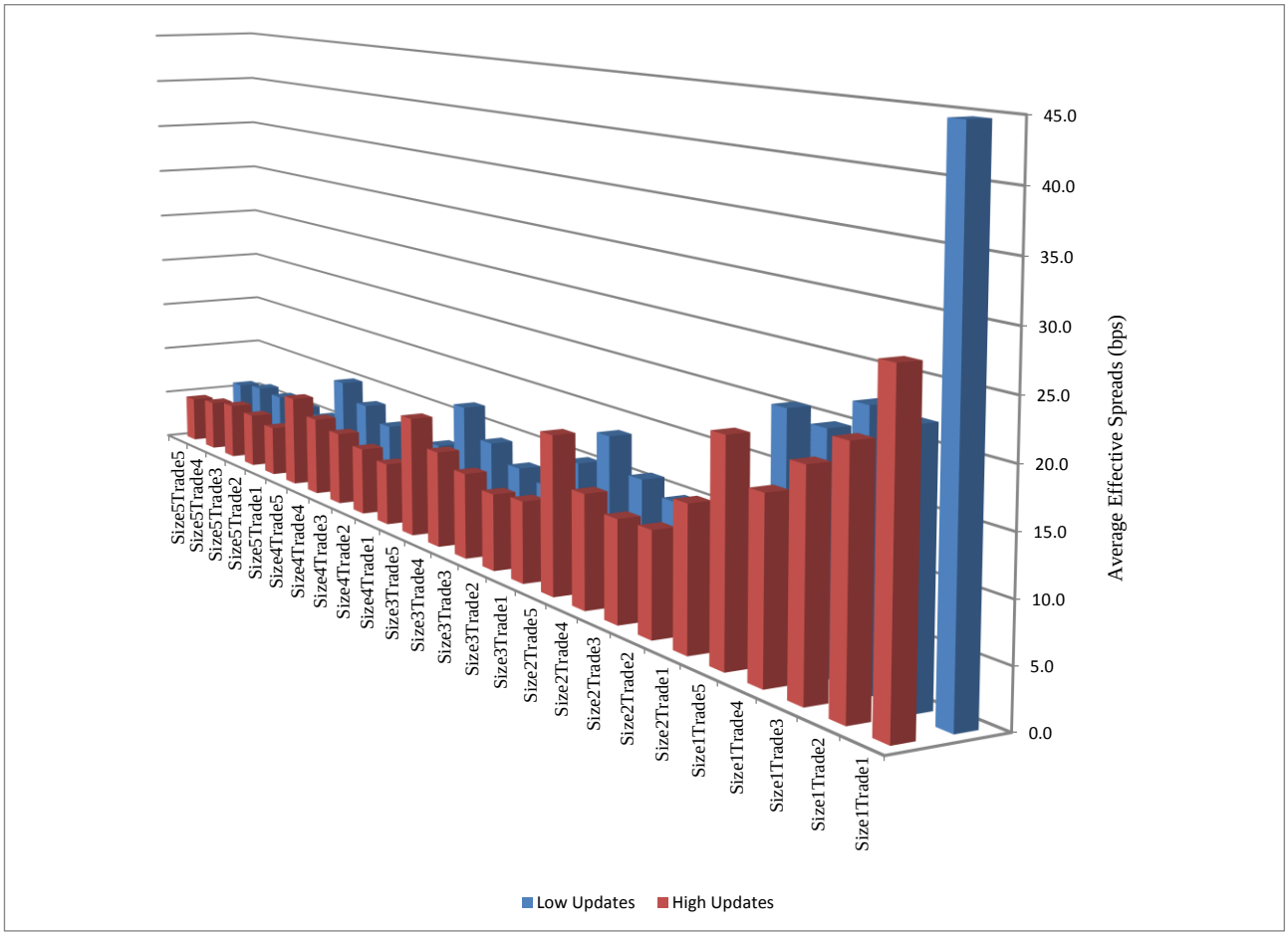

Figure 3: Stocks are sorted into size quintile based on prior month NYSE breakpoints, and within size groups into quintiles based on the number of trades in each half hour. For example, Size1Trade1 contains firms in the smallest size and trade quintile. Within these 25 (5x5) groups, stocks are further 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.

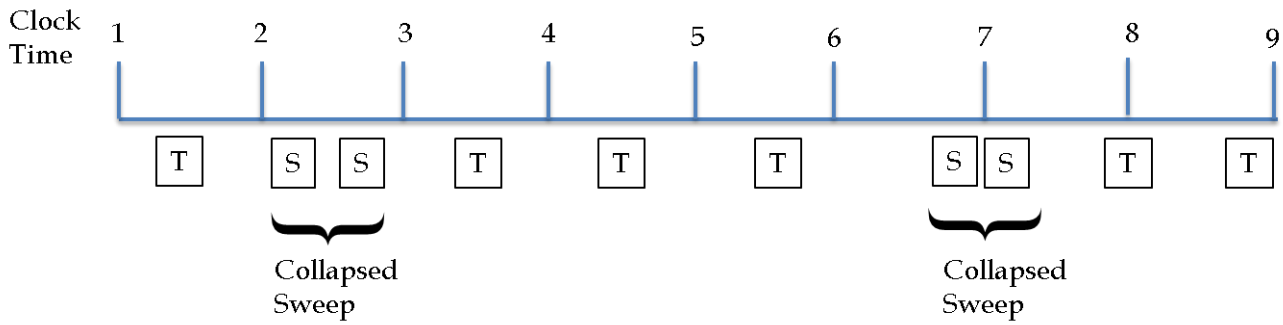

Figure 4: An illustration of collapsing consecutive sweeps. In this illustration, the time line is in seconds and the median time between trades in the prior month and the same half our interval is 1 second. Consecutive sweeps that take place with an inter-sweep time of 1 second are collapsed together.

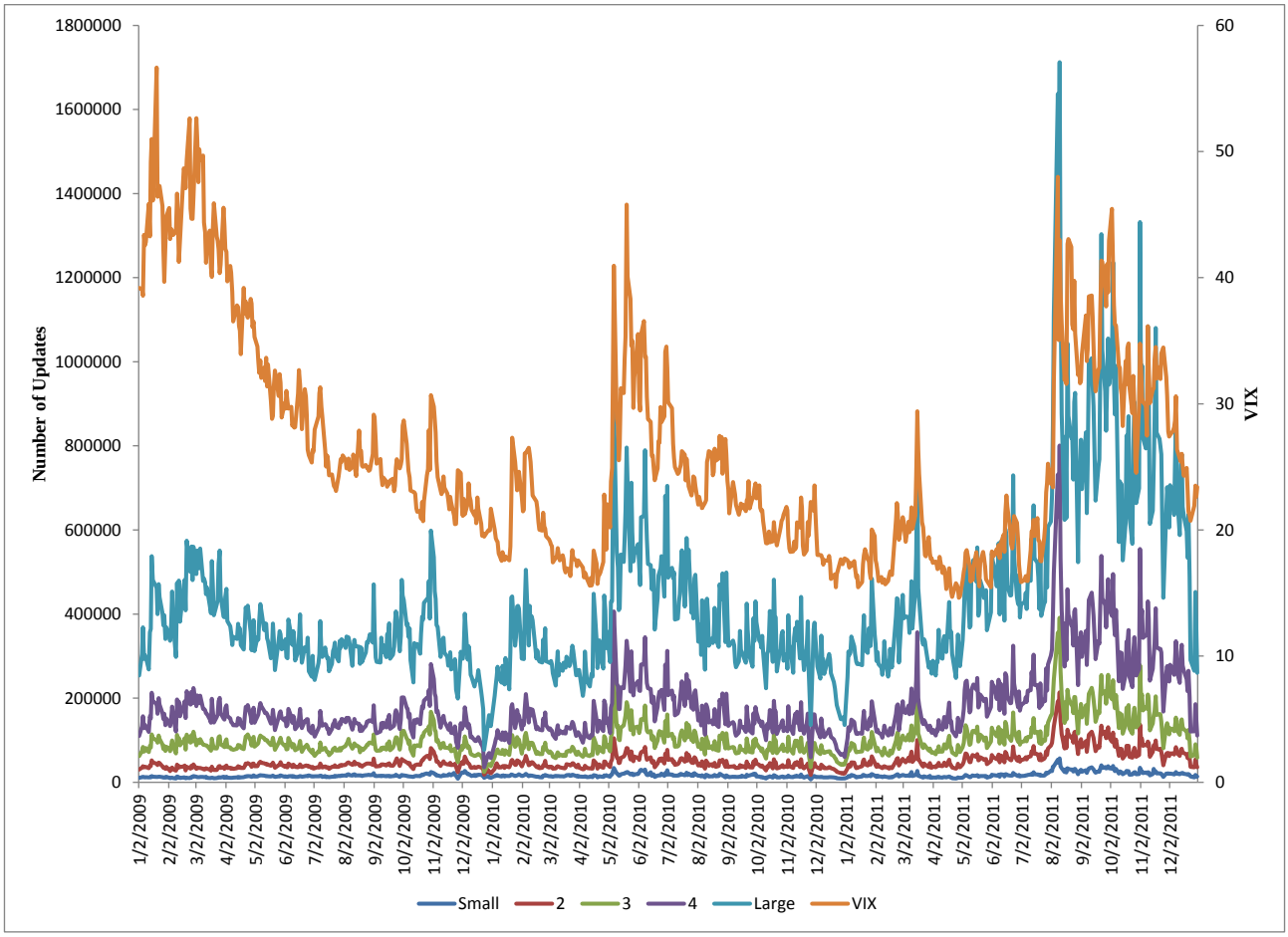

Figure 5: Each day, we sum the number of updates for each firm and then average across firms in size quintiles. Daily closing values of the VIX index are plotted using the right vertical axis.

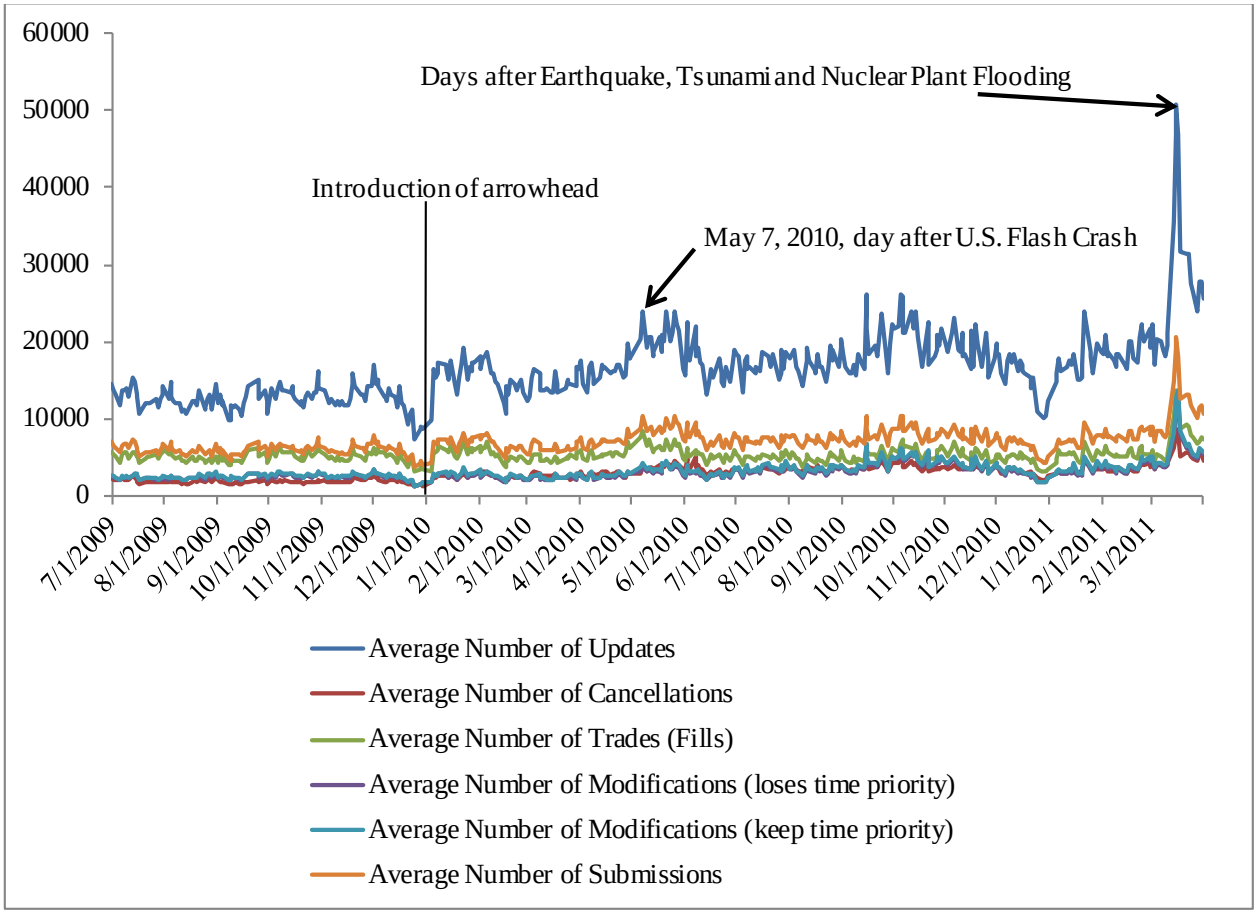

Figure 6: The figure shows the cross-sectional average of the number of updates, cancellations, trades, modifications that lose time priority, modifications that keep time priority and submissions between July 1, 2009 and March 31, 2011. The sample consists of the largest 300 stocks in the first section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

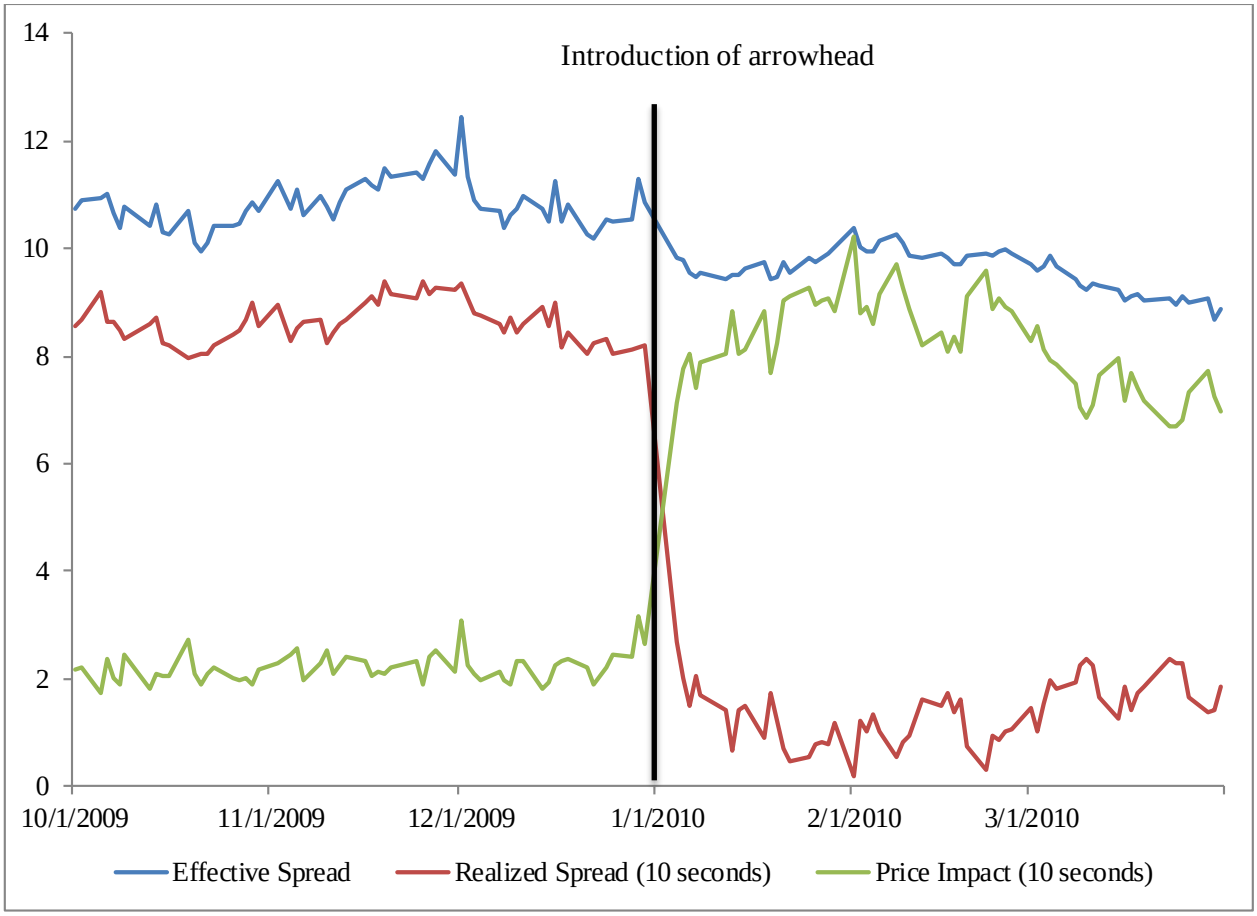

Figure 7: For each stock that does not experience a tick size change in the three months before and after the arrowhead introduction, we calculate the share-weighted average effective spread, realized spreads and price impact for each day. Realized spreads and price impact are based on quote midpoints 10 seconds after a trade. The figure shows the cross-sectional average on each day from October 1, 2009 to March 31, 2010.