Journal of Financial Economics 142 (2021) 314–337

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

# Bias in the effective bid-ask spread

Björn Hagströmer[1]

*Stockholm Business School, Stockholm University, Stockholm SE-10691, Sweden*

ARTICLE INFO

*Article history:*
Received 6 December 2019
Revised 29 August 2020
Accepted 27 September 2020
Available online 22 April 2021

*JEL classification:*
C15
G12
G20

*Keywords:*
Midpoint
Micro-price
Liquidity demand elasticity
Illiquidity
Rule 605

ABSTRACT

The effective bid-ask spread measured relative to the spread midpoint overstates the true effective bid-ask spread in markets with discrete prices and elastic liquidity demand. The average bias is 13%–18% for S&P 500 stocks in general, depending on the estimator used as benchmark, and up to 97% for low-priced stocks. Cross-sectional bias variation across stocks, trading venues, and investor groups can influence research inference. The use of the midpoint also undermines liquidity timing and trading performance evaluations, and can lead non-sophisticated investors to overpay for liquidity. To overcome these problems, the paper proposes new estimators of the effective bid-ask spread.



## 1. Introduction

The effective bid-ask spread is one of the most prevalent measures of market illiquidity, used in diverse applications ranging from the evaluation of market structure changes (e.g., Hendershott et al., 2011) and transaction cost measures (e.g., Hasbrouck, 2009), to asset pricing (e.g., Korajczyk and Sadka, 2008), corporate finance (e.g., Fang et al., 2009), and macroeconomics (e.g., Næs et al., 2011). In addition, the effective bid-ask spread has regulatory status in Rule 605 of the US Regulation National Market Systems (Reg NMS), which requires all exchanges to publish their execution costs on a monthly basis.

Conceptually, the effective bid-ask spread measures the cost of immediate execution, defined as twice the difference between the transaction price and the fundamental value. Whereas transaction prices are widely disseminated in financial markets, the fundamental value is unobservable. Empirical implementations of the effective bid-ask spread instead rely on the average of the best bid and ask prices, known as the "midpoint," as its benchmark (Blume and Goldstein, 1992; Lee, 1993). The use of the midpoint as a proxy for the fundamental value goes back to Demsetz (1968). I refer to the conceptual definition as the "effective spread" and to its conventional estimator as the "midpoint effective spread."

*E-mail address:* bjh@sbs.su.se

[1] Björn Hagströmer is at Stockholm Business School, Stockholm University, and a visiting researcher at the Swedish House of Finance. I thank the editor (Bill Schwert), an anonymous referee, Jonathan Brogaard, Petter Dahlström, Thierry Foucault, Jungsuk Han, Peter Hoffmann, Anqi Liu, Albert Menkveld, Lars Nordén, Olga Obizhaeva, Andreas Park, Angelo Ranaldo, Kalle Rinne, Ioanid Roşu, Patrik Sandås, Paul Schultz, Ingrid Werner, and Bart Yueshen, as well as seminar and conference participants at the Bank of America Global Quant Conference, Central Bank Workshop on Market Microstructure, CEPR-Imperial-Plato Market Innovator Conference, ECB, HEC Paris, Lund University, NBIM, the SEC Annual Conference on Financial Regulation, Stockholm Business School, the Swedish House of Finance, University of Luxembourg, and Warwick Frontiers of Finance Conference for helpful comments. I thank Nasdaq and Frank Hatheway for providing the HFT data. Nasdaq makes the data freely available to academics providing a project description and signing a nondisclosure agreement. I am grateful to the Jan Wallander Foundation and the Tom Hedelius Foundation for research funding. A previous version of the paper was circulated under the title *Overestimated Effective Spreads: Implications for Investors*.

https://doi.org/10.1016/j.jfineco.2021.04.018
0304-405X/

This paper challenges the use of the midpoint as benchmark for transaction cost measurement. I show that the midpoint effective spread overestimates the illiquidity of US equity markets. The bias varies systematically across stocks, trading venues, and investor groups, and it undermines liquidity timing and trading performance evaluations. I propose alternative estimators that mitigate the bias and can help non-sophisticated investors to reduce their execution costs.

The midpoint effective spread bias can be illustrated by a simple example. Consider a stock with a fundamental value of USD 25.0025 that has liquidity supplied at the nearest prices where trading is allowed, USD 25.00 and USD 25.01. The effective spread is then asymmetric. For trades executed at the bid price it is half a cent ($2 \times 0.25$ cents), whereas for ask-side trades it is three times higher, 1.50 cents ($2 \times 0.75$ cents). If investors factor in the cost asymmetry in their trading decisions, market orders are in this example more likely to arrive on the bid side than on the ask side. The effective spread is then, on average, smaller than the midpoint effective spread (which is one cent).

The problem with the midpoint is that the fundamental value of a security is a continuous variable, whereas observed prices are discrete. Gradual value changes are thus reflected in the midpoint only when they trigger a price change. The minimum incremental price change, known as the tick size and equal to one cent for most US stocks, also constrains the ability of market makers to quote prices symmetrically around the fundamental value (Anshuman and Kalay, 1998). The cost of immediacy for a buy market order thus often differs from that of a sell order. If traders respond to the cost asymmetry by trading more on the side of the market where the effective spread is tighter, as modeled by Goettler et al. (2005), the liquidity demand is elastic. This induces a positive bias in the midpoint effective spread.

*Approach.* I derive a model-free unbiasedness condition for effective spread estimators. The midpoint effective spread is unbiased only if the direction of trade is uncorrelated to the difference between the midpoint and the fundamental value. Zero correlation can be expected either if the market order flow is unaffected by subtick fluctuations in trading costs (i.e., if the liquidity demand elasticity is zero) or if investors are unable to infer fundamental value deviations from the midpoint. Otherwise, the expected midpoint effective spread overestimates the expected effective spread. Importantly, the expected bias is positive for both buy and sell trades and is thus not mitigated by averaging across a large set of trades.

I propose two alternative effective spread estimators that overcome the problem of discrete prices. For these, I rely on continuous fundamental value proxies that factor in the relative quantities posted at the best bid and ask prices, which I refer to as the "order book imbalance." The motivation is that the depth at a given price depends on a tradeoff between the liquidity suppliers' revenue at execution (the expected effective spread) and the costs of trading with informed traders (as in Glosten, 1994). If the bid-side spread is tight, market makers infer that the potential revenue is low and then quote relatively low bid-price quantities. Empirical evidence of such quoting schemes are provided by Kavajecz (1999) and Sandås (2001).

To preserve space in this introduction, I focus here on the results for an estimator called the "weighted midpoint effective spread." The weighted midpoint is a fundamental value proxy that is straightforward to compute from widely available data and is commonly applied in the financial industry. In the paper, I also report results for the "micro-price," an estimator proposed by Stoikov (2018), which adjusts the observed midpoint for expected future midpoint changes. The results for the two estimators are highly consistent.

My main empirical evaluation is based on a one-week sample (December 7–11, 2015) of trades and quotes in the S&P 500 index stocks. To study how the bias differs across investors, I also access a data set on US equity trading provided by Nasdaq, which flags all trades executed by high-frequency traders (HFTs). The HFT indicator allows me to see whether investors with higher market monitoring sophistication are better at coping with the midpoint effective spread bias. Finally, I expand the analysis to a 20-year sample (1996–2015) of all NYSE common stocks.

*Results.* The midpoint effective spread unbiasedness condition is not fulfilled in the data. I find that the direction of trade is significantly related to the fundamental value deviation from the midpoint. For example, when the midpoint is one basis point higher than the fundamental value (as in the example above), only 23% of all trades are buyer-initiated (paying the wide side of the spread). This leads to overestimation of the effective spread.

The weighted midpoint effective spread, which I use as benchmark in my evaluation, averages 2.84 basis points (bps) for the sample stocks. The midpoint effective spread is 3.22 bps. Though the nominal difference of 0.38 bps may seem small, in relative terms it is 13%. According to the micro-price effective spread, the bias is 18%. In the 20-year sample, I find that the midpoint effective spread bias is close to zero in 2001, then grows gradually until 2009, when it plateaus at around 17%, on average, across all stocks. Though the bias is most emphasized for large and liquid stocks, small stocks are also affected. The average bias for 2009–2015 is 18% for large-caps, 12% for mid-caps, and 7% for small-caps.

The economic significance of the bias may be gauged by studying its implications for academic research. I find that the bias influences applications based on cross-sectional liquidity comparisons, such as liquidity-sorted portfolios, trading venue rankings, and trading performance evaluations across investor groups.

The bias influences liquidity-sorted portfolios because it varies with price discreteness. Stocks with high "relative tick size" (tick size divided by share price) feature greater asymmetries between bid- and ask-side effective spreads (Anshuman and Kalay, 1998). In US equities priced above one dollar, where the tick size is fixed at one cent, variation in the relative tick size is driven entirely by the share price. I find that the lowest-priced S&P 500 stocks (below USD 15) have a bias of 52% on average (and 97% when benchmarking to the micro-price effective spread). The bias remains statistically significant for price levels up to USD 115, representing 76% of the S&P 500 trading volume.

Trading venue rankings are influenced because the bias varies across exchanges. For example, the average bias for trades at NYSE is 16%, whereas at Nasdaq BX it is only 7%. With the decentralized market structure prevailing in virtually all asset classes (Johnson, 2010), this variance in bias is important because venues are often compared to achieve best execution and to evaluate the merit of different market mechanisms. The Securities and Exchange Commission (SEC) motivates the Rule 605 reporting requirements by noting that they facilitate the individual investors' ability to assess execution quality across exchanges (Commission and Exchange, 2001).[2] I compare venue rankings across estimators and find that the midpoint effective spread rankings coincide with the benchmark in only 39% of the stock-days.

Finally, the bias feeds through to trading performance evaluations when investors differ in their monitoring of the fundamental value. In the Nasdaq HFT sample, I find that liquidity-supplying HFTs earn 15% *higher* effective spreads, relative to non-HFTs, when the weighted midpoint estimator is used. When instead the midpoint effective spread is applied, the opposite result holds: HFTs earn significantly *lower* rents in their liquidity supply (3% less than non-HFTs). The evidence shows that different effective spread estimators can yield diametrically opposite conclusions.

Can the alternative estimators of the effective spread improve trading performance? I simulate liquidity timing strategies of traders that differ only in their method of tracking the fundamental value. I find that investors who base their liquidity timing on the weighted midpoint not only outperform the midpoint trader in terms of effective spreads, but also achieve lower execution shortfall, which is a transaction price-based measure that does not rely on the accuracy of the fundamental value estimator. The evidence indicates that investors who focus on the midpoint effective spread estimator overpay for liquidity.

A key question when evaluating the importance of the results presented here is to what extent investors are aware of the bias. The mere existence of the midpoint effective spread bias shows that *some* investors are able to infer the fundamental value and adapt their order flow accordingly. If no one did, the orders would be randomly distributed between the bid and the ask sides, and the average effective spread would be unbiased. At the other extreme, if *all* investors observed the bid-ask spread asymmetries, the results documented here would serve to illuminate their trading decisions, but the practical implications would be limited. The results from the HFT analysis shows that the reality is somewhere in between, with systematic differences between investor groups. From a policy perspective, this makes the academic and regulatory recognition of the midpoint effective spread potentially problematic. The support of that measure may lull non-sophisticated investors into a false sense of confidence, which could amplify trading performance differences between investors.

To level the playing field, regulators or brokers could disseminate fundamental value estimates to market participants in real time. This would facilitate liquidity timing for the least sophisticated investors, who are otherwise unable to track sub-tick value fluctuations. For US equities, the fundamental value estimator could be based on the national best bid and offer (NBBO) data disseminated through the Consolidated Quote System. All data required for the weighted midpoint and the micro-price estimators, prices and volumes of the NBBO, are available continuously in that feed. Furthermore, the Rule 605 effective spread definition may be amended to avoid adverse effects of the bias.

*Contribution.* My findings add to the literature on the measurement of effective spreads, including the early work by Blume and Goldstein (1992), Lee (1993), and Petersen and Fialkowski (1994). The bias is consistent with the simulated limit order book market evidence by Goettler et al. (2005), and evidence for equity options by Muravyev and Pearson (2020).

Moreover, the results have implications for the liquidity measurement literature more generally. Roll (1984), Hasbrouck (2009), Corwin and Schultz (2012), and Abdi and Ranaldo (2017) develop effective spread proxies based on daily equity data, and Goyenko et al. (2009) and Jahan-Parvar and Zikes (2019) evaluate such estimators. Holden and Jacobsen (2014) show that the use of intraday data from the Monthly Trade and Quote (MTAQ) database results in distorted estimates of the effective spread. My findings imply that the benchmark used for all these illiquidity measurement evaluations, the midpoint effective spread, is itself a biased estimator.

The paper also contributes to the literature on the motives for initiating trades by submitting market orders. Sarkar and Schwartz (2009) report that market orders are more frequent on one side of the book when information asymmetries are high (e.g., ahead of merger news). Parlour (1998), Foucault et al. (2005), and Roşu (2009) model how traders choose their order types based on a tradeoff between the costs of crossing the spread with a market order, and the waiting costs associated with limit orders. Complementing their work, I show that the asymmetry of the bid- and ask-side effective spreads is an important determinant of market order submissions.

Finally, this paper contributes to the discussion about how price discreteness influences trading strategies and market quality. Budish et al. (2015) argue that time priority rules give high-frequency traders an advantage in trading securities with high tick size relative to share price. Yao and Ye (2018), and O'Hara et al. (2018) present empirical evidence in support of this conjecture. My finding that the effective bias is concentrated to such stocks adds another source of complexity facing non-sophisticated traders.

[2] The European Union has a similar rule. According to Directive 2014/65/EU in financial instruments (MiFID II), each trading venue and systematic internaliser should make midpoint effective bid-ask spread statistics available to the public. See RTS 27, Article 2, available at http://ec.europa.eu/finance/securities/docs/isd/mifid/rts/160608-rts-27_en.pdf.

## 2. Empirical framework

In this section, I derive a model-free condition for when effective spread estimators are unbiased, and discuss high-frequency proxies for the fundamental value of a security.[3]

### 2.1. Bias in effective spread estimators

In the presence of trading frictions, the transaction price $P$ typically differs from the fundamental value $X$. The effective spread quantifies the difference, and may be viewed as a premium paid for the service of immediacy in securities trading. The nominal effective spread is defined as

$$S = 2D(P - X), \tag{1}$$

where $D$ is a direction of trade indicator taking the value +1 for buyer-initiated trades, and -1 for seller-initiated trades. The multiplication by two is for consistency with the quoted bid-ask spread (defined below for a hypothetical roundtrip trade). For ease of exposition, I suppress stock and time subscripts for all variables in this section.

Because the fundamental value at the time of transaction is unobservable, the effective spread is estimated relative to a proxy, which I denote $\tilde{X}$. An effective spread estimator can then be defined as

$$\tilde{S} = 2D\left(P - \tilde{X}\right). \tag{2}$$

Various fundamental value proxies are distinguished with the superscript $v$, $\tilde{X}^v$. For example, I denote the midpoint $\tilde{X}^{mid}$. Similarly, an effective spread estimator utilizing the fundamental value estimator $v$ is denoted $\tilde{S}^v$. The midpoint effective spread, as defined by Blume and Goldstein (1992) and Lee (1993) as well as in the RegNMS Rule 605, is thus denoted $\tilde{S}^{mid}$.

An effective spread estimator is unbiased if the expected difference between the expressions in (1) and (2) is zero. The expected difference is

$$\mathrm{E}\left[\tilde{S} - S\right] = 2\mathrm{E}\left[D\left(X - \tilde{X}\right)\right], \tag{3}$$

implying that the effective spread estimator is unbiased if and only if $D$ and $(X - \tilde{X})$ are uncorrelated. This can be expected either if investors are unable to assess the sign of $(X - \tilde{X})$, or if the liquidity demand elasticity is zero.

Consider again the example in the introduction. When the fundamental value ($25.0025) is closer to the best bid price ($25.00) than to the best ask price ($25.01), the effective spread for sell market orders is tighter than that for buy market orders. If investors then submit more sell than buy market orders, there is a positive correlation between $\left(X - \tilde{X}^{mid}\right)$ and $D$. According to (3), such a correlation implies that the midpoint effective spread is overestimated.

### 2.2. Fundamental value estimators

*The midpoint.* The fundamental value of a security is an elusive but central concept in finance, and approximation methods vary widely. In market microstructure, the midpoint is the most common fundamental value proxy, with applications ranging from liquidity measurement (including the effective spread) to price discovery, realized volatility, and returns.

The midpoint is defined as

$$\tilde{X}^{mid} = \frac{P^A + P^B}{2}, \tag{4}$$

where $P^A$ and $P^B$ are the best bid and ask prices in the limit order book.

The appeal of the midpoint is arguably data availability and simplicity. Data on the best bid and ask prices are publicly available for many asset classes and market types (both auction and dealer markets) and in long time series. In markets where the quotes are valid until canceled, such as limit order book markets, midpoint observations are available continuously during trading hours. Furthermore, the midpoint is straightforward to compute in real time, and is easy to understand for all market participants.

The midpoint has, however, two important shortcomings. First, theoretical evidence shows that liquidity suppliers do not set their quotes symmetrically around the fundamental value when prices are discrete (Anshuman and Kalay, 1998) or when their inventory deviates from the preferred level (Hendershott and Menkveld, 2014). Second, a proxy of the fundamental value should ideally factor in expectations of future price changes (see, e.g., the discussion in Hasbrouck, 2002). The midpoint reflects contemporaneous prices only.

*The weighted midpoint.* The weighted midpoint is defined as

$$\tilde{X}^{wm} = \frac{P^B Q^A + P^A Q^B}{Q^A + Q^B}, \tag{5}$$

where $Q^A$ and $Q^B$ are the depths quoted at the best ask and bid prices, respectively.

The weighted midpoint is a continuous variable, which is appealing for a proxy of fundamental value. Relative to the midpoint, the additional data required to calculate the weighted midpoint are the quantities posted at the best bid and ask prices. Such data are available to investors through the NBBO feed, and to academics through the databases *Daily Trade and Quote* (DTAQ, distributed by NYSE) and *Tick History* (TH, distributed by Refinitiv). Finally, the linear specification makes real-time computation feasible.

Why the order book imbalance is potentially useful to track the fundamental value is best understood from a liquidity-supplier perspective. Although the bias considered here is driven in large part by the price sensitivity of liquidity demanders, the depths at the best quotes are determined by the distance to the fundamental value. In the model by Glosten (1994), the optimal depth in the limit order book is based on a tradeoff between the revenues expected from earning the effective spread, and the costs of trading with informed market orders. In a setting where liquidity demanders are potentially informed, and where the market order size depends on their marginal valuation and the terms of trade offered in the limit order book, Glosten (1994, Proposition 2) shows that the depth

[3] As discussed by Hasbrouck (2002), alternative terminology for the fundamental value include “efficient price,” “true price,” or “consensus price.”

posted at a given price level, in equilibrium, is increasing in the distance to the fundamental value.[4] That is, if the bid depth is lower than the ask depth, it indicates that the fundamental value is closer to the bid than to the ask price.

This conclusion is supported empirically by Sandås (2001), who analyzes Swedish stocks traded in a pure limit order book setting. Kavajecz (1999) shows that NYSE specialists shape their liquidity supply schedules following the same logic. Cont et al. (2014) document that their measure of changes in the order book imbalance can explain 65% of the variation in US equities midpoint changes, and Gould and Bonart (2016) find that the order book imbalance is useful in predicting the direction of the next midpoint change.

*The micro-price.* Stoikov (2018) recognizes the role of the order book imbalance, but criticizes the weighted midpoint for not being a martingale and for occasionally generating undesirable features. A counterintuitive example, he argues, is that a *lowered* best bid price can potentially lead the weighted midpoint to indicate a *higher* fundamental value. He proposes an alternative estimator, the micro-price, which incorporates expectations of future midpoint changes conditional on the state of the limit order book. The intuition is as follows.

Defining the current state as the combination of the midpoint, the quoted spread, and the order book imbalance, the probability of next-period combinations of the same three variables can be estimated using data on what happened when the market was in the same state historically. For any next-period state that does not imply a change in the midpoint, Stoikov (2018) repeats the procedure. In the limit, the result is a probability tree where all branch endpoints are associated with midpoint changes. The micro-price is the weighted average future midpoint implied by the probability tree.

Formally, the micro-price is given by

$$\tilde{X}^{mic} = \tilde{X}^{mid} + g\left(P^A - P^B, \frac{Q^B}{Q^B + Q^A}\right), \tag{6}$$

where $g(\cdot)$ is a function that adjusts the current midpoint for expected future midpoint changes. The value of this adjustment function depends on the quoted spread, defined as $QS = P^A - P^B$, and the order book imbalance, $I = Q^B/(Q^B + Q^A)$. The adjustment is determined by discretizing the two variables and treating combinations thereof as a finite state space. To evaluate the adjustment function at infinity, Stoikov (2018) analyzes the state space as a discrete time Markov chain with absorptive states. The absorptive states are given by midpoint changes of different magnitudes, and correspond to the branch endpoints in the probability tree. I provide extensive estimation details for the micro-price in Appendix A.[5]

The micro-price is based on the same data inputs as the weighted midpoint, but it is more complex to compute. The added benefit is that it is a martingale by construction. Furthermore, Stoikov (2018) reports that the micro-price outperforms the weighted midpoint in predicting future midpoint changes.

### 2.3. Alternative estimators of the effective spread

I apply the two alternative fundamental value proxies presented above to estimate the effective spread as defined in (2). I refer to the resulting estimators as the "weighted midpoint effective spread" ($\tilde{S}^{wm}$) and the "micro-price effective spread" ($\tilde{S}^{mic}$), respectively. As a group, I refer to them as "alternative effective spread estimators".

To my knowledge, none of the these estimators have been applied in the previous literature. Cartea et al. (2015, p. 71) suggest that the weighted midpoint could be a more economically meaningful benchmark than the midpoint when accounting for the effective spread in algorithmic trading, but they do not elaborate further on the issue.

Which estimator to use depends on the institutional setting, data availability, and computational resources. I offer a discussion about this choice in Section 8.

## 3. Data and sample

I use the TH database to access trades and quotes for US equities.[6] For sample selection purposes, I use stock characteristics available in monthly data from the Center for Research in Security Prices (CRSP). In addition, I use a database provided by Nasdaq, described below.

I consider three samples:

- The baseline sample includes one trading week (December 7–11, 2015) for the S&P 500 index stocks. During this sample period, the S&P 500 index consists of 506 stocks, all available in TH. I include trades from all relevant US national securities exchanges.[7] Trades in dark pools and over-the-counter markets are not included. I refer to this data set as the "S&P 500 sample."
- To study time-series and cross-sectional variation in the effective spread bias, I also consider a 20-year sample (Jan. 1, 1996–Dec. 31, 2015) of common stocks with their primary listing on the NYSE, NYSE Mkt, or ARCA.

[4] This holds in many, but not all, model specifications. Glosten (1994, p. 1137) presents a counterexample. Back and Baruch (2013) show that the equilibrium result requires sufficient adverse selection.

[5] Examples are available in the internet appendix (Section IA.A), which also includes a link to replication code for the micro-price as well as other calculations in this article.

[6] The TH database is not commonly used for US equity research but it is based on the same data sources as the DTAQ database. The trades come from the consolidated tape, and the quotes from the NBBO feed. For details on the TH sources and a comparison to the DTAQ data, see the internet appendix (Section IA.B).

[7] At the time of the S&P 500 sample, there were 12 national exchanges. Bats Global Holdings, Inc. (Bats; subsequently taken over by CBOE Holding, Inc.) owned four: BZX, BYX, EDGA, and EDGX. Nasdaq Inc. (Nasdaq) also held four: BX, PHLX, NASDAQ, and NASDAQ Intermarket. The latter two traded non-overlapping segments of stocks, and I report their activity jointly using the abbreviation NASD. Intercontinental Exchange, Inc. (ICE) owned three exchanges: NYSE, NYSE Mkt, and ARCA. The former two traded non-overlapping stocks, and I report their activity jointly using the abbreviation NYSE. Only one exchange was independent at the time, Chicago Stock Exchange (CSX; subsequently acquired by ICE). In TH, the corresponding exchange identifiers (in order of appearance in this footnote) are BTY, BAT, DEA, DEX, BOS, XPH, NAS, THM, NYS, ASE, PSE, and MID.

For a stock-month to be included in the sample, I require the average trade price to be greater than USD 5 and lower than USD 1000. Out of the stocks in the CRSP database that fulfil these criteria, 99.8% are also available in the TH database (see Appendix B for details on the matching methodology). Following Yueshen (2016), I restrict the time series of trading days to Wednesdays to limit the computational cost. As for the baseline sample, I use trades from all relevant exchanges (the set of relevant exchanges varies over time) and quotes from the official NBBO feed. I refer to this data set as the "time-series sample."

- To analyze differences across investor groups, I use a data set provided by Nasdaq, reporting all trades for 120 stocks along with a flag that indicates whether the active and the passive counterparty (or both) of a transaction is a high-frequency trader (HFT) or not (non-HFT). I use the latest trading week available in the data set, February 22–26, 2010, and refer to it as the "HFT sample." As the data set does not contain NBBO quotes, I match it to trades from TH, which are then straightforward to match to TH quotes.

I obtain the direction of trade indicator $D$ using the Lee and Ready (1991) algorithm, noting that Chakrabarty et al. (2015) show that the procedure performs well in a recent US equities sample. In the HFT sample, the direction of trade is directly observable. Reassuringly, all the conclusions of the HFT sample analysis remain unchanged when using the Lee and Ready (1991) algorithm instead of the observed variable.

The following screening is applied to all samples. To avoid opening and closing effects in the measurement of liquidity, I exclude trades executed in the first and the last five minutes of the trading day. I also remove block trades, defined as trades of at least 10,000 shares. Additional screens, excluding less than 0.01% of all trades, are described in Appendix B. Each trade observation contains information on the date, stock, time, price, volume, and trading venue. The S&P 500 sample contains 55.7 million trades, and the HFT sample holds 2.4 million trades. The time-series sample is much larger, with 4.5 billion trades.

The trades are matched to the last quote observation in force at the end of the preceding millisecond, as recommended by Holden and Jacobsen (2014). After the trade and quote matching, several screens are applied to exclude invalid quotes; see Appendix B.

The quotes contain the NBBO prices and depths, as well as the trading venue contributing the best quote on each side of the market. If there are several venues with quotes at the best price, the quote of the venue with the highest depth is reported, not the aggregate depth across venues. To use NBBO quotes for effective spread measurement is consistent with Rule 605.[8] The results of the paper do not change qualitatively if quoted depths are instead aggregated across trading venues (see the internet appendix, Section IA.C).

All spread measures in the paper are winsorized within each stock at the 1% and the 99% quantiles.

## 4. Main results

In this section, I confirm that the liquidity demand is elastic and quantify the resulting midpoint effective spread bias. I also show that the bias varies systematically in the cross-section of stocks and exchanges and that it has increased over time.

### 4.1. Liquidity demand elasticity

I assess the liquidity demand elasticity by investigating how market order arrivals depend on the "fundamental value deviation from the midpoint," defined as $\log \tilde{X}^{v} - \log \tilde{X}^{mid}$ and expressed in basis points (where log indicates natural logarithms). I categorize trades in the S&P 500 sample by the fundamental value deviation from the midpoint prevailing just before the trade. I create 21 trade categories using the following breakpoints: -2.1 bps, -1.9, ..., -0.1, +0.1, ..., +1.9, +2.1. The categories are labeled by the midpoint of their interval. For example, all trades where the fundamental value deviation from the midpoint lies within the interval (1.9,2.1] are put in the *+2.0 bps* bucket.[9]

Fig. 1, Panel (a), shows the frequency of buyer-initiated trades for each trade category. The black and gray lines correspond to the weighted midpoint and micro-price estimators of fundamental value, respectively. The null hypothesis is that market orders arrive independently of the midpoint deviation, as indicated by the dashed horizontal line.

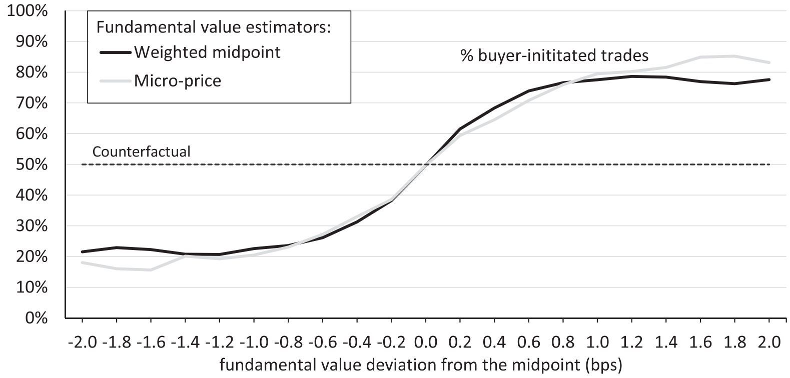
The results show that the probability of buyer-initiated trades tends to increase with the fundamental value deviation from the midpoint. The relation is monotonic for midpoint deviations that do not exceed one basis point. For example, consider the case when the midpoint deviates by -1 basis point from the fundamental value. This category corresponds to the example given in the introduction (with a midpoint deviation of -0.25 cents in a stock valued at USD 25.0025). For this case, I find that only 23% of all trades are buyer-initiated, according to the weighted midpoint proxy of fundamental value. For trades in the *+1.0 bps* category, in contrast, 78% of the trades are buyer-initiated. When the midpoint is close to the fundamental value, the split between buyer- and seller-initiated trades is even. Similar results hold when the micro-price is used as fundamental value estimator. Note that the symmetry around zero on the x-axis is a feature of the data, it is not imposed by the econometrician.

To formally assess the relation between the direction of trade and the fundamental value deviation from the midpoint, I estimate a probit model,

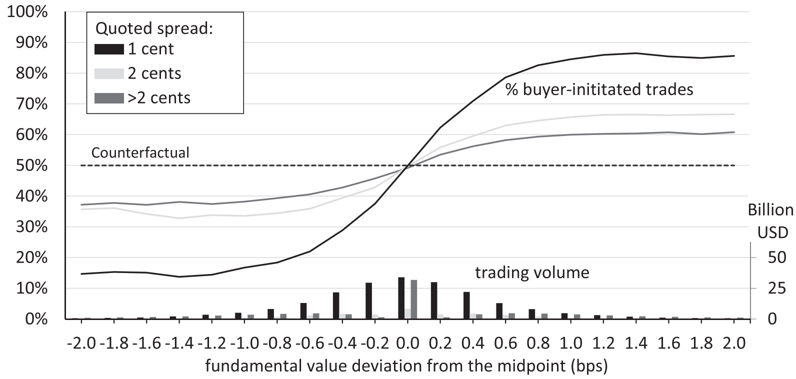
$$\Pr(Buy_t) = \underset{-0.43}{-0.00} + \underset{8.43}{0.45}\left(\log \tilde{X}_t^{wm} - \log \tilde{X}_t^{mid}\right) + \varepsilon_t. \quad (7)$$

[8] The technical details of the Rule 605 report requirements are in §240.11Ac1-5, available at https://www.sec.gov/rules/final/34-43590.htm. For the effective spread, see section (a)(2).

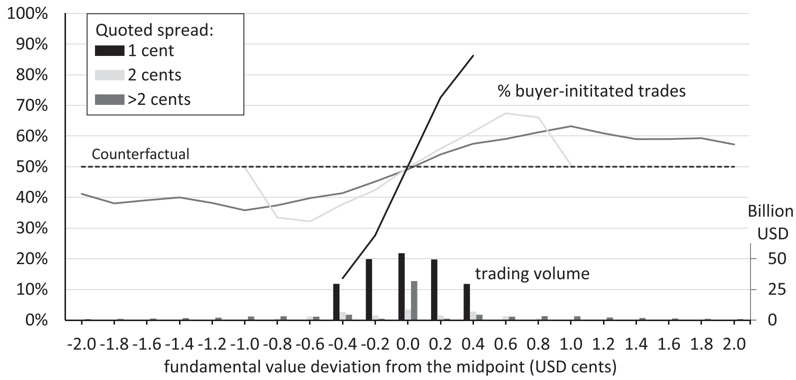
[9] The interval $(-2.1, 2.1]$ spans 93% of the sample trades when defined relative to the weighted midpoint. The corresponding number for the micro-price is 91%. The internet appendix, Section IA.D, shows results for the tails of the distribution.




(a) Fundamental value deviation in relative terms, across estimators




(b) Fundamental value deviation in relative terms, across spread levels




(c) Fundamental value deviation in nominal terms, across spread levels

**Fig. 1.** *Liquidity demand elasticity in the S&P 500 stocks.* This figure shows the frequency of buyer-initiated trades for different trade categories of the fundamental value deviation from the midpoint. The direction of trade is determined by the Lee and Ready (1991) algorithm. Panel (a) presents results for two alternative fundamental value proxies ($\tilde{X}^v$): the weighted midpoint and the micro-price (defined as in Section 2.2). The trade categories are based on the relative fundamental value deviation from the midpoint, defined as $log(\tilde{X}^v) - log(\tilde{X}^{mid})$, expressed in basis points. Trades are categorized by the breakpoints $-2.1$ bps, $-1.9, \ldots, -0.1, 0.1, \ldots, 1.9, 2.1,$ and the x-axis labels refer to the midpoint of each interval. Panel (b) presents results for three quoted spread levels: 1 cent, 2 cents, and more than 2 cents. All results in this panel are based on the weighted midpoint proxy for fundamental value. The trade categories are the same as in Panel (a). The panel also includes a bar chart of the aggregate dollar trading volumes for each trade category and each spread level, measured in billion USD on the right vertical axis. Panel (c) repeats the analysis of Panel (b), but with trade categories based on the nominal fundamental value deviation from the midpoint, defined as $\tilde{X}^{wm} - \tilde{X}^{mid}$ and expressed in USD cents. Trades are categorized by the breakpoints $-2.1$ cents, $-1.9, \ldots, -0.1, 0.1, \ldots, 1.9, 2.1$. The sample includes trades for all constituents of the S&P 500 index for the five trading days in the period December 7–11, 2015.

where the fundamental value is proxied by the weighted midpoint ($\tilde{X}_t^{wm}$), $t$ is a trade index, $Buy_t$ equals one for buyer-initiated trades and zero for seller-initiated trades, and $\varepsilon_t$ denotes the residuals. The estimation results, reported in (7), indicates a positive and significant relation between the direction of trade and the fundamental value deviation from the midpoint (the $z$-statistics, within parentheses, are based on standard errors that are clustered by stock, date, and trading venue, following Petersen, 2009). For the micro-price, the corresponding slope estimate is almost identical to that of the weighted midpoint, 0.44, with $z$-statistic 5.32.

The analysis in Section 2.1 shows that the midpoint effective spread is unbiased when either the liquidity demand elasticity is zero, or when investors are unable to infer the sign of the midpoint deviation; see Eq. (3). The empirical evidence strongly rejects the unbiasedness condition, implying that the midpoint effective spread is biased upwards.

Behind the averages reported in Fig. 1, of course, lies a range of trading strategies adopted by investors that are diverse in terms of sophistication and holding period. Indeed, the effective spread is often measured for individual strategies and clienteles as part of performance comparisons. What is important to note here is that the conclusion that the effective spread is biased does not depend on the underlying reason for the liquidity demand to be elastic.

For example, several papers analyze the optimal order choice by liquidity demanders in dynamic models of the limit order book (e.g., Parlour, 1998; Foucault et al., 2005; Roşu, 2009). They show that the choice between market and limit orders depends on the tradeoff between execution risk and waiting costs associated with passive trading, and the spread cost of active trading. In (Parlour, 1998) model, incoming traders assess the probability of a limit order execution by evaluating the depth at both the best bid and the best ask prices. Though that model poses an alternative story for the pattern observed inFig. 1, it does not alter the conclusion that such order flow regularities generate a bias in the midpoint effective spread.[10],[11]

Before quantifying the effective spread bias, I investigate the role of price discreteness in Fig. 1. Panel (b) disaggregates the trades by how constrained the best bid and ask prices are by the minimum tick size. Specifically, it shows separate probability curves for trades taking place when the nominal quoted spread equals one cent, two cents, and more than two cents. Furthermore, it reports trading volumes for each spread level and each trade category. All results in this panel are based on the weighted midpoint proxy for fundamental value.

The relation between the fundamental value deviation from the midpoint and the frequency of buyer-initiated trades is strongest for trades where the quoted spread is one cent. This category (the black line in the figure) features the S-shape that is familiar from Panel (a), but with a steeper slope.

The one-cent spread category is also highly dominant in terms of trading volume, representing 61% of the total. The two-cent spreads constitute 11%, and the remaining 28% executes in order books with larger spreads. The vertical bars in Panel (b) show that the trading volume (measured on the right vertical axis in billion USD) of the one-cent spread category is relatively dispersed across the trade categories. This result is consistent with Anshuman and Kalay (1998), who show that market makers quote more asymmetrically around the fundamental value when the tick size is binding.

Panel (c) of Fig. 1 repeats the analysis of Panel (b), but with the fundamental value deviation from the midpoint measured in cents instead of basis points. Note here that the maximum deviation for a one-cent spread is half a cent, and for a two-cent spread it can never exceed one cent. The results are mostly consistent with those of Panel (b), but with one notable exception. When the quoted spread is greater than one cent and the fundamental value deviation from the midpoint is high in absolute terms, the imbalance between buyer- and seller-initiated trades is relatively low. For two-cent spreads, for example, two-thirds of the trades in the *+0.8 cent* category are buyer-initiated, whereas the *+1.0 cent* category has a 50/50 split between buyers and sellers.

This counterintuitive result emerges in the tails of the order book imbalance distribution and represents very small trading volumes (in the case of two-cent spreads, 1.5% of the dollar volume falls in the $\pm 1.0$ *cent* bins). In the internet appendix, Section IA.D, I show that trades that cross the wide side of the spread in these bins are extraordinarily large. The pattern is consistent with traders utilizing an opportunity to execute large volume without splitting the order across venues or price levels.

### 4.2. The midpoint effective spread bias

Table 1, Panel (a), reports properties of the effective spread measured at the stock level. It contains effective spread estimates using either the midpoint, the weighted midpoint, or the micro-price as fundamental value proxy. In addition, the quoted spread, the trade price, and two measures of the aggregate trading volume are reported. Each spread observation is scaled by the prevailing midpoint and reported in basis points. All stock-level observations are dollar volume-weighted averages across all trades of the given stock, and the cross-sectional mean is, in turn, dollar volume-weighted across stocks.

The results show that the average midpoint effective spread is 3.22 bps. The midpoint effective spread is somewhat tighter than the quoted spread, typically due to executions obtaining price improvements or to hidden liquidity inside the NBBO. The cross-sectional distribution statis-

[10] Another aspect of Parlour (1998) model, as well as of the related models by Foucault et al. (2005) and Roşu (2009), is that the order book imbalance is generated by the patience of liquidity demanders, not the fundamental value. If that holds, the fundamental value proxies based on the order book imbalance are misleading. To verify that direction of trade is related to the fundamental value deviation from the midpoint, I obtain an alternative fundamental value proxy that does not rely on order book imbalance data, see the internet appendix, Section IA.E. Repeating the analysis in Fig. 1, the alternative proxy yields results that are highly consistent with those presented here.

[11] Section 6 holds a simulation of liquidity timing for investors who choose between market and limit orders.

tics in the table show that both the effective and the quoted spread are positively skewed, as the medians are somewhat lower than the means.

The alternative effective spread estimators proposed in this paper are, as expected, lower than the midpoint effective spread. The means of the weighted midpoint and the micro-price effective spreads are 2.84 bps and 2.73, respectively. I define the *Nominal bias* of the midpoint effective spread estimator as its difference to an alternative estimator of the effective spread, $\left(\tilde{S}^{mid} - \tilde{S}^{v}\right)$, and report it in basis points in Panel (b). The *Nominal bias* is on average 0.38 or 0.49 bps, depending on the estimator applied. The *t*-statistics show that the *Nominal bias* is statistically significant for both estimators.

It is also interesting to report the bias in relative terms. I define the *Relative average bias* as the average *Nominal bias* divided by the average effective spread. For the S&P 500 stocks, the *Relative average bias* amounts to 13% when the weighted midpoint effective spread is used as benchmark. When the midpoint effective spread is evaluated relative to the micro-price estimator, the average bias reaches 18%. I show below that the *Relative average bias* in the effective spread varies in the cross-section of stocks and venues, and that it has increased over time.

### 4.3. Variation across stocks

I expect the bias to be increasing with stock liquidity and decreasing with stock price. The reason for this is that liquid, low-priced, stocks in the US equity market are those where the pricing is most constrained by the tick size, which is fixed at one cent.[12] This leads to greater asymmetry between the bid-side and ask-side effective spreads.

To assess the relation between the overestimation and the relative tick size, I split the sample into trade price groups. The *USD10* group includes all trades in the USD 5.01–15 interval; the *USD20* group includes all trades in the USD 15.01–25 interval; and so on, with 10-dollar intervals for each price group. The category with the highest-priced trades considered is *USD190*, including trades in the USD 185.01–195 interval. In the S&P 500 sample, 98% of the trades fall within the price interval USD 5–195. Fig. 2, Panel (a), shows the effective spread relative to each of the three fundamental value proxies listed above. The midpoint estimator is plotted as a dark gray line, and the alternative estimators are plotted as black and light gray lines. The vertical bars show the average dollar trading volume of each share price group (measured on the right axis).

The share price groups from *USD30* to *USD120* span the lion's share of the trading activity (74% of the dollar volume and 77% of the trades). In that price interval, the effective spreads lie in the range 2.0–2.5 bps, on average, depending on the estimator. Stocks in the *USD10* and *USD20* categories have much higher spreads, which may be due to the fact that the tick size is more constraining for them than for higher-priced stocks.

Panel (b) of Fig. 2 shows the *Relative average bias* corresponding to each of the alternative effective spread estimators. The results support the notion that stocks with higher relative tick size have higher biases. The price groups *USD10* and *USD20* display average biases of 52% and 40% when benchmarking to the weighted midpoint estimator of the effective spread. When the micro-price is used instead as the benchmark, the midpoint effective spread bias tends to be even higher. For the lowest-priced stocks, the bias then reaches 97%.

Observations in Panel (b) that are significantly different from zero at the 95% level are marked by "+". The bias is positive and statistically significant for all price groups up to and including *USD110* (for both of the alternative effective spread estimators), corresponding to 76% of the dollar volume and 90% of the trades.[13]

Panel (c) of Fig. 2 zooms in on individual stocks and highlights the importance of liquidity. The most tick-constrained stocks, defined as stocks with a median quoted spread of one cent, are marked by black squares. It is clear from the figure that these stocks dominate the lowest share price groups, and that their bias is high relative to similarly priced but less tick-constrained stocks. Had the price group curve been drawn based on stocks where the median quoted spread exceeds one cent, it would have been flat and close to zero.

I evaluate below the extent to which the bias differences across stocks feed through to liquidity-sorted portfolios; see Section 5.2.

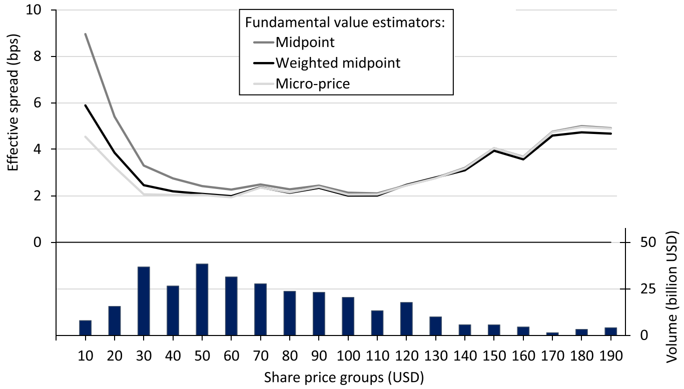
### 4.4. Variation across exchanges

The markets for securities trading in all asset classes are increasingly decentralized (Johnson, 2010). The S&P 500 index trading activity, for example, is split between index futures, ETFs, options, and the underlying stocks (which in turn trade at numerous exchanges as well as at alternative trading systems). The competitive markets for trading venues and security types foster innovation in market design. This motivates regulators, brokers, and academics to compare trading costs across platforms. If the midpoint effective spread bias varies across trading venues, such analyses are potentially altered. Examples of academic papers that compare execution quality between market mechanisms include Huang and Stoll (1996) and Bessembinder and Kaufman (1997). Aldrich and Lee (2018) emphasize the role of price discreteness in execution cost comparisons.

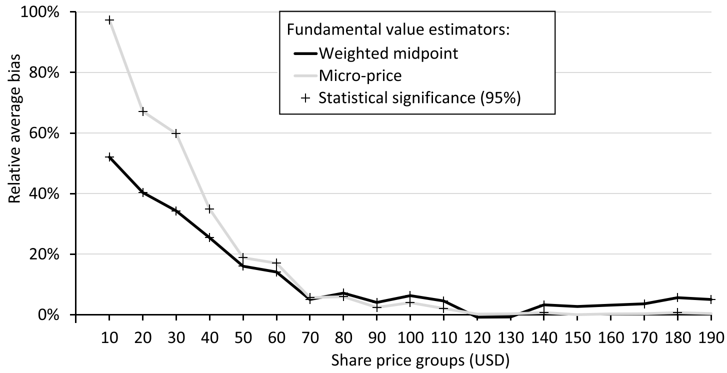
Current US equity market regulation embraces the midpoint as a fundamental value estimator. According to Rule 605 of RegNMS, all exchanges must publish monthly reports of their execution quality for each security traded. The Rule 605 reports include the average effective spread,

[12] Stocks priced below USD 1 have lower tick sizes, but there are no such stocks in the sample.

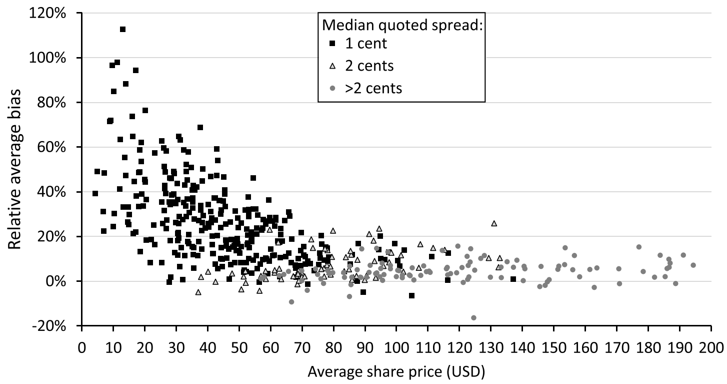
[13] The bias is also positive, around 5%, for stocks priced above USD 130 when the weighted midpoint is used as the fundamental value proxy. According to the micro-price, however, the bias is less than 1% for the high-priced stocks, which is in line with the expectation that the bias increases with price discreteness. This highlights an advantage of the micro-price relative to the weighted midpoint. If order book imbalances are less predictive of future price changes for high-priced stocks than for low-priced stocks, the midpoint adjustment function of the micro-price reflects that. For the weighted midpoint, in contrast, the midpoint adjustment is an invariant function of the quoted spread and the order book imbalance.




(a) Effective spreads




(b) *Relative average bias across estimators*




(c) *Relative average bias* across stocks, based on the weighted midpoint effective spread

**Fig. 2.** *Effective spread properties across trade price groups in the S&P 500 stocks.* Panel (a) shows the effective spread measured relative to the midpoint, the weighted midpoint, and the micro-price, averaged across stocks in the same trade price group. See Table 1 for variable definitions. It also includes a bar chart of the aggregate dollar trading volumes for each share price category, measured in billion USD on the right vertical axis. Each price group corresponds to a price interval of USD 10. For example, the *USD20* price group includes all trades priced higher than USD 15 and lower than or equal to USD 25. Panel (b) reports the *Relative average bias*, defined as in Table 1, for each alternative effective spread estimator and each price group. Observations that are significantly different from zero are marked by "+" (based on standard errors of residuals clustered by stock, date, and trading venue). Panel (c) presents the *Relative average bias* for individual stocks, measured using the weighted midpoint effective spread. The stocks are marked with different symbols depending on their median quoted spread (1 cent, 2 cents, and more than 2 cents). The sample includes all constituents of the S&P 500 index for the date interval December 7–11, 2015.

**Table 1**
*Effective spread properties in the S&P 500 stocks.*
This table shows descriptive statistics of liquidity and trading volume measures in Panel (a), and measures of the midpoint effective spread bias in Panel (b). The reported statistics are based on stock-level measures of each variable. Panel (a) includes the mean, the standard deviation, and the 5th, 25th, 50th, 75th, and 95th percentiles. The effective spread estimator $\tilde{S}^{\nu}$ is twice the difference between the trade price and the fundamental value, multiplied by the direction of trade indicator, and scaled by the midpoint. The superscript $\nu$ denotes the fundamental value estimator used and can be either the midpoint, the weighted midpoint, or the micro-price, defined as in Section 2.2. The effective spread is measured for each stock as the dollar-weighted average across all trades in the sample, excluding trades occurring in the first or last five minutes of the trading day, as well as block trades. The quoted spread *QS* is the difference between the national best ask and bid quotes just before each trade, divided by the midpoint and measured for each stock as the dollar-weighted average across trades. *Trade price* is the dollar-weighted average price across all trades for each stock. The mean reported for all the measures above is also dollar-weighted across stocks. The volume measures, *Number of trades* (measured in thousands) and *Dollar volume* (measured in million US dollars), are reported as equal-weighted averages across stocks. In Panel (b), the *Nominal bias* is the difference between the midpoint effective spread and $\tilde{S}^{\nu}$, reported in basis points. The $t$-statistic corresponding to the null that the value-weighted average *Nominal bias* is equal to zero, based on standard errors that are clustered by stock, date, and trading venue, is reported within parentheses. The *Relative average bias* is the average *Nominal bias* divided by the average effective spread $\tilde{S}^{\nu}$. The sample includes all constituents of the S&P 500 index for the date interval December 7–11, 2015.

(a) Liquidity and trading volume

| | | | Percentiles | | | | |
|---|---|---|---|---|---|---|---|
| | Mean | Std. dev. | 5th | 25th | 50th | 75th | 95th |
| Effective spread est., $\tilde{S}^{\nu}$ (bps) | | | | | | | |
| Midpoint | 3.22 | 2.18 | 1.64 | 2.35 | 3.02 | 4.39 | 7.78 |
| Weighted midpoint | 2.84 | 1.84 | 1.37 | 1.93 | 2.59 | 3.90 | 6.84 |
| Micro-price | 2.73 | 1.82 | 1.21 | 1.85 | 2.53 | 3.78 | 6.99 |
| Quoted spread, *QS* (bps) | 3.55 | 2.57 | 1.71 | 2.46 | 3.23 | 4.92 | 9.13 |
| *Trade price* (USD) | 119.27 | 101.06 | 16.36 | 37.74 | 59.60 | 94.23 | 186.43 |
| *Trade volume* (thousands) | 102.83 | 98.57 | 23.34 | 44.96 | 75.54 | 123.74 | 266.77 |
| *Dollar volume* (millions) | 687.92 | 874.92 | 128.11 | 271.88 | 424.31 | 745.60 | 2060.58 |

(b) Midpoint effective spread bias

| | *Nominal bias* | $t$-stat. | *Relative average bias* |
|---|---|---|---|
| Effective spread estimator, $\tilde{S}^{\nu}$ | | | |
| Weighted midpoint | 0.38 | 8.67 | 0.13 |
| Micro-price | 0.49 | 9.76 | 0.18 |

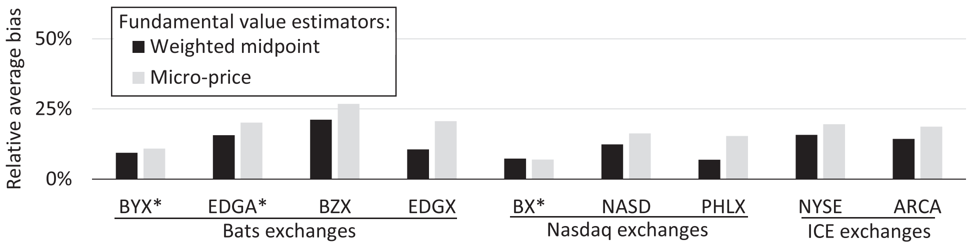



**Fig. 3.** *Midpoint effective spread bias across trading venues.* This figure shows the *Relative average bias*, defined as in Table 1, for each trading venue, and for each of the alternative effective spread estimators (based on the weighted midpoint and the micro-price). The sample includes all constituents of the S&P 500 index for the five trading days in the interval December 7–11, 2015. Exchanges that apply an inverted fee schedule are indicated by *. The exchanges are categorized by their corporate ownership. Exchange names corresponding to the three-letter abbreviations are spelled out in footnote 6.

defined the same way as the midpoint effective spread in this paper. The Commission and Exchange (2001, Section I) motivates the disclosure requirement that order routing decisions across trading venues must be "well-informed and fully subject to competitive forces." To the extent that traders use the Rule 605 data as the SEC intended, bias variation across trading venues can misdirect order routing decisions.

Fig. 3 displays the *Relative average bias* for each trading venue in the sample, and for each of the two alternative effective spread estimators. I exclude CSX (Chicago Stock Exchange), since it represents only 0.01% of the total trading volume. The remaining nine exchanges are categorized by their corporate ownership (they are all held by either Bats, Nasdaq, or ICE). The results uncover substantial differences across exchanges. For example, the bias for BX is around 7%. The corresponding statistic for trades executed at BZX lies in the 21%–27% range, depending on estimator. A similar but stronger results emerges when the analysis is constrained to low-priced stocks (below USD 50). The bias for BX is then in the 13%–20% range, whereas that of BZX is 47%–79% (see the internet appendix, Section IA.F).

What is the economics behind the bias variation across exchanges? A key distinguishing factor of modern equity exchanges is their fee schedules. Most venues subsidize liquidity suppliers by giving rebates to passively executed trades, and charge fees to actively executed trades (known as maker/taker fee schedules). Some venues, however, do the opposite, which is known as inverted fees. Sandås (2001) models the expected profits of liquidity suppliers similarly to Glosten (1994, discussed in Section 2.2), and adds order processing costs. In that setting, a maker rebate boosts the profit margin, which in equilibrium leads to higher quoted depth (as liquidity providers break even in expectation). Absent the rebate, or in the presence of a maker fee, the same depth level would not be sustainable. The consequence is that the asymmetry between bid-side and ask-side spreads is higher in maker/taker fee venues, which may lead to a higher bias. In Fig. 3, venues with inverted fees are indicated by an asterisk (*).[14] The results show that such venues have lower *Relative average bias*.

I explore below how the systematic bias variation across exchanges influences venue rankings; see Section 5.3.

### 4.5. Variation over time: NYSE stocks, 1996–2015

The sample used in the application above is limited to large-cap stocks and one trading week. To make sure that the findings are not sample-specific, and to spot trends over time, I expand the analysis to the time-series sample. There are, on average, 1471 stocks in each sample month, providing a rich heterogeneity in terms of, for example, liquidity and size.

I calculate the monthly midpoint effective spread and weighted midpoint effective spread for each stock in the time-series sample.[15] Fig. 4 presents the time series of the volume-weighted average *Relative average bias* as a solid gray line.[16]

Three patterns stand out. First, as indicated by the vertical lines, there are two tick size reforms in the sample: the change from one-eighth increments on the dollar to one-sixteenth increments on June 24, 1997, and the decimalization implemented at NYSE on January 28, 2001. Consistent with the price discreteness effects discussed in Section 4.3, these tick size reforms are followed by sharp declines in the midpoint effective spread bias. Considering the four weeks immediately before and after each event, I find that the midpoint effective spread falls by 25% in the 1997 reform and by 33% in the decimalization event. The corresponding numbers for the weighted midpoint estimator are 16% and 27%, respectively.[17] Even though the bias does not change the qualitative conclusion of the tick size reform analyses, the level of the effect is heavily influenced.

Second, for December 2015, the *Relative average bias* is 13%, which is the same as that of the S&P 500 sample reported in Table 1. Given the strong heterogeneity in terms of market capitalization in the time-series sample, the close correspondence is not clear ex ante. To shed further light on cross-sectional differences, Fig. 4 also displays volume-weighted average results for market capitalization terciles, labeled as large-caps, mid-caps, and small-caps (see the solid, dashed, and dotted dark lines, respectively). Throughout the sample period, the bias tends to be increasing in market capitalization. From January 2009 to December 2015, when the bias development is relatively flat, the *Relative average bias* for large-caps, mid-caps, and small-caps, is, on average, 18%, 12%, and 7%, respectively.

Third, from mid-2002 to early 2009, the *Relative average bias* displays an upward trend. After being slightly negative in 2001–2002, the average overestimation grows to 17% on average in 2009, where it then flattens out. The upward trend can partially be linked to liquidity improvements in the US equity market. From July 2002 to the dawn of the financial crisis in June 2007, the midpoint effective spread falls by 76%, from 13.3 bps to 3.2 bps. As the bid-ask spread shrinks, the tick size becomes more constraining, and the midpoint effective spread bias rises. Even though the financial crisis in 2007–2009 is associated with elevated illiquidity, the midpoint effective spread bias continues to increase. The reason is that prices fall more than the effective spreads rise, with the net effect being that price discreteness increases. I infer from these observations that there is no clear relation between the midpoint effective spread bias and the overall return of the market.

### 4.6. Economic significance

Establishing that the bias magnitude is high, 13%–18% in the S&P 500 sample, is not enough to claim that it is economically significant. A large bias may still be benign in many effective spread applications, as it is often the relative level of liquidity that is important. For example, Fig. 4 indicates clearly that the tick size changes in 1997 and 2001 influence the bias. But although accounting for the bias changes the magnitude of the effect, it does *not* change the inference that liquidity improves when the tick size is reduced.

When can the midpoint effective spread bias be expected to matter? The bias is more likely to influence the conclusion of applications that rely on the ordering of liquidity, rather than the level. In Section 5, I show that the bias can influence research inference in applications involving trading performance comparisons across investor groups, liquidity-sorted portfolio applications, and rankings of venues based on execution quality. In Section 6, I simulate trading strategies and show that investors' choices of fundamental value proxy directly influence their liquidity timing ability.

[14] I thank Shawn O'Donoghue for sharing information on the exchange fees that are applicable to my S&P 500 sample period. For data details, see O'Donoghue (2015).

[15] Due to its computational cost, I exclude the micro-price estimator from this analysis.

[16] The effective spread levels time series are plotted in the internet appendix, Section IA.H.

[17] Detailed analyses of liquidity effects of the tick size reforms mentioned here are provided by Goldstein and Kavajecz (2000) and Bessembinder (2003).

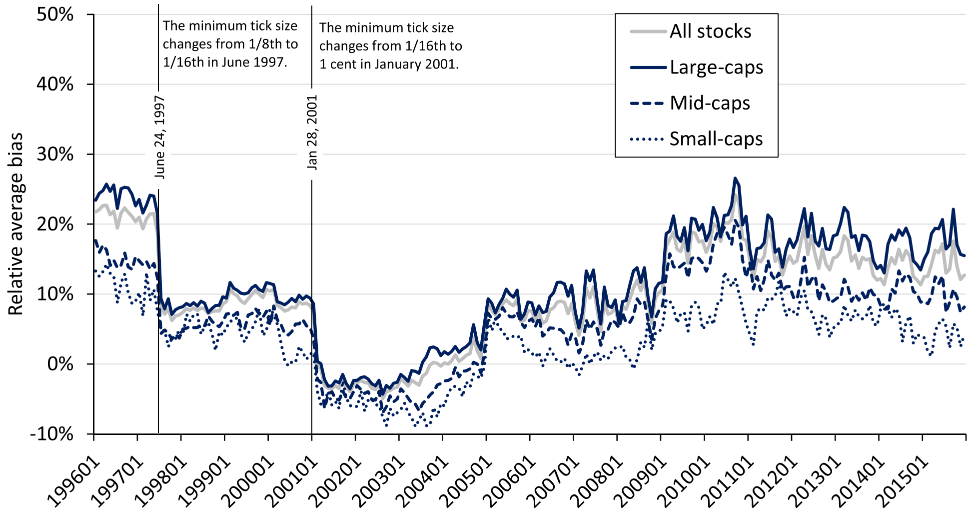



**Fig. 4.** *Effective spread bias in the time series and cross section.* This figure shows monthly time series of the midpoint effective spread bias, for all stocks as well as market capitalization segments. The large-caps, mid-caps, and small-caps segments are terciles formed in each month with the same number of stocks in each segment. The *Relative average bias* is defined as in Table 1. Cross-sectional averages are calculated using dollar volume weights. The sample includes common stocks with primary listings on the NYSE, NYSE Mkt, or ARCA, from January 1996 to December 2015.

## 5. Does the bias affect research inference?

### 5.1. Trading performance of HFTs vs. non-HFTs

To analyze how investors differ in their understanding of the bias, I consider the HFT sample. HFTs are known to invest heavily in technology in order to monitor and respond to information in real time (Brogaard et al., 2015; Commission and Exchange, 2001; Shkilko and Sokolov, 2020), and to have strong intraday market timing ability (Carrion, 2013). They are also able to predict price changes in the short term (Brogaard et al., 2014), which implies that they factor in probabilities of future midpoint changes in their analyses (similar to how the micro-price is constructed). The HFTs in this sample may thus be viewed as sophisticated in terms of market monitoring. The non-HFTs are more heterogeneous, with trade flows originating from retail and institutional investors as well as from proprietary traders that are not categorized as HFTs. On average, the non-HFTs are certainly less sophisticated than the HFTs in terms of limit order book monitoring.

I present descriptive statistics for the HFT sample in the internet appendix, Section IA.G. In summary, it features more dispersed liquidity relative to the S&P 500 sample, because the HFT sample is stratified across market capitalization segments. The *Relative average bias* in the HFT sample is relatively high. For the weighted midpoint and micro-price estimators, it is 27% and 51%, respectively, as compared to 13% and 18% in the S&P 500 sample.

The key feature of the HFT sample is that the data set flags four distinct trade categories: (i) trades where HFTs demand liquidity from other HFTs; (ii) trades where HFTs demand liquidity from non-HFTs; (iii) trades where non-HFTs demand liquidity from HFTs; and (iv) trades where non-HFTs demand liquidity from other non-HFTs.

In an analysis of the same data set but over a longer time period, Carrion (2013) concludes that HFTs consume liquidity when it is cheap and supply it when it is dear. To compare the effective spreads earned and paid by HFTs relative to those of non-HFTs, he estimates regressions that control for trade size and direction of trade. To see the effect of the bias on such trader group comparisons, I replicate his analysis below.

In line with Carrion (2013), I consider the model

$$\begin{aligned}\tilde{S}^{v}_{itn} = \alpha_{it} &+ \beta_1 HFT + \beta_2 (HFT \times Medium)\\ &+\beta_3 (HFT \times Large) + \beta_4 (HFT \times Buy)\\ &+\beta_5 Medium + \beta_6 Large + \beta_7 Buy + \varepsilon, \end{aligned} \tag{8}$$

where *HFT*, *Medium*, *Large*, and *Buy* are dummy variables, and the subscripts $i$, $t$, and $n$ denote stocks, dates, and trades, respectively. The model is estimated separately for active and passive trades. For passive trades, the *HFT* dummy equals one when an HFT supplies liquidity in the trade, and zero otherwise. For active trades, the *HFT* variable is based on who the liquidity demander is. The variables *Medium* and *Large* flag the trade size, with medium-sized trades defined to be in the range of 500–999 shares, and trades of 1000 shares or more being considered large. As above, *Buy* indicates if a trade is buyer-initiated (one) or seller-initiated (zero). The model includes stock and date fixed effects (denoted $\alpha_{it}$), and the standard errors are clustered by stock and date.[18] I estimate the model for each of the effective spread estimators. The results are presented in Table 2.

The *HFT* coefficient $\beta_1$ captures the difference between HFTs and non-HFTs. In active trades, the coefficient is sig-

[18] The reason that the standard errors in this section are not clustered on trading venue is that the sample contains trade observations from Nasdaq only.

**Table 2**
*Effective spread differences between HFTs and non-HFTs.*
This table shows results of regression analyses aimed at analyzing the difference in effective spreads between HFTs and non-HFTs. The regressions are repeated for the effective spread benchmarked to three different estimators of fundamental value: the midpoint, the weighted midpoint, and the micro-price (defined as in Section 2.2). The analysis is run separately for the active side of trades (the three leftmost columns) and the passive side (the three rightmost columns). The sample includes 120 stocks and covers five trading days in the interval February 22–26, 2010. Each column of the table presents estimates of the following OLS regression (following Carrion, 2013):

$$\tilde{S}^{v}_{itn} = \alpha_{it} + \beta_1 HFT + \beta_2(HFT \times Medium) + \beta_3(HFT \times Large) + \beta_4(HFT \times Buy) + \beta_5 Medium + \beta_6 Large + \beta_7 Buy + \varepsilon,$$

where *HFT, Medium, Large*, and *Buy* are dummy variables, and the subscripts $i$, $t$, and $n$ denote stocks, dates, and trades, respectively. *HFT* indicates if the trade involves an HFT on the side of trade analyzed. *Medium* indicates if the trade volume is larger than or equal to 500 shares but lower than 1000 shares. *Large* indicates if the trade volume is larger than or equal to 1000 shares. *Buy* indicates if the trade is buyer-initiated. The model accounts for stock and date fixed effects, and the standard errors are clustered by stock and date. For each coefficient estimate, $t$-statistics are provided within brackets, and statistical significance at the 95% and 90% confidence levels is indicated with ** and *, respectively. $R^2$ (full model) and $R^2$ (within) show the proportion of the variance explained by the full model and by the independent variables except the fixed effects $\alpha_{it}$, respectively.

| | Active side effective spread, $\tilde{S}^{v}$ | | | Passive side effective spread, $\tilde{S}^{v}$ | | |
|---|---|---|---|---|---|---|
| | Midpoint | Weighted midpoint | Micro-price | Midpoint | Weighted midpoint | Micro-price |
| *HFT* | −0.17** | −0.68** | −0.94** | −0.10** | 0.39** | 0.49** |
| | (0.05) | (0.08) | (0.11) | (0.03) | (0.05) | (0.06) |
| *HFT* × *Medium* | 0.03 | −0.20** | −0.41** | 0.08* | −0.12* | −0.10 |
| | (0.05) | (0.05) | (0.07) | (0.05) | (0.07) | (0.08) |
| *HFT* × *Large* | 0.03 | −0.30** | −0.63** | 0.21** | −0.07 | −0.07 |
| | (0.07) | (0.09) | (0.10) | (0.04) | (0.07) | (0.10) |
| *HFT* × *Buy* | 0.01 | 0.03 | 0.10 | 0.06 | 0.06 | 0.09** |
| | (0.04) | (0.04) | (0.06) | (0.04) | (0.04) | (0.04) |
| *Medium* | 0.04 | 0.34** | 0.46** | 0.02 | 0.36** | 0.38** |
| | (0.04) | (0.06) | (0.07) | (0.03) | (0.07) | (0.07) |
| *Large* | 0.09 | 0.54** | 0.77** | −0.001 | 0.53** | 0.64** |
| | (0.06) | (0.09) | (0.11) | (0.03) | (0.09) | (0.12) |
| *Buy* | −0.01 | −0.05 | −0.19** | −0.04 | −0.07* | −0.20** |
| | (0.04) | (0.04) | (0.06) | (0.03) | (0.04) | (0.04) |
| Observations | 2,184,112 | 2,184,112 | 2,184,112 | 2,184,112 | 2,184,112 | 2,184,112 |
| $R^2$ (full model) | 0.403 | 0.235 | 0.213 | 0.403 | 0.231 | 0.204 |
| $R^2$ (within) | 0.001 | 0.010 | 0.016 | <0.001 | 0.004 | 0.006 |

nificantly negative for all effective spread estimators (see the first three columns of Table 2). This evidence supports the view that HFTs take liquidity when it is cheap, and the magnitude of the difference depends on the estimator applied. To evaluate the coefficient estimates, it is useful to put them in relation to the average effective spread. For the midpoint effective spread, the average is 3.67 bps, implying that HFTs pay around 5% lower spreads than do non-HFTs ($-0.17/3.67 \approx -5\%$). The corresponding differences for the alternative effective spread estimators are much larger: 26% for the weighted midpoint and 44% for the micro-price. Consistent with the notion that HFTs monitor markets closely, this result shows that HFTs are better than non-HFTs at spotting when the fundamental value deviates from the midpoint, and that they adapt their liquidity demand accordingly.[19]

Turning to liquidity supply, the midpoint effective spread *HFT* estimate is significantly *negative* (−0.10), which is in stark contrast to the view that HFTs supply liquidity when it is expensive. But the same coefficient for the weighted midpoint and the micro-price effective spread is significantly *positive*. In relative terms (obtained the same way as for the active trades), the midpoint effective spread shows that HFTs earn 3% less than non-HFTs in their liquidity supply. The other two estimators indicate that HFTs charge 15% (weighted midpoint) and 23% (micro-price) wider spreads, relative to what non-HFTs earn for comparable trades.

There are two important conclusions to draw from the analysis of the HFT sample. First, the choice of estimator can change research inference. Using either of the alternative effective spread estimators, I confirm Carrion's (2013)) finding that HFTs supply liquidity when it is expensive. Using the midpoint effective spread estimator, however, the HFTs appear to earn significantly lower spreads than do non-HFTs. That is, different estimators yield diametrically opposed conclusions.

[19] The two $R^2$ statistics reported in Table 2 indicate that the proportion of variance explained by the models is dominated by the fixed effects. This is due to the high liquidity dispersion between securities in the HFT sample, which is stratified to contain equal numbers of large-cap, mid-cap, and small-cap stocks. That the full model $R^2$ is lower for the alternative estimator models than for the midpoint effective spread model is due to the fact that the dispersion *between* stocks is lower for the alternative estimators (for descriptive statistics, see the internet appendix, Section IA.G). The explained proportion of variance *within* stocks and dates is higher for the alternative effective spread models.

The second conclusion is that trader groups differ widely in their liquidity timing. Whereas the HFTs adapt their trading to the fundamental value as indicated by the order book imbalance, large groups of investors in the non-HFTs group appear to ignore fundamental value deviations from the midpoint. That investors differ in their ability to accurately measure liquidity can potentially drive a wedge between the execution costs of, for example, professional and private investors.

### 5.2. Liquidity portfolios

Liquidity-sorted portfolios are common in asset pricing and corporate finance studies. In this section, I evaluate the extent to which such portfolios are affected by the midpoint effective spread bias.

I form three sets of quintile portfolios: one based on the midpoint effective spread and one for each of the alternative effective spread estimators. The sorts are repeated for each of the five days in the S&P 500 sample. Following the methodology of Holden and Jacobsen (2014), I then calculate the percentage of stock-day observations where the midpoint effective spread allocates a stock to a lower or higher quintile compared to the portfolios of each of the alternative effective spread estimators.

I find that the stocks sorted by midpoint effective spread end up in the same quintile as the stocks sorted by the weighted midpoint effective spread in no more than 70% of the cases. The midpoint effective spread bias leads to 16% of the stock-days being allocated to a more illiquid portfolio than where they should be, and 15% being put into a less illiquid portfolio than where they should be.[20] Consistent with the results reported above, the portfolio difference is stronger when the midpoint estimator is benchmarked to the micro-price estimator of the effective spread. The portfolio allocations based on the midpoint effective spread are consistent with those of the micro-price effective spread in only 56% of the stock-days.

### 5.3. Trading venue rankings

Analyses of execution costs across trading venues are important to investors who choose where to trade, and to academics who evaluate the merit of competing market design features (e.g., Bessembinder and Kaufman, 1997; Huang and Stoll, 1997). The bias variation across exchanges identified in Section 4.4 implies that exchange rankings based on the midpoint effective spread are potentially misleading.

I compare venue rankings for the effective spread estimators based on the midpoint and the weighted midpoint, respectively. For each stock-date and each effective spread measure, I rank the exchanges on a scale from one to nine. A ranking of 1 indicates that a venue provides the tightest average effective spread, and a ranking of 9 shows that the venue has the worst execution quality for the given stock-date. Following Holden and Jacobsen (2014), I then compare the two effective spread estimators by computing their difference in rank for each stock-date. For example, if the exchange BZX is ranked 3 for a given stock-date in terms of the weighted midpoint effective spread, but has a rank of 5 in terms of the midpoint effective spread, the rank difference is −2. Because there are nine trading venues in the sample (as above, I exclude CSX from the analysis), the rank difference variable can potentially range from −8 to +8.

Table 3 presents the frequency of venue rank differences for each stock exchange in the sample. In Panel (a), comparing rankings based on the midpoint and the weighted midpoint, the "Average" column shows that the two effective spread estimators yield exactly the same ranking for a given venue in only 38.6% of all stock-days (see the row labeled "Same rank"). Of all ranking differences that are different from zero, almost half of the cases are off by more than one step. Even though rank differences of five steps or more in either direction are rare, it is notable that rank differences of the maximum eight steps exist. Such cases indicate that a venue that is ranked highest according to one effective spread estimator, is ranked lowest according to the other.

The inverted fee venues (BYX, EDGA, and BX, marked by "*" in the table) benefit from the midpoint effective spread bias in terms of higher exchange rankings. The rows marked "Lower rank" and "Higher rank" report the sum of rank differences below and above zero. According to these statistics, the inverted fee venues are more likely to be ranked artificially high than to be ranked artificially low. For example, BYX benefits from the bias in 30.2% of the stock-days, and suffers from the bias in only 21.0% of the rankings. The venues that apply maker/taker fees tend to have the opposite pattern.[21]

The bias in venue rankings is even stronger in low-priced stocks. For stocks priced below USD 50, only 30.8% of the cases record no difference between rankings based on the two effective spread estimators, compared to 38.6% for the full sample (for tabulated results, see Section IA.F in the internet appendix). The pattern of inverse fee venues benefiting from the bias at the expense of maker-taker fee venues is also stronger for low-priced stocks.

Panel (b) of Table 3 reproduces the analysis above using the micro-price effective spread estimator for the benchmark rankings. To conserve space, I include only the rows labeled "Lower rank", "Same rank", and "Higher rank" in this panel. The results are consistent with those of the weighted midpoint effective spread.

I conclude from this application that if investors base their order routing decision on the effective spreads reported by exchanges, they are potentially misdirected. The

[20] Note that the number of artificially low-ranked stocks does not need to match the number of artificially high-ranked stocks. For example, if a stock is assigned to the fifth quintile instead of the third it counts as one case of a higher quintile, while it potentially pushes two other stocks to lower quintiles (one from the fifth to the fourth, and one from the fourth to the third).

[21] There is one exception. In spite of running a maker-taker fee schedule, NASD has an overweight to "Higher rank" outcomes. For stocks priced below USD 50, however, the results for NASD conform to those of other maker-taker fee venues (see Section IA.F in the internet appendix).

**Table 3**
*Venue ranking differences across effective spread estimators.*
This table shows how effective spread venue rankings differ depending on the effective spread estimator used. For each stock-day and each effective spread estimator, venues are ranked based on the effective spread. Panel (a) reports the difference in rankings obtained when using the midpoint effective spread instead of the weighted midpoint effective spread. A positive (negative) rank difference indicates that a venue is ranked higher (lower) when the ranking is based on the midpoint effective spread, relative to the weighted midpoint effective spread. The columns report the distribution of rank differences for each exchange, as well as the average across venues. The venues are categorized by the exchange group and marked by "*" if they apply inverted fee schedules. The rows "Lower rank" and "Higher rank" report the sum of all negative and positive rank differences, respectively. The sample includes all constituents of the S&P 500 index for the five trading days in the interval December 7–11, 2015. In Panel (b), the midpoint effective spread ranking is compared in the same way, but with the micro-price effective spread as benchmark. The effective spread is defined as in Table 1. For interpretation of the exchange abbreviations, see footnote 6.

(a) Benchmarked to the weighted midpoint effective spread

| $Rank(\tilde{S}^{wm})-$ | Bats exchanges | | | | Nasdaq exchanges | | | ICE exch. | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $Rank(\tilde{S}^{mid})$ | BYX* | EDGA* | BZX | EDGX | BX* | NASD | PHLX | NYSE | ARCA | Average |
| -8 | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.1% | 0.4% | 0.0% | 0.1% |
| -7 | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.1% | 0.4% | 0.0% | 0.1% |
| -6 | 0.0% | 0.1% | 0.0% | 0.3% | 0.0% | 0.0% | 0.4% | 0.3% | 0.1% | 0.1% |
| -5 | 0.2% | 0.3% | 0.2% | 1.2% | 0.0% | 0.3% | 1.1% | 0.8% | 0.3% | 0.5% |
| -4 | 0.6% | 0.5% | 0.7% | 3.2% | 0.4% | 1.0% | 2.6% | 1.3% | 1.7% | 1.3% |
| -3 | 1.3% | 2.5% | 3.6% | 6.3% | 1.3% | 2.2% | 4.6% | 3.8% | 4.5% | 3.3% |
| -2 | 4.2% | 8.5% | 11.7% | 12.1% | 5.1% | 5.6% | 7.8% | 7.6% | 12.2% | 8.3% |
| -1 | 14.7% | 14.1% | 26.6% | 20.6% | 17.5% | 13.2% | 12.5% | 20.1% | 23.3% | 18.1% |
| Same rank | 48.8% | 32.8% | 35.6% | 30.0% | 40.7% | 36.7% | 42.6% | 43.4% | 36.5% | 38.6% |
| 1 | 19.9% | 19.0% | 14.8% | 15.5% | 13.4% | 25.8% | 10.7% | 15.2% | 13.5% | 16.4% |
| 2 | 6.5% | 11.2% | 5.1% | 6.3% | 8.4% | 10.9% | 6.6% | 4.3% | 5.4% | 7.2% |
| 3 | 1.7% | 5.7% | 1.4% | 2.8% | 5.4% | 3.4% | 4.6% | 1.5% | 1.9% | 3.2% |
| 4 | 1.0% | 3.6% | 0.3% | 1.2% | 3.2% | 0.8% | 3.2% | 0.6% | 0.4% | 1.6% |
| 5 | 0.6% | 1.2% | 0.0% | 0.6% | 2.3% | 0.0% | 2.2% | 0.2% | 0.1% | 0.8% |
| 6 | 0.4% | 0.4% | 0.0% | 0.0% | 1.5% | 0.0% | 0.5% | 0.0% | 0.0% | 0.3% |
| 7 | 0.1% | 0.0% | 0.0% | 0.0% | 0.7% | 0.0% | 0.3% | 0.0% | 0.0% | 0.1% |
| 8 | 0.0% | 0.0% | 0.0% | 0.0% | 0.1% | 0.0% | 0.1% | 0.0% | 0.0% | 0.0% |
| Lower rank | 21.0% | 26.0% | 42.8% | 43.7% | 24.3% | 22.3% | 29.2% | 34.7% | 42.1% | |
| Higher rank | 30.2% | 41.1% | 21.6% | 26.4% | 35.0% | 40.9% | 28.2% | 21.8% | 21.3% | |

(b) Benchmarked to the micro-price effective spread

| | Bats exchanges | | | | Nasdaq exchanges | | | ICE exch. | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | BYX* | EDGA* | BZX | EDGX | BX* | NASD | PHLX | NYSE | ARCA | Average |
| Lower rank | 11.9% | 20.3% | 34.9% | 39.7% | 14.0% | 21.0% | 28.7% | 27.5% | 35.7% | 26.0% |
| Same rank | 61.4% | 44.3% | 45.7% | 42.3% | 54.7% | 48.8% | 49.1% | 54.2% | 49.0% | 49.9% |
| Higher rank | 26.8% | 35.5% | 19.4% | 18.1% | 31.3% | 30.3% | 22.2% | 18.3% | 15.4% | 24.1% |

result is in sharp contrast with the regulator's ambition with Rule 605. Boehmer et al. (2006) provide empirical evidence showing that, at the time when the Rule 605 reports were introduced, they had a significant influence on venue market shares.

## 6. Does the bias affect traders?

The above analysis shows that the midpoint effective spread bias can alter the results of academic research. But does the bias matter to traders? What is the value of monitoring how the fundamental value deviates from the midpoint? In this section, I focus on how the bias influences liquidity timing. To do so, I define an ex ante measure of liquidity that may be viewed as the expected effective spread. I then document that investors who use the midpoint overlook a substantial proportion of the liquidity variation. Finally, I use simulations to analyze trading performance differences between investors who differ only in their choice of effective spread estimators.

### 6.1. *The* expected *effective spread*

Liquidity timing requires assessment of trading costs in real time, not ex post as done by the effective spread. The standard ex ante measure is the quoted spread, but that is defined for a hypothetical roundtrip trade and accordingly does not capture spread asymmetries. For the applications in this section, I instead introduce the "oneway quoted spread," defined for buyer-initiated trades as

$$OQS = 2\left(P^A - X\right). \qquad (9)$$

I adopt the same notational convention as above, with $\widetilde{OQS}^v$ being the oneway quoted spread estimator corresponding to the fundamental value estimator $\tilde{X}^v$.

The oneway quoted spread differs from the effective spread in that it can be observed continuously during market opening hours, not just at the time of trade. It may be thought of as the expected effective spread for a buy quantity that does not exceed the depth available at the best ask price.[22]

### 6.2. Liquidity variance decomposition

How much liquidity variation is overlooked by focusing on the midpoint estimator of the effective spread? The variance of the oneway quoted spread may be decomposed as

$$\mathrm{Var}(OQS) = \mathrm{Var}\left(\widetilde{OQS}\right) + \mathrm{Var}\left(\widetilde{OQS} - OQS\right) - 2\mathrm{Cov}\left(\widetilde{OQS}, \widetilde{OQS} - OQS\right). \quad (10)$$

The first right-hand-side component is the estimator variance, the second is the *Nominal bias* variance, and the third is the covariance of the estimator and its nominal bias.

I implement the decomposition to compare the midpoint oneway quoted spread variance to that of the weighted midpoint estimator. The results show that an investor who is viewing the market through the lens of the weighted midpoint oneway quoted spread faces a liquidity variance of 10.27 bps. According to the midpoint estimator, however, the variance is only 5.80 bps, implying that 44% of the total variation $(1 - 5.80/10.27 = 44\%)$ is overlooked. As the covariance term is close to zero (-0.35), the difference is mostly due to the variance in *Nominal bias* (4.82 bps). When benchmarking to the micro-price estimator, the midpoint effective spread underestimation of liquidity variation is somewhat lower, 34%.[23]

### 6.3. A simulation of liquidity timing

To see how the overlooked liquidity fluctuations influence trading performance, I simulate trading decisions of three hypothetical investors who choose between limit and market orders to fulfil their trading objective. The three investors differ in only one dimension: their assessment of the fundamental value. Whereas the first investor tracks the midpoint, the other two follow the weighted midpoint and the micro-price.

In each trading period, each investor has an objective to purchase one lot of a given stock.[24] At the beginning of the period, the investor submits a limit order at the best bid price. That limit order is kept until executed, or until the oneway quoted spread for an ask-side market order is lower than a given threshold. When the latter condition is fulfilled, the investor cancels the limit order and submits a market order. Note that even if the three investors set their liquidity threshold the same way, their market order timings potentially differ. Such differences are entirely due to the choice of fundamental value proxy used to estimate the oneway quoted spread.

In industry terminology, the dynamic choice of order aggressiveness makes the trading strategy akin to an opportunistic execution algorithm. The splitting of the trading interest into one lot per 10-minute period implies an ambition to track the time-weighted average price, known as a TWAP benchmark. For characterizations of execution algorithms, see Labadie and Lehalle (2010) and O'Hara (2015).

Taking time priority into account, the limit order submitted at the beginning of the period is considered to be executed only when the market order volume that has arrived at the limit bid price exceeds the number of lots available at that price when the order was posted.

I use the S&P 500 sample to simulate the trading strategies on each sample stock, for each 10-minute period. The turn of events in each period is the same as in the historical data, implying an assumption that other market participants are unaffected by the simulated order flow. The hypothetical investors are set to be slow relative to the HFTs analyzed above, monitoring the market no more than once per second. Specifically, I let the investors check the status of the order book at the first order book update in each second. The threshold for switching to a market order is set to the first quartile of the oneway quoted spread, measured in the previous trading day.

Table 4 presents the trading performance of the three hypothetical investors. In addition, it includes results for an aggressive trader who submits a market order immediately in the beginning of each period, and for a passive trader who never switches to a market order.

The first column reports the micro-price effective spread for each trader, calculated as an average across all trading periods and stocks, including both active and passive executions (that have negative spreads). As expected, the aggressive trader faces the widest spread (4.29 bps), and the passive trader pays a negative spread (-1.00 bps). The three liquidity timing investors do substantially better than the aggressive investor but also display great variation in terms of spreads paid. The midpoint investor pays 2.61 bps, compared to 2.26 bps and 1.93 bps for the other two. The second column reports the difference relative to the midpoint investor, which is used as the benchmark. The differences are statistically significant.

The result that investors who use the alternative fundamental value estimators achieve lower trading costs relies on the accuracy of the micro-price. An alternative trading performance metric, which is commonly applied in the financial industry, is the *Execution shortfall*. Defined as the log difference between the transaction price and the midpoint recorded in the beginning of the period, it is independent of the fundamental value at the time of trade.

The average *Execution shortfall* (displayed in the third column of Table 4) shows the same pattern as the micro-price effective spread: the highest average purchase price is paid by the aggressive trader (2.20 bps), followed by the midpoint investor (1.26 bps). The investors who monitor the weighted midpoint and the micro-price have significantly lower *Execution shortfall* than the midpoint in-

[22] The oneway quoted spread exceeds the expected effective spread if there is a positive probability of price improvements or hidden liquidity inside the spread. In the S&P500 sample, the frequency of trades inside the spread is 10.1%.

[23] A decomposition of the variance of the effective spread estimators yields similar results. The variance of the ask-side midpoint effective spread is 29% lower than that of the weighted midpoint estimator. When the micro-price effective spread is used as benchmark, the corresponding number is 27%.

[24] The trading objective is set to one lot per period in order to minimize price impact. Larger order sizes would require a more elaborate simulation environment.

**Table 4**
*Trading costs of simulated liquidity timing strategies.*
This table shows trading costs associated with five simulated trading strategies. Each strategy seeks to buy one lot per 10-minute period. The first three rows show results for strategies that submit a limit order at the best bid in the beginning of each period, and change to a buy market order only when the oneway quoted spread $\widetilde{OQS}^{v}$ (defined as the best ask price minus the fundamental value) is equal to or smaller than the first quartile of the oneway quoted spread in the previous trading day. These three strategies differ only in the fundamental value proxy used for liquidity measurement: midpoint, weighted midpoint, and micro-price (each defined as in Section 2.2). The fourth strategy, denoted *Immediate market orders*, submits a buy market order in the beginning of each period. The fifth strategy, denoted *Limit orders only*, never changes to a buy market order. The table reports four measures of execution quality for each strategy. The micro-price effective spread ($\tilde{S}^{mic}$) is defined as in Table 1 and averaged across active and passive trades, where the latter typically have negative effective spreads. *Execution shortfall* is the average log difference between the transaction price and the midpoint recorded in the beginning of the period. *Time to trade* is the average time elapsed from the beginning of the period to the transaction time. *Missed trades* is the fraction of periods where the strategy does not lead to a transaction. For each strategy, the table also reports how $\tilde{S}^{mic}$ and the *Execution shortfall* differ to the benchmark strategy, which is chosen to be the one timing the midpoint. Significance of the transaction cost differences is tested for using $t$-tests, with the test statistic reported within brackets and statistical significance at the 95% and 90% confidence levels indicated with ** and *, respectively. The standard errors of the statistical tests are clustered by stock and date. The sample includes all constituents of the S&P 500 index for the date interval December 7–11, 2015. The first day is used only for determining the market order threshold for the following day.

| | Micro-price effective spread $\tilde{S}^{mic}$ (bps) | | *Execution shortfall* (bps) | | *Time to* | *Missed* |
|---|---|---|---|---|---|---|
| Strategy | Mean | Diff. to midpoint strategy | Mean | Diff. to midpoint strategy | *trade* (seconds) | *trades* (%) |
| Timing the one-way quoted spread, $\widetilde{OQS}^{v}$ | | | | | | |
| Midpoint | 2.61 | | 1.26 | | 29.55 | 1.59 |
| Weighted midpoint | 2.26 | −0.36** | 1.08 | −0.18** | 23.39 | 0.71 |
| | | (−3.51) | | (−5.69) | | |
| Micro-price | 1.93 | −0.69** | 0.99 | −0.27** | 30.47 | 1.59 |
| | | (−11.00) | | (−12.24) | | |
| Immediate market orders | 4.29 | 1.68** | 2.20 | 0.94** | 0.38 | 0.00 |
| | | (6.47) | | (7.54) | | |
| Limit orders only | −1.00 | −3.61** | −2.15 | −3.41** | 65.32 | 14.25 |
| | | (−14.33) | | (−20.07) | | |

vestor, with differences of 0.18 bps and 0.27 bps, respectively. Though small in nominal terms, these numbers relative to the amount paid by the midpoint investor, imply 14% and 21% lower costs, respectively.

Zooming in on stocks with higher bias, the differences in liquidity timing performance are even stronger. For stocks priced below USD 50, the *Execution shortfall* for the midpoint strategy is 1.73 bps, whereas it lies in the 0.97–0.98 bps range for the traders who track the alternative fundamental value proxies (see the internet appendix, Section IA.F). In relative terms, the *Execution shortfall* is 43%–44% lower than what is paid by the midpoint trader.

The fact that the passive investor records negative trading costs for both the effective spread and the execution shortfall serves as a reminder of the well-known tradeoff between execution costs and opportunity costs (Perold, 1988). Passive traders can save on the cost of crossing the spread, but only at the risk of delayed or missed trades. Execution costs should thus be assessed jointly with opportunity costs. Such rigor is often not empirically feasible due to the lack of data on trading intentions, but in this simulation it is straightforward.

The two rightmost columns of Table 4 report the *Time to trade* and *Missed trades*. The former is measured as the time from the beginning of a period to the execution time (measured in seconds), and the latter is the fraction of periods where no execution is achieved. The trader who never crosses the spread misses 14% of the trades, and his average *Time to trade* exceeds one minute. The numbers for the other traders are much lower. Notably, the investors who track the midpoint and the micro-price have almost identical outcomes in this respect, with 1.59% missed trades and around 30 seconds of trade delay. The investor tracking the weighted midpoint estimator has lower opportunity costs.

The conclusion from this simulation is that investors who track one of the alternative spread estimators proposed in this article achieve lower trading costs. By focusing on the midpoint, investors overlook much of the variation, undermining their liquidity timing ability.

## 7. Bias in the price impact and the realized spread

The effective spread is frequently decomposed into the price impact and the realized spread (see, e.g., Bessembinder and Kaufman, 1997; Hendershott et al., 2011; Conrad et al., 2015). The former is a measure of how much the market maker is losing to liquidity demanders due to changes in the fundamental value. The latter is the effective spread net of price impact, which should cover all other costs and potentially leave the market maker with a profit. Both metrics require a fundamental value estimator, and, to my knowledge, all implementations to date rely on the midpoint. In this section, I investigate if the midpoint effective spread bias passes through to the price impact and realized spread.

The realized spread estimator is defined as $\widetilde{RS}_h^{v} = 2D_t\left(P_t - \tilde{X}_{t+h}^{v}\right)$, where $h$ is the evaluation horizon, which I set to either five minutes or ten seconds. Price impact is

**Table 5**
*Bias in trading performance evaluations.*
This table shows the results of effective spread decompositions using different fundamental value proxies. The realized spread $\widetilde{RS}_h^{\nu}$ is twice the difference between the trade price and the fundamental value estimator, $h$ seconds after the trade, multiplied by the direction of trade, and scaled by the midpoint. The superscript $\nu$ denotes the fundamental value estimator used and can be either the midpoint, the weighted midpoint, or the micro-price (defined as in Section 2.2). The price impact $\widetilde{PI}_h^{\nu}$ is twice the change in fundamental value $\nu$ over the $h$ seconds following a trade, multiplied by the direction of trade, and scaled by the midpoint. The realized spread and the price impact are measured for each stock as the dollar-weighted average across all trades in the sample. The *Nominal bias* for the realized spread is the difference between the midpoint realized spread and an alternative estimator of the realized spread, reported in bps. The $t$-statistic corresponding to the null that the value-weighted average *Nominal bias* is equal to zero, based on standard errors that are clustered by stock, date, and trading venue, is reported within parentheses. The *Relative average bias* for the realized spread is the average *Nominal bias* divided by the absolute mean of the alternative realized spread estimator. The *Nominal bias* and the *Relative average bias* for price impact are defined the same way. The sample includes the constituents of the S&P 500 index, covering the five trading days in the date interval December 7–11, 2015. Panels (a) and (b) hold results for the five-minute and ten-second horizons, respectively.

(a) Five-minute horizon

| | Mean (bps) | *Nominal bias* (bps) | $t$-stat. | *Relative average bias* |
|---|---|---|---|---|
| Price impact estimator, $\widetilde{PI}_{5min}^{\nu}$ | | | | |
| Midpoint | 3.87 | | | |
| Weighted midpoint | 3.48 | 0.39 | 8.67 | 10% |
| Micro-price | 3.36 | 0.50 | 9.91 | 13% |
| Realized spread estimator, $\widetilde{RS}_{5min}^{\nu}$ | | | | |
| Midpoint | −0.65 | | | |
| Weighted midpoint | −0.65 | −0.01 | −1.47 | −1% |
| Micro-price | −0.64 | −0.01 | −3.56 | −2% |

(b) Ten-second horizon

| | Mean (bps) | *Nominal bias* (bps) | $t$-stat. | *Relative average bias* |
|---|---|---|---|---|
| Price impact estimator, $\widetilde{PI}_{10sec}^{\nu}$ | | | | |
| Midpoint | 4.04 | | | |
| Weighted midpoint | 3.78 | 0.26 | 6.60 | 7% |
| Micro-price | 3.64 | 0.38 | 8.57 | 10% |
| Realized spread estimator, $\widetilde{RS}_{10sec}^{\nu}$ | | | | |
| Midpoint | −0.82 | | | |
| Weighted midpoint | −0.94 | 0.12 | 8.89 | 14% |
| Micro-price | −0.92 | 0.11 | 6.61 | 14% |

defined as $\widetilde{PI}_h^{\nu} = 2D_t\left(\tilde{X}_{t+h}^{\nu} - \tilde{X}_t^{\nu}\right)$. The volume-weighted averages for each estimator of fundamental value are presented in Table 5, along with their corresponding *Nominal bias* and *Relative average bias*, defined the same way as for the effective spread.[25] Panels (a) and (b) show the results for the five-minute and ten-second horizons, respectively.

I find that when trading performance is evaluated on a relatively long horizon, such as five minutes, the midpoint effective spread bias carries over almost entirely to the price impact. The price impact is overestimated by 0.39–0.50 bps, depending on the estimator applied as benchmark. That range is virtually the same as for the effective spread (0.38–0.49 bps, see Table 1). Because the price impact is higher, on average, than the effective spread, the bias is smaller in relative terms than for the effective spread (10%–13%). The $t$-statistic indicates that the midpoint price impact bias is statistically significant for both estimators. Similar to the results for the effective spread, the midpoint price impact bias is even stronger for low-priced stocks (see the internet appendix, Section IA.F).

The five-minute midpoint realized spread bias is -2% or less, depending on estimator. When the evaluation horizon is shorter (ten seconds, presented in Panel (b)), some of the midpoint effective spread nominal bias is picked up by the realized spread, but it remains more than twice as large for the price impact.

Notably, the price impact measure analyzed here is used by, for example, Goyenko et al. (2009) to evaluate price impact measures based on low-frequency data, such as the proxies proposed by Amihud (2002) and Pastor and Stambaugh (2003). My evidence indicates that the benchmark used in such evaluations is biased, and that the bias varies in the cross-section.

The midpoint price impact bias is also relevant to analyses of investors' ability to predict and trade ahead of price changes. For example, high price impact in active trading is often interpreted as a sign of informed trading (due either to private information or to an early response to public information). If that price impact is based on sub-tick value changes, however, that skill may reflect "nowcasting" (the ability to estimate the current fundamental value) as much as forecasting ability.

[25] The relative average bias is signed to have the same direction as the nominal bias, which is otherwise not the case when the average level is negative.

## 8. Effective spread estimator choice

The fundamental value of a security is an elusive concept, and the three estimators considered in this paper (and the alternative estimators discussed in the internet appendix, Sections IA.C and IA.E) are not an exhaustive set. A fundamental value estimator should factor in all the information available, implying that the estimators considered here rely on an implicit assumption that the limit order book reflects all relevant information. The proxies can potentially be refined by considering data on additional price levels in the order book, information generated at alternative trading systems, as well as trade and quote information from related securities and indices. Depending on the empirical context, alternative econometric approaches may improve the accuracy.

For cross-market comparisons, it is important that the fundamental value at any given instance is the same for all venues. This can be achieved by using the NBBO data, by aggregating quotes across venues, or by obtaining the efficient price variation from a time series model with several markets.

Among the estimators considered, I find the weighted midpoint to be a viable alternative to overcome the midpoint effective spread bias. The weighted midpoint is a linear function of the spread and the order book imbalance, which makes it feasible to estimate in real time for investors and regulators with access to the NBBO through the Consolidated Quote System. Academic researchers can easily calculate it using quote information available in DTAQ, MTAQ, or TH. Harris (2013) argues that the weighted midpoint estimator may be improved by accounting for exchange fees.

From a theoretical perspective, the most appealing estimator is perhaps the micro-price, because it is a martingale by construction. Empirically, Stoikov (2018) shows that the micro-price outperforms the weighted midpoint in predicting midpoint changes. Though the micro-price is relatively expensive to compute, this should not be viewed as an obstacle to real-time calculation, because the costly operations could be run before the market opens. Specifically, the midpoint adjustment function $g(QS, I)$ can be preestimated for all relevant combinations of the quoted spread and the order book imbalance, using historical data. For examples of such estimates, see the internet appendix, Section IA.A. What remains to do in real time is then to simply map the current quoted spread and order book imbalance to the relevant midpoint adjustment.

My findings complement those of Holden and Jacobsen (2014). They show that accurate estimation of the effective spread requires data from DTAQ, which is equivalent to TH data. I show that the DTAQ midpoint effective spread, which they use as their benchmark, is biased. It may be replaced by one of the alternative estimators discussed above. For analysts that are computationally and financially constrained, Holden and Jacobsen (2014) propose a second-best solution based on MTAQ data. That approach is straightforward to combine with the alternative effective spread estimators discussed here.

## 9. Conclusion

I show that the midpoint effective spread is a biased estimator of the effective spread. The bias varies systematically across stocks and trading venues, and undermines liquidity timing and trading performance evaluations. The bias is economically and statistically significant, robust across market capitalization segments and across continuous fundamental value estimators, and increasing over time.

Importantly, I find sizable differences across investor groups in the ability to gauge the fundamental value. The implication is that, whereas sophisticated investors understand the midpoint effective spread bias to a large extent, others do not. Regulators or brokers can bridge this gap by disseminating a more accurate fundamental value proxy. Furthermore, the Rule 605 reporting requirements could be amended to mitigate the bias.

The problem with using the midpoint as proxy for the fundamental value is application-specific. The fundamental value deviation from the midpoint is important for the effective spread in that it *influences* the order flow. Because the tight side of the spread attracts more flow than the wide side, the bias survives when averaging across trades. In other applications, such as measurement of returns and realized variance, I do not expect the fundamental value deviation from the midpoint to be problematic. The midpoint effective spread bias may, however, influence our understanding of price discovery, liquidity risk, the merit of low-frequency liquidity estimators, and liquidity premia in asset pricing. I leave these issues for future research.

## Appendix A. Micro-price estimation details

This appendix summarizes the micro-price estimation procedure outlined by Stoikov (2018) and specifies implementation details that deviate from his work. For empirical examples and programming implementation, see the internet appendix (Section IA.A) and the replication code, respectively.

The micro-price estimation essentially amounts to finding the midpoint adjustment $g(QS, I)$ for each combination of the quoted spread ($QS$) and the order book imbalance ($I$). To limit the number of states, the state variables are discretized. Furthermore, the midpoint adjustment is assumed to be independent of the midpoint level. Once the adjustment has been estimated for the full state space, the micro-price can be obtained at low computational cost by simply mapping prevailing spreads and order book imbalances to the appropriate midpoint adjustment, and inserting it in Eq. (6).

### A1. Sample

The input data for the micro-price estimation consists of NBBO quotes. No trade information is considered. I sample quotes at a 100 millisecond frequency, yielding 228,000 observations per trading day when the first and last five minutes are excluded ((6 h × 60 min + 20 min) × 60 s × 10 obs. per second).

The micro-price focuses on what the probable price change following a given quote is. This raises the concern that a trade matched to that quote influences the outcome. To avoid such a forward-looking bias, I base the estimation of $g(QS, I)$ on quotes from the previous week of each sample. That is, for the S&P 500 sample, I use data from November 30, to December 4, 2015. For the HFT sample, the previous week is February 16–19, 2010 (February 15, 2010 is a public holiday).

## A2. State space

The estimation procedure is based on the dynamics of the triplet $\left(\tilde{X}^{mid}, \bar{I}_\tau, \overline{QS}_\tau\right)$, where $\tau$ is a time index, and the bars above the variable names indicate that they are discrete state variable versions of the continuous variables $I_\tau$ and $QS_\tau$. The bar is omitted for $\tilde{X}^{mid}$ as no discretization is required for the spread midpoint. The quoted spread is also discrete in nature, and in the application by Stoikov (2018) the spread state variable is simply the number of ticks. My application, with hundreds of different stocks, requires a more flexible procedure for the discretization of spreads. I outline such a procedure below, along with an approach to define order book imbalance states differently across stocks and spread levels.

*Discretizing the quoted spread.* I refer to spread levels that are recorded in more than 1% of all quote observations as “common,” and spreads that are not common but that have a frequency exceeding 0.01% as “rare.” Even less frequent spread levels are disregarded. I form one state for each common spread level. One percent of the quote sample corresponds to more than 2000 quote observations, which I consider enough to estimate the midpoint adjustment function accurately. For the rare spreads, I do the following:

- If there are rare spreads that are lower than the lowest common spread level, I let them form a new state if they together constitute more than 1% of the quote sample. If they are less frequent than that, I include them in the lowest common spread state.
- If there are rare spreads that are higher than the highest common spread level, I let them form a new state if they together constitute more than 1% of the quote sample. If they are less frequent than that, I include them in the highest common spread state.
- If there are rare spreads that lie between two common spread levels, I include them in the closest lower common spread state.

*Discretizing the order book imbalance.* Recall the definition of order imbalance in Section 2.2:

$$I = \frac{Q^B}{Q^B + Q^A}, \tag{A.1}$$

where $Q^B$ and $Q^A$ represent the depths quoted at the best bid and ask prices, respectively. Because the order imbalance is a fraction of the quoted depth, I express the state bounds discussed below as fractions of integers, rather than in decimal form.

For each spread state, I form nine order imbalance states, as follows:

- States 1–4 are defined by the quartiles of order imbalance observations that are lower than or equal to 9/20 (if any).
- State 5 includes order imbalance observations that satisfy $9/20 < I_\tau \le 11/20$.
- States 6–9 are defined by the quartiles of order imbalance observations that are higher than 11/20 (if any).

By predefining the State 5 boundaries, I avoid putting a breakpoint at 1/2, which is a very common value in the data, representing a balanced order book. The quantile-defined breakpoints for all other states make the distribution of observations across states more uniform than with the equi-spaced boundaries used by Stoikov (2018). For the same reason, I allow the imbalance breakpoints to vary across spread states. Nevertheless, due to that imbalances cluster at certain fractions, there are infrequent cases in my sample where not all imbalance states are populated. In those cases, the midpoint adjustment can not be estimated for all states.[26]

## A3. Estimation

Stoikov's (2018) estimation procedure involves the following steps:

*1. Symmetrization.* I symmetrize the data such that for each observation ($\bar{I}_\tau$; $\overline{QS}_\tau$; $\bar{I}_{\tau+1}$; $\overline{QS}_{\tau+1}$; $dM$), where $dM$ is the midpoint change from $\tau$ to $\tau+1$, I add an observation that is mirrored in the imbalance dimension and has the opposite sign on $dM$ ($10-\bar{I}_t$; $\overline{QS}_t$; $10-\bar{I}_{t+1}$; $\overline{QS}_{t+1}$; $-dM$). The symmetrization of the input data ensures that the micro-price estimation converges. It also leads to symmetry in the $g\left(\overline{QS}, \bar{I}\right)$ estimates, i.e., $g\left(\overline{QS}, \bar{I}\right) = -g\left(\overline{QS}, 10-\bar{I}\right)$.

*2. Transition probability estimation.* The estimation procedure distinguishes transitory and absorbing states. Given the current state, a state is absorbing if it implies a midpoint change, and transitory otherwise. The micro-price estimation may be thought of as a probability tree where branches keep growing until they reach an absorbing state. The midpoint adjustment is then a probability-weighted average of midpoint changes associated with each branch. To analyze the probability tree, the next-period probability of each state, conditional on the current state, is required. The transition probabilities are assumed to equal the historically observed frequencies.

The transition probabilities between transitory states are captured by the square matrix $Q$. If there are $m$ spread states and $n$ imbalance states, $Q$ is an $mn \times mn$ matrix. For example, the top-left entry of $Q$ may show the probability of staying in the state $\left(\overline{QS}, \bar{I}\right) = (1, 1)$, with the spread midpoint unchanged.

The transition probabilities for changes from transitory states to absorbing states are recorded in two matrices, $T$ and $R$. The former is similar to $Q$, dimension $mn \times mn$,

[26] In rare cases, further adjustments are required. When there are no “common” spreads, I use deciles to form ten spread states. When the order imbalance interval [9/20,11/20] is unpopulated, I use only two order imbalance states, with 1/2 as the breakpoint. To limit the influence of outliers in the midpoint changes, I set midpoint changes exceeding the median spread of the highest spread state observations equal to that median. I set negative outliers equal to the negative of the same median.

in that it tracks the transition between spread-imbalance combinations, but it differs in that it only considers cases where there is also a change in the midpoint. For example, the top-left entry of $T$ may show the probability of staying in the state $(\overline{QS}, \bar{I}) = (1, 1)$ while the midpoint is changing. The matrix $R$, in turn, captures the magnitude of the midpoint change. Define a vector $K$ of all possible levels of non-zero midpoint changes. The dimension of $R$ is then $mn \times k$, where $k$ is the length of $K$.

*3. Computing the midpoint adjustment.* The one-period-ahead midpoint adjustment for each combination of $\overline{QS}$ and $\bar{I}$ is denoted $G^1$. Stoikov (2018) shows that $G^1 = (1 - Q)^{-1}RK$. To find the vector $G^*$, which is the expected midpoint change evaluated at infinity, Stoikov (2018) defines $B = (1 - Q)^{-1}T$ and shows that:

$$G^* = G^1 + \sum_{i=1}^{\infty} B^i G^1. \tag{A.2}$$

I consider ten iterations of the sum in (A.2), but the value of $G^*$ typically converges after 2–3 iterations.

## Appendix B. Data matching and screening

### B1. Matching CRSP and TH identifiers

To my knowledge, this is the first study that matches data from the CRSP and TH databases. The security identifier in TH is called the Reuters Instrument Code (RIC). The CRSP field with closest correspondence to the RICs is the ticker symbol at the primary exchange, `TSYMBOL`. For most stocks, `TSYMBOL` is identical to the RIC of the consolidated instrument in TH. In order to match all securities, however, the following adjustments are considered:[27]

- When `TSYMBOL` is empty, the CRSP field `TICKER` is used instead.
- Before January 1, 2012, share class information is not included in `TSYMBOL`. Then, when `TSYMBOL` cannot be matched to a RIC and the CRSP field `SHRCLS` is equal to `A` or `B`, I add a lowercase share class suffix (e.g., the `TSYMBOL` entry `AIS` is set to `AISa`).
- After January 1, 2012, `TSYMBOL` and `TICKER` differ when there is a share class suffix for `TSYMBOL`. I make the `TSYMBOL` share class suffix lowercase to match the TH identifier conventions (e.g., the `TSYMBOL` entry `VIAB` is set to `VIAb`). Other four-letter `TSYMBOL` entries are given a suffix `.K`, in line with TH consolidated instrument conventions (e.g., the `TSYMBOL` entry `ADGE` is set to `ADGE.K`).

### B2. Time stamps

The time stamps reported in the `Exchange Time` field in TH are given in milliseconds. For TH entries before October 23, 2006, however, the time stamps are given in seconds (I am unable to confirm whether this is the case for DTAQ as well). TH also assigns its own time stamps with microsecond granularity when the data are received by Refinitiv. The TH time stamps have higher granularity than the official time stamps, but they are subject to a reporting delay. I use `Exchange Time` when available at millisecond granularity and the internally assigned TH time stamps otherwise.

### B3. TH data screening

Each trade and quote observation in TH includes additional information in the `Qualifiers` field. I use that information to screen trades and quotes, using the following criteria:

(T1) Trades marked as regular, odd lots, or due to intermarket sweep orders are retained, unless any of the criteria (T2)–(T4) are satisfied. This screening utilizes the `[GVx_TEXT]` (where $x$ can be a number from 1 to 4) and `[LSTSALCOND]` information and excludes everything but the following entries: `@F_I` (where _ represents a space), `@__I`, `@F__`, `@___`, `_F__`, `_F_I`, and `___I`.

(T2) Trades with any of the following conditions indicated in the `[CTS_QUAL]` information are excluded: *derivatively priced* (`DPT`), *stock option related* (`SOT`), *threshold error* (`XSW`, `RCK`, `XO`), *out of sequence* (`SLD`), and *cross-trades* (`XTR`).

(T3) Trades with any of the following conditions indicated in the `[PRC_QL2]` information are excluded: *agency cross-trade* (`AGX`), *stock option trade* (`B/W`), *not eligible for last* (`NBL`), *derivatively priced* (`SPC`), and *stopped* (`STP`).

(T4) Trades flagged as corrected are excluded. Corrections are entered as separate observations in TH and linked by an order sequence number (`Seq.No.`) to the trade in question.

(Q1) Quotes marked as regular or as coinciding with changes in the limit up–limit down (LULD) price bands are retained, unless any of the criteria (Q2)–(Q4) are satisfied. This screening utilizes the `[PRC_QL_CD]` and `[PRC_QL3]` information and excludes everything but the following entries: `R__`, `___`, LPB, and RPB. For example, quotes with non-positive bid-ask spread, quotes that are associated with trading halts, and quotes marked as slow due to a liquidity replenishment point are thus excluded. Quotes coinciding with changes in the LULD price bands are retained because LULD limit updates do not influence the validity of the current quotes.

(Q2) Quotes marked as *non-executable* are excluded (`A`, `B`, or `C`, in the `[GV1_TEXT]` field).

(Q3) Quotes with non-regular conditions indicated by the `[CTS_QUAL]` information (taking the value `TH_`, `IND`, or `O__`) are excluded.

(Q4) Quotes where the bid-ask spread is either negative ("crossed"), zero ("locked"), or exceeding USD 5 are excluded.

The effects of the different screening criteria are presented in Table A.1. The trade screening criteria disqualify a negligible number of trades in each data set.

[27] The RICs in TH change over time. To track a security over time, a viable strategy is to access the CRSP time series, where the security identifier `PERMNO` is permanent. The time-varying `TSYMBOL` can then be matched to RICs as described here.

**Table A.1**
**Data screening statistics.**
This table shows the extent to which different screening criteria filter out trade observations (Panel (a)) and set quotes matched to trades to missing (Panel (b)).

(a) Trade screens

| Sample | (T1) | (T2) | (T3) | (T4) | All filters combined | Remaining # obs. |
|---|---|---|---|---|---|---|
| S&P 500 sample, Dec. 7–11, 2015 | < 0.01% | < 0.01% | < 0.01% | < 0.01% | < 0.01% | 55.7 million |
| HFT sample, Feb. 22–26, 2010 | < 0.01% | < 0.01% | < 0.01% | < 0.01% | < 0.01% | 2.4 million |
| Time series sample, Jan. 1, 1996–Dec. 31, 2015 | < 0.01% | < 0.01% | < 0.01% | < 0.01% | < 0.01% | 4,514.3 million |

(b) Quote screens

| Sample | (Q1) | (Q2) | (Q3) | (Q4) | All filters combined | Locked quotes |
|---|---|---|---|---|---|---|
| S&P 500 sample, Dec. 7–11, 2015 | 1.04% | < 0.01% | < 0.01% | 5.01% | 5.01% | 4.89% |
| HFT sample, Feb. 22–26, 2010 | 5.66% | < 0.01% | < 0.01% | 9.83% | 9.83% | 9.18% |
| Time series sample, Jan. 1, 1996–Dec. 31, 2015 | 0.14% | < 0.01% | < 0.01% | 5.90% | 6.04% | 5.39% |

Among the quote screening criteria, (Q2) and (Q3) each affect less than 0.01% of the observations. The criteria specified in (Q1) and (Q4), however, disqualify a substantial number of quotes. In particular, for the HFT sample, (Q1) and (Q4) affect 5.66% and 9.83% of the quote observations, respectively. Looking at the combination of those filters, however, it is clear that there is a strong overlap in the sense that quotes that are captured by (Q1) virtually always are also affected by (Q4). Upon closer inspection, the vast majority of the excluded quotes are locked, meaning that the bid and ask prices are equal. It is well-known that locked quotes are common in the NBBO data (Shkilko et al., 2008). Locked quotes cannot exist within an exchange. In the NBBO feed, however, they can appear due to that price changes are not simultaneous across venues, for example. Around 4.89% of all trades in the S&P 500 sample, 9.18% in the Nasdaq HFT sample, and 5.39% in the time-series sample are matched to such quotes (see the rightmost column of Table A.1, Panel (b)). The exclusion of locked quotes is consistent with Holden and Jacobsen (2014).

## References


Abdi, F., Ranaldo, A., 2017. A simple estimation of bid–ask spreads from daily close, high, and low prices. Rev. Financ. Stud. 30, 4437–4480.
Aldrich, E.M., Lee, S., 2018. Relative spread and price discovery. J. Empir. Financ. 48, 81–98.
Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. J. Financ. Mark. 5, 31–56.
Anshuman, V.R., Kalay, A., 1998. Market making with discrete prices. Rev. Financ. Stud. 11, 81–109.
Back, K., Baruch, S., 2013. Strategic liquidity provision in limit order markets. Econometrica 81, 363–392.
Bessembinder, H., 2003. Trade execution costs and market quality after decimalization. J. Financ. Quant. Anal. 38, 747–777.
Bessembinder, H., Kaufman, H.M., 1997. A cross-exchange comparison of execution costs and information flow for NYSE-listed stocks. J. Financ. Econ. 46, 293–319.
Blume, M. E., Goldstein, M. A., 1992. Displayed and effective spreads by market. In: Unpublished working paper, The Wharton School, University of Pennsylvania.
Boehmer, E., Jennings, R., Wei, L., 2006. Public disclosure and private decisions: equity market execution quality and order routing. Rev. Financ. Stud. 20, 315–358.
Brogaard, J., Hagströmer, B., Nordén, L., Riordan, R., 2015. Trading fast and slow: colocation and liquidity. Rev. Financ. Stud. 28, 3407–3443.
Brogaard, J., Hendershott, T., Riordan, R., 2014. High-frequency trading and price discovery. Rev. Financ. Stud. 27, 2267–2306.
Budish, E., Cramton, P., Shim, J., 2015. The high-frequency trading arms race: frequent batch auctions as a market design response. Q. J. Econ. 130, 1547–1621.
Carrion, A., 2013. Very fast money: high-frequency trading on the NASDAQ. J. Financ. Mark. 16, 680–711.
Cartea, A., Jaimungal, S., Penalva, J., 2015. Algorithmic and High-Frequency Trading. Cambridge University Press, Cambridge.
Chakrabarty, B., Pascual, R., Shkilko, A., 2015. Evaluating trade classification algorithms: bulk volume classification versus the tick rule and the Lee-Ready algorithm. J. Financ. Mark. 25, 52–79.
Commission, S., Exchange, 2001. Release no. 34–43590: disclosure of order execution and routing practices.
Conrad, J., Wahal, S., Xiang, J., 2015. High-frequency quoting, trading, and the efficiency of prices. J. Financ. Econ. 116, 271–291.
Cont, R., Kukanov, A., Stoikov, S., 2014. The price impact of order book events. J. Financ. Econom. 12, 47–88.
Corwin, S.A., Schultz, P., 2012. A simple way to estimate bid–ask spreads from daily high and low prices. J. Financ. 67, 719–760.
Demsetz, H., 1968. The cost of transacting. Q. J. Econ. 82, 33–53.
Fang, V.W., Noe, T.H., Tice, S., 2009. Stock market liquidity and firm value. J. Financ. Econ. 94, 150–169.
Foucault, T., Kadan, O., Kandel, E., 2005. Limit order book as a market for liquidity. Rev. Financ. Stud. 18, 1171.
Glosten, L.R., 1994. Is the electronic open limit order book inevitable? J. Financ. 49, 1127–1161.
Goettler, R.L., Parlour, C.A., Rajan, U., 2005. Equilibrium in a dynamic limit order market. J. Financ. 60, 2149–2192.
Goldstein, M.A., Kavajecz, K.A., 2000. Eighths, sixteenths, and market depth: changes in tick size and liquidity provision on the NYSE. J. Financ. Econ. 56, 125–149.
Gould, M.D., Bonart, J., 2016. Queue imbalance as a one-tick-ahead price predictor in a limit order book. Mark. Microstruct. Liq. 2, 1650006.
Goyenko, R., Holden, C., Trzcinka, C., 2009. Do liquidity measures measure liquidity? J. Financ. Econ. 92, 153–181.
O'Hara, M., Saar, G., Zhong, Z., 2018. Relative tick size and the trading environment. Rev. Asset Pricing Stud. 9, 47–90.
Harris, L., 2013. Maker–Taker pricing effects on market quotations. In: Unpublished working paper, University of Southern California, San Diego.
Hasbrouck, J., 2002. Stalking the "efficient price" in market microstructure specifications: an overview. J. Financ. Mark. 5, 329–339.
Hasbrouck, J., 2009. Trading costs and returns for US equities: estimating effective costs from daily data. J. Financ. 64, 1445–1477.
Hendershott, T., Jones, C.M., Menkveld, A.J., 2011. Does algorithmic trading improve liquidity? J. Financ. 66, 1–33.
Hendershott, T., Menkveld, A.J., 2014. Price pressures. J. Financ. Econ. 114, 405–423.
Holden, C.W., Jacobsen, S., 2014. Liquidity measurement problems in fast, competitive markets: expensive and cheap solutions. J. Financ. 69, 1747–1785.
Huang, R., Stoll, H., 1996. Dealer versus auction markets: a paired comparison of execution costs on NASDAQ and the NYSE. J. Financ. Econ. 41, 313–357.

Huang, R., Stoll, H., 1997. The components of the bid–ask spread: a general approach. Rev. Financ. Stud. 10, 995.
Jahan-Parvar, M., Zikes, F., 2019. When do low-frequency measures really measure transaction costs? Unpublished working paper. Board of Governors of the Federal Reserve System.
Johnson, B., 2010. Algorithmic Trading & DMA. 4Myeloma Press London.
Kavajecz, K.A., 1999. A specialist's quoted depth and the limit order book. J. Financ. 54, 747–771.
Korajczyk, R.A., Sadka, R., 2008. Pricing the commonality across alternative measures of liquidity. J. Financ. Econ. 87, 45–72.
Labadie, M., Lehalle, C. A., 2010. Optimal algorithmic trading and market microstructure. In: Unpublished working paper. HAL.
Lee, C., 1993. Market integration and price execution for NYSE-listed securities. J. Financ. 48, 1009–1038.
Lee, C., Ready, M., 1991. Inferring trade direction from intraday data. J. Financ. 46, 733–746.
Muravyev, D., Pearson, N., 2020. Option trading costs are lower than you think. Rev. Financ. Stud. 33, 4973–5014.
Næs, R., Skjeltorp, J.A., Ødegaard, B.A., 2011. Stock market liquidity and the business cycle. J. Financ. 66, 139–176.
O'Donoghue, S. M., 2015. The effect of maker-taker fees on investor order choice and execution quality in US stock markets. In: Unpublished working paper, Kelley School of Business, Indiana University, Bloomington.
O'Hara, M., 2015. High frequency market microstructure. J. Financ. Econ. 116, 257–270.
Parlour, C.A., 1998. Price dynamics in limit order markets. Rev. Financ. Stud. 11, 789–816.
Pastor, L., Stambaugh, R., 2003. Liquidity risk and stock returns. J. Polit. Econ. 11, 642–685.
Perold, A.F., 1988. The implementation shortfall: paper versus reality. J. Portf. Manag. 14, 4.
Petersen, M., Fialkowski, D., 1994. Posted versus effective spreads: good prices or bad quotes? J. Financ. Econ. 35, 269–292.
Petersen, M.A., 2009. Estimating standard errors in finance panel data sets: comparing approaches. Rev. Financ. Stud. 22, 435–480.
Roşu, I., 2009. A dynamic model of the limit order book. Rev. Financ. Stud. 22, 4601–4641.
Roll, R., 1984. A simple implicit measure of the effective bid–ask spread in an efficient market. J. Financ. 39, 1127–1139.
Sandås, P., 2001. Adverse selection and competitive market making: empirical evidence from a limit order market. Rev. Financ. Stud. 14, 705–734.
Sarkar, A., Schwartz, R.A., 2009. Market sidedness: insights into motives for trade initiation. J. Financ. 64, 375–423.
Shkilko, A., Sokolov, K., 2020. Every cloud has a silver lining: fast trading, microwave connectivity and trading costs. J. Financ. 75, 2899–2927.
Shkilko, A.V., Van Ness, B.F., Van Ness, R.A., 2008. Locked and crossed markets on NASDAQ and the NYSE. J. Financ. Mark. 11, 308–337.
Stoikov, S., 2018. The micro-price: a high-frequency estimator of future prices. Quant. Financ. 18, 1959–1966.
Yao, C., Ye, M., 2018. Why trading speed matters: a tale of queue rationing under price controls. Rev. Financ. Stud. 31, 2157–2183.
Yueshen, B. Z., 2016. Uncertain market making. In: Unpublished working paper, INSEAD.