A Simple Estimation of Bid-Ask Spreads from Daily Close, High, and Low Prices

Farshid Abdi
University of St. Gallen

Angelo Ranaldo
University of St. Gallen

We propose a new method to estimate the bid-ask spread when quote data are not available. Compared to other low-frequency estimates, this method utilizes a wider information set, namely, readily available close, high, and low prices. In the absence of end-of-day quote data, this method generally provides the highest cross-sectional and average time-series correlations with the TAQ effective spread benchmark. Moreover, it delivers the most accurate estimates for less liquid stocks. Our estimator has many potential applications, including an accurate measurement of transaction cost, systematic liquidity risk, and commonality in liquidity for U.S. stocks dating back almost one century. (*JEL* G15, G12, G20)

Received July 17, 2016; editorial decision May 23, 2017 by Editor Andrew Karolyi.

This paper provides a new method to accurately estimate the bid-ask spread based on readily available daily close, high, and low prices. Akin to the seminal *model* proposed by Roll (1984), the rationale of our estimator is the departure of the security price from its efficient value because of transaction costs. However, our estimator improves the Roll *measure* in two important respects: First, our method exploits a wider information set, namely, close, high, and low prices, which are readily available, rather than only close prices like in the Roll measure. Second, our estimator is completely independent of trade direction dynamics, unlike in the Roll measure, which relies on the occurrence of bid-ask bounces, and, consequently, relies on the assumption of serially independent trade directions that are equally likely.

We thank the editor Andrew Karolyi, an anonymous referee, Yakov Amihud, Allaudeen Hameed, Joel Hasbrouck, Robert Korajczyk, Asani Sarkar, Avanidhar Subrahmanyam, Paul Söderlind, and Jan Wrampelmeyer, as well as the participants of the 2017 AFA meetings in Chicago, 2013 CFE conference in London, and 2016 SFI research days in Gerzensee for comments and suggestions. All remaining errors are our own. We acknowledge financial support from the Swiss National Science Foundation (SNSF; grants 159418 and 154445). Parts of this paper were written while Abdi visited the Stern School of Business, New York University, whose hospitality is gratefully acknowledged. Supplementary data can be found on *The Review of Financial Studies* web site. Send correspondence to Angelo Ranaldo, Swiss Institute of Banking and Finance, University of St. Gallen, Unterer Graben 21, 9000 St. Gallen, Switzerland; telephone: +41712247010. E-mail: angelo.ranaldo@unisg.ch.

doi:10.1093/rfs/hhx084 Advance Access publication August 26, 2017

By virtue of its closed-form solution and straightforward computation, our method delivers very accurate estimates of effective spreads, both numerically and empirically. When quote data are unavailable, our estimator generally provides the highest cross-sectional and average time-series correlation with the effective spread based on Trade and Quotes (TAQ) data, which serve as the benchmark measure. Our estimator can be applied for a number of research purposes and to a variety of markets and assets because it is derived under very general conditions and is easy to compute.

Our estimation of the effective spread shares the theoretical framework with the Roll (1984) model, in which the efficient price of an asset follows a geometric Brownian motion. Within this framework, we follow three innovative steps to derive our simple estimator. First, we build a simple proxy for the efficient price using the mid-range, which we define as the mean of the daily high and low log-prices. The mid-range of every day represents (at least) one point in the continuous path of the efficient log-price process as half-spreads included in the high and low prices cancel out in the mid-range calculation. Moreover, the mean of two consecutive daily mid-ranges represents a natural proxy for the midpoint or efficient price at the time of the market close. In fact, the continuous efficient price path of day t ($t+1$) hits the mid-range before (after) the closing time on day t. Second, we calculate the squared distance between the close log-price and the midpoint proxy at the time of market close. We show that this squared distance is composed of the efficient-price variance and the squared effective spread at the closing time. As the third step, we derive an efficient-price variance estimator as a function of mid-ranges. The efficient-price variance is then removed from the squared distance between the close price and midpoint proxy (obtained in the previous step). The outcome is a simple measure for the proportional spread, $Spread = 2\sqrt{E[(c_t - \eta_t)(c_t - \eta_{t+1})]}$, in which c is the daily close log-price and η is the daily mid-range, that is, the average of daily high and low log-prices. This simple closed-form solution resembles the Roll's autocovariance measure. However, instead of the autocovariance of consecutive close-to-close price returns like in the *Roll* measure, our estimator relies on the covariance of close-to-mid-range returns around the same close price.

One might use low-frequency bid-ask spread measures, instead of the more sophisticated high-frequency measures, to achieve the following goals: (a) measuring bid-ask spreads in the absence of quote data and (b) benefit from the computational savings. Measuring bid-ask spreads when quote data are unavailable is essential, because the access to quote data, even at daily frequency, is limited to certain securities, markets, and (recent) periods.[1] The computational benefits from using low-frequency measures are also substantial because of the overwhelming size of intraday quote data, and time-consuming

[1] For example, end-of-day bid and ask quotes are missing in the CRSP data set from 1942 to 1992.

data handling and filtering techniques.[2] An *approximation* of intraday bid-ask spreads with end-of-day quotes[3] provides accurate measures and computational savings (Chung and Zhang 2014; Fong, Holden, and Trzcinka 2017). However, the availability of end-of-day quote data for the last 75 years of U.S. stocks is limited to a recent period, that is, from 1993 onwards. As TAQ data are also available for this time period, end-of-day quotes are mostly helpful for the purpose of saving computational time. Thus, needs for an accurate measurement of bid-ask spreads when (intraday) quote data are unavailable remain unmet. The previous literature overcomes this issue by employing price data to *estimate* the effective spread.[4] Starting with the Roll (1984) measure (hereafter *Roll*), a number of models have been proposed. Hasbrouck (2004, 2009) proposes a Gibbs sampler Bayesian estimation of the Roll model (hereafter *Gibbs*). Lesmond, Ogden, and Trzcinka (1999) introduce an estimator based on zero returns (*LOT*). Compared with *Roll*, estimating the *LOT* measure is computationally intensive since it relies on optimizing the maximum likelihood function for every single month to get the monthly estimates. Following the same line of reasoning, Fong, Holden, and Trzcinka (2017) develop a new estimator (*FHT*) that simplifies existing *LOT* measures. Holden (2009), jointly with Goyenko, Holden, and Trzcinka (2009), introduces the Effective Tick measure based on the concept of price clustering (*EffTick*). By taking their difference, the high and low prices have been traditionally used to proxy volatility (e.g., Garman and Klass 1980; Parkinson 1980; Beckers 1983). More recently, Corwin and Schultz (2012) use them to put forward an original estimation method for transaction costs (*HL*). Assuming the high (low) price being buyer- (seller-) initiated, they decompose the observed price range into two parts: efficient price volatility and bid-ask spread. To cover a wide range of applications, we perform our analysis across various sample periods, including the 1993–2015 period, in which end-of-day quoted spreads are also available, to compare spread estimates to the accurate TAQ effective spread benchmark, and from 1926 onwards to embrace the entire price data history of U.S. stock markets.

This paper contributes to the literature by providing a new estimation method of transaction costs *jointly* based on close, high, and low prices. The rationale of our model is to bridge the two above-mentioned estimation methodologies, that is, the long-established approach based on close prices originated from Roll (1984) and the more recent one relying on high and low prices (Corwin and Schultz 2012). In doing so, our model has four main advantages over the previous estimation methods. First, the joint utilization of the daily high, low,

[2] The key advantages of using daily data, including large computational time savings, are comprehensively discussed by Holden, Jacobsen, and Subrahmanyam (2014).

[3] The use of end-of-period quotes, at frequencies lower than daily, goes back to Stoll and Whaley (1983).

[4] Rather than approximating and estimating transaction costs, an alternative approach to measuring illiquidity is to use proxies for the price impact, in particular the Amihud (2002) illiquidity measure.

and close prices allows our model to benefit from the richest readily available information set of price data.[5] Second, unlike Roll (1984), our measure does not rely on bid-ask bounces and, therefore, is independent of trade direction time-series dynamics of close prices. Third, unlike Corwin and Schultz's (2012) *HL* estimator, our model neither needs to violate Jensen's inequality in order to construct the closed-form estimator nor does it need ad hoc adjustments for nontrading periods, such as weekends, holidays, and overnight closings. Finally, our estimates using the mid-range and close price are only marginally sensitive to the number of trades per day, whereas the high-low estimator proposed by Corwin and Schultz (2012) further underestimates effective costs when the daily number of trades are lower, that is, when stocks (and markets) are less liquid.

We empirically test our method by using daily CRSP data to estimate bid-ask spreads and compare the monthly estimates to TAQ data, which serves as the benchmark to compute the effective spread. As recommended by Holden and Jacobsen (2014), we use *Daily* (Millisecond) TAQ data to enhance the precision of our analysis. Thus, the availability of the Daily TAQ data naturally defines our main sample period, which spans from October 2003 to December 2015, that is, 147 months. Then, we assess the performance of our method by comparing bid-ask spread estimates with the *Monthly* TAQ data between January 1993 and September 2003 thus extending our analysis to 23 years of TAQ data, that is, from the beginning of 1993 to the end of 2015. As emphasized in the literature, for example, by Goyenko, Holden, and Trzcinka (2009), the decision criteria for selecting the best estimator depends on the particular application of the estimates. To cover the widest range of possible applications, we use three different criteria to gauge the quality of the estimators: cross-sectional correlation, time-series correlation, and prediction errors. To ensure a comprehensive assessment, we consider the average correlations for all the available stocks, as well as for subsamples, based on a variety of criteria, including shorter time periods, primary exchanges (NYSE, AMEX, and NASDAQ), market capitalization, and the magnitude of bid-ask spreads.

Several clear results emerge from our study. First, the closing percentage quoted spread is generally the most accurate monthly spread proxy according to the above-mentioned criteria. This is generally true when end-of-day quote data are available, i.e. from 1993 onwards, except for the predecimalization era in U.S. stock markets that dates before 2001.[6] During the 1993–2000 period, end-of-day spreads provide the highest average time-series correlations compared

[5] Unlike the availability of close, high, and low prices, the availability of open prices is subject to additional limitations. For example, open prices are missing in the CRSP data between July 1962 and June 1992.

[6] NYSE (NASDAQ) decimalization started for few of the listed stocks in August 2000 (March 2001), followed by wider implementation in the next months and completion in January 2001 (April 2001).

to the TAQ effective spreads, whereas our estimates show highest average cross-sectional correlations and lowest estimation errors.

Second, our estimator provides the most accurate estimates in the absence of quote data, making it the best choice for applications that rely on longer time horizons, going back beyond 1993. Compared with other bid-ask spread estimators that do not rely on quote data (the *HL*, *Roll*, *Gibbs*, *EffTick*, and *FHT* measures), it provides the highest cross-sectional correlation with the intraday effective spread. On a monthly basis, the average cross-sectional correlation of our estimates with the Daily TAQ effective spreads is 0.74, whereas the other estimators range from 0.37 to 0.65. The analysis of Monthly TAQ data from 1993 to 2003 delivers consistent results, that is, our estimates have the highest average cross-sectional correlation of 0.86, whereas those of other estimators range from 0.61 to 0.83. These results are consistent whether correlations are taken for estimates in levels or in changes, and across subperiods. When breaking down the cross-section of stocks into quintiles based on companies' size and effective spread size, our estimator provides the highest cross-sectional correlations for small to medium market capitalizations and for a medium to large effective spread size. This can be seen as a suitable characteristic because accurate estimates of transaction costs are particularly needed for less liquid securities.

Third, in the absence of end-of-day quotes, our estimator also delivers the highest average time-series correlations with the effective spread benchmark. Compared with other estimators, it provides the highest average time-series correlations over the entire sample period, across two out of three market venues (AMEX and NASDAQ), for small to medium market capitalizations, and for a medium to large effective spread size.

Finally, in the absence of end-of-day quotes, our estimates generally exhibit the lowest prediction errors in terms of root-mean-square errors (RMSEs) when compared with the TAQ benchmark. The overall evidence suggests that our estimates are the best available option (a) in the absence of quote data, according to all three criteria or (b) according to two out of the three criteria, when end-of-day quote data are less accurate, that is, during the predecimalization era.

A natural question is whether our estimator provides additional information beyond that contained in the other estimators. To answer this question, we measure partial correlations between our estimates and the TAQ benchmark, while controlling for *HL*, *Roll*, *Gibbs*, *EffTick*, and *FHT* estimates. We find that the average partial cross-sectional and partial time-series correlations for our estimates are significantly positive for the entire sample, for every primary exchange, and for every effective-spread quintile. Average partial correlations are especially higher for quintiles with a medium to large effective spread size; that is, our estimator provides even more additional explanatory power for less liquid stocks. These results are in line with our numerical analysis that document the marginal sensitivity of our estimates to the number of trades per day, whereas Corwin and Schultz's (2012) method produces substantially smaller estimates of transaction costs for less-frequently traded stocks.

An accurate measurement of transaction costs is important for at least two applications: First, to analyze how and to what extent transaction costs erode asset returns (e.g., Amihud and Mendelson 1986). To illustrate the potential application in this respect, we compute estimates of bid-ask spreads for NYSE (AMEX) stocks for the period from 1926 (1962) through 2015. Then we discuss the reliability of our estimator in describing the developments of transaction costs over long time spans and for large cap and small cap stocks. Second, investors demand a premium for liquidity risk, that is, the chance that liquidity disappears when it is needed to trade. To comprehend this issue, it is necessary to obtain accurate estimates of transaction costs for individual stocks, stock portfolios, and the whole market. Through the lens of the liquidity-adjusted capital asset pricing model (LCAPM), proposed by Acharya and Pedersen (2005), we analyze which model provides accurate estimates of systematic liquidity risk, that is, estimates close to those based on the TAQ effective spreads. We show that our model precisely captures all the different components of systematic liquidity risk in the cross-section of the market, in particular the component originated by comovements of liquidity of individual stocks and that of the whole market, that is, commonality in liquidity, as well as negative covariations between stock returns and illiquidity. Overall, our model provides more accurate estimates for (liquidity) systematic risk than do the *Roll* and *HL* estimators, and it can be used to analyze commonality in liquidity and return-liquidity covariations. Our estimator has many potential applications in areas other than asset pricing, including corporate finance, risk management, and other important research areas that need an accurate measure of trading costs over long periods.

1. The Estimator

We first explain our model in theory, and then, provide details for its best use in practice.

1.1 Model

Our model relies on assumptions similar to those made in the Roll (1984) model. We assume that the efficient price follows a geometric Brownian motion (GBM) and the observed price at each time point can be either buyer initiated or seller initiated. To keep the notation concise, we directly implement the model on *log-price*, and the superscript e refers to efficient prices. Equation (1) shows how the observed market price and efficient price at the closing time are related. The random variable c_t represents the observable close log-price, and the random variable c_t^e represents the efficient log-price at the closing time. The random variable q_t is the trade direction indicator, and s is the relative spread, which we aim to estimate. In line with Roll (1984), we assume that trade directions are independent of the efficient price.

$$c_t = c_t^e + q_t \frac{s}{2}, q_t = \pm 1 \tag{1}$$

For the sake of convenience, we temporarily make two assumptions. However, our estimator is robust to the relaxation of these assumptions as shown in the appendix. Like in Corwin and Schultz (2012), the first assumption is that the high price (h_t) to be buyer initiated ($q_t^h = 1$), and the daily low price (l_t) to be seller initiated ($q_t^l = -1$). Equations (2) and (3) represent these points.

$$h_t = h_t^e + \frac{s}{2}, \tag{2}$$

$$l_t = l_t^e - \frac{s}{2}. \tag{3}$$

This assumption likely holds for frequent trades on a continuous efficient price path, which allows both buyer-initiated and seller-initiated trades to occur when the *efficient* price process is near its high (low) values. In such circumstances, a non-zero spread size make buyer- (seller-) initiated trades higher (lower) than the ones of opposite direction, increasing the chance to *select* the buyer- (seller-) initiated trades as the high (low) trade prices. It is worth stressing that this assumption seems to be supported by real data.[7] Moreover, our results are robust to relaxation of this assumption analytically and numerically. Our results analytically hold when we relax this assumption by allowing trade directions of high and low prices being stochastic and independent of the efficient price process (see Appendix C). Furthermore, when we relax this assumption in our numerical simulations, our estimator still outperforms its competitors when the trades are less frequently observed (increasing the chance to violate Equations (2) and (3)).

The second simplifying assumption that we make is that the efficient-price movement during nontrading periods is zero. As we show analytically in Appendix B and later in numerical simulations, our results are also robust to the relaxation of this assumption. We start with defining mid-range and then derive our estimator using the mid-range.

Definition 1. We define the mid-range as the average of daily high and low log-prices:

$$\eta_t \equiv \frac{(l_t + h_t)}{2} \tag{4}$$

One can replace the efficient high and low log-prices with the observed values since the spreads cancel out.

[7] Using Daily TAQ data between October 2003 and December 2015 and an algorithm similar to Lee and Ready (1991), we observe that around 90% (91%) of stocks-days include high (low) prices that are above (below) the quote midpoints. The Internet Appendix provides more details.

Proposition 1. Assuming that the efficient price follows a continuous path (in our case a GBM):

(i) The mid-range of observed prices coincides with mid-range of efficient price:

$$\eta_t = \frac{(l_t^e + h_t^e)}{2} \tag{5}$$

(ii) η_t represents at least one point in the efficient-price process. In other words, the efficient price hits η_t at least once during the day.

(iii) A straightforward and unbiased proxy for the end-of-day midquote of day t is the average of mid-ranges of the same day and the next day, since the end of the day midquote of day t occurs between the time at which η_t and η_{t+1} are hit. As shown in Equation (6), this proxy is unbiased:

$$E\left[c_t^e - \frac{(\eta_t + \eta_{t+1})}{2}\right] = 0. \tag{6}$$

Proposition 2. The squared distance between close log-price of day t and the proposed mid-point proxy includes two components: bid-ask spread component and efficient price variance component Equation (7) shows this relation:

$$E\left[\left(c_t - \frac{(\eta_t + \eta_{t+1})}{2}\right)^2\right] = s^2/4 + (1/2 - k_1/8)\sigma_e^2, \quad k_1 \equiv 4\ \ln(2). \tag{7}$$

Garman and Klass (1980), Parkinson (1980), and Beckers (1983) use the value of k_1 for the purpose of estimating volatility using the daily price range. Here, rather than using the range, we take the average of high and low prices and use it as an efficient price proxy. Proofs for Propositions 2 and 3 are available in Appendix A. The effective half-spread, by definition, is the distance between the price and the contemporaneous midquote. We interpret Equation (7) to be a characterization of the standard definition of the effective half-spread, that is, when the unobservable midpoint is proxied by the average mid-ranges. We argue that the average of the consecutive mid-ranges of days t and $t+1$ is a natural proxy for the midquote or the efficient price at the closing time of day t since the mid-range of day t occurs before the closing time and the mid-range of the next day occurs after it. As expressed in Equation (7), the squared distance between the close price and the proxy for the midquote contains two components: the squared effective half-spread and the transitory variance. The squared effective spread term represents the squared distance between the observed close price and the midquote at the time of market close. The transitory variance term represents the squared distance between the midquote at the close time and its approximation, that is, the average of two consecutive mid-ranges. Figure 1 provides a graphical illustration of the two components of the dispersion measure introduced in Equation (7) in the framework of the

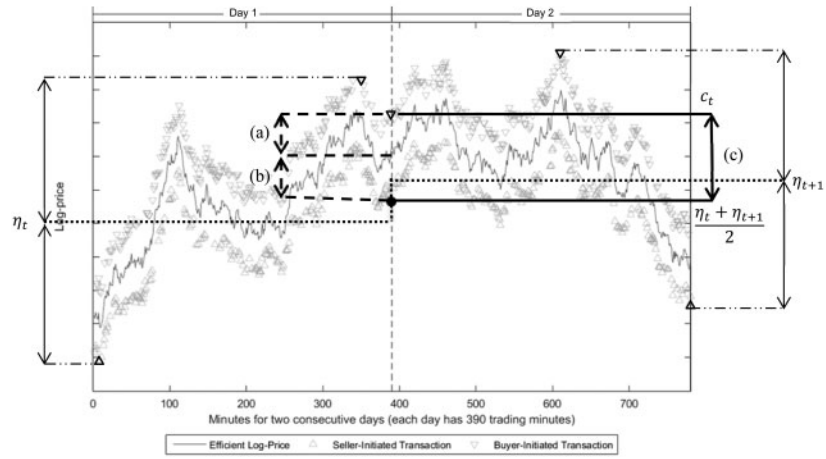

Figure 1
The schematic decomposition of the distance between closing price and average mid-ranges
The log-price process is simulated with one-minute increments for the duration of two days of working hours. Each working day consists of 390 minutes, with one trade at the end of every minute. The input of our model, which consists of daily price data, is represented by the five thicker triangulars. Four triangulars represent the two high and low prices for days t and $t+1$, and one represents the close price at day t. The figure provides a simple illustration that the distance between c_t and $(\eta_t + \eta_{t+1})/2$, shown as (c) in the picture, can be decomposed into two components: (a) the distance between close price and the unobserved efficient close price, that is, the effective half-spread and (b) the distance between efficient close price and the midquote proxy.

Roll (1984) model. The figure illustrates that the distance between the close price and the average of the two consecutive mid-ranges reflects two quantities, namely, the effective spread and the intraday efficient-price variation (σ_e^2). As the next step, we propose a way to compute a measure of intraday volatility, which we will remove from the dispersion between the close price and the midquote proxy.

Proposition 3. The variance of changes in mid-ranges is a linear function of efficient price variance. Equation (8) provides the accurate relation:

$$E\left[(\eta_{t+1} - \eta_t)^2\right] = (2 - k_1/2)\sigma_e^2, \quad k_1 \equiv 4\ \ln(2). \tag{8}$$

Since the mid-ranges are both independent of the spread, their difference only reflects the volatility of the efficient-price path. We also perform several numerical simulations to assess the quality of the estimate of the efficient price volatility in Proposition 3. We find two main results: First, the estimated efficient price volatility implied by our model closely follows the "true" efficient price volatility. Second, our volatility estimate is less sensitive to the trading frequency. In other words, it is still accurate and less biased than the high-low volatility estimates, even for a very low frequency of trades. This is a favorable property of our volatility estimates compared to the use of price range, which, as shown in Garman and Klass (1980) and Beckers (1983), is

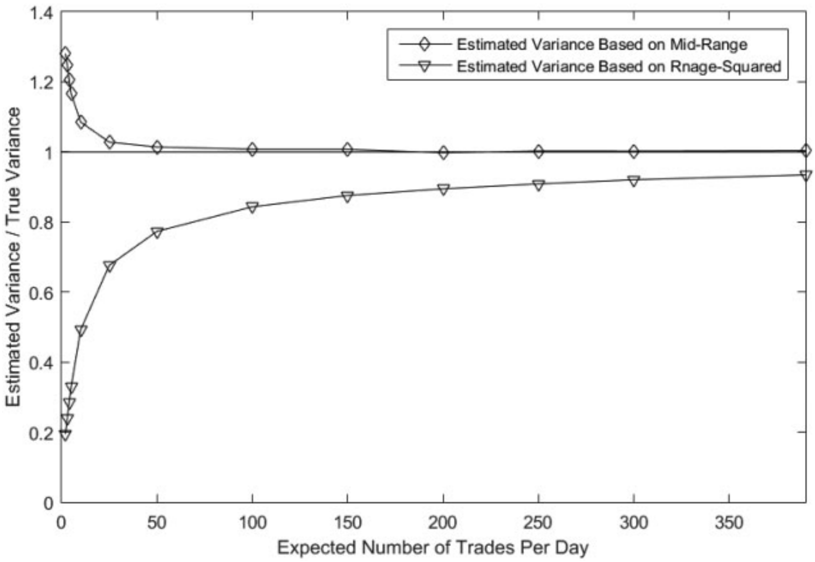

Figure 2
Sensitivity of variance estimates to the number of daily trades
The figure shows the relative bias of variance estimates, using ranges and mid-ranges of a simulated discrete random walk to estimate the variance, and the sensitivity of the bias to the expected number of trades per day. We simulate a random walk for 210,000 days, with 390 one-per-minute trades, and a daily volatility of 3%. Each trade has certain chance of being observed, allowing the expected number of trades specified in the horizontal axis, ranging from 2 to 390. The variance based on the mid-range is calculated as $\sigma_\eta^2 = 1/(2-2\log(2))\ E[(\eta_t - \eta_{t-1})^2]$, and the range-based variance is calculated as $\sigma_{h-l}^2 = 1/(4\log(2))E[(h_t - l_t)^2]$ Expected values are estimated by using the means of a sample of 210,000 day simulations. The estimation outputs are divided by the preassigned variance of 0.03^2 in order to be comparable with 1.

considerably biased if the trades are observed less frequently. Figure 2 illustrates the explained simulation results. By its accurate estimation of efficient price variance, Proposition 3 provides us with a way to remove the efficient price variance part introduced in Proposition 2.

Theorem 1. The squared effective spread can be estimated as shown in Equation (9):

$$s^2 = 4E\left[\left(c_t - (\eta_t + \eta_{t+1})/2\right)^2\right] - E\left[(\eta_{t+1} - \eta_t)^2\right] = 4E[(c_t - \eta_t)(c_t - \eta_{t+1})] \quad (9)$$

Proof of Theorem 1: Multiplying both sides of Equation (7) by four, subtracting Equation (8), and simplifying the outcome expression leads to Equation (9).

Interestingly, the estimator derived in Theorem 1 resembles the *Roll* autocovariance measure, in which $c_{t+1}(c_{t-1})$ is replaced with $\eta_{t+1}(\eta_t)$. However, this simple and intuitive formulation leads to some important improvements. Hereafter, we compare our estimator to the *Roll* and *HL* measures.

Unlike Roll (1984), the derivation of Equation (9) does not need to rely on additional restrictive assumptions on the serial independence of trades and equal likelihood of buyer-initiated and seller-initiated close price, which do not find empirical support.[8] Compared to the *HL* estimator (Corwin and Schultz 2012), our model should perform better for at least three reasons: First, it benefits from the richer readily available information set of price data, i.e. the daily high, low, and close prices. Second, unlike Corwin and Schultz's (2012) *HL* estimator, our model is robust to the price movements in nontrading periods, such as weekends, holidays, and overnight price changes. Therefore, it does not rely on ad-hoc overnight price adjustments.[9] Finally, by relying on the average of high and low prices instead of the price range, our model is less sensitive to the number of observed trades per day. This is a key advantage that we will analyze numerically and empirically in the next Sections.

1.2 Dealing with negative estimates

We aim to use the model to estimate effective spreads for every month-stock. One can estimate the expectation term in Equation (9) by using the sample "moment," that is, a simple average of the two-day values, and, then, by taking the squared root of the outcome to get the spread estimate. However, because of the estimation errors, the estimation of the right-hand side expression in Equation (9) might become negative. Three ways to deal with this issue have been suggested in the previous literature (e.g., Corwin and Schultz 2012): (1) set negative monthly estimates to zero, and then calculate the spread (2), set negative two-day estimates to zero and then take the average of the two-day calculated spreads, or (3) remove negative estimates and just calculate the spread for positive estimates and take their average. Numerical simulations and empirical comparisons with the TAQ data indicate that the first two approaches provide better outcomes, both in terms of bias and estimation errors. We call the first approach the *monthly corrected* estimate and the second one the *two-day corrected* version. Equations (10) and (11), respectively, show the way we calculate the two versions.

$$\hat{s}_{monthly\ corrected} = \sqrt{\max\left\{4\frac{1}{N}\sum_{t=1}^{N}(c_t-\eta_t)(c_t-\eta_{t+1}),0\right\}}, \tag{10}$$

$$\hat{s}_{two\text{-}day\ corrected} = \frac{1}{N}\sum_{t=1}^{N}\hat{s}_t, \quad \hat{s}_t = \sqrt{\max\{4(c_t-\eta_t)(c_t-\eta_{t+1}),0\}}. \tag{11}$$

where N shows the number of days in the month and $\hat{s}_t$ refers to the two-day estimates. As shown in Equation (11), to calculate the two-day corrected

[8] Hasbrouck and Ho (1987) and Choi, Salandro, and Shastri (1988), among others, find a serial dependence in the trade directions, and Harris (1989) and McInish and Wood (1990) show that close prices are more likely to be buyer initiated than seller initiated.

[9] Appendix B provides the proof.

version, we follow three steps. First, we calculate estimates of squared spreads over two-day periods. If the two-day estimates are negative, then we set them to zero. Second, we take their square roots. Finally, we average them over a month. This way of taking average of two-day estimates after removing negative values is similar to the correction method applied by Corwin and Schultz (2012). Although the two-day correction approach increases the bias because of setting more negative values to zero compared with the monthly corrected version, it provides better results in terms of higher correlation with the high-frequency benchmark (Corwin and Schultz 2012).

The better association of the two-day corrected version with real data can be explained by some restrictive assumptions in the Roll (1984) model, which our estimator also relies on, in particular the constant spread and volatility. First, the monthly corrected estimate hinges on $E(s^2)$, which consists of the squared mean, plus the variance of bid-ask spreads. This is larger than the squared mean when the spread is not constant. With the use of a two-day period for the spread estimation, we isolate a single incident of a close-price transaction, and therefore, no assumption on the distribution of the spread over consecutive days is needed. Second, the two-day time window is more inclined to capturing transient price patterns, such as heteroscedasticity and volatility clustering.

1.3 Other spread estimators that use daily data

Here, we shortly review the most common methods for bid-ask spread estimation, which we empirically analyze in the next sections, and summarize in Table 1. For the sake of completeness, we include the average of the end of the day CRSP *quoted* spreads, which generally provide accurate *approximation* of bid-ask spreads (Chung and Zhang 2014; Fong, Holden, and Trzcinka 2017). However, the main interest of this paper is to compare *estimation* methods based on *price* data when quote data are not available.

Roll (1984) initiated the use of price data for bid-ask spread estimation. To return a nonnegative spread, the first-order autocovariance of the price changes must be negative. However, Roll (1984) finds positive estimated autocovariances for several stocks, even over a one-year sample period. Harris (1990) finds out that the positive estimated autocovariances are occurring when the spreads tend to be smaller. This motivates the common practice of replacing the positive autocovariances with zero to get a zero spread estimate.

Hasbrouck (2004, 2009) develops a Gibbs sampler Bayesian estimator to overcome the negative spread estimates. Using annual estimates, Hasbrouck (2009) shows that the spreads originated from the *Gibbs* method have higher correlations with the high-frequency benchmark. Following Corwin and Schultz (2012), among others, we perform our empirical analysis on a monthly basis.[10]

[10] Joel Hasbrouck has kindly provided the SAS codes for the Gibbs sampler estimator on his personal Web page. We modify the codes by altering the estimation windows from stock-years into stock-months. We only consider

Table 1
Other bid-ask estimation methods using daily data

Label	Inputs	Description
Roll	Close price	$Roll = 2\sqrt{\max\{-cov(\Delta c_{t+1}, \Delta c_t), 0\}}$, where c is close log-price
Gibbs	Close price	Gibbs sampler Bayesian estimation of spreads by setting a nonnegative prior density for the spreads
EffTick	Close price	$EffTick = \frac{\sum_{j=1}^{J} \hat{\gamma}_j S_j}{\bar{P}}$, $S_j = \$1/8, \$1/4, \$1/2, \$1$, $\hat{\gamma}_j = \begin{cases} \text{Min}\left[\text{Max}\left\{U_j, 0\right\}, 1\right], & j=1 \\ \text{Min}\left[\text{Max}\left\{U_j, 0\right\}, 1 - \sum_{k=1}^{j-1} \hat{\gamma}_k\right], & j=2,\dots,J \end{cases}$, $U_j = \begin{cases} 2F_j, & j=1 \\ 2F_j - F_{j-1}, & j=2,\dots,J-1, \\ F_j - F_{j-1}, & j=J \end{cases}$ $F_j = \frac{N_j}{\sum_{j=1}^{J} N_j}$, where N_j is the number of prices divisible by S_j
FHT	Close price	$FHT = 2\sigma N^{-1}\left(\frac{1+\text{Zeros}}{2}\right)$, $\text{Zeros} = \frac{\text{ZRD}}{\text{TD+NTD}}$, where *ZRD*, *TD*, and *NTD* are respectively number of days with zero returns, total number of trading days, and number of nontrading days. $N^{-1}(.)$ refers to the inverse normal cumulative distribution function
HL	High price and low price	$HL = \frac{1}{N}\sum_{t=1}^{N} \hat{s}_t$, $\hat{s}_t = \max\left\{\frac{2(e^{\alpha_t}-1)}{1+e^{\alpha_t}}, 0\right\}$, $\alpha_t = \frac{\sqrt{2\beta_t}-\sqrt{\beta_t}}{3-2\sqrt{2}} - \sqrt{\frac{\gamma_t}{3-2\sqrt{2}}}$, $\beta_t = 1/2\left[\left(h^*_{t+1} - l^*_{t+1}\right)^2 + \left(h^*_t - l^*_t\right)^2\right]$, $\gamma_t = \left(\max\{h^*_{t+1}, h^*_t\} - \min\{l^*_{t+1}, l^*_t\}\right)^2$, where h^* and l^* are respectively high, and low log-prices, adjusted for overnight price movements N shows the number of two-day estimates in the month and $\hat{s}_t$ refers to the two-day spread estimate
CRSP_S	Bid and ask quotes	$CRSP\_S = \overline{\frac{Ask_t - Bid_t}{M_t}}$, $M_t = \frac{Ask_t + Bid_t}{2}$, where *Ask* and *Bid* are CRSP bid and ask quotes. Zero bid-ask spreads, and the ones higher than 50% are discarded before taking the monthly average

This table summarizes the bid-ask estimators used in this paper, which are Roll (Roll 1984), Hasbrouck (Gibbs; 2004, 2009), Holden, jointly with Goyenko, Holden, and Trzcinka (EffTick; 2009, 2009), Fong, Holden, and Trzcinka (FHT; 2017), and Corwin and Schultz (HL; 2012). We include the CRSP end-of-day bid-ask spreads (*CRSP_S*), a measure based on quote data, like in Chung and Zhang (2014).

Fong, Holden, and Trzcinka (2017) develop an estimator, named *FHT*, which relies on the assumption that price movements that are smaller than the bid-ask spread will be unobservable and are reflected in the days with zero returns. They argue that the measure simplifies the LOT measure developed by Lesmond, Ogden, and Trzcinka (1999) and it performs very well in estimating liquidity

---

stock-months in which there are at least 12 days with trades. As he already noted on his Web page, the monthly estimator is less accurate than is the annual version because of the weight of the prior density in the outputs.

of the global equity market to the extent that it becomes one of the most accurate measures.

Holden (2009), jointly with Goyenko, Holden, and Trzcinka (2009), develops a proxy for the effective spread based on observable price clustering. Larger spreads are associated with larger effective tick sizes. The steps to calculate their EffTick measure are shown in Table 1.

More recently, Corwin and Schultz (2012) develop an estimator based on daily high and low prices. They argue that high (low) prices are almost always buyer (seller) initiated. Therefore, the daily price range reflects both the efficient price volatility and its bid-ask spread. They build their model on the comparison of one- and two-day price ranges. The latter should twice reflect the variance of the former, but they should have the same bid-ask spread. This reasoning gives a nonlinear system of two equations with two unknowns that does not have a general closed-form solution. The authors provide an approximate closed-form solution at the cost of neglecting Jensen's inequality.

2. Numerical Simulations

In this section, we perform several numerical simulations under different settings. For ease of comparison, we define the setting of simulations similar to that in Corwin and Schultz (2012). We compare two versions of our measure, labeled *CHL,* with the *HL* and *Roll* estimates, that is, the monthly corrected and the two-day corrected versions.[11]

Panel A of Table 2 shows the results for the near-ideal settings. For each relative spread under analysis, we perform 10,000 time simulations for 21-day months of the price process. Each day consists of 390 minutes in which trades are observable. We simply draw from $M_t = M_{t-1} e^{z\sigma/\sqrt{390}}$, $P_t = M_t e^{q_t s/2}$, $z \sim N(0,1)$, where M_t and P_t represent the efficient price and observed transaction price at time t, respectively. We set the daily standard deviation of efficient-price return, σ to be 3%. q_t can be equally likely -1 or +1 for every individual observed trade, relaxing the assumption of buyer- (seller-) initiated high (low) prices. We report both the bias and the estimation errors, in terms of RMSEs, in the table. The results showed in panel A are twofold: First, both *CHL* and *HL* show considerably lower estimation errors compared to the *Roll*. Second, although the *CHL* monthly corrected estimates tend to be less-biased than the two-day corrected version, they do not show very different estimation errors.

[11] Shane Corwin has kindly provided the SAS codes for the HL estimator on his personal Web site. The code produces several versions of spread estimates. We consider two of them in our simulations. The first version, named MSPREAD_0, is calculated by setting two-day negative estimates to zero and then taking the monthly average. The second version, named XSPREAD_0, is calculated by directly setting the negative monthly averaged estimates to zero. Although the second version produces less-biased results in some simulation cases, Corwin and Schultz (2012) advocate the former method, which is better associated with the TAQ benchmark.

Table 2
Estimated bid-ask spreads from simulations

	Bias					RMSEs				
	CHL		HL		Roll	CHL		HL		Roll
	2-day	Month	2-day	Month		2-day	Month	2-day	Month	
A. Near-ideal conditions										
0.5% spread	0.7%	0.2%	0.9%	0.1%	0.7%	0.8%	0.8%	1.0%	0.5%	1.5%
1.0% spread	0.3%	0.0%	0.8%	0.0%	0.3%	0.5%	0.8%	0.8%	0.6%	1.5%
3.0% spread	−0.6%	−0.1%	0.2%	−0.1%	−0.4%	0.8%	0.7%	0.5%	0.6%	1.9%
5.0% spread	−0.7%	0.0%	0.0%	−0.1%	−0.4%	0.9%	0.6%	0.6%	0.6%	2.2%
8.0% spread	−0.4%	0.0%	−0.2%	−0.2%	−0.5%	0.7%	0.5%	0.6%	0.6%	2.7%
B. Each trade is visible with a chance of 10% (average of 39 trades Per day)										
0.5% spread	0.7%	0.2%	0.6%	−0.3%	0.7%	0.8%	0.8%	0.7%	0.4%	1.5%
1.0% spread	0.3%	−0.1%	0.3%	−0.5%	0.3%	0.5%	0.8%	0.5%	0.7%	1.5%
3.0% spread	−0.6%	−0.1%	−0.3%	−0.8%	−0.4%	0.8%	0.8%	0.6%	1.0%	1.9%
5.0% spread	−0.7%	−0.1%	−0.7%	−0.8%	−0.5%	0.9%	0.6%	0.9%	1.0%	2.2%
8.0% spread	−0.5%	−0.1%	−0.9%	−0.9%	−0.4%	0.8%	0.6%	1.1%	1.1%	2.6%
C. Each trade is visible with a chance of ≈0.5% (average of two trades per day)										
0.5% spread	0.6%	0.3%	0.0%	−0.4%	1.0%	0.8%	1.0%	0.3%	0.5%	2.3%
1.0% spread	0.2%	−0.1%	−0.5%	−0.9%	0.6%	0.6%	1.0%	0.6%	0.9%	2.2%
3.0% spread	−0.9%	−0.9%	−1.9%	−2.6%	−0.4%	1.3%	1.7%	2.0%	2.6%	2.8%
5.0% spread	−1.8%	−1.3%	−3.1%	−3.8%	−1.0%	2.1%	2.1%	3.2%	4.0%	3.8%
8.0% spread	−2.7%	−1.8%	−4.6%	−5.2%	−1.6%	3.2%	2.7%	4.8%	5.5%	5.4%
D. Random spreads										
0.5% Spread	0.7%	0.2%	0.9%	0.1%	0.7%	0.8%	0.8%	1.0%	0.5%	1.6%
1.0% Spread	0.4%	0.0%	0.7%	0.0%	0.4%	0.5%	0.9%	0.8%	0.6%	1.5%
3.0% Spread	−0.3%	0.4%	0.1%	−0.2%	0.0%	0.6%	0.8%	0.6%	0.7%	1.9%
5.0% Spread	−0.3%	0.7%	−0.3%	−0.5%	0.4%	0.9%	1.1%	0.8%	0.9%	2.3%
8.0% Spread	−0.3%	1.2%	−1.0%	−1.1%	0.7%	1.2%	1.6%	1.5%	1.6%	3.1%
E. Rare trades, random spreads, and overnight price movements, all together										
0.5% spread	0.7%	0.4%	0.0%	−0.4%	1.2%	0.9%	1.1%	0.4%	0.5%	2.6%
1.0% spread	0.3%	0.1%	−0.4%	−0.9%	0.9%	0.7%	1.1%	0.6%	0.9%	2.6%
3.0% spread	−0.8%	−0.6%	−1.9%	−2.5%	−0.1%	1.3%	1.7%	2.0%	2.6%	3.1%
5.0% spread	−1.5%	−0.8%	−3.1%	−3.8%	−0.4%	2.1%	2.3%	3.2%	4.0%	4.3%
8.0% spread	−2.5%	−1.1%	−4.8%	-5.5%	−0.6%	3.4%	3.1%	5.0%	5.8%	6.0%

Each simulation consists of 10,000 21-day months of stock prices, and each day consists of 390 minutes. For each minute, the trajectory of a geometric Brownian motion with daily volatility of 3% and a constant relative spread with the values mentioned in the table is simulated. The labels in the first row refer to the estimators from the following models: ours (CHL), Corwin and Schultz's (HL; 2012), and Roll's (Roll; 1984). 2-day and month refer to the two-day corrected and monthly corrected versions, in which two-day or monthly negative estimates are set to zero. We run the simulations in five separate scenarios. Panel A shows the results in the near-ideal situation. Panel B shows the results when trades in each minute are observable with only a 10% chance. Panel C shows the results when on average only two trades are observed per day. That is, trades in each minute are observable with around 0.5% chance. For both monthly and two-day corrected estimates, every two-day input that included a day with no trade or only one trade is discarded. Panel D shows the results when the spreads of each day are uniformly distributed between zero and twice the nominal average value. Panel E encompasses the "imperfections" in scenarios C, D, and adding an "overnight" price change with 50% of the standard deviation of the daily price change. The overnight adjustment procedure for HL estimates is as the same as that used in Corwin and Schultz (2012).

2.1 Less-frequently observed trades

Using a similar setting, we now consider a certain chance to observe each of the one-minute trades, which can introduce a downward bias in estimating the variance using the range (Garman and Klass 1980; Beckers 1983). We already confirmed its effect on volatility estimates in the previous section. Here, we aim to assess how the environment of infrequent trades affects bid-ask spread estimates. As the downward bias is larger for the cases with less-frequently observed trades, we design two separate settings. In panel B of Table 2, each per-minute trade has a 10% chance of being observed, allowing an average of 39 trades per day. In panel C, each trade only has a chance of 2/390≈0.5% of being observed, allowing an average of two trades per day. This implies that sometimes there are no transactions or only one trade per day meaning identical high and low prices, and zero range. To avoid these cases, we discard any two-day period that includes a nontrading day or a day with zero price range, and calculate the spreads for the rest of the two-day periods in the sample.

Three clear results emerge from this analysis. First, under the most challenging circumstances in panel C, *HL* estimates are always more severely downward biased compared to the *CHL* estimates. Second, comparing panels A and B, it is clear that even a small reduction of number of trades per day leads to a significant change in levels for the *HL* estimates, but not for *CHL* estimates. Finally, in both settings of moderate and low number of observed trades per day, *CHL* estimates tend to have lower estimation errors when the effective spreads are large, which represent the less liquid stocks.

To visualize the estimates' sensitivity to the number of trades per day, we perform several simulations allowing different number of observed trades per day, with averages between 390 and 2 trades (as in the Table 2). Figure 3 shows the *CHL* and *HL* two-day corrected estimates for these simulations in which the bid-ask spread is set to be 1%. Each one-minute trade is observed with a certain chance, which is set in a way that allows the average number of trades per day being the values shown in the horizontal axis. The figure illustrates two main findings: First, the *CHL* estimates are only marginally sensitive to the number of observed trades per day, and only for the very low number of trades per day, say below five trades per day. The opposite applies for *HL* estimates. From 5 to 390 trades per day, the *HL* estimates range from 74 to 175 bps (instead of the 100-bps "true" proportional spread), whereas the C*HL* estimates remain in a narrow range from 127 to 132 bps. The steepness of the *HL* curve in the figure illustrates the high sensitivity of the *HL* estimates to the number of trades per day, especially below 100 trades per day. Perhaps a more important concern is the direction of this sensitivity, which entails that the *HL* estimates indicate a considerably narrower spread when fewer transactions take place, contrary to common wisdom that the occurrence of fewer trades indicates more illiquid stocks or markets.

To have a better sense of the actual number of trades per day, we look into the Daily TAQ consolidated data set between October 2003 and December 2015 and count how many regular trades for the U.S. common stocks are recorded between 9:30 a.m. to 4:00 p.m. EST. We refer to the landmark of 100 trades represented by the dotted line in Figure 3. We find out that not only 25% of

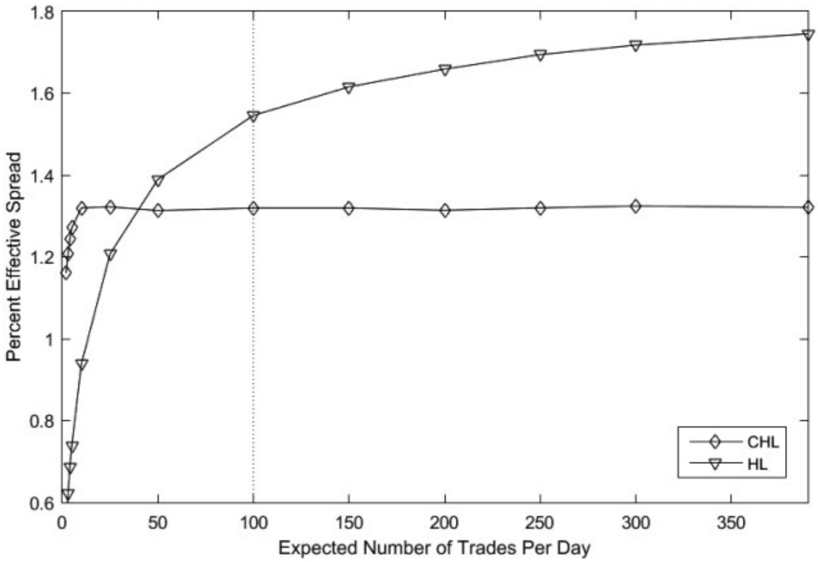

Figure 3
Sensitivity of bid-ask spread estimates to the number of daily trades
This figure shows the estimates from our model (CHL) and the one proposed by Corwin and Schultz (HL; 2012) for a simulated price process. For every expected number of trades between 2 and 390, specified in the horizontal axis, we simulate 10,000 months of 21-day price evolution, in which the unobservable efficient price has a daily volatility of 3%, and one trade in every 390 minuts. Each of 390 trades are equally likely above (below) the efficient price process by half-spread, and are observed with a certain chance, allowing average number of trades specified in the horizontal axis. The simulations are performed using a constant spread of 1%.

stock-days in our sample include less than 100 trades, but also theses stock-days belong to 77% of stocks in the sample. These numbers suggest that the *HL* estimates' sensitivity to the daily number of trades can be a broader issue that goes way beyond a limited number of illiquid stocks.

2.2 Random spreads

The settings in panel D of Table 2 are the same as those used in panel A, except that the spreads are no longer constant. By considering various spread sizes ($a\%$ spreads), the spreads for each day are randomly drawn from a uniform distribution with the range of $(0,2a)$. We find two interesting results: First, comparing panels A and D of Table 2, we see that the biases of *CHL* two-day corrected estimates change the least amongst the other estimates, which means that they are the least sensitive to the release of the assumption of constant spreads. At the same time *HL* estimates tend to considerably decrease by making the spreads random, allowing a –1% bias for the 8% average spreads. Second, in most of the cases of panel D, the *CHL* two-day corrected estimates show lower estimation errors compared to the *HL* ones.

2.3 All imperfections together

In panel E of Table 2 we report the simulation results in which we include different imperfections at the same time, namely observing average of two trades per day, random spreads as specified before, and an "overnight" price change corresponding to a half standard deviation of daily price returns to panel A. The overnight characterization represents more general nontrading periods, such as weekends, holidays, and overnight closings.[12] Two clear results emerge. First, *CHL* estimates are less biased and more accurate for medium to large spread values. Second, although the *CHL* monthly corrected estimates tend to show lower bias than the *CHL* two-day corrected estimates, the two-day corrected estimates are more accurate in terms of estimation errors for four out of the five spread levels. *HL* two-day corrected estimates are also more accurate than the *HL* monthly estimates, confirming Corwin and Schultz (2012) results. For this reason, we analyze the two-day corrected *CHL* estimates in the next sections.

3. A Comparison of Spread Estimates from Daily Data Using the TAQ Benchmark

We now turn to the analysis and comparison of the main estimation methods of transaction costs specified in the literature, using the TAQ effective spreads as the benchmark. We conduct the main analysis using Daily TAQ data, as recommended by Holden and Jacobsen (2014), between October 2003 and December 2015, and follow up with a robustness check using Monthly TAQ benchmark between January 1993 and September 2003 at the end of the section. Using CRSP daily data, we estimate the effective spreads for common stocks listed in the main three stock markets in the United States, namely, NYSE, AMEX, and NASDAQ. In addition to our estimator (*CHL*), we estimate the spreads originating from the following estimators: *Roll* (Roll 1984), *Gibbs* (Hasbrouck 2009), *EffTick* (Holden 2009; Goyenko, Holden, and Trzcinka 2009), *HL* (Corwin and Schultz 2012), and *FHT* (Fong, Holden, and Trzcinka 2017). We also include CRSP average end-of-day spreads using a more recent sample of 1993 onwards, in which end-of-day quote data are available. In the following analysis, we use the two-day corrected version for our estimator and for the *HL* measure, as recommended by Corwin and Schultz (2012).

To calculate our *CHL* measure, we do the following: (1) we keep the previous daily high, low, and close prices on those days when a stock does not trade, or has a zero price range; (2) we use the two-day corrected version; that is, we set negative two-day estimates of squared spreads to zero and then take the square

[12] We perform additional numerical simulations reported in the Internet Appendix. These include overnight price movements, and the relaxation of the assumption of equal likelihood of buyer-initiated and seller-initiated trades. The trade direction imbalance highly affects the Roll estimates but the effect on *CHL* and *HL* estimates is marginal.

roots and average over the month; and (3) we discard estimates for months in which there are less than 12 applicable days.[13,14]

To calculate the *HL* estimates, we exactly follow Corwin and Schultz (2012). More specifically, (1) we keep the previous daily high and low prices on those days when a stock does not trade, or has a zero price range, and, for the days with zero range, we adjust the high and low prices of previous day in the ad-hoc way explained in their paper. (2) we perform the ad-hoc overnight adjustment as explained in their paper; (3) we use the two-day corrected version; that is, we set negative two-day estimates to zero; and (4) we discard stock-months with less than 12 two-day estimates. We then calculate the other measures and merge all the estimations. We finally discard stock-months in which (1) any of the estimates produce a missing value, (2) a stock split or enormous distribution occurred, (3) a change of the primary exchange occurred, or (4) a stock has a time-series of less than six monthly estimates.[15,16]

We construct the main high-frequency benchmark for our analysis by calculating the effective spread from Daily TAQ data. Equation (12) defines the proportional effective spread at time t. As recommended by Holden and Jacobsen (2014), we use Daily TAQ data, with milliseconds time stamps, instead of the Monthly TAQ data. In fact, the authors show that in fast, competitive markets of today, the Daily TAQ granularity is more precise, whereas the usage of Monthly TAQ data might lead to incorrect statistical inferences.

$$ES_t = \frac{2|P_t - M_t|}{M_t}, \quad M_t = \frac{B_t + A_t}{2}. \tag{12}$$

The time span of the data set, covering 147 months of Daily TAQ data, starts in October 2003 and ends in December 2015. To calculate the effective spread from Daily TAQ data, we closely apply the procedure explained by Holden and Jacobsen (2014).[17] More precisely, we first clean up the National Best Bid and Offer (NBBO) data set by removing any best bid (ask) in which the bid-ask spread is above five dollars and the bid (ask) is more than 2.5 dollars above (below) the previous midpoint. We also remove any quotes from the

---

[13] An applicable day is defined as one with a closing price, high price, low price, price range, and volume above zero. Inclusion or exclusion of the volume criterion does not visually change any outcomes. It is also possible and accurate to replace missing η_{t+1} values, for the two-day estimates in which no trade occurs on day $t+1$, with readily available mid-quotes. However, to have a fair comparison with other estimates, we refrain from using midquotes in our estimates. In favor of the Corwin and Schultz (2012) estimates, we keep using the midquotes for their nontrading days and overnight price adjustments.

[14] As we merge the estimates in the next step, this filter will be applied to other estimates as well. Therefore, all the estimates will have similar quality in terms of the selected months-stocks.

[15] We discard stock-months in which the cumulative price adjustment factor (cfacpr) changes more than 20%.

[16] For example, the *Gibbs* estimator's code returns errors for the few stock-months in which the price remains constant for most of the days in the month, because the initial trade directions used in the simulations are calculated as sign of daily returns.

[17] To calculate the effective spreads using Daily TAQ data, we use the same SAS codes kindly provided by Craig Holden on his Web site. We add additional criteria to keep the trades/quotes records with no symbol suffixes.

Table 3
Summary statistics for different estimators

	N	Mean	Median	SD	$\rho(.,ES_{i,t})$	$\% \leqslant 0$
Effective spread	579,872	0.82%	0.27%	1.41%	1.000	0.00
CHL - Two-day	579,872	1.39%	1.02%	1.30%	0.745	0.00
CHL - Monthly	579,872	1.24%	0.74%	1.85%	0.680	33.96
HL - Two-day	579,872	1.21%	0.93%	1.03%	0.660	0.00
HL - Monthly	579,872	0.58%	0.31%	0.87%	0.625	24.25
Roll	579,872	1.50%	0.72%	2.56%	0.454	42.81
Gibbs	579,872	2.13%	1.47%	2.96%	0.397	0.00
EffTick	579,872	2.03%	0.64%	4.81%	0.419	27.44
EffTick – Alt. incr.	579,872	0.25%	0.07%	0.72%	0.514	0.00
FHT	579,872	0.26%	0.00%	0.69%	0.436	61.61
CRSP_S	579,872	0.82%	0.21%	1.61%	0.957	0.30

This table provides the main summary statistics for the pooled sample of the main estimators considered in this paper. The column labeled N refers to the number of stock-months of estimates in the sample. The column labeled $\rho(.,ES_{i,t})$ refers to the correlation of different estimates with the TAQ effective spread benchmark. The row labels refer to the TAQ effective spread benchmark (effective spread), our estimator (CHL), and the estimators proposed by Corwin and Schultz (HL; 2012), Roll (Roll; 1984), Hasbrouck (Gibbs; 2009), Holden (EffTick; 2009), and Fong, Holden, and Trzcinka (FHT; 2017). For the sake of completeness, we include the CRSP end-of-day bid-ask spreads (*CRSP_S*) as motivated by Chung and Zhang (2014). For calculating the CHL estimates, we replace the missing high, low, and close price with the previous days' values. We then discard monthly estimates for the months with fewer than 12 trading days (that is, days with positive high, low, and close price, as well as positive volume). The *HL* estimates are exactly calculated like in Corwin and Schultz (2012); that is, (1) missing daily high and low prices are replaced with those of previous days, (2) overnight adjustments are applied, and (3) monthly estimates with fewer than 12 two-day estimates are discarded. We merge the results of different estimators and discard stock-months in which any of the estimates are missing. We compute two versions of the *HL* (*CHL*) estimator, that is, the two-day corrected and monthly corrected versions labeled two-day and monthly. In the two-day corrected version for *HL* (*CHL*), we set each negative two-day spread (squared spread) to zero, and then the spreads (square roots of estimated squared spreads) are averaged within a month. The monthly corrected *HL* estimates are calculated by averaging all the two-day spreads within the month and then setting negative monthly averages to zero. The monthly *CHL* estimates are calculated like those in Equation (10). The *Roll* estimates are calculated by setting positive monthly autocovariance estimates to zero. The zeros reported for *EffTick* estimates reflect the months in which none of the prices are divisible by the base-eight denomination increments, and presumably reflect smaller spreads. We consider a second variant of *EffTick* measure (*EffTick – Alt. incr.*) by using the tick sizes of 1¢, 5¢, 10¢, 25¢, 50¢, or $1.00 as our sample time span lies after the decimalization of stock markets.

consolidated quotes (CQ) file if the spread is more than five dollars. Second, we merge the CQ and NBBO (cleaned) data to construct a complete official NBBO data set. Third, we match trades with constructed official NBBO quotes one millisecond before them.[18] In addition to the above-mentioned filters, we discard all trades outside the market opening hours and with proportional effective spreads above 40%. We compute the dollar-weighted average for intraday proportional effective spreads to obtain the average daily spreads. Then we take the average of daily spreads to construct the monthly benchmark.

The final step in the data preparation is to link the CRSP and Daily TAQ using CUSIPs in the TAQ master files.[19] This matching strategy allows us to cover 98% of stock-months estimates from the CRSP. We provide the summary statistics for the estimates in Table 3. As we compare the pooled data in Table 3,

18 Starting from July 27 2015, Daily TAQ timestamps are provided in microseconds, and, we match trades with the official NBBO quotes one microsecond before them.

19 We use the monthly master files, which cover a longer portion of our sample. For 2015, however, we rely on daily master files because monthly master files are not available after 2014.

both the mean and standard deviation convey valuable information about the explanatory power of the estimators. The mean provides a simple measure for the level or size of the estimated transaction costs, and the standard deviation gives information about the time-series and cross-sectional dispersion of spread estimates around the mean. We also include overall correlations of estimates with the TAQ effective spreads benchmark, confirming the better association of two-day corrected estimates over monthly corrected estimates, with the benchmark. Running a pooled regression of TAQ effective spreads on the *CHL* two-day corrected estimates, $ES_{i,t} = a + b\ CHL_{i,t} + \varepsilon_{i,t}$ we obtain the values of -0.29%, 0.8053, and 56% respectively for a, b and R^2, whereas the same regression on the *CHL* monthly corrected estimates delivers the values of 0.18%, 0.5169, and 46% respectively for a, b and R^2. Although the sample means shown in Table 3 suggest that *CHL* monthly corrected estimates are slightly less biased, the two-day corrected estimates are better associated with the TAQ benchmark by showing a higher R^2 and a slope coefficient closer to one. Therefore, following Corwin and Schultz (2012), we use the two-day corrected estimates for our analysis in the rest of the paper.

Calculating *EffTick* estimates, we observe some stock-months in which none of the prices are divisible by the base-eight denomination increments. Since this is likely because their spread is smaller than the base-eight denomination increments, we set the estimates for these stock-months to zero. To address this issue in a more comprehensive way, we also consider a second variant of the *EffTick* measure using the tick sizes of 1¢, 5¢, 10¢, 25¢, 50¢, or \$1.00 as our sample time span lies after the decimalization of stock markets. In Table 3, this second variant is labeled *EffTick – Alt. Incr*. Clearly, this variant underestimates both the mean and the variations of the effective spread. This is why we consider the original base-eight denomination variant in the next sections.

The summary statistics in Table 3 suggest that the end-of day quote spread is generally the best low-frequency measure in terms of correlation with the TAQ effective spread benchmark. The *CHL* is the best measure not relying on quotes data followed by the *HL* estimator.

To provide empirical support to the numerical analysis of the previous section, we perform two subsampling attempts with respect to the results of Table 3. First, to show that the bias sensitivity in bid-ask spread estimation to the average number of trades per day, we sort the stock-months of the estimates into 10 decile groups based on monthly averages of the number of trades per day. We then measure the average bias of *CHL* (*HL*) estimates as the average of the difference between *CHL* (*HL*) estimates and the TAQ effective spreads, for every decile. Figure 4 shows the average biases for the decile groups. In line with the simulation patterns in Figure 3, *CHL* (*HL*) estimates are less (more) sensitive to the number of trades per day. *HL* estimates show an average bias ranging from -146 to 78 bps, whereas the average bias of *CHL* estimate range from -80 to 84 bps. Compared with *HL*, the negative bias of *CHL* is substantially smaller for the less liquid deciles, confirming the numerical simulation results. In the

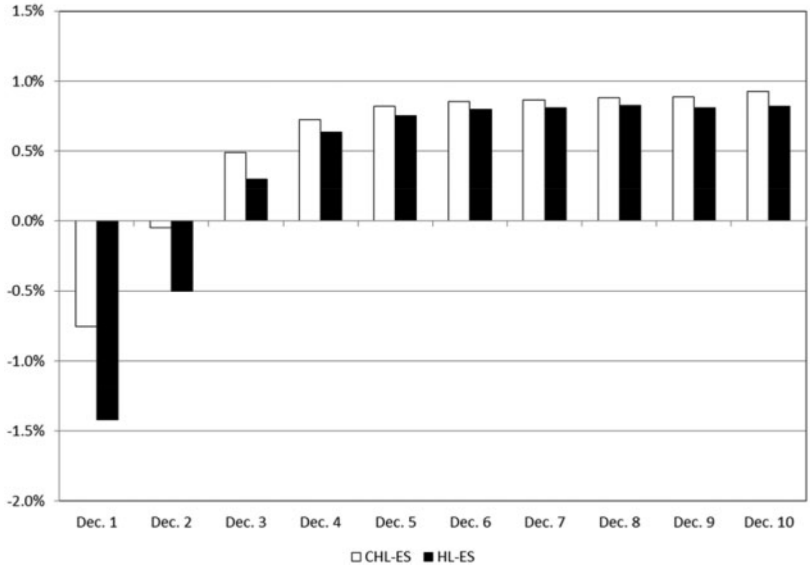

Figure 4
Estimation bias and average daily number of trades
The labels in the legend refer to the TAQ effective spreads (*ES*), estimators from our method (*CHL*), and Corwin and Schultz's estimates (HL; 2012). We group 579,872 stock-months into ten deciles sorted by the average number of daily trades in the month. For every decile, we measure the difference between average *CHL* (*HL*) estimates with the average *ES* estimates.

second subsampling attempt, we group stocks into five quintiles sorting them by their average number of daily trades during the sample period. Table 4 shows the correlation coefficients between different estimates and the TAQ effective spread benchmark for each quintile and for the entire sample. As expected, in the absence of quote data the *CHL* estimates have the highest correlation with the TAQ benchmark for the entire sample, as well as for the first three quintiles representing stocks less frequently traded.[20]

As a decomposition of the standard deviations reported in Table 3, we also compute the cross-sectional standard deviation of the estimates on a monthly basis to assess how well the estimators' dispersion follows that of the TAQ benchmark across time. Figure 5 shows the results for some estimators. It is clearly evident that the cross-sectional dispersions from our estimator most closely track that of the benchmark.

We now turn to identifying which criteria should be used to assess the measurement performance of the effective spread estimators. As stressed by

20 We also observe that *CHL* has the highest correlation with the Amihud price impact measure, which reflects another dimension of market liquidity. This holds for the entire sample, and for each of the five quintiles, sorted by average number of daily trades. Moreover, the correlations are higher for less-frequently traded quintiles. The Internet Appendix provides the results.

Table 4
Correlations for quintiles based on the average number of trades

	N	CHL	HL	Roll	Gibbs	EffTick	FHT	CRSP_S
Full sample	579,872	**0.745**	0.660	0.454	0.397	0.419	0.436	0.957
ANTD quintile 1	77,978	**0.820**	0.762	0.572	0.685	0.371	0.367	0.936
ANTD quintile 2	103,920	**0.785**	0.721	0.459	0.459	0.406	0.450	0.945
ANTD quintile 3	110,083	**0.701**	0.677	0.357	0.334	0.444	0.444	0.943
ANTD quintile 4	130,725	0.616	**0.627**	0.282	0.251	0.485	0.430	0.931
ANTD quintile 5	157,166	0.529	**0.557**	0.246	0.240	0.519	0.463	0.912

The table shows the correlation coefficients between different monthly estimates and the TAQ effective spread benchmark. We group the stocks into five quintiles sorting them by their average number of trades per day during the sample period. The daily number of trades is counted using TAQ consolidated trades data for trades that occur between 9:30 and 16:00 and have a positive price and volume. The first four quintiles are constructed of 1,392 stocks, and the fifth is constructed of 1,393 stocks. The labels in the first row refer to our estimator (CHL) and the estimators proposed by Corwin and Schultz (HL; 2012), Roll (Roll; 1984), Hasbrouck (Gibbs; 2009), Holden (EffTick; 2009), Fong, Holden, and Trzcinka (FHT; 2017), and Chung and Zhang (CRSP_S; 2014). N refers to the number of stock-months of estimates for the entire sample, as well as for each quintile. To compare estimators in the absence of quote data, we exclude the CRSP_S, and an asterisk would indicate numbers not significantly different from the estimator with the highest correlation marked in bold, using Fisher's z-test to compare the correlation coefficients.

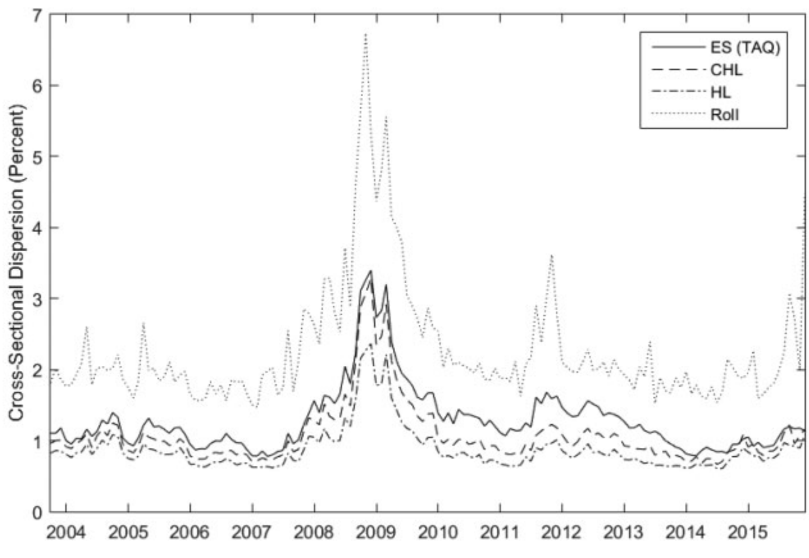

Figure 5
Cross-sectional dispersion of monthly spread estimates
This figure shows the standard deviations of spread estimates across stocks for each month from October 2003 to December 2015. In addition to the effective spread based on the Daily TAQ data, the labels refer to our estimator (CHL) and the estimators proposed by Corwin and Schultz (HL; 2012) and Roll (Roll; 1984).

Goyenko, Holden, and Trzcinka (2009), the choice of the best estimator, depending on the specific application, should be based on different criteria. For the sake of completeness, our analysis encompasses the three main criteria used in the literature: cross-sectional correlation, time-series correlation, and

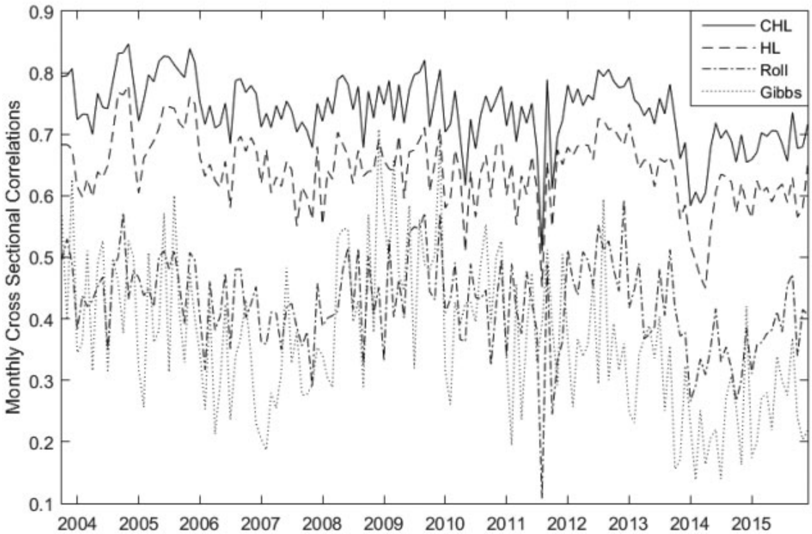

Figure 6
Cross-sectional correlation of monthly spread estimates
This figure shows the cross-sectional correlation between model-implied percentage spread estimates and effective spreads from the Daily TAQ data for each month from October 2003 to December 2015. The labels refer to our estimator (CHL) and the estimators proposed by Corwin and Schultz (HL; 2012), Roll (Roll; 1984), and Hasbrouck (Gibbs; 2009).

prediction errors. In addition, we test whether the average partial correlations of our estimates with the effective spread benchmark, controlling for other estimates, are positive. Doing so, we can test whether our estimates provide additional explanatory power that cannot be explained by combination of other estimates. In the following part, we analyze the accuracy of the estimators applying the explained set of criteria that should support a complete assessment and cover a wide range of applications.

3.1 Cross-sectional correlations

For each month, we calculate the correlation of the estimates with TAQ effective spreads that serve as the benchmark. Figure 6 shows the development across time for the cross-sectional correlations of the spread benchmark with the *CHL*, *HL*, *Roll*, and *Gibbs* estimators. It is clearly discernable that our *CHL* estimator provides the highest cross-sectional correlation for each month. The results in panel A of Table 5 confirm that end-of-day quote data provide the most accurate spread estimates and, in the absence of quote data, the *CHL* estimates have the highest time-series average cross-sectional correlation over the entire sample and across all subperiods. We apply the approach proposed by Goyenko, Holden, and Trzcinka (2009) to perform the statistical inferences to assess whether the average correlations are significantly different. More specifically, to compare the average correlation of the two estimators, we compute the

Table 5
Average cross-sectional correlations with the TAQ benchmark

	N	CHL	HL	Roll	Gibbs	EffTick	FHT	CRSP_S
A. Average cross-sectional correlations with effective spreads for monthly estimates								
Full period	3,944.7	**0.738**	0.642	0.424	0.369	0.409	0.426	0.959
2003–2007	4,380.5	**0.762**	0.664	0.435	0.378	0.458	0.522	0.963
2008–2011	3,870.8	**0.736**	0.635	0.428	0.442	0.391	0.401	0.959
2012–2015	3,555.5	**0.715**	0.625	0.409	0.288	0.374	0.349	0.956
B. Average cross-sectional correlations with changes in effective spreads for monthly estimates								
Full period	3,895.7	**0.298**	0.284	0.114	0.093	0.026	0.037	0.666
2003–2007	4,266.2	**0.328**	0.306	0.128	0.093	0.029	0.066	0.659
2008–2011	3,765.9	**0.304**	0.292	0.121	0.122	0.028	0.036	0.643
2012–2015	3,461.3	**0.245**	0.239*	0.086	0.057	0.018	0.002	0.684
C. Analysis across different markets								
NYSE	1,337.1	0.495*	0.481*	0.213	0.239	**0.500**	0.405	0.919
AMEX	297.1	**0.735**	0.644	0.452	0.522	0.318	0.435	0.947
NASDAQ	2,310.5	**0.710**	0.588	0.404	0.334	0.368	0.348	0.955
D. Analysis across market capitalization								
Size quintile 1	595.4	**0.690**	0.564	0.380	0.377	0.239	0.304	0.942
Size quintile 2	678.3	**0.568**	0.388	0.339	0.277	0.216	0.200	0.935
Size quintile 3	755.8	**0.457**	0.331	0.212	0.155	0.251	0.237	0.914
Size quintile 4	860.4	**0.423**	0.398	0.165	0.145	0.349	0.282	0.887
Size quintile 5	1,055.0	0.439	**0.494**	0.153	0.178	0.456*	0.340	0.856
E. Analysis across effective spread size								
ES quintile 1	1,035.0	0.385	**0.458**	0.129	0.175	0.351	0.159	0.651
ES quintile 2	841.4	0.375	**0.419**	0.141	0.159	0.339	0.220	0.752
ES quintile 3	753.2	**0.400**	0.388*	0.148	0.147	0.331	0.280	0.843
ES quintile 4	715.9	**0.497**	0.390	0.246	0.204	0.248	0.251	0.897
ES quintile 5	599.1	**0.693**	0.579	0.415	0.465	0.225	0.275	0.918

This table shows the average cross-sectional correlations between estimates of transaction costs and the TAQ benchmark for each month. The monthly correlations are averaged over the specified sample periods. The labels in the first row refer to our estimator (CHL) and the estimators proposed by Corwin and Schultz (HL; 2012), Roll (Roll; 1984), Hasbrouck (Gibbs; 2009), Holden (EffTick; 2009), Fong, Holden, and Trzcinka (FHT; 2017), and Chung and Zhang (CRPS_S; 2014). N is the average number of stocks per month. To compare estimators in the absence of quote data, we exclude the CRSP_S and an asterisk indicates numbers not significantly different from the estimator with the highest correlation marked in bold in every row. We test our hypotheses on the time series of pairwise difference in correlations for two estimators and assess whether the mean is significantly different from zero. We adjust for any potential time-series autocorrelation by using Newey-West (1987) standard errors with four lags autocorrelation. The size quintiles are sorted by increasing market capitalization at the last observed period for each individual stock. The spread quintiles are sorted by increasing average effective spreads during the whole sample period.

pairwise difference of their cross-sectional correlations with the benchmark at each month. We then test if the average value for this time series is significantly different from zero, while adjusting for the autocorrelations using Newey-West (1987) standard errors with four lags. To compare estimators not relying on quote data, we exclude the CRSP spreads and an asterisk indicates numbers not significantly different from the estimator with the highest correlation marked in bold in every row. The findings in Table 5 indicate that the time-series average cross-sectional correlation coefficients of our estimator are statistically higher than other measures that are not relying on quote data.

We substantiate the previous analysis by examining the cross-sectional correlations in first differences, that is, taking the changes in monthly (estimated) spreads. Panel B of Table 5 shows the time-series average of cross-sectional correlations for the changes. As expected, average correlations based on changes in spreads are lower than those based on spread levels. However, as for the correlation in levels, the average correlation in first differences of our estimator with the benchmark is the highest and statistically different from the other estimates.

Next, we perform a subsampling analysis of the cross-sectional correlation for levels of effective spreads across three dimensions: market venues, market capitalization, and effective spread size. First, we identify the three primary exchanges in which the stocks are listed using the CRSP exchange codes, that is, NYSE, AMEX, and NASDAQ. Second, we examine whether our results from the cross-sectional analysis depend on firm size. To do this, we decompose the entire sample into five quintiles by the firm's market capitalization value for each individual stock at the last observed period. Third, we consider whether our findings are sensitive to the magnitude of transaction costs. As before, we form five quintiles according to the average effective spread size over the entire sample period. The results of these three subsampling analyses are reported in panels C, D, and E of Table 5, respectively.

Three main findings arise: First, our estimator provides the best results for stocks listed on the AMEX and NASDAQ. Though, when only NYSE stocks are considered, it shares the highest cross-sectional correlation with the *HL* and *EffTick* measures. This is in line with the previous simulation results highlighting the relative accuracy of *CHL* estimates for less-liquid stocks. Second, our estimator significantly outperforms the other measures across all market capitalizations, except for the fifth quintile (Quintile 5), which includes the largest capitalization. Third, our estimator performs significantly better than the other estimators for stocks traded with medium and large transaction costs (from Quintiles 3 to 5 sorted by smallest to largest effective spreads). In sum, our estimator provides the overall highest cross-sectional correlations with the effective spread benchmark in the absence of quote data. Its estimates are particularly accurate for stocks with lower liquidity, proxied by small-medium market capitalizations and effective spreads of medium and large magnitude.

3.2 Time-series correlations

As the second criterion, we analyze stock-by-stock time-series correlations between the different spread estimates and the TAQ effective spread. We first calculate the time-series correlation between bid-ask spread estimates and the effective spread benchmark for each individual stock and each individual estimator. Then we compute the average of these time-series correlations across all sample stocks for each individual estimator. To compare the average correlations originating from different estimates, we use paired t-test.

Table 6 shows the main results. Similar to Table 5, Table 6, panel A (B), shows the average time-series correlations for the levels (changes) of effective spreads and stocks are sorted by exchanges, market capitalization, and spread

Table 6
Average time-series correlations for spread estimates of individual stocks compared to the TAQ benchmark

	N	CHL	HL	Roll	Gibbs	EffTick	FHT	CRSP_S
A. Average time-series correlations with effective spreads: Monthly estimates								
Full period	7,210	**0.518**	0.510	0.242	0.330	0.310	0.181	0.739
2003–2007	5,652	**0.393**	0.377	0.140	0.247	0.252	0.124	0.614
2008–2011	4,783	**0.611**	0.604	0.317	0.436	0.267	0.150	0.757
2012–2015	4,406	0.314	**0.325**	0.106	0.175	0.148	0.072	0.614
B. Average time-series correlations with changes in effective spreads: Monthly estimates								
Full period	7,124	0.287*	**0.290**	0.115	0.166	0.050	0.024	0.452
2003–2007	5,574	0.256*	**0.258**	0.096	0.167	0.040	0.012	0.386
2008–2011	4,727	0.340	**0.351**	0.146	0.211	0.064	0.034	0.470
2012–2015	4,331	0.190	**0.196**	0.066	0.102	0.016	0.003	0.388
C. Analysis across different markets								
NYSE	2,174	0.430	**0.454**	0.181	0.279	0.305	0.133	0.622
AMEX	831	**0.557**	0.505	0.261	0.408	0.291	0.264	0.822
NASDAQ	4,587	**0.540**	0.525	0.258	0.333	0.305	0.183	0.770
D. Analysis across market capitalization								
Size quintile 1	1,442	**0.695**	0.650	0.381	0.514	0.332	0.334	0.900
Size quintile 2	1,442	**0.571**	0.522	0.291	0.391	0.310	0.190	0.836
Size quintile 3	1,442	**0.478**	0.469	0.207	0.282	0.305	0.153	0.752
Size quintile 4	1,442	0.427	**0.453**	0.163	0.224	0.302	0.134	0.650
Size quintile 5	1,442	0.417	**0.456**	0.167	0.241	0.300	0.092	0.555
E. Analysis across effective spread size								
ES quintile 1	1,442	0.409	**0.444**	0.160	0.242	0.269	0.066	0.502
ES quintile 2	1,442	0.420	**0.468**	0.160	0.228	0.308	0.129	0.654
ES quintile 3	1,442	0.460	**0.471**	0.190	0.262	0.337	0.200	0.786
ES quintile 4	1,442	**0.588**	0.532	0.282	0.361	0.328	0.243	0.856
ES quintile 5	1,442	**0.711**	0.634	0.417	0.559	0.309	0.265	0.895

The labels in the first row refer to our estimator (CHL) and the estimators proposed by Corwin and Schultz (HL; 2012), Roll (Roll; 1984), Hasbrouck (Gibbs; 2009), Holden (EffTick; 2009), Fong, Holden, and Trzcinka (FHT; 2017), and Chung and Zhang (CRSP_S; 2014). N is the number of stocks in the subsamples with, at least, six months of estimates. The averages are computed across stocks. To compare estimators in the absence of quote data, we exclude the CRSP_S and an asterisk indicates numbers not significantly different from the estimator with the highest correlation marked in bold in every row. We use a paired t-test for the statistical inferences. The size quintiles are sorted by increasing market capitalization at the last observed period for each individual stock. The spread quintiles are sorted by increasing average effective spreads during the whole sample period.

size in panels C, D, and E, respectively. End-of-day quoted spreads show the highest average time-series correlations. In the absence of end-of-day quotes, our model provides the highest average time-series correlation for monthly spreads of the overall sample and for two out of three subperiods. The only exception is the 2012–2015 subperiod, in which the *HL* measure has a 0.01 higher correlation. For changes of spreads, the *HL* method generates the highest time-series correlation. Our estimator provides the second-highest time-series correlations, except for the 2008–2011 subperiod, which, in statistical terms, is not significantly lower than the *HL* one. The remaining parts of Table 6 suggest that (1) our estimators outperform the others for stocks listed on the AMEX and NASDAQ, whereas the *HL* has the highest time-series correlation for NYSE stocks; (2) our measure (the *HL* measure) performs best for small- and medium-sized (large-sized) firms; and (3) our measure (the *HL* measure) performs best

Table 7
Prediction errors

	N	CHL	HL	Roll	Gibbs	EffTick	FHT	CRSP_S
A. RMSEs, breakdown for different periods, and across different markets								
Full period	3,944.7	**0.0104**	0.0107	0.0221	0.0289	0.0441	0.0130	0.0043
2003–2007	4,380.5	**0.0084**	0.0086	0.0182	0.0250	0.0368	0.0101	0.0030
2008–2011	3,870.8	**0.0141**	0.0141*	0.0291	0.0317	0.0551	0.0175	0.0062
2012–2015	3,555.5	**0.0089**	0.0094	0.0192	0.0302	0.0408	0.0117	0.0037
NYSE	1,337.1	0.0089	0.0077	0.0162	0.0231	0.0170	**0.0030**	0.0012
AMEX	297.1	**0.0115**	0.0124	0.0286	0.0253	0.0994	0.0190	0.0062
NASDAQ	2,310.5	**0.0111**	0.0118	0.0238	0.0316	0.0436	0.0154	0.0050
B. RMSEs, excluding stock-months with zero estimates								
Full period	648.4	**0.0115**	0.0127	0.0261	0.0201	0.0771	0.0168	0.0059
2003–2007	819.2	**0.0089**	0.0096	0.0214	0.0173	0.0623	0.0124	0.0039
2008–2011	617.8	**0.0156**	0.0173	0.0347	0.0245	0.0965	0.0227	0.0089
2012–2015	497.5	**0.0101**	0.0114	0.0226	0.0186	0.0733	0.0155	0.0050
NYSE	125.3	0.0101	0.0085	0.0224	0.0185	0.0345	**0.0042**	0.0017
AMEX	80.8	**0.0111**	0.0123	0.0288	0.0204	0.1421	0.0177	0.0064
NASDAQ	442.4	**0.0119**	0.0137	0.0262	0.0202	0.0681	0.0185	0.0064

We measure the accuracy of different monthly estimates by computing their root-mean-squared errors (RMSEs), as well as mean absolute errors (MAEs) with respect to the TAQ benchmark. Prediction errors are calculated every month and then averaged through the months in the sample. N is the average number of stocks per month. The labels refer to our estimator (CHL) and the estimators proposed by Corwin and Schultz (HL; 2012), Roll (Roll; 1984), Hasbrouck (Gibbs; 2009), Holden (EffTick; 2009), Fong, Holden, and Trzcinka (FHT; 2017), and Chung and Zhang (CRSP_S; 2014). To compare estimators in the absence of quote data, we exclude the CRSP_S and an asterisk indicates numbers not significantly different from the estimator with the lowest average prediction error marked in bold in every row. We test our hypotheses on the time series of pairwise difference in prediction errors for two estimators and assess whether the mean is significantly different from zero. We adjust for any potential time-series autocorrelation by using Newey-West (1987) standard errors with four lags autocorrelation

when stocks are traded with large (small) effective spreads.[21] The time-series correlation analysis confirms the previous findings that our estimator generally provides the most accurate estimates of effective costs, especially for less liquid stocks.

3.3 Prediction errors

A straightforward way to assess the accuracy of the bid-ask spread estimation is to observe how far an estimate, that is, the model prediction of the effective spread, is from the TAQ effective spread benchmark. We measure this by RMSEs of monthly estimates with respect to the TAQ effective benchmark at the same period.[22] In line with Goyenko, Holden, and Trzcinka (2009), we calculate the prediction errors every month and then average them through the time.

We report the results in two separate settings in Table 7, and like in Tables 4, 5, and 6, we focus on the comparison of the estimators not relying on quote data.

[21] As an additional test, which we report in the Internet Appendix, we construct equally weighted portfolios of stocks and then compare the correlation of the estimated portfolios' spread to that of the high-frequency benchmark. The estimated spreads of market-wide portfolio show a time-series correlation of 0.965 with the ones of the TAQ benchmark.

[22] We repeat this analysis using mean-absolute errors (MAEs) and, confirming the results of this section, find out that for the entire sample *CHL* estimates have the lowest MAEs compared with other estimates. The Internet Appendix provides the results.

In panel A, we include the entire sample, including the zero estimates for all measures to compare the overall accuracy of estimates. In panel B, we exclude the stock-months in which *Roll*, *EffTick*, or *FHT* estimates are zero to compare the accuracy nonzero estimates. In both settings, end-of-day quoted spreads show lowest RMSEs. However, in absence of end-of-day quotes, our estimator (*CHL*) provides the lowest RMSEs compared with other estimators across the entire sample, as well as AMEX and NASDAQ listed stocks. The difference between average RMSEs of our estimates and the other estimates is also significant, using Newey-West (1987) standard errors with four lags to test whether the time-series of pairwise difference of RMSEs is statistically different from zero.

3.4 Partial correlations

Since our *CHL* estimates jointly use close, high, and low prices, which are also partially used in other estimators discussed in the paper, it is worth testing whether our estimates include any additional information in explaining the bid-ask spreads beyond the *combination* of other estimates.[23] We measure this additional explanatory power in terms of average partial correlations. More specifically, by setting the partitioned regression framework of Equation (13), we examine the ability of *CHL* to predict the effective spread benchmark, whereas the predictive power of the other estimates is already taken into account.

$$ES_{i,t}=\alpha+\boldsymbol{\beta}\ \boldsymbol{EST}_{i,t}+\gamma CHL_{i,t}+\varepsilon_{i,t}, \tag{13}$$

$ES_{i,t}$ represents the TAQ effective spread for stock i in month t, and $\boldsymbol{EST}_{i,t}$ is a vector of other estimates including *HL* and *Roll*. Using the Frisch-Waugh-Lovell theorem, we first regress $ES_{i,t}$ and $CHL_{i,t}$ on $\boldsymbol{EST}_{\mathbf{i,t}}$ and then calculate the correlation between the orthogonal complements yielded from of above regressions, that is, $\epsilon_{ES_{i,t}|\boldsymbol{EST}_{i,t}}$, and $\epsilon_{CHL_{i,t}|\boldsymbol{EST}_{i,t}}$. To calculate time-series (cross-sectional) partial correlations, we perform the above regressions in the time-series (cross-sectional) dimension for every stock (month) and average the calculated partial correlations across stocks (months).

Table 8 shows the average partial correlations calculated, while controlling for different set of estimates in the following order: we control for *HL* (third column titled "CHL|HL"), we add *Roll* (fourth column titled "CHL|HL, Roll"), and we move forward by adding other estimators to the set of controls (adding *Gibbs* in the fifth column, *EffTick* in the sixth column, and *FHT* in the last column). Panel A of Table 8 shows the average partial cross-sectional correlations testing whether they are different from zero by using

[23] We also consider comparing the correlation of *CHL* estimates and the effective spread benchmark, with the ones from combination of other estimates. To do so, we combine other estimates both by taking their simple average and using their first principal component. As reported in the Internet Appendix, our estimates show the highest time-series and cross-sectional correlation with the effective spread benchmark.

Table 8
Partial correlations

	N	CHL | HL	CHL | HL, Roll	CHL | HL, Roll, Gibbs	CHL | HL, Roll, Gibbs, EffTick	CHL | HL, Roll, Gibbs, EffTick, FHT
A. Average partial cross-sectional correlations with the TAQ benchmark						
All stocks, levels	3,944.7	**0.478**	**0.455**	**0.450**	**0.439**	**0.430**
All stocks, changes	3,895.7	**0.159**	**0.155**	**0.151**	**0.150**	**0.149**
NYSE	1,337.1	**0.188**	**0.202**	**0.190**	**0.166**	**0.159**
AMEX	297.1	**0.473**	**0.436**	**0.412**	**0.408**	**0.405**
NASDAQ	2,310.5	**0.496**	**0.469**	**0.465**	**0.456**	**0.450**
ES quintile 1	1,035.0	**0.042**	**0.067**	**0.056**	**0.048**	**0.045**
ES quintile 2	841.4	**0.078**	**0.095**	**0.089**	**0.074**	**0.070**
ES quintile 3	753.2	**0.159**	**0.175**	**0.176**	**0.164**	**0.157**
ES quintile 4	715.9	**0.338**	**0.325**	**0.330**	**0.326**	**0.320**
ES quintile 5	599.1	**0.472**	**0.429**	**0.402**	**0.402**	**0.401**
B. Average partial time-series correlations for spread estimates of individual stocks compared to the TAQ benchmark						
All stocks, levels	5,964	**0.219**	**0.229**	**0.199**	**0.186**	**0.183**
All stocks, changes	5,896	**0.132**	**0.137**	**0.120**	**0.119**	**0.119**
NYSE	1,895	**0.117**	**0.147**	**0.120**	**0.111**	**0.111**
AMEX	591	**0.325**	**0.308**	**0.260**	**0.250**	**0.245**
NASDAQ	3,711	**0.251**	**0.253**	**0.226**	**0.212**	**0.207**
ES quintile 1	1,325	**0.088**	**0.122**	**0.100**	**0.095**	**0.094**
ES quintile 2	1,235	**0.091**	**0.126**	**0.109**	**0.098**	**0.097**
ES quintile 3	1,156	**0.166**	**0.186**	**0.171**	**0.154**	**0.153**
ES quintile 4	1,159	**0.338**	**0.332**	**0.302**	**0.280**	**0.271**
ES quintile 5	1,089	**0.454**	**0.408**	**0.343**	**0.332**	**0.327**

We calculate the partial correlation between the daily TAQ effective spread and our estimates (CHL) removing the effects explained by other estimates, that is, $\rho(\epsilon_{ES_{i,t}|\mathbf{EST}_{i,t}}, \epsilon_{CHL_{i,t}\mathbf{EST}_{i,t}})$. $\mathbf{EST}_{i,t}$ includes a constant, Corwin and Schultz (HL; 2012), Roll (Roll; 1984), Hasbrouck (Gibbs; 2009), Holden (EffTick; 2009), and Fong, Holden, and Trzcinka (FHT; 2017) estimates. The spread quintiles are sorted by increasing average effective spreads during the whole sample period. In panel A, N refers to the average number of stocks per month, and, in panel B, N refers to the number of stocks in the subsamples with at least 24 months of estimates. Panel A shows the average partial cross-sectional correlations. The bold numbers are significantly different from zero using a 5% two-tailed confidence interval. The statistical test for the average of cross-sectional correlations is based on Newey-West (1987) standard errors with four lags autocorrelation. Panel B shows the average partial time-series correlations, as the average of partial time-series correlations for individual stocks. The bold numbers are significantly different from zero using a t-test for the average of time-series correlations. To avoid overfitting in calculating the partial time-series correlations, we discard the stocks with fewer than 24 months of estimates.

Newey-West (1987) standard errors with four lags in the time-series of monthly-estimated cross-sectional correlations. All average cross-sectional correlations are significantly different from zero and positive, indicating that *CHL* has some additional explanatory power, not already included in any overidentified models, in predicting the effective spread. For instance, the average partial cross-sectional correlation of *CHL* and TAQ effective spreads after controlling for *HL*, *Roll*, *Gibbs*, *EffTick*, and *FHT* is 0.430 for the entire sample and 0.159, 0.405, and 0.450 for NYSE, AMEX, and NASDAQ stocks, respectively. Another interesting result is that the additional explanatory ability of *CHL* is larger for less liquid stocks as indicated by the increasing partial correlations from quintiles 1 to 5 in rows 8 to 12. All these findings remain consistent when average partial time-series correlations are considered (panel B of Table 8).

To show that the additional explanatory ability of *CHL* is related to illiquidity rather than to volatility, we double sort the stocks by these two properties. First, we construct illiquidity terciles by sorting the stocks by average effective spreads across the entire sample. Then we construct volatility terciles within every illiquidity tercile by sorting stocks according to their daily price volatility across the entire sample. We then calculate average partial cross-sectional and time-series correlations with the TAQ effective spread benchmark for the nine groups controlling for the explanatory power of *HL* and *Roll*. Panel A (B) of Figure 7 shows the average partial cross-sectional (time-series) correlations for the nine groups. It delivers two main messages: First, correlations are considerably higher for the illiquid terciles corroborating the previous findings. Second, there is no discernable pattern in terms of volatility within the three illiquidity terciles, suggesting that illiquidity rather than volatility explains the additional explanatory power of *CHL*.

All in all, in the absence of end-of-day quotes our estimates generally show the highest average time-series and cross-sectional correlations, as well as the lowest RMSEs, with respect to the Daily TAQ benchmark. Moreover, the estimates include additional information in explaining the TAQ benchmark that cannot be explained by the other bid-ask spread estimates. As showed in Table 9, the results are confirmed when we repeat the analysis for the period of January 1993 to September 2003 using the Monthly TAQ effective spreads.[24] Over this sample period, our estimates have even higher (lower) average cross-sectional correlations (estimation errors) than end-of-day quotes. The subsampling across time shows that this mainly occurs in the two subsamples before 2001 suggesting that end-of-day quote data are less accurate in the predecimalization era of U.S. stock market.[25]

4. Other Applications

Well-performing estimators of transaction costs can be applied in a variety of research areas. To illustrate their potential uses, we propose two simple applications. The first example is a description of the historical spread estimates for stocks listed on NYSE (AMEX) from 1926 (1962) to 2015. In the second example, the spread estimates are applied to measure systematic risks originating from liquidity issues.

[24] See the Internet Appendix for more details on the construction of Monthly TAQ benchmark and additional analysis.

[25] Intuitively, when tick sizes are larger, measuring end-of-day spreads produces larger estimation variance, and, consequently, larger estimation errors. For example, when the tick size is large enough that the spread size is only two (one) ticks wide, observing either the end of day bids or asks one tick further than the intraday value causes a 50% (100%) measurement error.

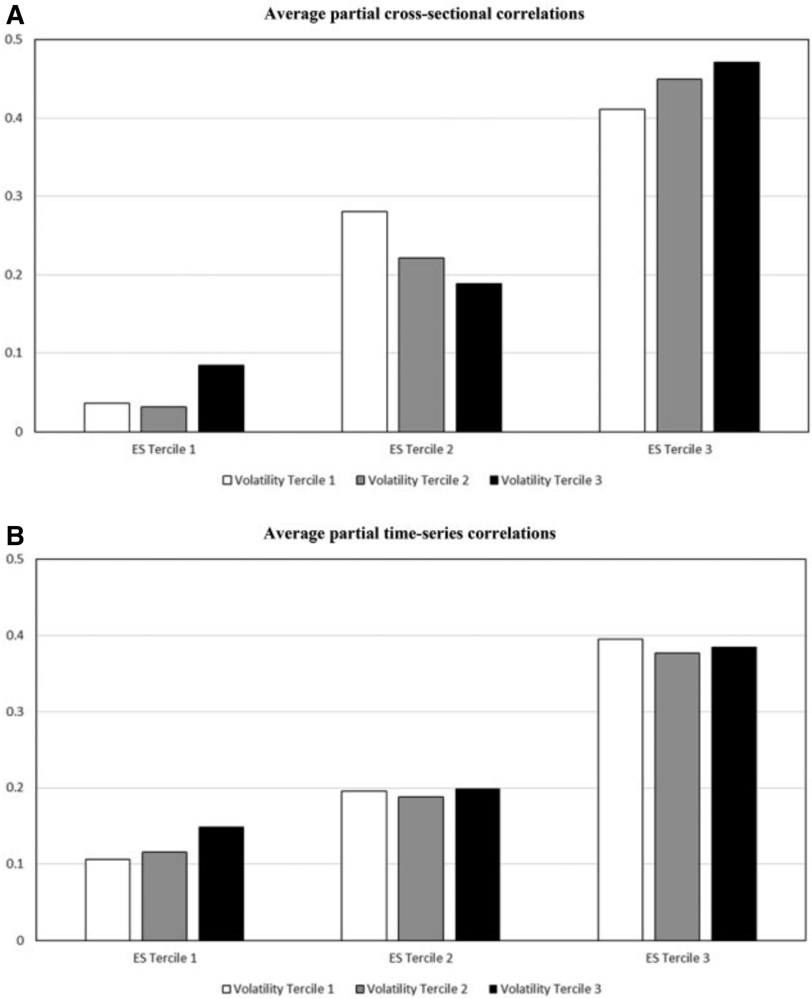

Figure 7
Average partial correlations after controlling for HL and Roll
We split the stocks sample into three illiquidity terciles by sorting them with their average effective spread during the sample period. Then we break down each illiquidity tercile into three volatility terciles using the daily volatility of the stocks during the sample period. The partial correlations are the correlations between the residuals of regressing TAQ effective spreads and our estimates (CHL) on Corwin and Schultz's (HL; 2012) and Roll's (Roll; 1984) estimates.

4.1 Estimating historical spreads for U.S. stocks

By using the close, high, and low price data from CRSP and the methodology explained above, we calculate the estimates of the bid-ask spreads based on our model. Specifically, we use the price values from previous days for the days with missing price values and construct the two-day corrected version of our estimates. We finally discard stock-months with fewer than 12 trading days.

Table 9
Comparison with the monthly TAQ benchmark, January 1993–September 2003

	N	CHL	HL	Roll	Gibbs	EffTick	FHT	CRSP_S
A. Average cross-sectional correlations with the TAQ benchmark								
All stocks, levels	5,009.2	**0.861**	0.833	0.605	0.713	0.637	0.644	0.846
All stocks, changes	4,925.6	**0.471**	0.460	0.206	0.266	0.194	0.153	0.578
1993–1995	3,922.4	**0.812**	0.808*	0.609	0.747	0.562	0.607	0.787
1996–2000	5,830.4	**0.890**	0.869	0.620	0.737	0.728	0.684	0.836
2001–2003	4,701.5	**0.860**	0.795	0.572	0.635	0.552	0.614	0.927
NYSE	1,578.4	0.810*	0.808*	0.453	0.629	**0.812**	0.755	0.856
AMEX	353.1	**0.929**	0.918	0.651	0.846	0.788	0.743	0.850
NASDAQ	4,925.6	**0.471**	0.460	0.206	0.266	0.194	0.153	0.578
B. Average time-series correlations for spread estimates of individual stocks								
All stocks, levels	10,783	**0.586**	0.580	0.280	0.445	0.464	0.402	0.778
All stocks, changes	10,676	**0.401**	0.400*	0.155	0.285	0.221	0.116	0.586
1993–1995	6,137	0.496	0.504	0.229	0.379	**0.527**	0.243	0.634
1996–2000	9,130	0.558	0.575	0.252	0.415	**0.590**	0.350	0.777
2001–2003	5,805	**0.568**	0.539	0.211	0.446	0.273	0.249	0.761
NYSE	2,754	0.343	0.353	0.095	0.286	**0.460**	0.408	0.584
AMEX	1,078	**0.644**	0.622	0.285	0.505	0.600	0.350	0.649
NASDAQ	7,744	**0.645**	0.632	0.326	0.482	0.450	0.383	0.863
C. Root-mean-square errors w.r.t TAQ benchmark								
All stocks	5,009.2	**0.0142**	0.0151	0.0309	0.0252	0.0466	0.0225	0.0224
1993–1995	3,922.4	**0.0205**	0.0217	0.0313	0.0233	0.0420	0.0288	0.0266
1996–2000	5,830.4	**0.0113**	0.0121	0.0296	0.0237	0.0398	0.0198	0.0250
2001–2003	4,701.5	**0.0126**	0.0135	0.0328	0.0299	0.0639	0.0205	0.0133
NYSE	1,578.4	0.0065	**0.0059**	0.0168	0.0161	0.0182	0.0074	0.0213
AMEX	353.1	**0.0118**	0.0156	0.0319	0.0186	0.0522	0.0238	0.0466
NASDAQ	3,077.7	**0.0223**	0.0234	0.0408	0.0341	0.0591	0.0326	0.0208

This table compares different estimates with the monthly TAQ (MTAQ) benchmark between January 1993 and September 2003. The labels in the first row refer to our estimator (CHL) and the estimators proposed by Corwin and Schultz (HL; 2012), Roll (Roll; 1984), Hasbrouck (Gibbs; 2009), Holden (EffTick; 2009), Fong, Holden, and Trzcinka (FHT; 2017), and Chung and Zhang (CRSP_S; 2014). The spread quintiles are sorted by increasing average effective spreads during the whole sample period. To compare estimators in the absence of quote data, we exclude the CRSP_S and an asterisk indicates numbers not significantly different from the estimator with the highest correlation (lowest RMSE) marked in bold in every row. In panel A, N refers to the average number of stocks per month. Cross-sectional correlations are calculated per month and averaged across the sample. We test our hypotheses on the time series of pairwise difference in correlations for two estimators and assess whether the mean is significantly different from zero. In panel B, N refers to the number of stocks in the subsamples with at least six months of estimates. Time-series correlations are calculated for each individual stock and then averaged across assets. We use a paired t-test for the statistical inferences. In panel C, N refers to the average number of stocks per month. RMSEs are calculated for every month and then averaged through time. We test our hypotheses on the time series of pairwise difference in prediction errors for two estimators and assess whether the mean is significantly different from zero. We adjust for any potential time-series autocorrelation by using Newey-West (1987) standard errors with four lags autocorrelation. An asterisk indicates numbers not significantly different from the highest correlation marked in bold in every row of panels A and B, and from the estimator with the lowest average prediction error marked in bold in every row in panel C.

Figure 8 shows the time development of the estimated spreads computed for three equally weighted portfolios: the smallest and largest market capitalization deciles, as well as the entire stocks sample. The spreads originated from our model display relatively stable variation over time. Reassuringly, this also applies to the smallest market capitalization decile. In contrast, Corwin and Schultz (2012) document that the spread estimates generated by their model display considerable variation over time, and these are extraordinarily high during the Great Depression, in which the market-wide average estimates of

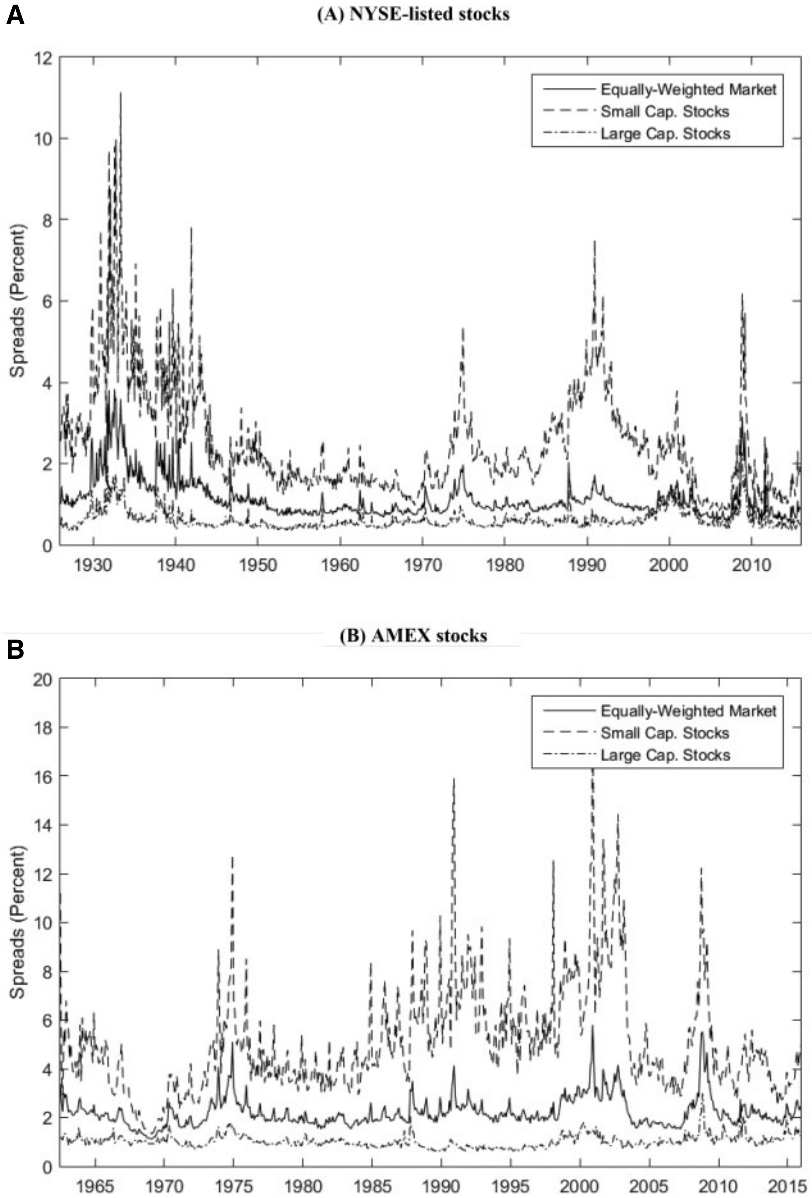

Figure 8

Time-series evolution of estimated spread, calculated as equally weighted portfolios of stocks

This figure shows the monthly historical developments of spread estimates from our model. Small cap and large cap portfolios are represented by the first and last decile of stocks sorted by market capitalization at the end of each month. Panel A (B) shows the estimates for stocks listed on the NYSE (AMEX) between 1926 (1962) and 2015.

the effective spreads are as high as 20% for NYSE stocks and 50% for small cap stocks. Instead, panel A (B) of Figure 8 shows that our estimates for the NYSE (AMEX) stocks evolve pretty steadily across every decade, remaining within an economically reasonable range; that is, the market-wide estimated effective spread does not exceed 4% (6%) for NYSE (AMEX) stocks. Moreover, the average estimated effective spread for the small cap stocks listed on the NYSE (AMEX) does not exceed 12% (19%) during the entire sample.

The results in this subsection suggest that our estimator can be used in various research areas across many types of markets and assets, including less actively-traded ones. This is especially true for researchers interested in the ability of an estimator to capture the temporal evolution of spreads over long time spans that predate quote data or international markets without quote data.

4.2 Estimating systematic liquidity risk

The results presented in Section 3 show that the spread estimates from our model closely follow the effective spread benchmark, suggesting that our estimator can be adopted for gauging transaction costs and liquidity. Another crucial application of spread estimates is liquidity *risk*. As liquidity risk is not diversifiable, its accurate measurement is crucial for at least two purposes: first, to identify and gauge systematic risk stemming from illiquidity issues, and, second, to perform effective asset and risk management. The collapse of Long-Term Capital Management L.P. (LTCM) and more recent experiences during the last financial crises are vivid examples of incorrect consideration of liquidity risk.

Acharya and Pedersen (2005) propose a liquidity-adjusted capital asset pricing model (LCAPM) in which expected returns in time t for stock i (r_t^i) net of transaction costs (s_t^i) are explained by the risk-free interest rate, and expected market returns (r_t^M) net of market transaction costs (s_t^M). Then the systemic risk of return net of trading costs is decomposed into four components:

$$\beta_1 = \frac{cov\left(r_{t+1}^i, r_{t+1}^M\right)}{var\left(r_{t+1}^M - s_{t+1}^M\right)}, \quad \beta_2 = \frac{cov\left(s_{t+1}^i, s_{t+1}^M\right)}{var\left(r_{t+1}^M - s_{t+1}^M\right)},$$

$$\beta_3 = \frac{cov\left(r_{t+1}^i, s_{t+1}^M\right)}{var\left(r_{t+1}^M - s_{t+1}^M\right)}, \quad \beta_4 = \frac{cov\left(s_{t+1}^i, r_{t+1}^M\right)}{var\left(r_{t+1}^M - s_{t+1}^M\right)} \tag{14}$$

Whereas β_1 represents the standard market beta, β_2, β_3, and β_4 capture important aspects of systematic risk due to liquidity issues. β_2 measures the commonality of liquidity with the market-wide liquidity and is expected to be positive (Chordia, Roll, and Subrahmanyam 2000). Higher β_2 translates into less liquid stocks in times of market illiquidity. Huberman and Halka (2001), and Hasbrouck and Seppi (2001) document the presence of a systematic, time-varying component of liquidity that comoves with the liquidity of individual stocks. Kamara, Lou, and Sadka (2008) show important implications of the cross-sectional variation of commonality in liquidity, including the decline over time of diversification benefits against aggregate liquidity shocks by holding large-cap stocks. Karolyi, Lee, and van Dijk (2012) study the commonality in liquidity across 40 countries and over two decades, and suggest that commonality in liquidity is better explained by the demand-side determinants. β_3 is typically negative as market liquidity tends to dry up when stock prices decline. Pastor and Stambaugh (2003) show that investors demand a premium for the sensitivity of stock returns to aggregate liquidity shocks. Watanabe and Watanabe (2008) document that aggregate liquidity is priced and the liquidity risk premium is twice as high as the value premium in high-beta states. β_4 is also expected to be negative as the liquidity of individual stocks tend to decrease in downturn markets. Hameed, Kang, and Viswanathan (2010) provide empirical evidence of significant increases of bid-ask spreads when the stock market experiences large negative returns.

The above-mentioned literature points to the importance of an accurate measurement of different dimensions of liquidity risk and its variation in the cross section of stocks. By using the effective spread estimates of Section 3, we calculate the four systematic risk components of Equation (15) for each stock in our sample based on the Daily TAQ effective spreads, as well as the *Roll* estimates, the *HL* estimates, and our estimates. In addition to the filtrations explained in Section 3, we discard stocks with fewer than 30 months of data and the stock-months in which the monthly CRSP return is missing. Following Asparouhova, Bessembinder, and Kalcheva (2010, 2013), we use a gross return-weighted portfolio of all the stocks to construct the market return and market liquidity to avoid biases calculating portfolio returns.

To assess the quality of the estimates for systematic risk, we compare them to those based on the TAQ effective spreads. In other words, we gauge how well the liquidity risk estimates generated by the *Roll*, the *HL*, and our model are associated to those obtained from Daily TAQ data, for the cross section of US stock market. Table 10 shows the cross-sectional correlations between the liquidity risks estimates generated from different estimators, and the ones estimated using the Daily TAQ benchmark.

Because the correlations for β_1 are close to one as a result of the secondary importance of transaction costs to compute standard betas, they are not tabulated Thus, we concentrate our analysis on β_2 β_3, and β_4 because they are more influenced by transaction costs. Overall, the results for all stocks, shown in panel A of Table 10, clearly indicate the superiority of our estimator to capture the cross-sectional estimation of systematic liquidity risks. The correlation between *CHL* estimates and the benchmark spreads for β_2 β_3, and β_4 are pretty high, that is, 0.830, 0.935, and 0.755, respectively. The same correlation coefficients for the *HL* and the *Roll* estimators are lower, especially for β_2 and β_4, which are around 0.1 lower for the *HL* and around 0.2 lower for the *Roll*. In many of the cases, these differences are statistically significant using a two-tailed Fisher's z-test with a 5% significance level. The same picture generally holds

Table 10
Cross-sectional correlations of estimated systematic liquidity risks with the ones of TAQ benchmark

		$\rho(\beta_2^{ES},\beta_2^{Estimates})$			$\rho(\beta_3^{ES},\beta_3^{Estimates})$			$\rho(\beta_4^{ES},\beta_4^{Estimates})$		
	N	CHL	HL	Roll	CHL	HL	Roll	CHL	HL	Roll
A. Cross-section of estimated systematic risks: All stocks										
Full period	5,547	**0.830**	0.744	0.652	**0.935**	0.919	0.838	**0.755**	0.673	0.459
2003–2007	4,119	**0.501**	0.457	0.170	**0.670**	0.658*	0.407	**0.532**	0.439	0.214
2008–2011	3,574	**0.736**	0.604	0.354	**0.971**	0.970*	0.896	**0.557**	0.428	0.296
2012–2015	3,268	**0.325**	−0.032	0.078	**0.545**	0.327	0.298	**0.571**	0.510	0.243
B. Cross-section of estimated systematic risks considering liquidity shocks of AR(2) model										
Full period	5,433	**0.524**	0.446	0.158	0.856	**0.879**	0.563	**0.530**	0.266	0.234
2003–2007	4,010	**0.285**	0.152	−0.078	0.808*	**0.812**	0.595	**0.364**	0.147	0.064
2008–2011	3,654	**0.398**	0.265	−0.031	0.893	**0.913**	0.627	**0.520**	0.107	0.142
2012–2015	3,231	**0.137**	0.126*	0.080	0.724	**0.752**	0.398	**0.295**	0.221	0.148
C. Analysis across different markets										
NYSE	1,777	**0.697**	0.681*	0.443	**0.935**	0.925	0.844	0.670*	**0.675**	0.244
AMEX	515	**0.883**	0.825	0.689	**0.925**	0.918*	0.824	**0.787**	0.706	0.524
NASDAQ	3,429	**0.855**	0.782	0.692	**0.934**	0.914	0.835	**0.775**	0.692	0.460
D. Analysis across market capitalization										
Size quintile 1	1,032	**0.870**	0.790	0.726	**0.945**	0.936*	0.842	**0.793**	0.712	0.525
Size quintile 2	1,009	**0.722**	0.577	0.558	**0.920**	0.885	0.822	**0.612**	0.533	0.328
Size quintile 3	1,062	**0.632**	0.510	0.363	**0.939**	0.928*	0.847	**0.516**	0.423	0.220
Size quintile 4	1,161	**0.570**	0.487	0.314	**0.921**	0.905	0.832	**0.551**	0.530*	0.208
Size quintile 5	1,283	**0.444**	0.431*	0.265	**0.922**	0.910	0.825	0.444*	**0.500**	0.098
E. Analysis across effective spread size										
ES quintile 1	1,278	0.566*	**0.567**	0.353	**0.898**	0.888*	0.768	0.568*	**0.618**	0.238
ES quintile 2	1,153	0.708*	**0.719**	0.369	**0.935**	0.919	0.846	0.603	**0.662**	0.147
ES quintile 3	1,059	**0.713**	0.694*	0.365	**0.939**	0.927	0.840	**0.659**	0.651*	0.244
ES quintile 4	1,074	**0.789**	0.724	0.507	**0.924**	0.899	0.827	**0.746**	0.712*	0.384
ES quintile 5	983	**0.884**	0.821	0.765	**0.943**	0.934*	0.849	**0.811**	0.744	0.526

We calculate the components of systematic risk implied by the LCAPM model (Acharya and Pedersen 2005) by using the daily TAQ effective spreads, *Roll* model estimates (Roll; 1984), the *HL* estimates (Corwin and Schultz; 2012), and the estimates from our model (labeled *CHL*). N refers to the number of stocks. The table reports the cross-sectional correlation of betas based on *Roll*, *HL*, and *CHL* estimates ($\beta_i^{Estimates}$), with betas based on the TAQ effective spreads (β_i^{ES}). We discard stocks with fewer than 30 months of effective spread estimates. Betas are calculated for the spreads in levels and the residuals of AR(2) regressions in panels A and B, respectively. Panels C, D, and E show the results from subsampling analyses across exchanges (NYSE, AMEX, and NASDAQ), market capitalization, and spread size. In panel D, the size quintiles are sorted by increasing market capitalization at the last observed period for each individual stock. In panel E, the spread quintiles are sorted by increasing average effective spreads during the whole sample period. An asterisk indicates values not significantly different from that with the higher correlation marked in bold for every set of values. The statistical inferences are performed using Fisher's z-test.

when we perform the subsampling analysis across the 2003–2007, 2008–2011, and 2012–2015 subperiods.[26]

Following Acharya and Pedersen (2005), we analyze liquidity innovations generated from an AR(2) model. The analysis of liquidity in innovations, rather than by levels, helps us control for the persistence in the transaction cost process, thereby capturing the unexpected component of transaction costs.

[26] We reiterate the analysis using 25 portfolios sorted by illiquidity level like in Acharya and Pedersen (2005). As reported in the Internet Appendix, the results are fully consistent.

The results in panel B of Table 10 confirm the high accuracy of *CHL* estimates to gauge systematic liquidity risks using spread innovations. The correlation coefficients between the estimates of β_2, β_3, and β_4 from our model and the TAQ spreads are 0.524, 0.856, and 0.530, whereas the same correlations for *HL* estimates are 0.446, 0.879, and 0.266, and those for the *Roll* estimates are 0.158, 0.563, and 0.234, respectively. The subsampling analysis across shorter periods delivers consistent results, confirming that *CHL* estimates provide systematic risk estimates that follow the ones of the TAQ benchmark more closely, no matter if transaction costs are in levels or innovations.

Like in Section 3, we reiterate the subsampling analysis across primary exchange, market capitalization, and effective spread size (panels C, D, and E).[27] Overall, our estimator outperforms the other measures when stocks are grouped by market venues, market capitalization, and spread size. The only exceptions are β_4 for the NYSE and the largest capitalization quintile, for which the correlation between *HL* estimator and TAQ benchmark is higher, but the difference is not statistically significant. When stocks are subsampled from smallest to largest transaction costs, our estimator (the *HL* estimator) performs better across less (more) liquid stocks.

5. Conclusion

Building on the seminal model proposed by Roll (1984), we have derived a new way to estimate bid-ask spreads using price data. Compared with the *Roll* measure, our model has two important benefits: First, it takes advantage of a richer information set of daily close, high, and low prices, whereas the *Roll* measure solely relies on the close prices. Thereby, our model improves estimation accuracy. From the high and low prices, we can compute the mid-range, that is, the mean of the daily high and low log-prices, that proxies the efficient price. Second, our estimator is fully independent of order-flow dynamics, and therefore it does not rely on bid-ask bounces, as the original *Roll* measure does. Our method of estimating effective spreads is straightforward, is easy to compute, and has an intuitive closed-form solution that resembles the *Roll* measure. Whereas the *Roll* measure relies on the covariance of consecutive close-to-close price returns, our estimator relies on the covariance of close-to-mid-range returns around the same close price.

We tested our method numerically and empirically by using Trade and Quotes (TAQ) data. The simulation analysis shows that considering all imperfections together (i.e., infrequent trading, inconstant spreads, and nontrading periods), our model provides more accurate estimates than those from the high-low estimator proposed by Corwin and Schultz (2012) and the *Roll* model for less liquid securities, for which transaction costs and liquidity issues are of

27 To facilitate comparisons, we use the same quintile groups like in Section 3. However, here we remove a few more stocks that have fewer than 30 months of data.

much more concern. In the empirical analysis, the effective spread computed with TAQ data serves as the benchmark for our comparative considerations. When end-of-day quote data are available, that is, from 1993 onwards, the closing percentage quoted spread generally represents the most accurate low-frequency spread proxy. This is especially true across the post-decimalization era in the U.S. stock market from 2001 onwards, whereas before it, the closing percentage quoted spread (our estimator) outperforms the other estimators in terms of average time-series correlations (average cross-sectional correlations and lowest estimation errors).

On the other hand, when quote data are unavailable, our estimator is the most accurate one. Assessed against other estimates, it generally provides the highest cross-sectional and average time-series correlation with the TAQ effective spread benchmark, as well as the smallest prediction errors. We also have documented the additional explanatory ability of our estimates that systematically goes beyond that of other estimates. This additional predictive ability is especially larger for less liquid stocks. The numerical and empirical analyses suggest that our estimates are stable and much less sensitive to the number of trades per day, whereas the Corwin and Schultz (2012) high-low estimates produce substantially smaller spread estimates for lower number of trades per day, that is, for more illiquid stocks. The ability of our estimator to provide much more accurate spread estimates for less liquid stocks is a suitable characteristic because accurate estimates of transaction costs are particularly needed for less liquid securities and markets.

To illustrate some potential applications, we reconstructed the historical development of our spread estimates for stocks listed on NYSE (AMEX) from 1926 (1962) through 2015. These patterns display relatively stable variation over time and remain within an economically meaningful range, even for small-cap stocks. Then we estimated the components of systematic liquidity risk like in the liquidity-adjusted capital asset pricing model (LCAPM), which was postulated by Acharya and Pedersen (2005). The overall result is that our estimator provides accurate estimations of the systematic liquidity, in the sense that systematic risk betas based on our estimates are the closest to those of the TAQ benchmark and that our model generally outperforms other models in estimating systematic risk originating from commonality in liquidity and covariation between stock returns and illiquidity.

Our estimator has many potential applications for future research. It should be useful for researchers who work in asset pricing, corporate finance, risk management, and other important research areas and need a simple but accurate measure of trading costs over long periods. Our model could be suitably applied to many securities, including those traded over-the-counter or in emerging markets, for which data are of limited quality or availability.

Appendix A. Proof of Propositions 2 and 3

We first derive two propositions A1 and A2 that we need for the proofs.

Proposition A1. Under the model assumptions, Equation (A1) holds:

$$E\left[\left(h_t^e - c_t^e\right)\left(c_t^e - l_t^e\right)\right] = (2\log(2)-1)\sigma_e^2. \tag{A1}$$

Proof of proposition A1: to prove Proposition A1, we use the two following equations from Garman and Klass (1980):

$$E\left[\left(h_t^e - l_t^e\right)^2\right] = 4\log(2)\sigma_e^2, \tag{A2}$$

$$E\left[\left(h_t^e - c_t^e\right)^2\right] = E\left[\left(l_t^e - c_t^e\right)^2\right] = \sigma_e^2. \tag{A3}$$

Plugging (A2) and (A3) into (A1) leads to the proof

$$E\left[\left(h_t^e - c_t^e\right)\left(c_t^e - l_t^e\right)\right] = \frac{1}{2}E\left[\left(h_t^e - c_t^e + c_t^e - l_t^e\right)^2 - \left(h_t^e - c_t^e\right)^2 - \left(c_t^e - l_t^e\right)^2\right] = (2\log(2)-1)\sigma_e^2. \tag{A4}$$

Proposition A2. Under the model assumptions, Equation (A5) holds:

$$E\left[\left(c_t^e - h_t^e\right)^2\right] = E\left[\left(c_t^e - h_{t+1}^e\right)^2\right]. \tag{A5}$$

Proposition A2 is the result of the symmetry of Brownian motion in forward-looking and backward-looking expressions. More specifically, the distance between the efficient close price of day t and the efficient high (low) price of the same day is equal to the distance between the efficient open price and the efficient high (low) price of the next day:

$$E\left[\left(c_t^e - h_t^e\right)^2\right] = E\left[\left(o_{t+1}^e - h_{t+1}^e\right)^2\right]. \tag{A6}$$

Proof of Proposition A2: We assume no overnight price movements, so the efficient close price of day t and the efficient open price of day $t+1$ are identical and therefore, replacing o_{t+1}^e with c_t^e leads to the proof.

Proof of Proposition 2

Now we use the two propositions for the proof of Proposition 2 of the paper. The stepwise proof is as follows:

$$E\left[\left(c_t - (\eta_t + \eta_{t+1})/2\right)^2\right] = E\left[\left(c_t^e + q_t s/2 - \eta_t/2 - \eta_{t+1}/2\right)^2\right] \tag{A7}$$

$$= E\left[q_t^2 s^2/4 + 1/4\left(c_t^e - \eta_t\right)^2 + 1/4\left(c_t^e - \eta_{t+1}\right)^2 + 1/2\left(c_t^e - \eta_{t+1}\right)\left(c_t^e - \eta_t\right)\right.$$

$$\left. + qs/4\left(c_t^e - \eta_{t+1}\right) + qs/4\left(c_t^e - \eta_t\right)\right], \tag{A8}$$

$$= s^2/4 + 1/2E\left[\left(c_t^e - h_t^e/2 - l_t^e/2\right)^2\right], \tag{A9}$$

$$= s^2/4 + (1/2)E\left[(1/4)\left(c_t^e - h_t^e\right)^2 + (1/4)\left(c_t^e - l_t^e\right)^2 + (1/2)\left(c_t^e - l_t^e\right)\left(c_t^e - h_t^e\right)\right], \tag{A10}$$

$$= s^2/4 + \sigma_e^2/4 - (\log(2)/2 - 1/4)\sigma_e^2 = s^2/4 + (1/2 - \log(2)/2)\sigma_e^2. \tag{A11}$$

Equation (A7) is the result of the definition of the Roll (1984) model. Equation (A9) is the result of Proposition A2 and finally, we derive Equation (A11) using Proposition A1.

Proof of Proposition 3
The proof for Proposition 3 of the paper is similar to that of Proposition 2:

$$E\left[(\eta_{t+1}-\eta_t)^2\right]=E\left[\left(\eta_{t+1}-c_t^e+c_t^e-\eta_t\right)^2\right], \tag{A12}$$

$$=2E\left[(c_t^e-\eta_t)^2\right]=2E\left[\left(c_t^e-h_t^e/2-l_t^e/2\right)^2\right], \tag{A13}$$

$$=2E\left[(1/4)\left(c_t^e-h_t^e\right)^2+(1/4)\left(c_t^e-l_t^e\right)^2+(1/2)\left(c_t^e-l_t^e\right)\left(c_t^e-h_t^e\right)\right], \tag{A14}$$

$$=(2-2\log(2))\sigma_e^2. \tag{A15}$$

Equation (A13) is the result of Proposition A2, and, finally, we derive Equation (A15) using Proposition A1.

Appendix B. Proof of Robustness to Nontrading Periods

To prove the robustness of our estimator to nontrading periods, we repeat the logical steps followed in the paper by including the nontrading period in the efficient price variance. We then show that this term cancels out when we derive the outcome expression.

Definition B1. The nontrading period (e.g., overnight) efficient-price variance is defined as follows:

$$\sigma_{Nontrading}^2=E\left[\left(o_{t+1}^e-c_t^e\right)^2\right]. \tag{B1}$$

Proposition B1. If we consider a price movement during nontrading periods with the variance of $\sigma_{Nontrading}^2$, Equation (B2) holds:

$$E\left[\left(c_t-(\eta_t+\eta_{t+1})/2\right)^2\right]=s^2/4+(1/2-\log(2)/2)\sigma_e^2+1/4\sigma_{Nontrading}^2. \tag{B2}$$

Proof of Proposition B1: The proof is similar to the proof of Proposition 2, which is explained in Appendix A. The only difference arises because the distance between efficient close price of day t and the efficient high (low) price of day $t+1$ is higher than the distance between efficient close price of day t and the efficient high (low) price at the same day. Therefore, Equation (A5) no longer holds, and, instead, Equation (B3) shows the link between the two quantities. Using Equation (B3) and following the steps of the proof in in Appendix A leads to the proof of Proposition B1.

$$E\left[(c_t^e-h_{t+1}^e)^2\right]=E\left[\left(c_t^e-o_{t+1}^e+o_{t+1}^e-h_t^e\right)^2\right]=\sigma_{Nontrading}^2$$
$$+E\left[\left(o_{t+1}^e-h_t^e\right)^2\right]=\sigma_{Nontrading}^2+E\left[\left(c_t^e-h_t^e\right)^2\right]. \tag{B3}$$

Proposition B2. If we consider a price movement during nontrading periods (e.g., overnight) with the variance of $\sigma_{Nontrading}^2$, Equation (B4) holds:

$$E\left[(\eta_{t+1}-\eta_t)^2\right]=(2-2\log(2))\sigma_e^2+\sigma_{Nontrading}^2. \tag{B4}$$

Proof of Proposition B2: The proof is very similar to the proof of Proposition B1.

Proof of robustness to nontrading periods

When calculating s^2 using the two equations of proposition B1 and B2, the nontrading variance terms cancel out, and the result is identical to Equation (9):

$$s^2 = 4E[(c_t - \eta_t)(c_t - \eta_{t+1})]. \tag{B5}$$

Therefore, the spread estimates are independent of price movements during nontrading periods.

Appendix C. Relaxing the Assumption of the Buyer- (Seller-) Initiated High (Low) Prices

By relaxing the assumption of buyer- (seller-) initiated high (low) price, we obtain Equations (C1) and (C2) as a generalization of settings expressed in Equations (2) and (3).

$$h_t = h_t^e + q_t^h \frac{s}{2}, q_t^h = \pm 1, \tag{C1}$$

$$l_t = l_t^e + q_t^l \frac{s}{2}, q_t^l = \pm 1. \tag{C2}$$

Compared to Equations (2) and (3), here we allow the trade direction of high and low prices to be stochastic and independent of the efficient price process. The midrange η_t is the same as the one used in Definition 1 of the paper, that is, the average of observed high and low log-prices.

Proposition C1. Theorem 1 still holds if the assumptions of buyer-(seller-) initiated high (low) prices are replaced with the following assumptions:

(1) The trade directions of high and low prices are independent of the ones of previous day.
(2) The trade directions of high and low prices are independent of the ones for close prices.
(3) The chance of high price being buyer-initiated is equivalent to the chance of low price being seller-initiated.[28] This symmetry between the two trade directions is specified more formally in Equation (C3).

$$E\left[q_t^h\right] = -E\left[q_t^l\right] \tag{C3}$$

Proof of Proposition C1: Starting from the right-hand side of Equation (9) and replacing the observed close, high and low prices with the right-hand sides of Equations (1), (C1), and (C2). Using the assumptions that the efficient price path and trade directions are independent of each other, and the expected symmetry in efficient log-price movements, one can derive Equation (C4):

$$4E[(c_t - \eta_t)(c_t - \eta_{t+1})] = s^2\ E\left[q_t^2 + \frac{1}{4}\left(q_t^h + q_t^l\right)\left(q_{t+1}^h + q_{t+1}^l\right) - \frac{1}{2} q_t \left(q_t^h + q_t^l + q_{t+1}^h + q_{t+1}^l\right)\right]. \tag{C4}$$

Then, using the assumptions in Proposition C1, the expectation term in the right-hand side of Equation (C4) reduces to $E\left[q_t^2\right]$, which is equal to Equation (9) of the paper. It is important to note that q_t^h and q_t^l refer to the trade direction of *observed* rather than *efficient* high (low) prices. Hence, Equation (C3) does not necessarily impose dependence between trade directions and efficient price values. More specifically, while trade directions can be independent of the efficient price path, the high (low) *observed* trades might more often reflect buyer- (seller-) initiated trades because these trades are more likely to be *selected* as high (low) observed prices.

[28] As shown in the Internet Appendix, the analysis of Daily TAQ data provides empirical support to this assumption.

References

Acharya, V. V., and L. H. Pedersen. 2005. Asset pricing with liquidity risk. *Journal of Financial Economics* 77:375–410.

Amihud, Y. 2002. Illiquidity and stock returns: Cross section and time-series effects. *Journal of Financial Markets* 5:31–56.

Amihud, Y., and H. Mendelson. 1986. Asset pricing and the bid-ask spread. *Journal of Financial Economics* 17:223–49.

Asparouhova, E. N., H. Bessembinder, and I. Kalcheva. 2010. Liquidity biases in asset pricing tests. *Journal of Financial Economics* 96:215–37.

———. 2013. Noisy prices and inference regarding returns. *Journal of Finance* 68:665–714.

Beckers, S. 1983. Variance of security price returns based on high, low, and closing prices. *Journal of Business* 56:97–112.

Choi, J. Y., D. Salandro, and K. Shastri. 1988. On the estimation of bid-ask spread: Theory and evidence. *Journal of Financial and Quantitative Analysis* 23:219–29.

Chordia, T., R. Roll, and A. Subrahmanyam. 2000. Commonality in liquidity. *Journal of Financial Economics* 56:3–28.

Chung, K. H., and H. Zhang. 2014. A simple approximation of intraday spreads using daily data. *Journal of Financial Markets* 17:94–120.

Corwin, S. A., and P. Schultz. 2012. A simple way to estimate bid-ask spreads from daily high and low prices. *Journal of Finance* 67:719–59.

Fong, K., C. W. Holden, and C. A. Trzcinka. 2017. What are the best liquidity proxies for global research? *Review of Finance* Forthcoming.

Garman, M. B., and M. J. Klass. 1980. On the estimation of security price volatilities from historical data. *Journal of Business* 53:67–78.

Goyenko, R. Y., C. W. Holden, and C. A. Trzcinka. 2009. Do liquidity measures measure liquidity? *Journal of Financial Economics* 92:153–81.

Hameed, A., W. Kang, and S. Viswanathan. 2010. Stock market declines and liquidity. *Journal of Finance* 65:257–93.

Harris, L. E. 1989. A day-end transaction price anomaly. *Journal of Financial and Quantitative Analysis* 24: 29–45.

———. 1990. Statistical properties of the Roll serial covariance bid/ask spread estimator. *Journal of Finance* 45:579–590.

Hasbrouck, J. 2004. Liquidity in the futures pits: Inferring market dynamics from incomplete data. *Journal of Financial and Quantitative Analysis* 39:305–26.

———. 2009. Trading costs and returns for US equities: the evidence from daily data. *Journal of Finance* 64:1445–77.

Hasbrouck, J. and Thomas S. Y. Ho. 1987. Order arrival, quote behavior, and the return-generating process. *Journal of Finance* 42:1035–48.

Hasbrouck, J., and D. J. Seppi. 2001. Common factors in prices, order flows, and liquidity. *Journal of Financial Economics* 59:383–411.

Holden, C. W. 2009. New low-frequency liquidity measures. *Journal of Financial Markets* 12:778–813.

Holden, C. W., and S. Jacobsen. 2014. Liquidity measurement problems in fast, competitive markets: expensive and cheap solutions. *Journal of Finance* 69:1747–85.

Holden, C. W., S. Jacobsen, and A. Subrahmanyam. 2014. The empirical analysis of liquidity, *Foundations and Trends in Finance* 8:263–365.

Huberman, G., and D. Halka. 2001. Systematic liquidity. *Journal of Financial Research* 24:161–78.

Kamara, A., X. Lou, and R. Sadka. 2008. The divergence of liquidity commonality in the cross-section of stocks. *Journal of Financial Economics* 89:444–66.

Karolyi, G. A., K.-H. Lee, and M. A. van Dijk. 2012. Understanding commonality in liquidity around the world. *Journal of Financial Economics* 105:82–112.

Lee, C. M. C., and M. J. Ready. 1991. Inferring trade direction from intraday data. *Journal of Finance* 46:733–46.

Lesmond, D. A., J. P. Ogden, and C. A. Trzcinka. 1999. A new estimate of transaction costs. *Review of Financial Studies* 12:1113–41.

McInish, T. H., and R. A. Wood. 1990. An analysis of transactions data for the Toronto Stock Exchange: Return patterns and the end-of-day effect. *Journal of Banking and Finance* 14:441–58.

Newey, W. K., and K. D. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55:703–8.

Parkinson, M. 1980. The extreme value method for estimating the variance of the rate of return. *Journal of Business* 53:61–65.

Pastor, L., and R. F. Stambaugh. 2003. Liquidity risk and expected stock returns. *Journal of Political Economy* 111:642–85.

Roll, R. 1984. A simple implicit measure of the effective bid-ask spread in an efficient market, *Journal of Finance* 39:1127–39.

Stoll, H., and R. E. Whaley. 1983. Transaction costs and the small firm effects. *Journal of Financial Economics* 12:57–79.

Watanabe, A., and M. Watanabe. 2008. Time-varying liquidity risk and the cross section of stock returns. *Review of Financial Studies* 21:2449–86.