Available online at www.sciencedirect.com

ScienceDirect

Journal of Financial Markets 12 (2009) 778–813

Journal of FINANCIAL MARKETS

www.elsevier.com/locate/finmar

New low-frequency spread measures

Craig W. Holden*

Kelley School of Business, Indiana University, 1309 E. Tenth St., Bloomington IN 47405-1701, USA

Available online 22 July 2009

Abstract

I develop new spread proxies that pick up on three attributes of the low-frequency (daily) data: (1) price clustering, (2) serial price covariance accounting for midpoint prices on no-trade days, and (3) the quoted spread that is available on no-trade days. I develop and empirically test two different approaches: an integrated model and combined models. I test both new and existing low-frequency spread measures relative to two high-frequency benchmarks (percent effective spread and percent quoted spread) on three performance dimensions: (1) higher individual firm correlation with the benchmarks, (2) higher portfolio correlation with the benchmarks, and (3) lower distance relative to the benchmarks. I find that on all three performance dimensions the new integrated model and the new combined model do significantly better than existing low-frequency spread proxies.

© 2009 Elsevier B.V. All rights reserved.

JEL classification: C15; G12; G20

Keywords: Liquidity; Effective spread; Transaction cost; Asset pricing; Market efficiency

1. Introduction

In a classic and influential paper, Roll (1984) develops a simple proxy for the effective spread using price data only. Lesmond et al. (1999) and Hasbrouck (2004) develop additional proxies for the effective spread using low-frequency (daily) data. Amihud (2002) and Pastor and Stambaugh (2003) develop low-frequency liquidity measures that perhaps might be viewed as proxies for price impact, more than for the effective spread. Collectively, these low-frequency spread proxies allow the study of liquidity over relatively long periods of time and across countries. This is helpful to the asset pricing literature,

*Tel.: +1 812 855 3383; fax: +1 812 855 5875.
E-mail address: cholden@indiana.edu

doi:10.1016/j.finmar.2009.07.003

because recent studies suggest that liquidity is a priced risk factor. This is also helpful to recent studies in the market efficiency and corporate finance literatures, which utilize spread proxies for the cost of trade by stock, by time period, and across countries. Is it possible to create new low-frequency spread proxies that perform better than the existing low-frequency spread proxies? In this paper, better performance is primarily evaluated relative to two high-frequency benchmarks (percent effective spread and percent quoted spread) and on three dimensions: (1) higher individual firm correlation with the benchmarks, (2) higher portfolio correlation with the benchmarks, and (3) lower distance (tracking error) relative to the benchmarks. I find that the answer is "yes" on all three dimensions.

Spread proxies can be constructed from daily data going back more than 80 years in the United States and for various time spans in countries around the world. For U.S. equity markets, the Center for Research in Security Prices (CRSP) provides five key daily stock variables: prices, returns adjusted for splits and dividends, volume, high/ask, and low/bid.[1] These five variables are available for all NYSE/AMEX firms from December 31, 1925 to the present and for all NASDAQ firms from December 14, 1972 to the present.

High-performing low-frequency spread measures would be very helpful to the asset pricing literature. Chordia et al. (2000), Sadka (2003), Acharya and Pedersen (2005), Fujimoto (2004), Hasbrouck (2009), and others show that in recent U.S. experience various liquidity measures vary systematically and are priced; Bekaert et al. (2007) provide similar evidence for emerging markets where liquidity concerns may be more pronounced. Spread proxies going back in time and/or across countries are needed to determine whether or not these asset pricing relationships hold up across time and space.

High-performing low-frequency spread measures would be very helpful to the market efficiency and corporate finance literatures. De Bondt and Thaler (1985), Jegadeesh and Titman (1993), Jegadeesh and Titman (2001), Chan et al. (1996), Rouwenhorst (1998), and many others have found trading strategies that appear to generate significant abnormal returns. Correctly scaled spread proxies over time and/or across countries are needed to determine if these trading strategies are truly profitable net of a relatively precise measure of cost of trading. Similarly, Dennis and Strickland (2003), Kalev et al. (2003), Cao et al. (2004), Lipson and Mortal (2004a), Schrand and Verrecchia (2004), Lesmond et al. (2005), and many others examine the impact of corporate finance events on stock liquidity. Helfin and Shaw (2000), Lipson and Mortal (2004b), Lerner and Schoar (2004), and many others examine the influence of liquidity on capital structure, security issuance form, and other corporate finance decisions. Spread proxies over time would expand the potential sample size of this literature. Spread proxies across countries would greatly extend the potential diversity of international corporate finance environments that this literature could analyze.

This paper develops new, low-frequency spread measures that pick up on three attributes of the daily data. One attribute is price clustering—the higher likelihood for trade prices to be on rounder increments. One can directly observe the frequency of various price clusters (odd eighths, odd quarters, etc. on a fractional price grid and off-pennies, off-nickels, off-dimes, etc.[2] on a decimal price grid) and use this information to infer the

[1] High/ask means the highest trade price on a trading day or the closing ask price on a non-trading day. Similarly, low/bid means the lowest trade price on a trading day or the closing bid price on a non-trading day.

[2] Off-pennies are penny price points that are not nickels, dimes, or any higher clusters, namely where the last digit of the price is 1, 2, 3, 4, 6, 7, 8, or 9. Off-nickels are nickel price points that are not dimes, quarters, or any higher clusters, namely where the last two digits of the price are 05, 15, 35, 45, 55, 65, 85, or 95. And so on.

effective spread. The second attribute is serial covariance of price changes. I extend the Roll framework of serial covariance to account for no-trade days in which the reported price is the closing midpoint. The third attribute is the high/ask and low/bid variables available in the CRSP stock data. These variables directly supply the quoted spread on no-trade days.

I develop and empirically test two different approaches to incorporating these three attributes. First, I develop an integrated model that directly includes these attributes. The base integrated model, which I call "*Holden*," combines two attributes: price clustering and serial correlation. The expanded integrated model, which I call "*Holden*2," combines all three attributes. Second, I develop combined models, which I call *Multi-Factor* models, that are linear combinations of simpler one-attribute or two-attribute models.[3] I show theoretically that *Multi-Factor* models have the potential to diversify away some imperfectly-correlated error terms.

Then, I test the new, low-frequency spread measures against the existing low-frequency spread measures (Hasbrouck Gibbs, LOT Mixed, LOT Y-split, Pastor and Stambaugh, Roll, and Zeros).[4] All proxies are compared to two high-frequency benchmarks: (1) percent effective spread and (2) percent quoted spread. Both of these benchmarks are computed from the NYSE's Trade and Quote (TAQ) dataset for 400 firms from 1993 to 2005. Percent effective spread is a volume-weighted average based on every trade and corresponding BBO[5] quote in the stock-month. Percent quoted spread is a time-weighted average based on every BBO quote in the stock-month.

I test on three performance dimensions. First, I compute the correlation of each spread proxy with each benchmark based on individual firms. Second, I create an aggregate spread measure for each proxy and benchmark based on an equally weighted portfolio across all 400 firms. Then, I compute the pure time-series correlation of each aggregate spread proxy with each aggregate benchmark. Third, I compute the average root mean squared error for each spread proxy compared to each benchmark.

I find that on all three performance dimensions with regard to both benchmarks the new integrated model *Holden*2 does significantly better than existing low-frequency spread proxies. I find that on all three performance dimensions with regard to both benchmarks the new combined model *Multi-Factor*2 does significantly better than existing low-frequency spread proxies, except for one tie. Summarizing six tests (three performance dimensions *X* two benchmarks), the combined model *Multi-Factor*2 does significantly better than the integrated model *Holden*2 on four out of six tests.

I also find that these new proxies are robust by size quintiles, price quintiles, and tick size regime. Consistently, *Holden*2 is the best integrated model and *Multi-Factor*2 is the best combined model. Across all size, price, and tick size regime comparisons, *Multi-Factor*2 is the most frequent winner and *Holden*2 is the second most frequent winner. Finally, I compare the proxies to low-frequency spread benchmarks and find that new proxies consistently outperform existing proxies.

[3] Out of the infinite number of possible versions, I test very simple combinations (50–50% weightings) of the one-attribute or two-attribute models that perform well on their own. A particularly successful combination, called "*Multi-Factor*2," is defined as 50% * *Extended Roll*2 + 50% * *No Trade Quoted Spread*. The latter two models are explained later in the text.

[4] I do not test against the Amihud (2002) measure, since it is really a proxy for price impact, rather than spread.

[5] BBO means the best bid and offer. It is the highest bid price and lowest ask available for a given stock at a moment in time.

An empirical companion paper, Goyenko et al. (2009), performs extensive testing of monthly and annual liquidity proxies. It tests earlier-developed proxies, *Holden*, *Effective Tick*, and *Effective Tick*2, and existing proxies from the prior literature. In their monthly effective spread comparisons, they found that *Holden*, *Effective Tick*, and LOT Y-split significantly outperformed all other spread proxies then in existence. In this paper, later-developed proxies, *Holden*2, *Multi-Factor*1, *Multi-Factor*2, and other newly developed proxies are tested for the first time. I find that *Holden*2 and *Multi-Factor*2 outperform the earlier-developed proxies, *Holden*, *Effective Tick*, and LOT Y-split, and outperform existing proxies from the prior literature.

Hasbrouck (2009) tests *annual* estimates of four liquidity proxies using U.S. data. Lesmond (2005) tests *quarterly* estimates of five liquidity proxies using data from 31 emerging countries. The vast majority of asset pricing, market efficiency, or corporate finance literature that examines the role of liquidity is based on *monthly* (or finer) data. Goyenko et al. (2009) and this paper contribute to our knowledge of potential spread proxies by testing *monthly* estimates.

The paper is organized as follows. Section 2 develops the integrated Holden model. Section 3 develops both single-attribute models and combined multi-factor models and shows that multi-factor models can diversify away some imperfectly correlated measurement error. Section 4 describes the data and empirically tests both new and existing low-frequency spread measures on three performance dimensions. Section 5 concludes. Three appendices contain technical details.

2. The integrated approach

2.1. Setup of the Holden model

I begin by develop an integrated model that directly incorporates multiple attributes of the daily data. Called the Holden model, it is constructed on top of the general microstructure framework of Huang and Stoll (1997) and uses their notation as much as possible. Huang and Stoll allow the spread to be generated by all three traditional microstructure components (adverse selection, inventory risk, and order processing costs). Their model includes as special cases the covariance spread models (Roll, 1984; Stoll, 1989; George et al., 1991) and the trade indicator spread models (Glosten and Harris, 1988; Madhavan et al., 1997).

I modify the Huang and Stoll model in two ways. First, I allow the effective spread S_t to change from day to day. This modification follows the precedent of Madhavan et al. (1997), who model and estimate a trade indicator spread model in which the effective spread and its components change from hour to hour, and it implements one of the Huang and Stoll (1997) suggestions for future research. Second, I eliminate their parameter for the inventory risk component (their β). Inventory risk is unobservable in daily data and so I allow the inventory risk component to contribute to an error term.

Theoretical reasons why the spread might change from day-to-day include: (1) anticipated information events (e.g., earnings announcements, dividend announcements, etc.) which change the ex ante amount of adverse selection (Lee et al., 1993), (2) changes in the mix of patient versus impatient traders causing variation in the aggressiveness of undercutting the existing spread (Foucault et al., 2005), (3) stochastic volatility, which causes time-variation in the inventory cost component and time-variation in the noise/

signal ratio for privately informed traders, which changes the ex ante amount of adverse selection (Harris, 2003, p. 313). Empirical evidence that observed spreads (and spread components) in fact do vary, even on an hour-by-hour basis, include Chan et al. (1995), Madhavan et al. (1997), Chung et al. (2003), and Chung and Zhao (2003).

Let V_t be the unobservable fundamental value of the stock at the end of day t. It is assumed to evolve as follows:

$$V_t = V_{t-1} + \tfrac{1}{2}\alpha S_{t-1} Q_{t-1} + \varepsilon_t \tag{1}$$

where α is the percentage of the half-spread attributable to adverse selection, Q_{t-1} is a buy/sell/no-trade indicator on day $t-1$, and ε_t is a serially uncorrelated public information shock on day t.

Let $\mu/2$ be the probability of a closing trade at the ask, $\mu/2$ be the probability of closing trade at the bid, and $1-\mu$ be the probability of a no-trade day where the reported price is the closing midpoint. Then the buy/sell/no-trade indicator Q_t is given by[6]

$$Q_t \equiv \begin{cases} +1 & \text{Closing trade is a buy (prob} = \mu/2) \\ 0 & \text{No-trade day; reported price is the closing midpoint (prob} = 1-\mu) \\ -1 & \text{Closing trade is a sell (prob} = \mu/2). \end{cases} \tag{2}$$

Let M_t be the unobserved bid–ask midpoint at the end of day t. It is determined by the fundamental value of the stock plus inventory effects as given by

$$M_t = V_t + \omega_t, \tag{3}$$

where ω_t is cumulative inventory effect of all prior trades.

Taking the first difference of Eq. (3) and combining it with Eq. (1) yields the daily change in midpoint

$$\Delta M_t = \tfrac{1}{2}\alpha S_{t-1} Q_{t-1} + \varepsilon_t + \Delta\omega_t, \tag{4}$$

where Δ is the change operator.

Let P_t be the observed trade price at the end of day t. It is determined by

$$P_t = M_t + \tfrac{1}{2} S_t Q_t + \eta_t(S_t), \tag{5}$$

where the error term $\eta_t(S_t)$ accounts for rounding the trade price to the same discrete price cluster as the spread S_t. For example, when $S_t = \frac{1}{4}$, the price is rounded to the nearest $\$\frac{1}{4}$ increment (e.g., $\$\frac{1}{4}, \$\frac{1}{2}, \$\frac{3}{4}$, whole dollar). Then $E[\eta_t(S_t)] = 0$, because rounding is equally likely to be up or down.

Combining Eqs. (4) and (5) yields the key price change process

$$\Delta P_t = \tfrac{1}{2} S_t Q_t - (1-\alpha)\tfrac{1}{2} S_{t-1} Q_{t-1} + e_t, \tag{6}$$

where by definition $e_t \equiv \varepsilon_t + \Delta\eta_t + \Delta\omega_t$. For simplicity, e_t is assumed to be normally distributed with a mean $\bar{e}$ and a standard deviation σ_e. Eq. (6) is the same as Haung and Stoll's equation (5), except that spreads change from day-to-day and the inventory risk component has been moved to the error term.

[6] In Huang and Stoll (1997), $Q_t = 0$ means that a trade happens at the midpoint, rather than a no-trade day where the reported price is the closing midpoint. When limited to daily data, the former is not observable, whereas the later is observable when the daily volume variable is zero.

Define H_t as the half-spread on date t, given by

$$H_t \equiv \tfrac{1}{2}S_t Q_t. \tag{7}$$

Substitute Eq. (7) into Eq. (6) and solve for the error term implied by the price change and the half-spreads

$$e_t = \Delta P_t - (H_t - (1-\alpha)H_{t-1}). \tag{8}$$

2.2. Price clustering

An interesting attribute of the daily data is price clustering. Price clustering is defined by there being a higher likelihood of trade prices on rounder increments. On a fractional price grid, whole dollars are rounder than half dollars, which are rounder than quarter dollars, which are rounder that eighths of a dollar. On a decimal price grid, whole dollars are rounder than quarters, which are rounder than dimes, which are rounder than nickels, which are rounder than pennies.

Harris (1991) documents that price clustering is remarkably persistent over time and across stocks. He finds extensive price clustering in CRSP daily closing prices during 1963–1987. He even finds significant price clustering in NYSE closing prices from 1854, which is the earliest date with *typeset* price records. Harris offers the theoretical explanation that price clustering reduces the negotiation costs between two potential traders by avoiding frivolous price increments that waste time and by reducing the amount of information that needs be exchanged.

The intuition for incorporating price clustering in the model comes from Christie and Schultz (1994), who emphasize the connection between observed price clusters and the spread. For example, if trade prices are exclusively on even eighths increments ($\$\frac{1}{4}, \$\frac{1}{2}, \$\frac{3}{4}, \$1$), then the bid–ask spread must be $\$\frac{1}{4}$ (or larger). Whereas, if trade prices are half of the time on odd eighths increments ($\$\frac{1}{8}, \$\frac{3}{8}, \$\frac{5}{8}, \$\frac{7}{8}$), then the likely bid–ask spread is $\$\frac{1}{8}$. In their application, the much greater avoidance of odd eighth trade prices in NASDAQ stocks, compared to a matched sample of NYSE stocks, provided evidence of implicit collusion by NASDAQ dealers to maintain abnormally large spreads.

In my application, the frequency with which closing prices occur in clusters of *special* prices (e.g., odd $\frac{1}{8}$s, odd $\frac{1}{4}$s, odd $\frac{1}{2}$s, and whole dollars) can be used to infer the effective spread. There are similar clusters of *special* prices on a decimal price grid (e.g., off pennies, off nickels, off dimes, off quarters, and whole dollars) that can be used to infer the effective spread in that case. This section develops the Holden model for any fractional price grid, where it is simpler and easier to see the intuition. Appendix A provides the modifications required to accommodate any decimal or fractional price grid.

Price clustering is included by assuming that trading is conducted in two steps. First, traders decide (explicitly or implicitly) what price cluster to use on a particular day, so as to minimize negotiation costs (Harris, 1991). On a fractional price grid, they choose whether to use eighths, quarters, halves, or wholes. On a decimal price grid, they choose whether to use pennies, nickels, dimes, quarters, or dollars. Second, the traders negotiate a particular price from the chosen price cluster.

Following the price cluster/spread connection (Christie and Schultz), I assume that the effective spread on a particular day equals the increment of the price cluster on that particular day. In other words, the choice of one is equivalent to the choice of the other.

I model the effective spread on a particular day S_t as a random drawn from a set of possible effective spreads $s_j, j = 1, 2, \ldots, J$ with corresponding probabilities $\gamma_j, j = 1, 2, \ldots, J$. I follow the convention that the possible effective spreads $s_1, s_2, \ldots, s_J$ are ordered from smallest to largest. For example on a $\$\frac{1}{8}$ price grid, S_t is modeled as having a probability γ_1 of $s_1 = \$\frac{1}{8}$ spread, γ_2 of $s_2 = \$\frac{1}{4}$ spread, γ_3 of $s_3 = \$\frac{1}{2}$ spread, and γ_4 of $s_4 = \$1$ spread.[7]

Further I assume that the closing price is uniformly distributed on the possible price increments for a given effective spread size. For example on date t, if $S_t = s_1 = \$\frac{1}{8}$, then the closing trade price is uniformly distributed on the eight possible price increments ($\$\frac{1}{8}, \$\frac{1}{4}, \$\frac{3}{8}, \ldots, \$1$). On date $t+1$, if $S_{t+1} = s_2 = \$\frac{1}{4}$, then the closing trade price is uniformly distributed on the four rounder price increments ($\$\frac{1}{4}, \$\frac{1}{2}, \$\frac{3}{4}, \$1$). And so on.

Let C_t be the observable price/midpoint cluster on day t. Clusters $C_t = 1, 2, \ldots, J$ are the clusters of special *prices* that correspond to possible effective spreads $s_1, s_2, \ldots, s_J$. Clusters $C_t = J+1, J+2, \ldots, 2J$ are the clusters of special *midpoints* that correspond to possible effective spreads $s_1, s_2, \ldots, s_J$.

Fig. 1 illustrates this set-up on a $\$\frac{1}{8}$ price grid. The first part of the tree is the effective spread nodes S_t. There are four possible nodes: $s_1 = \$\frac{1}{8}$ spread, $s_2 = \$\frac{1}{4}$ spread, $s_3 = \$\frac{1}{2}$ spread, and $s_4 = \$1$ spread that happen with probabilities $\gamma_1, \gamma_2, \gamma_3$, and γ_4, respectively. From any of these nodes, the next part of the tree is the buy/sell/midpoint indicator Q_t. It can take on values $+1$ (Buy), -1 (Sell), and 0 (Midpoint) with probabilities $\mu/2$, $\mu/2$, and $1-\mu$, respectively. From any of these second-level nodes, the next part of the tree is the observable price/midpoint clusters C_t. Consider the first, second-level node representing $S_t = s_1 = \$\frac{1}{8}$ spread and $Q_t = +1$ (Buy). From this node, there are four feasible clusters: $C_t = 1$ Odd $\$\frac{1}{8}$ prices, $C_t = 2$ Odd $\$\frac{1}{4}$ prices, $C_t = 3$ Odd $\$\frac{1}{2}$ prices, and $C_t = 4$ Whole dollar prices. Given the assumed uniform distribution, the probabilities of these four clusters are $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, and $\frac{1}{8}$, respectively. Consider the third, second-level node representing $S_t = s_1 = \$\frac{1}{8}$ spread and $Q_t = 0$ (Midpoint). From this node, there is only one feasible cluster: $C_t = 5$ Odd $\$\frac{1}{16}$ midpoints. The rest of the setup tree is analogous.

2.3. The likelihood function

The ultimate goal is to estimate a spread proxy. One version of the model, *Holden*, is the weighted-average percent effective spread computed using the estimated spread probabilities $\hat{\gamma}_j$ as follows:

$$Holden = \frac{\sum_{j=1}^{J} \hat{\gamma}_j s_j}{\bar{P}}, \tag{9}$$

where $\bar{P}$ is the average trade price over the time period of aggregation. Naturally, the estimated spread probabilities must sum to one

$$\sum_{j=1}^{J} \hat{\gamma}_j = 1. \tag{10}$$

[7] A $\$\frac{3}{8}$ spread (or a $\$\frac{5}{8}$ spread, etc.) results in using all eight possible price increments ($\$\frac{1}{8}, \$\frac{1}{4}, \$\frac{3}{8}, \ldots, \$1$) uniformly. This cannot be empirically distinguished from a $\$\frac{1}{8}$ spread, which also results in using all eight possible price increments uniformly. Therefore, I have chosen to model only those possible effective spreads that can be empirically distinguished from each other.

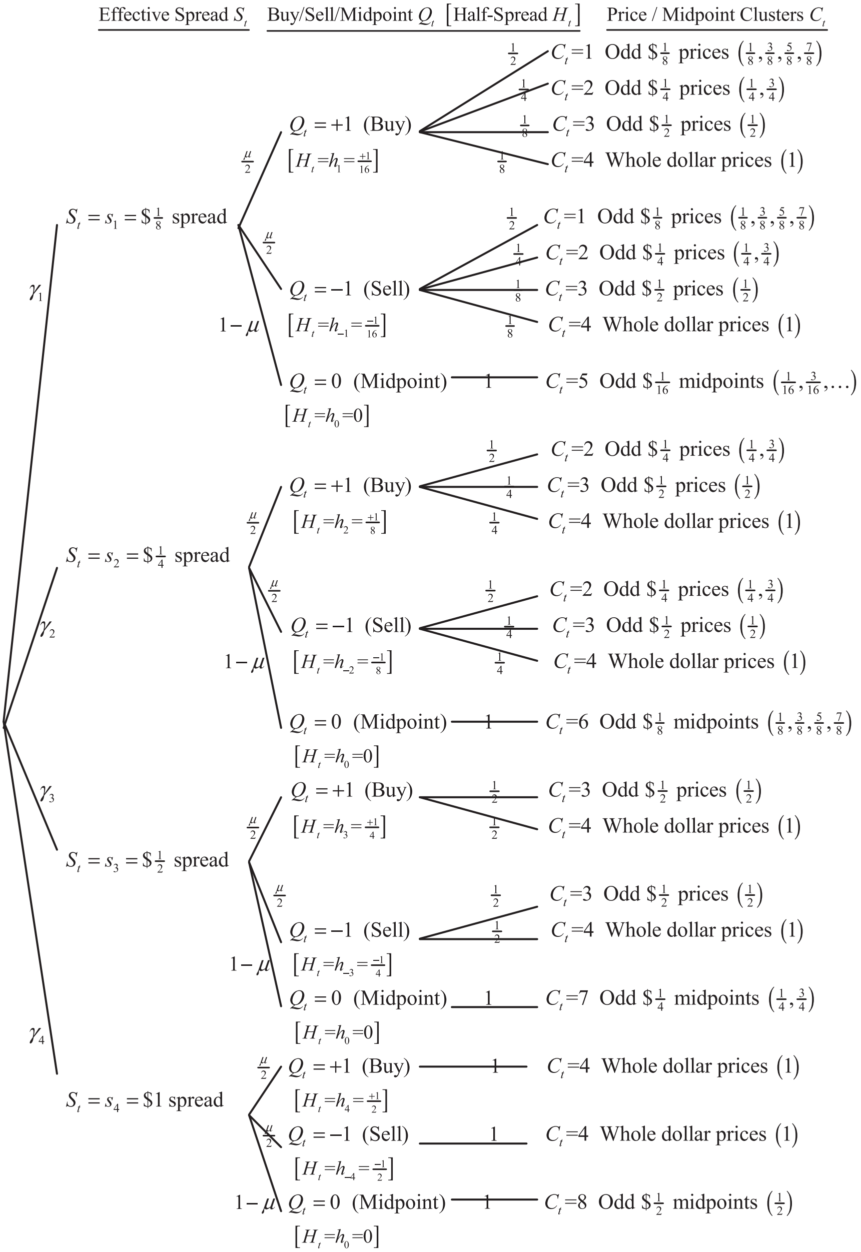
Effective Spread S_t
Buy/Sell/Midpoint Q_t [Half-Spread H_t]
Price / Midpoint Clusters C_t
γ_1
γ_2
γ_3
γ_4
$S_t = s_1 = \$\frac{1}{8}$ spread
$S_t = s_2 = \$\frac{1}{4}$ spread
$S_t = s_3 = \$\frac{1}{2}$ spread
$S_t = s_4 = \$1$ spread
$\frac{\mu}{2}$
$1-\mu$
$Q_t = +1$ (Buy)
$[H_t = h_1 = \frac{+1}{16}]$
$Q_t = -1$ (Sell)
$[H_t = h_{-1} = \frac{-1}{16}]$
$Q_t = 0$ (Midpoint)
$[H_t = h_0 = 0]$
$[H_t = h_2 = \frac{+1}{8}]$
$[H_t = h_{-2} = \frac{-1}{8}]$
$[H_t = h_3 = \frac{+1}{4}]$
$[H_t = h_{-3} = \frac{-1}{4}]$
$[H_t = h_4 = \frac{+1}{2}]$
$[H_t = h_{-4} = \frac{-1}{2}]$
$\frac{1}{2}$
$\frac{1}{4}$
$\frac{1}{8}$
1
C_t=1 Odd $\$\frac{1}{8}$ prices $(\frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8})$
C_t=2 Odd $\$\frac{1}{4}$ prices $(\frac{1}{4}, \frac{3}{4})$
C_t=3 Odd $\$\frac{1}{2}$ prices $(\frac{1}{2})$
C_t=4 Whole dollar prices (1)
C_t=5 Odd $\$\frac{1}{16}$ midpoints $(\frac{1}{16}, \frac{3}{16}, \ldots)$
C_t=6 Odd $\$\frac{1}{8}$ midpoints $(\frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8})$
C_t=7 Odd $\$\frac{1}{4}$ midpoints $(\frac{1}{4}, \frac{3}{4})$
C_t=8 Odd $\$\frac{1}{2}$ midpoints $(\frac{1}{2})$

Fig. 1. Setup of the Holden model on a $\$\frac{1}{8}$ price grid.

An interesting attribute of the CRSP stock database comes from two variables: high/ask and low/bid. On trading days, these variables report the high and low trade prices for the day. On no-trade days, these variables report the closing ask and bid. Thus on every no-trade day, the CRSP stock database provides the quoted spread $QS_t \equiv A_t - B_t$, where A_t is the closing ask price on day t and B_t is the closing bid price on day t.

A second version of the model, *Holden*2, utilizes the quoted spread from any no-trade days that exist. Specifically, this version uses the weighted-average effective spread on trading days and the average quoted spread on no-trade days as follows:

$$Holden2 = \frac{\mu\sum_{j=1}^{J}\hat{\gamma}_j s_j + \begin{cases} (1-\mu)\dfrac{1}{NTD}\sum_{t=1}^{NTD} QS_t & \text{When } NTD > 0, \\ 0 & \text{When } NTD = 0 \end{cases}}{\bar{P}}, \tag{11}$$

where *NTD* is the number of no-trade days in the estimation time interval. Thus, *Holden*2 *integrates* all three attributes: price clustering, serial covariance accounting for no-trade midpoints, and the no-trade quoted spread.

On three successive trading days, we observe a price triplet (P_t, P_{t+1}, P_{t+2}), which uniquely corresponds to a price cluster triplet (C_t, C_{t+1}, C_{t+2}). Define H as the set of all half-spread triplets (H_t, H_{t+1}, H_{t+2}) that are feasible given the observed price cluster triplet.[8] Using three prices at a time allows the serial covariance of the price changes to be picked up, but avoids the combinatoric explosion of feasible half-spread combinations that would result if all observations were used at the same time.

The parameters to be estimated are: (1) the probability of a trading day μ, (2) all-but-the-highest of the spread probabilities $\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_{J-1}$, (3) the mean of the error variable $\bar{e}$, (4) the standard deviation of the error variable σ_e, and (5) the percentage of the spread due to adverse selection α. For a given a set of parameter values $(\mu, \hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_{J-1}, \bar{e}, \sigma_e, \alpha)$, the likelihood of the price triplet is

$$= \sum_{(H_t, H_{t+1}, H_{t+2}) \in H} \left\{ \begin{array}{l} Pr(C_t) \cdot Pr(C_{t+1}) \cdot Pr(C_{t+2}) \cdot Pr(H_t|C_t) \cdot Pr(H_{t+1}|C_{t+1}) \cdot Pr(H_{t+2}|C_{t+2}) \\ \cdot n(P_{t+1} - H_{t+1} - (P_t - (1-\alpha)H_t)) \cdot n(P_{t+2} - H_{t+2} - (P_{t+1} - (1-\alpha)H_{t+1})) \end{array} \right\}, \tag{12}$$

where $n()$ is the normal density with a mean of $\bar{e}$ and a standard deviation of σ_e. The price cluster probabilities $Pr(C_t)$ and the half-spread conditional probabilities $Pr(H_t|C_t)$ are given in Appendix B.

Taking the log of the likelihood of a given price triplet, the overall likelihood function becomes the sum of the log likelihoods of all price triplets in the time period of aggregation

$$\sum_{t=1}^{T-2} Ln(Pr(P_t, P_{t+1}, P_{t+2}|\mu, \hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_{J-1}, \bar{e}, \sigma_e, \alpha)), \tag{13}$$

where T is the number of days in the time period of aggregation. The likelihood function is estimated using overlapping three-day windows to maximize the amount of information extracted.

The likelihood function is maximized subject to constraints using a nonlinear, numerical optimizer, which continues to iterate until five consecutive iterations lead to no further increase in the objective function out to eight-digit accuracy and all constraints are meet

[8] For example, suppose that the price $P_t = \$25\frac{1}{8}$ is an odd eighth that corresponds to price cluster $C_t = 1$. For this price cluster there is only feasible spread $S_t = \$\frac{1}{8}$. Thus, there are only two feasible values of the signed half-spreads $H_t \in \{\$\frac{1}{16}, -\$\frac{1}{16}\}$. Similarly, P_{t+1} and P_{t+2} imply the feasible values of the signed half-spreads H_{t+1} and H_{t+2}. Taking all combinations of the feasible values on each day, yields the set of feasible half-spread triplets.

with eight-digit precision. The constraints are that $\mu, \hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_J, \sigma_e, \alpha$ are greater than or equal to zero and the constraints that $\mu, \hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_J, \alpha$ are less than or equal to one. The constraints $\gamma_J \geq 0$ and $\gamma_J \leq 1$ can be expressed as a function of parameters to be estimated as $1 - \sum_{j=1}^{J-1} \hat{\gamma}_j \geq 0$ and $1 - \sum_{j=1}^{J-1} \hat{\gamma}_j \leq 1$.

3. The combined approach

This section develops pure single-attribute models based on: (1) price clustering, (2) serial covariance accounting for no-trade midpoints, and (3) no-trade quoted spread. Then it develops combined models that are linear combinations of the simpler models and shows that these combined models have the potential to diversify away some imperfectly correlated error terms.

3.1. *A pure price clustering model*

The first step is to collect the empirical frequency of each price cluster. Let N_j be the observed number of special trade prices corresponding to the jth spread $(j = 1, 2, \ldots, J)$ from positive-trade days. For example on a $\$\frac{1}{8}$ price grid, N_1 through N_4 are the observed number of odd $\frac{1}{8}$ prices, the number of odd $\frac{1}{4}$ prices, the number of odd $\frac{1}{2}$ prices, and the number of whole dollar prices, respectively.

Let N_{J+j} be the observed number of no-trade midpoints corresponding to the jth spread $(j = 1, 2, \ldots, J)$ from no-trade days. On a $\$\frac{1}{8}$ price grid, N_5 through N_8 are the number of odd $\frac{1}{16}$ midpoints, the number of odd $\frac{1}{8}$ midpoints, the number of odd $\frac{1}{4}$ midpoints, and the number of odd $\frac{1}{2}$ midpoints, respectively. Thus, the complete frequency distribution spans $2J$ events, with N_1 through N_J representing special trade prices and N_{J+1} through N_{2J} representing no-trade midpoints.

Let F_j and F_{J+j} be the observed probabilities of special trade prices and no-trade midpoints, respectively, corresponding to the jth spread $(j = 1, 2, \ldots, J)$. These empirical probabilities are easily computed as

$$F_j = \frac{N_j}{\sum_{j=1}^{2J} N_j} \quad \text{for } j = 1, 2, \ldots, 2J. \tag{14}$$

Again, the complete set of probabilities includes F_1 through F_J representing special trade prices and F_{J+1} through F_{2J} representing no-trade midpoints.

To get the intuition of a price clustering model, consider what happens when the spread is $\$\frac{1}{8}$. When there is a trade, half of the time the price is an odd $\frac{1}{8}$ and half of the time it is an even $\frac{1}{8}$. When there is no trade, the midpoint is strictly an odd $\frac{1}{16}$. Intuitively, to compute the probability of a $\$\frac{1}{8}$ spread, we need to double the probability of trade odd $\frac{1}{8}$ prices (in order to account for trade even $\frac{1}{8}$ prices) and add the probability of no-trade odd $\frac{1}{16}$ midpoints.[9]

[9] Trade prices on odd $\frac{1}{8}$ *are* used to infer the probability of a $\$\frac{1}{8}$ spread, because they could only have been generated by a $\$\frac{1}{8}$ spread under the assumptions of the model. Trade prices on even $\frac{1}{8}$ are *not* used to infer the probability of a $\$\frac{1}{8}$ spread, because they could have been generated by a $\$\frac{1}{4}$ spread or higher. The probability of the no-trade odd $\frac{1}{16}$ midpoints is *not* doubled, because there is a one-to-one mapping between odd $\frac{1}{16}$ midpoints and the $\$\frac{1}{8}$ spread. Note that a $\$\frac{1}{4}$ spread generates odd $\frac{1}{8}$ midpoints, a $\$\frac{1}{2}$ spread generates odd $\frac{1}{4}$ midpoints, and a \$1 spread generates odd $\frac{1}{2}$ midpoints. Thus, on a fractional price grid there is no overlap between the midpoints generated by any of the spreads.

Let U_j be the unconstrained probability of the jth spread ($j = 1, 2, \ldots, J$). Following the intuition above, the unconstrained probability[10] of the effective spread is

$$U_j = \begin{cases} 2F_j + F_{J+j}, & j = 1, \\ 2F_j - F_{j-1} + F_{J+j}, & j = 2, 3, \ldots, J-1\,, \\ F_j - F_{j-1} + F_{J+j}, & j = J. \end{cases} \tag{15}$$

Possible effective spreads larger than the smallest face the problem of overlapping price increments. For example, even $\frac{1}{8}$ prices can be generated by both a $\$\frac{1}{8}$ spread and a $\$\frac{1}{4}$ spread. Intuitively, to compute the probability of a $\$\frac{1}{4}$ spread, we need to do the same thing as the smallest, but also subtract off the probability of those even$\frac{1}{8}$ prices that were generated by a $\$\frac{1}{8}$ spread. The Jth spread is also a little different, because there is no larger spread above it. This implies that there are no price increments (no even \$1 increments) that are not used because they might have been generated by some larger spread. Hence, F_J the probability of the Jth spread is *not* doubled.

A price clustering model assumes a higher frequency on rounder increments and in large samples this will undoubtedly occur. But in small samples it is possible that *reverse* price clustering may be realized (i.e., a lower frequency on rounder increments). Reverse price clustering unintentionally causes the unconstrained probability of one or more effective spread sizes to go above 1 or below 0. Thus, constraints are added to generate proper probabilities. Let $\hat{\gamma}_j$ be the *constrained* probability of the jth spread ($j = 1, 2, \ldots, J$). It is computed in order from smallest to largest as follows:

$$\hat{\gamma}_j = \begin{cases} \text{Min}[\text{Max}\{U_j, 0\}, 1], & j = 1, \\ \text{Min}[\text{Max}\{U_j, 0\}, 1 - \sum_{k=1}^{j-1} \hat{\gamma}_k], & j = 2, 3, \ldots, J. \end{cases} \tag{16}$$

One version of a pure price clustering model, called *Effective Tick*3, is simply a probability-weighted average of each effective spread size divided by the average price

$$\textit{Effective Tick3} = \frac{\sum_{j=1}^{J} \hat{\gamma}_j s_j}{\bar{P}}. \tag{17}$$

The idea behind the name is that the minimum tick size tells you the ticks that are *allowed*, but the effective tick size tells you the weighted-average tick size that is *effectively used*.

The effective tick model is very flexible. Possible effective spreads (s_j's) can easily be added based on smaller increments (such as $\$\frac{1}{256}$ or $\$\frac{1}{128}$ in a fractional world or subpennies in a decimal world), added based on larger increments, or deleted. Similarly, it is trivial to scale the possible effective spreads and cluster definitions up or down by orders of magnitude if a particular stock/time period trades in a different price range (i.e., Berkshire Hathaway).[11] The observed price increments in the data usually suggest what possible effective spreads are relevant for a particular stock/time period.

A simpler version, called *Effective Tick*,[12] is the same except that it only uses trade prices. In other words, it throws away the no-trade midpoints ($N_{J+1} = N_{J+2} = \cdots = N_{2J} = 0$). Another simple version, called *Effective Tick*2, is the same as *Effective Tick*, except that it

[10] Appendix A provides the required modification to accommodate any decimal or fractional price grid.

[11] Equivalently, the raw prices/midpoints can be scaled down (e.g., divided by 1,000 for Berkshire Hathaway) and then the estimated effective spread can be scaled back up (multiplied by 1,000).

[12] Effective Tick and Effective Tick2 are jointly developed by this paper and Goyenko et al. (2009).

uses all trade and non-trade days. In other words, it treats all prices as if they were trade prices. These versions might be useful if trade prices are substantially more informative than no-trade midpoints or for non-U.S. databases where no-trade midpoints are not available.

The final version, called *Effective Tick*4, integrates two attributes of the daily data: price clustering and the No-Trade Quoted Spread. Specifically, it uses the price clustering information from trading days and the no-trade quoted spread information from no-trade days as follows:

$$Effective\ Tick4 = \frac{\hat{\mu}(Effective\ Tick1) + \begin{cases} (1-\hat{\mu})\dfrac{1}{NTD}\sum_{t=1}^{NTD} PQS_t & \text{When } NTD > 0, \\ 0 & \text{When } NTD = 0 \end{cases}}{\bar{P}}, \tag{18}$$

where $\hat{\mu}$ is the estimated probability of a trading day given by

$$\hat{\mu} = \frac{TD}{TD + NTD} \tag{19}$$

and where TD and NTD are the number of trading days and no-trade days, respectively.

3.2. A pure serial covariance model including no-trade midpoints

Another interesting attribute of the daily data is the serial covariance of observed price changes. Roll (1984) uses the serial covariance to develop a proxy of the effective spread. He assumes that a security trades half of time at the bid and half of time at the ask. The bouncing of the security's price back and forth between the bid and ask creates negative serial covariance. Roll's famous formula writes the bid–ask spread as a simple function of the serial covariance.

I extend the Roll model by considering the possibility of a no-trade day where the reported price is the closing midpoint. To illustrate this extension, consider the simple case in which the midpoint is constant. That is, suppose that no fundamental information comes out about the firm and liquidity providers do not have to contend with adverse selection or inventory concerns. Fig. 2 provides a schematic of nine possible paths of the reported closing price given that the date $t-1$ closing price was at the bid.

On any date, let μ be the probability of a trading day and $1-\mu$ be the probability of a no-trade day. Maintaining all of the assumptions of the Roll framework, including that the

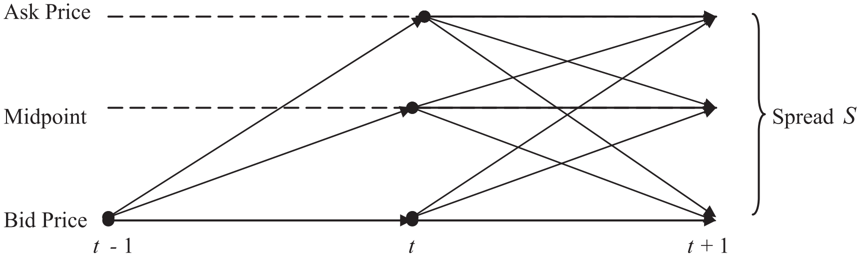

Fig. 2. Possible paths of the reported closing price.

Table 1
Probability distribution conditional on the date $t{-}1$ price.

		P_{t-1} is at the bid					P_{t-1} is at the midpoint					P_{t-1} is at the ask		
			ΔP_t					ΔP_t					ΔP_t	
		0	$+\frac{S}{2}$	$+S$			$-\frac{S}{2}$	0	$+\frac{S}{2}$			$-S$	$-\frac{S}{2}$	0
	$-S$	0	0	$\frac{\mu^2}{4}$		$-S$	0	0	$\frac{\mu^2}{4}$		$-S$	0	0	$\frac{\mu^2}{4}$
	$-\frac{S}{2}$	0	$\frac{\mu(1-\mu)}{2}$	$\frac{\mu(1-\mu)}{2}$		$-\frac{S}{2}$	0	$\frac{\mu(1-\mu)}{2}$	$\frac{\mu(1-\mu)}{2}$		$-\frac{S}{2}$	0	$\frac{\mu(1-\mu)}{2}$	$\frac{\mu(1-\mu)}{2}$
ΔP_{t+1}	0	$\frac{\mu^2}{4}$	$(1-\mu)^2$	$\frac{\mu^2}{4}$	ΔP_{t+1}	0	$\frac{\mu^2}{4}$	$(1-\mu)^2$	$\frac{\mu^2}{4}$	ΔP_{t+1}	0	$\frac{\mu^2}{4}$	$(1-\mu)^2$	$\frac{\mu^2}{4}$
	$+\frac{S}{2}$	$\frac{\mu(1-\mu)}{2}$	$\frac{\mu(1-\mu)}{2}$	0		$+\frac{S}{2}$	$\frac{\mu(1-\mu)}{2}$	$\frac{\mu(1-\mu)}{2}$	0		$+\frac{S}{2}$	$\frac{\mu(1-\mu)}{2}$	$\frac{\mu(1-\mu)}{2}$	0
	$+S$	$\frac{\mu^2}{4}$	0	0		$+S$	$\frac{\mu^2}{4}$	0	0		$+S$	$\frac{\mu^2}{4}$	0	0

bid and ask price are equally likely, then on any given date the bid and ask each have a probability of $\mu/2$ and the midpoint has a probability of $1-\mu$.

Let S be a constant spread. Let P_t be the reported closing price on date t. Table 1 shows the probability distribution of the subsequent price changes $\Delta P_t \equiv P_t - P_{t-1}$ and $\Delta P_{t+1} \equiv P_{t+1} - P_t$ conditional on the date $t-1$ price.

From any of the three possible starting points (bid, midpoint, or ask) for P_{t-1}, there are always three possible immediate price changes ΔP_t and five possible subsequent price changes ΔP_{t+1}. Given the independence of each date's price moves, all of the probabilities are identical in the three boxes in Table 1. The only difference in the three boxes is the immediate price changes ΔP_t.

Given that the initial probabilities for P_{t-1} being a bid, midpoint, or ask are $\mu/2$, $1-\mu$, $\mu/2$, respectively, then Table 2 reports the combined joint distribution of ΔP_t and ΔP_{t+1}.

From Table 2, it is easy to compute the serial covariance of the joint distribution as

$$Cov(\Delta P_t, \Delta P_{t+1}) = -\frac{\mu S^2}{4}. \tag{20}$$

Solving for S yields

$$S = 2\sqrt{\frac{-Cov(\Delta P_t, \Delta P_{t+1})}{\mu}}. \tag{21}$$

This is identical to Roll's formula, except for the probability of a trading day μ in the denominator. In Roll's original framework, $\mu = 1$. Substituting this into the equation above yields Roll's formula.

Mirroring Roll's approach, the extended formula above was derived in the simple case of no innovations in the fundamental value of the security. In Roll's appendix, he drops that assumption and allows innovations in the fundamental value. He maintains the assumptions that: (1) markets are efficient and (2) innovations in the fundamental value of the security are independent of buy/sell realizations. The former assumption implies that the covariance between the fundamental value innovation on one date and the fundamental value innovation on another date must be zero. The later assumption implies that the covariance between the fundamental value innovation on one date and the

Table 2
Combined joint distribution of successive price changes.

				ΔP_t		
		$-S$	$-\frac{S}{2}$	0	$+\frac{S}{2}$	$+S$
	$-S$	0	0	$\frac{\mu^3}{8}$	$\frac{\mu^2(1-\mu)}{4}$	$\frac{\mu^3}{8}$
	$-\frac{S}{2}$	0	$\frac{\mu^2(1-\mu)}{4}$	$\frac{\mu(\mu-1)(\mu-2)}{4}$	$\frac{\mu(\mu-1)(\mu-2)}{4}$	$\frac{\mu^2(1-\mu)}{4}$
ΔP_{t+1}	0	$\frac{\mu^3}{8}$	$\frac{\mu(\mu-1)(\mu-2)}{4}$	$\frac{\mu^3}{4}+(1-\mu)^3$	$\frac{\mu(\mu-1)(\mu-2)}{4}$	$\frac{\mu^3}{8}$
	$+\frac{S}{2}$	$\frac{\mu^2(1-\mu)}{4}$	$\frac{\mu(\mu-1)(\mu-2)}{4}$	$\frac{\mu(\mu-1)(\mu-2)}{4}$	$\frac{\mu^2(1-\mu)}{4}$	0
	$+S$	$\frac{\mu^3}{8}$	$\frac{\mu^2(1-\mu)}{4}$	$\frac{\mu^3}{8}$	0	0

buy/sell realization on another date must be zero. So, the only non-zero source of serial covariance is the bid/ask bounce. Thus, the basic formula that Roll derives in the simple case, also applies in the more general case when innovations in the fundamental value are permitted.

The identical logic applies in this extended framework. The extended formula shown in Eq. (6), which was derived in the simple case, also applies in the more general case when innovations in the fundamental value are permitted.

Another issue which is important for the extended Roll model is splits and dividends. Consider a stock with a closing price of \$60 on one day, does a 2-for-1 split to open at \$30 the next day, and then closes at \$29. The raw, close-to-close price change is −\$31. If the raw price change is included in the serial covariance computation, it will greatly distort the result. In this case, −\$30 of the price change was due to the 2-for-1 split and only the remaining price change of −\$1 was due to new information and/or bid/ask/midpoint bounce. Thus, it is appropriate to remove the portion of the price change due to splits or dividends and use the remaining price change to compute the serial covariance.

Let ar_t be the adjusted return on date t, which accounts for splits and dividends. For the U.S., the CRSP stock database includes adjusted returns.[13] Let AP_t be the adjusted price on date t, which accounts for splits and dividends. Given the adjusted price, the adjusted return is simply $ar_t = AP_t/AP_{t-1} - 1$. Define the adjusted price change ΔP_t^* as

$$\Delta P_t^* = ar_t \cdot P_{t-1}, \tag{22}$$

where P_{t-1} is the *unadjusted* price on date $t-1$. Note that the adjusted price change ΔP_t^* is *not* the same thing as the change in the adjusted price $AP_t - AP_{t-1}$. To see why the adjusted price change ΔP_t^* is better to use, suppose that a firm splits its stock from \$60 to \$30 on March 10. It is likely that the stock's effective spread on March 10–31 is smaller than its effective spread on March 1–9 and that its effective spread for all of March is a weighted average of before and after. The purpose of computing the adjusted price change ΔP_t^* is to remove the \$30 drop on the split date, but also to allow the smaller price changes generated by its smaller effective spread from March 10 to March 31 to influence the extended Roll estimate for all of March. Removing the price change due to splits and dividends is potentially helpful not only to the extended Roll model, but also for the Hasbrouck Gibbs sampler measure described in Hasbrouck (2004).

Another important issue is price changes generated by systematic versus idiosyncratic forces. After controlling for splits and dividends, the adjusted returns of an individual stock can be decomposed into three parts: (1) systematic value innovations in the individual stock generated by systematic innovations in the market or in broad macro factors, (2) idiosyncratic value innovations in the individual stock generated by firm-specific innovations, and (3) bid/ask/midpoint bounce. From a "signal extraction" perspective, the bid/ask/midpoint bounce is the signal to be extracted and the systematic value innovations and idiosyncratic value innovations are both noise terms. In principle, it should be useful to remove the systematic value innovations, because the residuals that are left will have a higher signal-to-noise ratio.

[13] For the rest of the world, Datastream, Bloomberg, and other international data vendors provide daily *adjusted prices* for all stocks on exchanges around the world.

The empirical procedure is to perform a "market model" regression[14]

$$ar_t - r_f = \alpha + \beta(r_{mt} - r_f) + z_t, \tag{23}$$

where r_f is the daily riskfree rate, α and β are the regression coefficients, r_{mt} is the value-weighted market return on date t, and z_t is the regression residual. Then use the residual to compute the idiosyncratic adjusted price change ΔP_t^{**} as

$$\Delta P_t^{**} = z_t \cdot P_{t-1}. \tag{24}$$

Both versions of the extended Roll models use this idiosyncratic adjusted price change ΔP_t^{**}. The first version, called *Extended Roll*1, uses zero when the serial covariance is positive as follows:

$$Extended\ Roll1 = \begin{cases} \dfrac{2\sqrt{\dfrac{-Cov(\Delta P_t^{**}, \Delta P_{t+1}^{**})}{\hat{\mu}}}}{\bar{P}} & \text{when } Cov(\Delta P_t^{**}, \Delta P_{t+1}^{**}) < 0, \\ 0 & \text{when } Cov(\Delta P_t^{**}, \Delta P_{t+1}^{**}) > 0. \end{cases} \tag{25}$$

The second version, called *Extended Roll*2, substitutes *Effective Tick* when the serial covariance is positive. That is, set

$$Extended\ Roll2 = \begin{cases} \dfrac{2\sqrt{\dfrac{-Cov(\Delta P_t^{**}, \Delta P_{t+1}^{**})}{\hat{\mu}}}}{\bar{P}} & \text{when } Cov(\Delta P_t^{**}, \Delta P_{t+1}^{**}) < 0 \quad \text{and} \\ EffectiveTick & \text{when } Cov(\Delta P_t^{**}, \Delta P_{t+1}^{**}) > 0. \end{cases} \tag{26}$$

3.3. A pure no-trade quoted spread model

Eckbo and Norli (2002) use the low-frequency high/ask and low/bid variables to create a market-wide liquidity measure. Specifically, for each month they identify all stocks that meet the following three conditions: (1) 10 or more trading days in the month, (2) one or more no-trading days in the month, and (3) a stock price between \$1 and \$1,000. Then, for each identified stock they compute the average value of the percent quoted spread over all no-trading days. Finally, they exclude the two most extreme observations at both ends of the cross-section and aggregate the remaining stocks into a market-wide liquidity measure by computing an equally weighted average.[15]

Building on the spirit of Eckbo and Norli, I create a pure no-trade quoted spread proxy for *individual* stock-months. The main downside to using the no-trade bid and ask prices is that they are only available a small portion of the time. In my sample, only 26% of the 62,100 stock-months contain one or more no-trade days. The other 74% of the stock-months have trades on every day of the month and thus, do not provide no-trade bid and ask prices.

[14] As an alternative methodology, the market model regression could be replaced by a factor model regression. That is, individual stock adjusted returns could be regressed on any asset pricing factors (e.g., the Fama-French, 1993 three factors). Again, the purpose of removing value innovations due to systematic factors is to achieve greater precision in parsing the bid/ask/midpoint bounce from the idiosyncratic value innovations.

[15] Recently, Corwin and Schultz (2009) develop another measure of liquidity using high/ask and bid/low.

Recall that *NTD* is the number of no-trade days in a given estimation time interval. Define the No-Trade Quoted Spread as

$$No\ Trade\ Quoted\ Spread = \begin{cases} \left(\frac{1}{NTD}\sum_{t=1}^{NTD} QS_t\right)\Big/\bar{P} & \text{When } NTD>0, \\ 0 & \text{When } NTD = 0. \end{cases} \tag{27}$$

3.4. Combinations

So far in this section, I have developed three pure single-attribute models based on: (1) price clustering, (2) serial covariance accounting for no-trade midpoints, and (3) no-trade quoted spread.[16] Now, I am ready to define combined models that are linear combinations of the simpler models. I call them multi-factor models. Intuitively, we can think of single-attribute models as capturing the truth plus different error terms. The potential advantage of a multi-factor model is to diversify away some of the imperfectly correlated error terms. This potential advantage is formally demonstrated below.

Let Y be the true value of the spread benchmark. Assume that Y is a random variable given by $Y = \bar{Y} + \varepsilon_Y$, where $\bar{Y}$ is a constant mean and ε_Y is normally distributed error term $\varepsilon_Y \sim N(0, \sigma_Y^2)$. Let X and Z be two spread proxies based upon two different attributes. We can model these attributes by assuming that each spread proxy represents the truth plus its own noise. That is, $X = Y + \varepsilon_X$ and $Z = Y + \varepsilon_Z$, where $\varepsilon_X \sim N(0, \sigma_X^2)$ and $\varepsilon_Z \sim N(0, \sigma_Z^2)$. Both error terms are assumed to be independent of ε_Y, but are allowed to have a non-zero correlation with each other $\rho_\varepsilon = Corr(\varepsilon_X, \varepsilon_Z)$. Without loss of generality, the spread proxy X is assumed to be better than the spread proxy Z, in the sense of X being more precise than Z ($\sigma_X^2 < \sigma_Z^2$). It is straightforward to compute the correlation between the spread proxy X and the spread benchmark Y as $\rho_{XY} = \sigma_Y / \sqrt{\sigma_Y^2 + \sigma_X^2}$.

Define a multi-factor model L as a linear combination of the two spread proxies

$$L = w_x X + w_z Z, \tag{28}$$

where w_x and w_z are constants. Then the correlation between the multi-factor model L and the spread benchmark Y is $\rho_{LY} = \sigma_Y / \sqrt{\sigma_Y^2 + w_X^2\sigma_X^2 + w_Z^2\sigma_Z^2 + 2w_X w_Z \sigma_X \sigma_X \rho_\varepsilon}$.

Let G be the gain (improvement) in the correlation of the multi-factor model L over the better spread proxy X

$$G \equiv \rho_{LY} - \rho_{XY} = \frac{\sigma_Y}{\sqrt{\sigma_Y^2 + w_X^2\sigma_X^2 + w_Z^2\sigma_Z^2 + 2w_X w_Z \sigma_X \sigma_X \rho_\varepsilon}} - \frac{\sigma_Y}{\sqrt{\sigma_Y^2 + \sigma_X^2}}. \tag{29}$$

By inspection, the multi-factor gain is positive ($G>0$) when the multi-factor model L denominator is less than the better spread proxy X denominator, $w_X^2\sigma_X^2 + w_Z^2\sigma_Z^2 + 2w_X w_Z \sigma_X \sigma_X \rho_\varepsilon < \sigma_X^2$. From this comparison, it is clear that a higher error variance of the second best proxy tends to reduce the multi-factor gain ($\partial G/\partial\sigma_Z^2 < 0$), whereas a lower error correlation tends to increase the multi-factor gain ($\partial G/\partial\rho_\varepsilon < 0$). In other words, a multi-factor model correlation can be greater than the better individual proxy correlation, when the diversified away measurement error more than offsets the higher error variance

[16] Other spread proxies in the literature are based on other attributes. The LOT measures are based on a model of informed versus uninformed trading. Zero return measures are based on the relative lack of return innovations in stocks with a large percentage spread. And so on.

of the second best proxy. And the more different the two factors are ($\rho_\varepsilon \downarrow$), then the larger the gain from diversification is ($G \uparrow$).

Multi-factor models could take on an infinite number of possibilities. In the next section, two specific multi-factor models are empirically analyzed on the U.S. data. Both multi-factor models are simple 50–50% combinations of the simpler models:

$$\begin{aligned}&\textit{Multi-Factor}1 = (1/2)\cdot \textit{Effect Tick}4 + (1/2)\cdot \textit{Extended Roll}2,\\&\textit{Multi-Factor}2 = (1/2)\cdot \textit{Extended Roll}2 + (1/2)\cdot \textit{No Trade Quoted Spread}.\end{aligned}\tag{30}$$

4. An empirical comparison on three performance dimensions

I compute the spread benchmarks from the NYSE TAQ data from 1993 to 2005. Because of the enormous size of the TAQ data, a random sample is selected. Following the methodology of Hasbrouck (2009), a stock must meet five criteria to be eligible: (1) it has to be a common stock, (2) it has to be present on the first and last TAQ master file for the year, (3) it has to have the NYSE, AMEX, or NASDAQ as the primary listing exchange, (4) it does not change primary exchange, ticker symbol, or cusip over the year, and (5) has to be listed in CRSP. Four hundred stocks are randomly select each year from the universe of eligible stocks in 1993. Rolling forward, if any of the 1993 selections is not eligible in 1994, then a replacement is randomly drawn from the universe of eligible stocks in 1994. This process continues rolling forward over a 13-year span. Thirty stock-months are lost because there are an insufficient number of prices (2 or less) to compute all of the measures. An additional 270 stock-months have suspicious bid or ask prices that yield a spread wider than \$1.00 or are missing entirely from TAQ (despite being on CRSP). Thus, the final sample size is 62,100 stock-months.

4.1. *Full sample results*

Using the full sample, Table 3 compares the performance of the new low-frequency spread proxies developed in this paper and existing low-frequency spread proxies (Hasbrouck Gibbs, LOT Mixed, LOT Y-split, Pastor and Stambaugh, Roll, and Zeros) from the prior literature. Appendix C provides the formulas for the existing proxies. For the *Effective Tick* and *Holden* measures, a fractional grid accounting for price increments as small as $\$\frac{1}{64}$ was used from 1/93 to 1/01 for NYSE stocks and from 1/93 to 3/01 for NASDAQ stocks. A decimal grid was used thereafter. The benchmarks are the volume-weighted, percent effective spread and the time-weighted, percent quoted spread.

First consider Panel A, which is the joint time-series cross-sectional correlation of each low-frequency proxy with two benchmarks, percent effective spread and percent quoted spread, based on individual firms (62,100 stock-months). Checking the two versions of the integrated model, the highest joint correlations come from *Holden*2 (0.785 with percent effective spread and 0.865 with percent quoted spread). The highest joint correlation from the existing low-frequency spread proxies is *Hasbrouck Gibbs* (0.740 and 0.787 with the two benchmarks, respectively). Immediately below each correlation is a list of the other proxies that a given proxy is insignificantly different from. When a "*" appears below a correlation, then that proxy's correlation is statistically significantly different from all other correlations in the same row. *Holden*2 is significantly different from any other

Table 3
Low-frequency spread proxies compared to high-frequency percent effective and quoted spreads using the full sample.

The high-frequency benchmark percent effective spread is a volume-weighted average based on every trade and corresponding BBO quote in the NYSE TAQ database for a sample firm-month. The high-frequency benchmark percent quoted spread is a time-weighted average based on every BBO quote in the sample firm-month. All low-frequency spread proxies are calculated from CRSP daily stock data for a sample firm-month. The sample spans 1993–2005 inclusive and consists of 400 randomly selected stocks with annual replacement of stocks that do not survive, resulting in 62,100 firm-months. Bold numbers are statistically different from zero at the 1% level when $N = 62{,}100$ or at the 5% level when $N = 156$. * (**) means that the proxy's correlation / prediction error is statistically significantly different from all of the other correlations/prediction errors in the same row at the 1% (5%) level.

	New low-frequency spread proxies											Existing Low-Frequency Spread Proxies					
	Holden (H)	Holden 2 (H2)	Effective Tick (ET)	Effective Tick2 (ET2)	Effective Tick3 (ET3)	Effective Tick4 (ET4)	Extended Roll1 (ER1)	Extended Roll2 (ER2)	No Trade Quoted Spread (NTQS)	Multi-Factor1 (MF1)	Multi-Factor2 (MF2)	Pastor and Stam-baugh (P&S)	Roll (R)	Has-brouck Gibbs (HG)	LOT Mixed (LM)	LOT Y-split (LYS)	Zeros (Z)
Panel A. Joint Time-series cross-sectional correlations with high-frequency benchmarks based on individual firms ($N = 62{,}100$ stock-months)																	
% Effective spread	**0.704**	**0.785**	**0.656**	**0.642**	**0.739**	**0.777**	**0.669**	**0.706**	**0.500**	**0.665**	**0.805**	**−0.147**	**0.627**	**0.740**	**0.621**	**0.697**	**0.453**
Insignificantly different from	**ER2**	*	*	*	**HG**	*	**MF1**	**H**	*	**ER1**	*	*	**LM**	**ET3**	**R**	*	*
% Quoted spread	**0.763**	**0.865**	**0.704**	**0.690**	**0.794**	**0.851**	**0.727**	**0.764**	**0.567**	**0.733**	**0.877**	**−0.117**	**0.675**	**0.787**	**0.677**	**0.750**	**0.539**
Insignificantly different from	**ER2**	*	*	*	*	*	*	**H**	*	*	*	*	**LM**	*	**R**	*	*
Panel B. Pure time-series correlations with high-frequency benchmarks based on an equally weighted portfolio (*$N = 156$ portfolio-months*)																	
% Effective spread	**0.951**	**0.953**	**0.941**	**0.939**	**0.946**	**0.959**	**0.955**	**0.954**	**0.921**	**0.962**	**0.971**	**−0.366**	**0.925**	**0.905**	**0.722**	**0.931**	**0.874**
Insignificantly different from	**H2** **ER1** **ER2**	**H** **ER1** **ER2**	**ET2** **ET3** **ER1** **ER2** **R**	**ET** **ER1** **ER2** **NTQS** **R** **LY**	**ET** **ER1** **ER2**	**ER1** **ER2** **MF1**	**ET** **ET2** **ET3** **ET4** **H** **H2** **ER2** **MF1**	**ET** **ET2** **ET3** **ET4** **H** **H2** **ER1** **MF2**	**ET2** **R** **HG** **LY**	**ET4** **ER1** **ER2**	**	**	**ET** **ET2** **NTQS** **LY**	**Z** **NTQS**	**	**ET2** **R** **NTQS**	**HG**

% Quoted spread	0.985	0.989	0.979	0.978	0.985	0.990	0.928	0.931	0.950	0.977	0.977	−0.359	0.876	0.861	0.758	0.977	0.955
Insignificantly different from	ET3	ET4	ET2 MF1 MF2 LY	ET MF1 MF2 LY	H	H2	ER2	ER1	Z	ET ET2 MF2 LY	ET ET2 MF1 LY	**	HG	R	**	ET ET2 MF1 MF2	NTQS
Panel C. Average root mean squared prediction error of high frequency benchmarks (*N* = 156 *months*)																	
%Effective spread	0.0286	0.0234	0.0311	0.0316	0.0289	0.0261	0.0317	0.0293	0.0434	0.0290	0.0226	4.8372	0.0322	0.0287	0.0606	0.0342	0.1610
Insignificantly different from	HG ET3 ER2 MF1	MF2	ET2 ER1 ER2 MF1 R	ER ER1 MF1 R	H ER2 MF1 HG	MF1	ET ET2 MF1 R	H ET ET3 MF1 HG	**	H ET ET2 ET3 ET4 ER1 ER2 R HG	H2	**	ET ET2 ER1 MF1	H ET3 ER2 MF1	**	**	**
%Quoted spread	0.0312	0.0240	0.0336	0.0346	0.0325	0.0252	0.0313	0.0284	0.0384	0.0234	0.0215	4.8370	0.0339	0.0329	0.0554	0.0330	0.1555
Insignificantly different from	ER1 NTQS	MF1	ER1 NTQS R HG LY	NTQS R	ER1 NTQS HG LY	MF1	H ET ET3 HG LY	**	H ET ET2 ET3 R HG LY	H2 ET4 MF2	MF1	**	ET ET2 NTQS HG LY	ET ET3 ER1 NTQS R LY	**	ET ET3 ER1 NTQS R HG	

proxies. Specifically, its correlation with percent effective spread is significantly higher than the corresponding correlation of any of the existing proxies and the same is true of its correlation with respect to percent quoted spread.

Next consider the single-attribute models. Their correlations are bit lower. *Effective Tick*3, *Extended Roll*2, and the *No-Trade Quoted Spread* have the highest correlation with both benchmarks in their attribute categories. The two-attribute model *Effective Tick*4 (which integrates price clustering and the no-trade quoted spread) has higher correlations with both benchmarks.

Finally, consider the combined models. *Multi-Factor*2 has a joint correlation of 0.805 with percent effective spread and 0.877 with percent quoted spread. These correlations are significantly higher than any of the other new or old proxies. *Multi-Factor*2 includes all three attributes, because it combines *Effective Tick*4 (which integrates price clustering and the no-trade quoted spread) and *Extended Roll*2 (based on serial covariance accounting for no-trade midpoints).

Next, consider Panel B. For each proxy and benchmark, I create an aggregate spread measure based on an equally weighted portfolio across all 400 firms. This panel reports the pure time-series correlation of each aggregate spread proxy with each aggregate benchmark over 156 monthly observations. All of the portfolio correlations in Panel B are much higher than their individual firm counterparts in Panel A. This indicates that aggregating across assets causes a substantial amount of the measurement error to be diversified away. This is consistent with Goyenko et al. (2009) and Hasbrouck (2009), who also find much higher portfolio correlations.

First consider the two integrated models. *Holden* and *Holden*2 have time-series correlations with percent effective spread of 0.951 and 0.953, respectively. These two correlations are statistically indistinguishable from each other. *Holden*2 has the highest time-series correlation with percent quoted spread at 0.989. The integrated measures have significantly higher time-series correlations with both benchmarks than any of the existing proxies.

Turning to the single-attribute models, *Effective Tick*3, *Extended Roll*1, and the *No-Trade Quoted Spread* have relatively high time-series correlations with both benchmarks. The two-attribute model *Effective Tick*4 has even higher time-series correlations with both benchmarks.

Finally, consider the combined models. *Multi-Factor*2 has a time-series correlation with percent effective spread of 0.971. *Multi-Factor*2 is significantly higher than any of the other new or old proxies. Regarding a time-series correlation with percent quoted spread, *Multi-Factor*1, *Multi-Factor*2, and LOT Y-split are tied at 0.977. But all three are edged out by *Holden*2 and *Effective Tick*4 at 0.989 and 0.990, respectively, which are significantly higher time-series correlations than any of the existing proxies.

Finally, Panel C examines the distance (tracking error) between each proxy and the benchmarks. I compute the cross-sectional root mean squared error (RMSE) between each proxy and the benchmarks across all stocks in a given month and then average the cross-sectional RMSE over all 156 months.

Checking the integrated models, the lowest RMSEs come from *Holden*2 (0.0234 and 0.0240 with the two benchmarks, respectively). The lowest RMSEs from the existing low-frequency spread proxies come from *Hasbrouck Gibbs* (0.287 and 0.0329). So, *Holden*2 has significantly lower RMSEs with both benchmarks than any of the existing proxies.

Turning to the one-attribute models, *Effective Tick*3, *Extended Roll*2, and the *No-Trade Quoted Spread* have the lowest RMSEs with both benchmarks in their attribute categories. The two-attribute model *Effective Tick*4 has lower RMSEs with both benchmarks.

Finally, consider the combined models. *Multi-Factor*2 has the lowest RMSEs (0.0226 and 0.0204). Both RMSEs are statistically indistinguishable from Holden2, but significantly lower than rest of the new proxies or any of the old proxies.

To summarize the Table 3 full sample results, I find that on all three performance dimensions with regard to both benchmarks, the new integrated model *Holden*2 does significantly better than existing low-frequency spread proxies. I also find that on all three performance dimensions with regard to both benchmarks, the new combined model *Multi-Factor*2 does significantly better than existing low-frequency spread proxies, except for one tie. Summarizing six tests (three performance dimensions *X* two benchmarks), the combined model *Multi-Factor*2 does significantly better than the integrated model *Holden*2 on four out of six tests.

4.2. Size, price, and tick size regime results

In the next three tables, I examine the robustness of these results by size quintile, by price quintile, and by tick size regime. Table 4 breaks out the sample by size quintile. The joint time-series cross-sectional correlations for individual firms (Panel A) are generally better for small firms and worse for larger firms. Similarly, the time-series correlations for portfolios (Panel B) are generally better for small firms and worse for larger firms. By contrast, the average RMSEs (Panel C) are quite consistently better for larger firms. In Panels A and B, the *Roll* and *Hasbrouck Gibbs* correlations fall sharply for larger firms, consistent with the finding of Goyenko et al. (2009). In the same family of single attribute proxies, *Extended Roll*1 and *Extended Roll*2 do better, but are not great. In nearly all cases, *Holden*2 is the best integrated model and *Multi-Factor*2 is the best combined model. Across all comparisons (rows) in Table 4, *Multi-Factor*2 is the most frequent winner.

Table 5 breaks out the sample by price quintile. In Panels A and B, most of the proxies perform about the same across different price quintiles. The notable exception is that *Roll* and *Hasbrouck Gibbs* (and to a lesser extent *Extended Roll*1 and *Extended Roll*2), drop off sharply for high price firms. In Panel C, the average RMSEs are quite consistently better for high-priced firms. In the majority cases, *Holden*2 is the best integrated model and *Multi-Factor*2 is the best combined model. Across all comparisons (rows) in Table 5, *Multi-Factor*2 is the most frequent winner.

Table 6 breaks out the sample by tick size regime. The joint time-series cross-sectional correlations for individual firms (Panel A) worsen modestly in the decimal era. The time-series correlations for portfolios (Panel B) are the same or improve modestly in the decimal era. Average root mean squared errors (Panel C) improve modestly in the decimal era. In nearly all cases, *Holden*2 is the best integrated model and *Multi-Factor*2 is the best combined model. Across all comparisons (rows) in Table 6, *Multi-Factor*2 is the most frequent winner.

To summarize Tables 4–6, I find that the best-performing new proxies are robust by size quintiles, by price quintiles, and by tick size regime. Consistently, *Holden*2 is the best integrated model and *Multi-Factor*2 is the best combined model. Across all comparisons (rows) in Tables 4–6, *Multi-Factor*2 is the most frequent winner and *Holden*2 is the second most frequent winner.

Table 4
Low-frequency spread proxies compared to high-frequency percent effective and quoted spreads by size quintiles.

The high-frequency benchmark percent effective spread is a volume-weighted average based on every trade and corresponding BBO quote in the NYSE TAQ database for a sample firm-month. The high-frequency benchmark percent quoted spread is a time-weighted average based on every BBO quote in the sample firm-month. All low-frequency spread proxies are calculated from CRSP daily stock data for a sample firm-month. The sample spans 1993–2005 inclusive and consists of 400 randomly selected stocks with annual replacement of stocks that do not survive, resulting in 62,100 firm-months. Bold numbers are statistically different from zero at the 1% level when $N = 62{,}100$ or at the 5% level when $N = 156$.

	New low-frequency spread proxies											Existing low-frequency spread proxies					
	Holden	Holden 2	Effective Tick	Effective Tick2	Effective Tick3	Effective Tick4	Extended Roll1	Extended Roll2	No Trade Quoted Spread	Multi-Factor1	Multi-Factor2	Pastor and Stambaugh	Roll	Has-brouck Gibbs	LOT Mixed	LOT Y-split	Zeros
Panel A. Joint time-series cross-sectional correlations with high-frequency benchmarks based on individual firms ($N = 62{,}100$ *stock-months*)																	
% Effective spread																	
Size 1 (smallest)	**0.652**	**0.763**	**0.588**	**0.573**	**0.684**	**0.750**	**0.642**	**0.670**	**0.433**	**0.580**	**0.782**	**−0.179**	**0.609**	**0.792**	**0.627**	**0.675**	**0.349**
Size 2	**0.548**	**0.629**	**0.519**	**0.508**	**0.642**	**0.612**	**0.555**	**0.598**	**0.439**	**0.665**	**0.687**	**−0.172**	**0.525**	**0.658**	**0.460**	**0.577**	**0.339**
Size 3	**0.613**	**0.672**	**0.604**	**0.597**	**0.652**	**0.676**	**0.462**	**0.530**	**0.400**	**0.645**	**0.666**	**0.046**	**0.436**	**0.482**	**0.398**	**0.523**	**0.345**
Size 4	**0.626**	**0.631**	**0.608**	**0.607**	**0.648**	**0.621**	**0.272**	**0.345**	**0.172**	**0.485**	**0.505**	**0.049**	**0.271**	**0.277**	**0.302**	**0.477**	**0.328**
Size 5 (largest)	**0.620**	**0.699**	**0.618**	**0.601**	**0.661**	**0.703**	**0.240**	**0.310**	**0.418**	**0.477**	**0.479**	**−0.012**	**0.171**	**0.191**	**0.294**	**0.541**	**0.335**
% Quoted spread																	
Size 1 (smallest)	**0.714**	**0.844**	**0.635**	**0.621**	**0.737**	**0.825**	**0.703**	**0.734**	**0.488**	**0.643**	**0.858**	**−0.134**	**0.669**	**0.849**	**0.689**	**0.722**	**0.434**
Size 2	**0.620**	**0.771**	**0.579**	**0.565**	**0.724**	**0.739**	**0.644**	**0.685**	**0.632**	**0.815**	**0.807**	**−0.152**	**0.587**	**0.739**	**0.533**	**0.661**	**0.481**
Size 3	**0.709**	**0.820**	**0.682**	**0.666**	**0.748**	**0.792**	**0.536**	**0.605**	**0.596**	**0.802**	**0.769**	**−0.016**	**0.473**	**0.508**	**0.479**	**0.638**	**0.523**
Size 4	**0.710**	**0.735**	**0.708**	**0.704**	**0.749**	**0.714**	**0.281**	**0.355**	**0.366**	**0.567**	**0.547**	**0.058**	**0.256**	**0.222**	**0.332**	**0.571**	**0.504**
Size 5 (largest)	**0.698**	**0.822**	**0.730**	**0.708**	**0.786**	**0.845**	**0.198**	**0.273**	**0.562**	**0.490**	**0.489**	**0.006**	**0.113**	**0.134**	**0.342**	**0.633**	**0.495**

Panel B. Pure time-series correlations with high-frequency benchmarks based on an equally-weighted portfolio ($N = 156$ *portfolio-months*)

% Effective spread																	
Size 1 (smallest)	**0.939**	**0.940**	**0.930**	**0.926**	**0.937**	**0.945**	**0.959**	**0.957**	**0.862**	**0.939**	**0.961**	**−0.333**	**0.945**	**0.963**	**0.790**	**0.929**	**0.847**
Size 2	**0.914**	**0.924**	**0.901**	**0.901**	**0.926**	**0.932**	**0.935**	**0.937**	**0.900**	**0.936**	**0.945**	**−0.378**	**0.938**	**0.936**	**0.708**	**0.905**	**0.845**
Size 3	**0.931**	**0.926**	**0.914**	**0.913**	**0.918**	**0.929**	**0.899**	**0.920**	**0.816**	**0.928**	**0.948**	**−0.010**	**0.864**	**0.811**	**0.668**	**0.887**	**0.810**
Size 4	**0.882**	**0.882**	**0.854**	**0.855**	**0.855**	**0.875**	**0.480**	**0.640**	**0.651**	**0.847**	**0.857**	**0.035**	**0.336**	**0.365**	**0.487**	**0.809**	**0.727**
Size 5 (largest)	**0.856**	**0.856**	**0.818**	**0.817**	**0.802**	**0.825**	**0.528**	**0.651**	**0.410**	**0.848**	**0.861**	**−0.028**	**0.371**	**0.435**	**0.582**	**0.779**	**0.576**
% Quoted spread																	
Size 1 (smallest)	**0.974**	**0.980**	**0.967**	**0.965**	**0.972**	**0.982**	**0.963**	**0.964**	**0.894**	**0.958**	**0.985**	**−0.313**	**0.948**	**0.977**	**0.838**	**0.964**	**0.925**
Size 2	**0.966**	**0.976**	**0.949**	**0.950**	**0.977**	**0.975**	**0.949**	**0.954**	**0.954**	**0.981**	**0.975**	**−0.339**	**0.946**	**0.937**	**0.746**	**0.958**	**0.939**
Size 3	**0.970**	**0.984**	**0.969**	**0.965**	**0.981**	**0.986**	**0.864**	**0.883**	**0.935**	**0.973**	**0.958**	−0.040	**0.799**	**0.694**	**0.702**	**0.959**	**0.949**
Size 4	**0.982**	**0.983**	**0.984**	**0.984**	**0.983**	**0.984**	**0.300**	**0.486**	**0.826**	**0.840**	**0.818**	0.014	0.134	0.113	**0.541**	**0.966**	**0.944**
Size 5 (largest)	**0.965**	**0.979**	**0.974**	**0.972**	**0.979**	**0.985**	0.098	**0.257**	**0.609**	**0.582**	**0.603**	−0.039	−0.078	−0.008	**0.562**	**0.961**	**0.894**

Panel C. Average root mean squared prediction error of high frequency benchmarks ($N = 156$ *months*)

% Effective spread																	
Size 1 (smallest)	**0.0519**	**0.0443**	**0.0575**	**0.0585**	**0.0526**	**0.0463**	**0.0516**	**0.0474**	**0.0793**	**0.0527**	**0.0383**	**10.437**	**0.0520**	**0.0499**	**0.0980**	**0.0622**	**0.2151**
Size 2	**0.0278**	**0.0249**	**0.0293**	**0.0298**	**0.0278**	**0.0257**	**0.0304**	**0.0276**	**0.0377**	**0.0247**	**0.0222**	**0.9819**	**0.0305**	**0.0264**	**0.0631**	**0.0328**	**0.1950**
Size 3	**0.0166**	**0.0152**	**0.0175**	**0.0177**	**0.0176**	**0.0157**	**0.0229**	**0.0207**	**0.0255**	**0.0159**	**0.0147**	**0.1667**	**0.0235**	**0.0177**	**0.0445**	**0.0202**	**0.1503**
Size 4	**0.0092**	**0.0092**	**0.0099**	**0.0099**	**0.0099**	**0.0097**	**0.0199**	**0.0186**	**0.0157**	**0.0115**	**0.0111**	**0.0278**	**0.0211**	**0.0130**	**0.0356**	**0.0125**	**0.1090**
Size 5 (largest)	**0.0039**	**0.0037**	**0.0041**	**0.0041**	**0.0040**	**0.0038**	**0.0142**	**0.0137**	**0.0066**	**0.0069**	**0.0071**	**0.0266**	**0.0163**	**0.0107**	**0.0237**	**0.0061**	**0.0852**
% Quoted spread																	
Size 1 (smallest)	**0.0587**	**0.0432**	**0.0638**	**0.0658**	**0.0613**	**0.0453**	**0.0516**	**0.0466**	**0.0693**	**0.0420**	**0.0371**	**10.437**	**0.0565**	**0.0594**	**0.0875**	**0.0615**	**0.2046**
Size 2	**0.0300**	**0.0241**	**0.0319**	**0.0326**	**0.0308**	**0.0254**	**0.0296**	**0.0262**	**0.0331**	**0.0191**	**0.0215**	**0.8293**	**0.0313**	**0.0307**	**0.0679**	**0.0318**	**0.1937**
Size 3	**0.0177**	**0.0153**	**0.0188**	**0.0192**	**0.0192**	**0.0162**	**0.0228**	**0.0202**	**0.0239**	**0.0140**	**0.0149**	**0.1681**	**0.0242**	**0.0200**	**0.0410**	**0.0194**	**0.1461**
Size 4	**0.0089**	**0.0086**	**0.0094**	**0.0094**	**0.0096**	**0.0091**	**0.0192**	**0.0176**	**0.0146**	**0.0106**	**0.0108**	**0.0275**	**0.0206**	**0.0137**	**0.0337**	**0.0112**	**0.1070**
Size 5 (largest)	**0.0036**	**0.0032**	**0.0035**	**0.0036**	**0.0035**	**0.0032**	**0.0145**	**0.0139**	**0.0060**	**0.0071**	**0.0072**	**0.0260**	**0.0167**	**0.0113**	**0.0236**	**0.0055**	**0.0849**

Table 5
Low-frequency spread proxies compared to high-frequency percent effective and quoted spreads by price quintiles.

The high-frequency benchmark percent effective spread is a volume-weighted average based on every trade and corresponding BBO quote in the NYSE TAQ database for a sample firm-month. The high-frequency benchmark percent quoted spread is a time-weighted average based on every BBO quote in the sample firm-month. All low-frequency spread proxies are calculated from CRSP daily stock data for a sample firm-month. The sample spans 1993–2005 inclusive and consists of 400 randomly selected stocks with annual replacement of stocks that do not survive, resulting in 62,100 firm-months. Bold numbers are statistically different from zero at the 1% level when $N = 62{,}100$ or at the 5% level when $N = 156$.

	New low-frequency spread proxies											Existing low-frequency spread proxies					
	Holden	Holden 2	Effective Tick	Effective Tick2	Effective Tick3	Effective Tick4	Extended Roll1	Extended Roll2	No Trade Quoted Spread	Multi-Factor1	Multi-Factor2	Pastor and Stambaugh	Roll	Has-brouck Gibbs	LOT Mixed	LOT Y-split	Zeros
Panel A. Joint time-series cross-sectional correlations with high-frequency benchmarks based on individual firms ($N = 62{,}100$ *stock-months*)																	
% Effective spread																	
Price 1 (lowest)	**0.635**	**0.744**	**0.583**	**0.562**	**0.686**	**0.730**	**0.626**	**0.656**	**0.427**	**0.573**	**0.769**	**−0.157**	**0.588**	**0.766**	**0.601**	**0.661**	**0.413**
Price 2	**0.637**	**0.748**	**0.589**	**0.580**	**0.668**	**0.742**	**0.598**	**0.637**	**0.645**	**0.738**	**0.750**	**−0.208**	**0.534**	**0.728**	**0.479**	**0.624**	**0.359**
Price 3	**0.587**	**0.716**	**0.545**	**0.508**	**0.584**	**0.711**	**0.531**	**0.573**	**0.627**	**0.717**	**0.708**	**−0.176**	**0.415**	**0.568**	**0.376**	**0.554**	**0.409**
Price 4	**0.650**	**0.780**	**0.637**	**0.606**	**0.639**	**0.781**	**0.533**	**0.565**	**0.725**	**0.766**	**0.737**	**−0.178**	**0.390**	**0.459**	**0.398**	**0.617**	**0.519**
Price 5 (Highest)	**0.606**	**0.835**	**0.606**	**0.569**	**0.630**	**0.835**	**0.465**	**0.501**	**0.751**	**0.742**	**0.709**	**−0.209**	**0.294**	**0.283**	**0.405**	**0.670**	**0.448**
% Quoted spread																	
Price 1 (lowest)	**0.703**	**0.824**	**0.634**	**0.612**	**0.746**	**0.802**	**0.688**	**0.721**	**0.485**	**0.640**	**0.846**	**−0.123**	**0.649**	**0.828**	**0.667**	**0.709**	**0.492**
Price 2	**0.752**	**0.875**	**0.706**	**0.696**	**0.780**	**0.863**	**0.689**	**0.731**	**0.768**	**0.866**	**0.867**	**−0.155**	**0.607**	**0.798**	**0.555**	**0.713**	**0.477**
Price 3	**0.691**	**0.880**	**0.659**	**0.621**	**0.713**	**0.874**	**0.592**	**0.634**	**0.782**	**0.851**	**0.824**	**−0.141**	**0.452**	**0.613**	**0.453**	**0.679**	**0.554**
Price 4	**0.718**	**0.884**	**0.719**	**0.683**	**0.731**	**0.888**	**0.561**	**0.596**	**0.841**	**0.852**	**0.807**	**−0.156**	**0.405**	**0.458**	**0.438**	**0.680**	**0.611**
Price 5 (Highest)	**0.646**	**0.902**	**0.663**	**0.619**	**0.692**	**0.909**	**0.465**	**0.499**	**0.858**	**0.797**	**0.738**	**−0.164**	**0.282**	**0.258**	**0.414**	**0.703**	**0.531**

Panel B. Pure time-series correlations with high-frequency benchmarks based on an equally weighted portfolio ($N = 156$ *portfolio-months*)																	
% Effective spread																	
Price 1 (lowest)	**0.927**	**0.930**	**0.923**	**0.919**	**0.932**	**0.940**	**0.951**	**0.948**	**0.849**	**0.930**	**0.957**	**−0.247**	**0.926**	**0.954**	**0.767**	**0.909**	**0.841**
Price 2	**0.921**	**0.927**	**0.903**	**0.900**	**0.914**	**0.936**	**0.933**	**0.937**	**0.902**	**0.941**	**0.947**	**−0.292**	**0.935**	**0.935**	**0.702**	**0.896**	**0.814**
Price 3	**0.934**	**0.935**	**0.909**	**0.901**	**0.901**	**0.935**	**0.894**	**0.923**	**0.910**	**0.951**	**0.956**	**−0.306**	**0.822**	**0.850**	**0.621**	**0.885**	**0.830**
Price 4	**0.970**	**0.972**	**0.965**	**0.965**	**0.960**	**0.969**	**0.788**	**0.848**	**0.912**	**0.964**	**0.957**	**−0.326**	**0.600**	**0.614**	**0.654**	**0.953**	**0.925**
Price 5 (highest)	**0.898**	**0.892**	**0.877**	**0.877**	**0.854**	**0.872**	**0.486**	**0.588**	**0.611**	**0.842**	**0.827**	−0.103	**0.344**	**0.363**	**0.591**	**0.831**	**0.660**
% Quoted spread																	
Price 1 (lowest)	**0.968**	**0.976**	**0.962**	**0.959**	**0.972**	**0.979**	**0.953**	**0.954**	**0.877**	**0.955**	**0.981**	**−0.236**	**0.926**	**0.964**	**0.813**	**0.955**	**0.922**
Price 2	**0.981**	**0.986**	**0.967**	**0.968**	**0.980**	**0.985**	**0.942**	**0.946**	**0.948**	**0.980**	**0.976**	**−0.267**	**0.925**	**0.916**	**0.760**	**0.968**	**0.930**
Price 3	**0.987**	**0.987**	**0.977**	**0.977**	**0.980**	**0.986**	**0.851**	**0.888**	**0.957**	**0.972**	**0.965**	**−0.248**	**0.765**	**0.750**	**0.689**	**0.967**	**0.936**
Price 4	**0.974**	**0.986**	**0.975**	**0.976**	**0.982**	**0.987**	**0.719**	**0.783**	**0.955**	**0.952**	**0.933**	**−0.318**	**0.505**	**0.500**	**0.651**	**0.972**	**0.966**
Price 5 (highest)	**0.936**	**0.979**	**0.950**	**0.945**	**0.949**	**0.985**	0.148	**0.268**	**0.846**	**0.651**	**0.614**	−0.109	−0.013	0.012	**0.534**	**0.933**	**0.902**
Panel C. Average root mean squared prediction error of high frequency benchmarks ($N = 156$ months)																	
% Effective spread																	
Price 1 (lowest)	**0.0499**	**0.0453**	**0.0552**	**0.0559**	**0.0489**	**0.0474**	**0.0525**	**0.0480**	**0.0817**	**0.0534**	**0.0387**	**8.8370**	**0.0527**	**0.0489**	**0.1039**	**0.0623**	**0.1950**
Price 2	**0.0282**	**0.0240**	**0.0301**	**0.0307**	**0.0300**	**0.0245**	**0.0307**	**0.0280**	**0.0359**	**0.0237**	**0.0220**	**2.7498**	**0.0314**	**0.0274**	**0.0571**	**0.0315**	**0.1721**
Price 3	**0.0182**	**0.0151**	**0.0195**	**0.0202**	**0.0199**	**0.0155**	**0.0224**	**0.0208**	**0.0230**	**0.0162**	**0.0149**	**0.4588**	**0.0230**	**0.0176**	**0.0418**	**0.0208**	**0.1734**
Price 4	**0.0124**	**0.0096**	**0.0132**	**0.0137**	**0.0136**	**0.0100**	**0.0175**	**0.0165**	**0.0149**	**0.0110**	**0.0103**	**0.3141**	**0.0184**	**0.0131**	**0.0340**	**0.0142**	**0.1417**
Price 5 (highest)	**0.0084**	**0.0059**	**0.0088**	**0.0091**	**0.0090**	**0.0061**	**0.0148**	**0.0142**	**0.0096**	**0.0084**	**0.0080**	**0.1096**	**0.0173**	**0.0134**	**0.0237**	**0.0087**	**0.1012**
% Quoted spread																	
Price 1 (Lowest)	**0.0534**	**0.0426**	**0.0583**	**0.0600**	**0.0542**	**0.0450**	**0.0517**	**0.0463**	**0.0727**	**0.0422**	**0.0362**	**8.8362**	**0.0553**	**0.0558**	**0.0946**	**0.0598**	**0.1871**
Price 2	**0.0308**	**0.0235**	**0.0328**	**0.0338**	**0.0333**	**0.0244**	**0.0301**	**0.0270**	**0.0317**	**0.0214**	**0.0191**	**2.7507**	**0.0328**	**0.0313**	**0.0517**	**0.0308**	**0.1665**
Price 3	**0.0222**	**0.0153**	**0.0235**	**0.0245**	**0.0243**	**0.0160**	**0.0223**	**0.0204**	**0.0193**	**0.0151**	**0.0133**	**0.4602**	**0.0248**	**0.0219**	**0.0377**	**0.0207**	**0.1673**
Price 4	**0.0171**	**0.0117**	**0.0179**	**0.0185**	**0.0184**	**0.0122**	**0.0187**	**0.0175**	**0.0133**	**0.0122**	**0.0106**	**0.3158**	**0.0209**	**0.0178**	**0.0320**	**0.0160**	**0.1373**
Price 5 (highest)	**0.0115**	**0.0074**	**0.0119**	**0.0122**	**0.0121**	**0.0076**	**0.0156**	**0.0149**	**0.0086**	**0.0093**	**0.0084**	**0.1105**	**0.0188**	**0.0160**	**0.0231**	**0.0103**	**0.0989**

4.3. Low-frequency benchmark results

From 1993 to the present, the CRSP stock database includes closing bid and ask prices for all stocks. Table 7 compares the new and existing low-frequency spread proxies to the low-frequency percent effective spread and to the low-frequency percent quoted spread. The low-frequency benchmark percent effective spread is the daily average of the percent effective spread based on the last trade price of the day and the closing bid and ask prices for the sample firm-month. The low-frequency benchmark percent quoted spread is the daily average of the percent quoted spread based on the closing bid and ask prices for the sample firm-month.

In Panel A, the joint time-series cross-sectional correlations for individual firms are generally much higher with regard to the low-frequency percent quoted spread than with regard to low-frequency percent effective spread. In both cases, *Holden*2 is the overall winner and *Multi-Factor*2 is the best combined model. In Panel B, the time-series correlations for portfolios tend to be modestly higher with regard to the low-frequency percent quoted spread than with regard to low-frequency percent effective spread, but all of the correlations are in the high 0.80's or 0.90's for all of the new proxies. Surprisingly, the winner with regard to low-frequency percent quoted spread is *No Trade Quoted Spread* and the winner with regard to low-frequency percent effective spread is *Effective Tick*4.

Panel C tells an interesting story. The winner with regard to low-frequency percent quoted spread is *Multi-factor*1 and the new proxies' average RMSEs with respect to the low-frequency percent quoted spread are in the range [0.0255,0.0421]. This range is the same order of magnitude as the new proxies' average RMSEs with respect to the *high-frequency* percent quoted spread in Table 3, Panel C. By contrast, the winner with regard to low-frequency percent effective spread is *Zeros* and the new proxies' average RMSEs are in the range [0.5357,0.5462], which is 10–20 times greater! This raises the question of whether the new proxies have suddenly fallen flat with regard to low-frequency percent effective spread or whether something is different about low-frequency percent effective spread?

One problem with low-frequency percent effective spread is that last trade price is not necessarily synchronous with the closing quotes. The closing bid–ask quotes are measured at 4:00 p.m., but the last trade may have taken place minutes or even hours earlier when the prevailing bid–ask quotes may have been very different. The mismatch of 4:00 p.m. quotes and an earlier trade price adds noise to measurement of low-frequency percent effective spread. To check on this possibility, Table 8 reports the descriptive statistics for all four spread benchmarks. The means and medians of the four spread benchmarks are relatively close. The standard deviations of high-frequency percent quoted spread, high-frequency percent effective spread, and low-frequency percent quoted spread are relatively close, spanning the range [0.0402, 0.0540]. However, the standard deviation of low-frequency percent effective spread is 0.6885, more than 12 times larger! Looking at the correlations, high-frequency percent quoted spread, high-frequency percent effective spread, and low-frequency percent quoted spread have high correlations with each other (in the 0.80's). However, the correlations of low-frequency percent effective spread with the other three spread benchmarks are much lower (in the 0.40's and 0.50's). The very high standard deviation and low correlations suggest that low-frequency percent effective spread is a very noisy benchmark. The much larger average RMSEs in Table 7, Panel C and the lower correlations in Table 7, Panels A and B relative to low-frequency percent

Table 6
Low-frequency spread proxies compared to high-frequency percent effective and quoted spreads by tick size regime.

The high-frequency benchmark percent effective spread is a volume-weighted average based on every trade and corresponding BBO quote in the NYSE TAQ database for a sample firm-month. The high-frequency benchmark percent quoted spread is a time-weighted average based on every BBO quote in the sample firm-month. All low-frequency spread proxies are calculated from CRSP daily stock data for a sample firm-month. The sample spans 1993–2005 inclusive and consists of 400 randomly selected stocks with annual replacement of stocks that do not survive, resulting in 62,100 firm-months. Bold numbers are statistically different from zero at the 1% level when $N = 62,100$ or at the 5% level when $N = 156$.

	New low-frequency spread proxies											Existing low frequency spread proxies					
	Holden	Holden 2	Effective Tick	Effective Tick2	Effective Tick3	Effective Tick4	Extended Roll1	Extended Roll2	No Trade Quoted Spread	Multi-Factor1	Multi-Factor2	Pastor and Stambaugh	Roll	Has brouck Gibbs	LOT Mixed	LOT Y-split	Zeros
Panel A. Joint time-series corss-sectional correalations with high-frequency benchmarks based on indidual firms ($N = 62,100$ stock-months)																	
% Effective spread																	
$1/8 tick size	**0.711**	**0.826**	**0.656**	**0.635**	**0.754**	**0.817**	**0.778**	**0.806**	**0.765**	**0.839**	**0.859**	**−0.176**	**0.714**	**0.851**	**0.648**	**0.726**	**0.415**
$1/16 tick size	**0.711**	**0.726**	**0.668**	**0.669**	**0.742**	**0.714**	**0.571**	**0.629**	**0.537**	**0.734**	**0.762**	**−0.075**	**0.519**	**0.634**	**0.518**	**0.680**	**0.430**
$.01 tick size	**0.610**	**0.682**	**0.557**	**0.541**	**0.642**	**0.677**	**0.482**	**0.520**	**0.227**	**0.345**	**0.648**	**−0.113**	**0.519**	**0.549**	**0.613**	**0.561**	**0.410**
% Quoted spread																	
1/8 tick size	**0.750**	**0.886**	**0.682**	**0.670**	**0.793**	**0.874**	**0.840**	**0.869**	**0.856**	**0.913**	**0.923**	**−0.134**	**0.774**	**0.902**	**0.693**	**0.757**	**0.448**
$1/16 tick size	**0.744**	**0.798**	**0.701**	**0.677**	**0.756**	**0.785**	**0.635**	**0.692**	**0.624**	**0.814**	**0.839**	**−0.059**	**0.565**	**0.685**	**0.553**	**0.700**	**0.498**
$.01 tick size	**0.683**	**0.806**	**0.622**	**0.589**	**0.693**	**0.790**	**0.547**	**0.585**	**0.277**	**0.401**	**0.740**	**−0.103**	**0.577**	**0.607**	**0.661**	**0.623**	**0.455**
Panel B.Pure time-series correlations with high-frequency benchmarks based on an equally-weighted portfolio (54, 43, and 57 portfolio-months, respectively)																	
% Effective Spread																	
$1/8 tick size	**0.949**	**0.964**	**0.941**	**0.949**	**0.955**	**0.967**	**0.926**	**0.941**	**0.914**	**0.952**	**0.968**	**−0.250**	**0.927**	**0.954**	**0.338**	**0.932**	**0.835**
$1/16 tick size	**0.867**	**0.784**	**0.866**	**0.864**	**0.832**	**0.773**	**0.821**	**0.845**	**0.430**	**0.909**	**0.934**	−0.118	**0.801**	**0.752**	**0.270**	**0.780**	0.024
$.01 tick size	**0.929**	**0.927**	**0.926**	**0.930**	**0.939**	**0.927**	**0.937**	**0.943**	**0.740**	**0.904**	**0.945**	**−0.227**	**0.930**	**0.910**	**0.914**	**0.910**	**0.769**

Table 6 (*continued*)

	New low-frequency spread proxies											Existing low frequency spread proxies					
	Holden	Holden 2	Effective Tick	Effective Tick2	Effective Tick3	Effective Tick4	Extended Roll1	Extended Roll2	No Trade Quoted Spread	Multi-Factor1	Multi-Factor2	Pastor and Stambaugh	Roll	Hasbrouck Gibbs	LOT Mixed	LOT Y-split	Zeros
% Quoted spread																	
$1/8 tick size	**0.963**	**0.979**	**0.940**	**0.951**	**0.972**	**0.977**	**0.932**	**0.944**	**0.948**	**0.944**	**0.974**	**−0.246**	**0.936**	**0.946**	**0.370**	**0.933**	**0.886**
$1/16 tick size	**0.843**	**0.929**	**0.835**	**0.836**	**0.896**	**0.924**	**0.510**	**0.535**	**0.745**	**0.843**	**0.803**	**−0.288**	**0.453**	**0.403**	**0.313**	**0.827**	**0.428**
$.01 tick size	**0.985**	**0.985**	**0.982**	**0.977**	**0.983**	**0.987**	**0.963**	**0.969**	**0.740**	**0.918**	**0.984**	**−0.187**	**0.944**	**0.944**	**0.978**	**0.977**	**0.877**
Panel C. Average root mean squared prediction error of high frequency benchmarks (54, 43, and 57 months, respectively)																	
% Effective spread																	
$1/8 tick size	**0.0364**	**0.0287**	**0.0405**	**0.0413**	**0.0356**	**0.0298**	**0.0331**	**0.0303**	**0.0435**	**0.0314**	**0.0249**	**9.0838**	**0.0359**	**0.0359**	**0.0893**	**0.0468**	**0.2419**
$1/16 tick size	**0.0243**	**0.0250**	**0.0261**	**0.0263**	**0.0247**	**0.0263**	**0.0336**	**0.0309**	**0.0513**	**0.0294**	**0.0223**	**1.8450**	**0.0338**	**0.0265**	**0.0636**	**0.0301**	**0.1679**
$.01 tick size	**0.0246**	**0.0221**	**0.0261**	**0.0265**	**0.0256**	**0.0224**	**0.0288**	**0.0272**	**0.0370**	**0.0265**	**0.0207**	**3.2074**	**0.0272**	**0.0237**	**0.0316**	**0.0257**	**0.0800**
% Quoted spread																	
$1/8 tick size	**0.0467**	**0.0329**	**0.0507**	**0.0519**	**0.0477**	**0.0345**	**0.0362**	**0.0322**	**0.0399**	**0.0253**	**0.0289**	**9.0838**	**0.0434**	**0.0498**	**0.0779**	**0.0474**	**0.2296**
$1/16 tick size	**0.0266**	**0.0235**	**0.0282**	**0.0292**	**0.0281**	**0.0249**	**0.0319**	**0.0286**	**0.0481**	**0.0248**	**0.0198**	**1.8447**	**0.0334**	**0.0286**	**0.0597**	**0.0306**	**0.1629**
$.01 tick size	**0.0203**	**0.0160**	**0.0218**	**0.0225**	**0.0215**	**0.0166**	**0.0263**	**0.0248**	**0.0294**	**0.0204**	**0.0159**	**3.2070**	**0.0252**	**0.0204**	**0.0312**	**0.0215**	**0.0807**

Table 7
Low-frequency spread proxies compared to low-frequency percent effective spread and to low-frequency percent quoted spread.

The low-frequency benchmark percent effective spread is the daily average of the percent effective spread based on the last trade price of the day and the closing bid and ask prices from the CRSP daily stock database for the sample firm-month. The low-frequency benchmark percent quoted spread is the daily average of the percent quoted spread based on the closing bid and ask prices from the CRSP daily stock database for the sample firm-month. All low-frequency spread proxies are calculated from CRSP daily stock data for a sample firm-month. The sample spans 1993–2005 inclusive and consists of 400 randomly selected stocks with annual replacement of stocks that don't survive, resulting in 62,100 firm-months. Bold numbers are statistically different from zero at the 1% level when $N = 62{,}100$ or at the 5% level when $N = 156$.

	New low frequency spread proxies											Existing low-frequency spread proxies					
	Holden	Holden 2	Effective Tick	Effective Tick2	Effective Tick3	Effective Tick4	Extended Roll1	Extended Roll2	No Trade Quoted Spread	Multi-Factor1	Multi-Factor2	Pastor and Stambaugh	Roll	Has-brouck Gibbs	LOT Mixed	LOT Y-split	Zeros
Panel A. Joint time-series cross-sectional correlations with low-frequency benchmarks based on individual firms ($N = 62{,}100$ stock-months)																	
Low-freq. % eff. spread	**0.318**	**0.542**	**0.297**	**0.261**	**0.357**	**0.536**	**0.363**	**0.372**	**0.468**	**0.428**	**0.491**	**−0.084**	**0.235**	**0.371**	**0.296**	**0.390**	**0.519**
Low-freq. % quo. spread	**0.717**	**0.876**	**0.669**	**0.658**	**0.763**	**0.863**	**0.648**	**0.689**	**0.636**	**0.744**	**0.841**	**−0.103**	**0.604**	**0.714**	**0.624**	**0.703**	**0.505**
Panel B. Pure time-series correlations with low-frequency benchmarks based on an equallyweighted portfolio ($N = 156$ portfolio-months)																	
Low-freq. % eff. spread	**0.907**	**0.928**	**0.904**	**0.898**	**0.899**	**0.932**	**0.870**	**0.874**	**0.935**	**0.910**	**0.919**	**−0.296**	**0.829**	**0.812**	**0.701**	**0.884**	**0.883**
Low-freq. % quo. spread	**0.984**	**0.988**	**0.983**	**0.982**	**0.976**	**0.989**	**0.938**	**0.949**	**0.945**	**0.982**	**0.985**	**−0.323**	**0.898**	**0.895**	**0.763**	**0.966**	**0.951**
Panel C. Average root mean squared prediction error of low-frequency benchmarks ($N = 156$ months)																	
Low-freq. % eff. spread	**0.5441**	**0.5385**	**0.5448**	**0.5462**	**0.5462**	**0.5383**	**0.5362**	**0.5358**	**0.5372**	**0.5357**	**0.5368**	**3.4012**	**0.5392**	**0.5421**	**0.5349**	**0.5432**	**0.4998**
Low-freq. % quo. spread	**0.0392**	**0.0284**	**0.0411**	**0.0421**	**0.0401**	**0.0293**	**0.0389**	**0.0360**	**0.0349**	**0.0255**	**0.0285**	**4.8374**	**0.0415**	**0.0416**	**0.0576**	**0.0390**	**0.1540**

Table 8
Benchmark descriptive statistics.

High-frequency percent effective spread is a volume-weighted average based on every trade and corresponding BBO quote in the NYSE TAQ database for a sample firm-month. High-frequency percent quoted spread is a time-weighted average based on every BBO quote in the sample firm-month. Low-frequency percent effective spread is the daily average of the percent effective spread based on the last trade price of the day and the closing bid and ask prices from the CRSP daily stock database for the sample firm-month. Low-frequency percent quoted spread is the daily average of the percent quoted spread based on the closing bid and ask prices from the CRSP daily stock database for the sample firm-month. The sample spans 1993–2005 inclusive and consists of 400 randomly selected stocks with annual replacement of stocks that do not survive, resulting in 62,100 firm-months.

	High-frequency effective spread	High-frequency quoted spread	Low-frequency effective spread	Low-frequency quoted spread
Mean	0.0285	0.0331	0.3079	0.0366
Median	0.0160	0.0177	0.0162	0.0204
Standard deviation	0.0402	0.0463	0.6885	0.0540

Correlations	High-frequency effective spread	High-frequency quoted spread	Low-frequency effective spread	Low-frequency quoted spread
High-frequency effective spread	1.000			
High-frequency quoted spread	0.880	1.000		
Low-frequency effective spread	0.460	0.536	1.000	
Low-frequency quoted spread	0.805	0.893	0.537	1.000

effective spread are consistent with this interpretation. This suggests that we should not place much weight on the low-frequency percent effective spread results and should put much more weight on the low-frequency percent quoted spread results.

To summarize Table 7 focusing on the low-frequency percent quoted spread results, I find that new proxies consistently outperform existing proxies. In Panel A, the best new proxy has a joint correlation of 0.876 versus the best existing proxy with 0.714. In Panel B, the best new proxy has a time-series correlation of 0.989 versus the best existing proxy with 0.966. In Panel C, the best new proxy has an RMSE of 0.0255 versus the best existing proxy with 0.0390.

5. Conclusion

I develop new spread proxies that are computed from low-frequency (daily) data. First, I develop an integrated model, the Holden model, which directly includes three attributes of the daily data: price clustering, serial covariance accounting for no-trade midpoints, and the no-trade quoted spread. Second, I develop combined models, the Multi-Factor models, that are linear combinations of simpler one-attribute or two-attribute models. I show theoretically that Multi-Factor models have the potential to diversify away some imperfectly-correlated error terms. Next, I empirically test both new and existing low-frequency spread measures on three performance dimensions: (1) higher individual firm correlation with effective or quoted spread, (2) higher portfolio correlation with effective or quoted spread, and (3) lower distance (tracking error) relative to effective or quoted spread.

I find that on all three performance dimensions with regard to both high-frequency benchmarks, the new integrated model *Holden*2 does significantly better than existing low-frequency spread proxies. I find that on all three performance dimensions with regard to both benchmarks, the new combined model *Multi-Factor*2 does significantly better than existing low-frequency spread proxies, except for one tie. Summarizing six tests (three performance dimensions X two benchmarks), the combined model *Multi-Factor*2 does significantly better than the integrated model *Holden*2 on four out of six tests.

I also find that these new proxies are robust by size quintiles, price quintiles, and tick size regime. Consistently, *Holden*2 is the best integrated model and *Multi-Factor*2 is the best combined model. Across all size, price, and tick size regime comparisons, *Multi-Factor*2 is the most frequent winner and *Holden*2 is the second most frequent winner. Finally, I compare the proxies to low-frequency spread benchmarks and find that new proxies consistently outperform existing proxies.

Acknowledgments

I thank Avanidhar Subrahmanyam (the editor), an anonymous referee, Diego Agudelo, Ruslan Goyenko, Sam Henkel, Stacy Jacobsen, Chuck Trzcinka, and seminar participants at Indiana University. I am solely responsible for any errors.

Appendix A. Modifications to accommodate any decimal or fractional price grid

Holden model: Let A_j and A_{J+j} be the *total* number of trade prices and no-trade midpoints, respectively, corresponding to the jth spread ($j = 1, 2, \ldots, J$). Let D_{jk} be the number of special price increments for the jth spread ($j = 1, 2, \ldots, J$) that *overlap* the price increments of the kth spread.

The probabilities of the trade price clusters are

$$Pr(C_t = j) = \sum_{k=1}^{j} \gamma_k \mu \frac{D_{jk}}{A_k}, \quad j = 1, 2, \ldots, J. \tag{31}$$

Similarly, the probabilities of the no-trade midpoint clusters are

$$Pr(C_t = J + j) = \sum_{k=1}^{j} \gamma_k (1 - \mu) \frac{D_{J+j,k}}{A_{J+k}}, \quad j = 1, 2, \ldots, J. \tag{32}$$

The conditional probability of a half-spread given a particular price cluster is

$$Pr(H_t = h_k | C_t = j) = \frac{\gamma_{|k|} \left(\frac{\mu}{2}\right) \frac{D_{j|k|}}{A_{|k|}}}{Pr(C_t = j)}, \quad k \neq 0, \ \ k \leq j \text{ and } j = 1, 2, \ldots, J \tag{33}$$

and

$$Pr(H_t = h_0 | C_t =)j = 0, \quad j = 1, 2, \ldots, J. \tag{34}$$

Similarly, the conditional probability of a half-spread given a particular midpoint cluster is

$$Pr(H_t = h_0 | C_t = J + j) = 1, \quad j = 1, 2, \ldots, J \tag{35}$$

and

$$Pr(H_t = h_k | C_t = J + j) = 0, \quad k \neq 0,\ k \leq j \text{ and } j = 1, 2, \ldots, J. \tag{36}$$

Effective Tick: Let B_j and B_{J+j} be the number of special prices and special midpoints, respectively, corresponding to the *j*th spread ($j = 1, 2, \ldots, J$). The unconstrained probability of the *j*th spread is

$$U_j = \begin{cases} \left(\dfrac{A_1}{B_1}\right) F_1 + \left(\dfrac{A_{J+1}}{B_{J+1}}\right) F_{J+1}, & j = 1, \\ \left(\dfrac{A_j}{B_j}\right) F_j - \sum\limits_{k=1}^{j-1} \left(\dfrac{O_{jk}}{B_k}\right) F_k + \left(\dfrac{A_{J+j}}{B_{J+j}}\right) F_{J+j} - \sum\limits_{k=1}^{j-1} \left(\dfrac{O_{J+j,k}}{B_{J+k}}\right) F_{J+k}, & j = 2, 3, \ldots, J. \end{cases} \tag{37}$$

Detailed decimal and fractional examples are available at: www.kelley.iu.edu/cholden/examples.pdf.

Appendix B. Price cluster probabilities and half-spread conditional probabilities

The probability of a trade price cluster is obtained by summing the probabilities of all paths in Fig. 2 tree that lead to a particular cluster. The probabilities of the trade price clusters are

$$Pr(C_t = j) = \begin{cases} \sum\limits_{k=1}^{j} \gamma_k \mu \left(\frac{1}{2}\right)^{j-k+1}, & j = 1, 2, \ldots, J-1, \\ \sum\limits_{k=1}^{J} \gamma_k \mu \left(\frac{1}{2}\right)^{j-k}, & j = J. \end{cases} \tag{38}$$

Similarly, the probabilities of the no-trade midpoint clusters are

$$Pr(C_t = J + j) = \gamma_j (1 - \mu), \quad j = 1, 2, \ldots, J. \tag{39}$$

Let $h_j\, j = -J, -J+1, \ldots, +J$ be the feasible half-spread values.[17] The conditional probability of a half-spread given a particular price cluster is obtained by starting at a particular price cluster and determining the probability of each feasible half-spread from that cluster. The conditional probability of a half-spread given a particular price cluster is

$$Pr(H_t = h_k | C_t = j) = \begin{cases} \dfrac{\gamma_{|k|} \left(\dfrac{\mu}{2}\right) \left(\dfrac{1}{2}\right)^{j-k+1}}{Pr(C_t = j)} & \text{when } k \neq 0 \text{ and } j = 1, 2, \ldots, J-1, \\ \dfrac{\gamma_{|k|} \left(\dfrac{\mu}{2}\right) \left(\dfrac{1}{2}\right)^{j-k}}{Pr(C_t = j)} & \text{when } k \neq 0 \text{ and } j = J, \end{cases} \tag{40}$$

[17] More specifically, when $Q_t = +1$, then the possible half-spreads $h_1, h_2, \ldots, h_J$ are given by $h_j = \frac{1}{2} s_j$. When $Q_t = 0$, then the only possible half-spread is $h_0 = 0$. When $Q_t = -1$, then the possible half-spreads $h_{-1}, h_{-2}, \ldots, h_{-J}$ are given by $h_{-j} = \frac{-1}{2} s_j$.

and

$$Pr(H_t = h_0 | C_t = j) = 0, \quad j = 1, 2, \ldots, J. \tag{41}$$

Similarly, the conditional probability of a half-spread given a particular midpoint cluster is

$$Pr(H_t = h_0 | C_t = J + j) = 1, \quad j = 1, 2, \ldots, J \tag{42}$$

and

$$Pr(H_t = h_k | C_t = J + j) = 0 \quad k \neq 0 \text{ and } j = 1, 2, \ldots, J. \tag{43}$$

Appendix C. Existing low-frequency spread proxies

All dollar spread proxies below are converted to percent spread proxies by dividing by the average price $\bar{P}$ (see Table C1).

Table C1

Reference	Proxy
Hasbrouck (2004)	Hasbrouck Gibbs = $2c$, where the half-spread c, the variance of the public information shock σ_ε^2, the latent buy/sell/no-trade indicators $Q = \{Q_1, Q_2, \ldots, Q_T\}$, and the latent “efficient prices” $V = \{V_1, V_2, \ldots, V_T\}$ are estimated numerically using a Gibbs sampler.
Lesmond et al. (1999)	$LOT\ Mixed = \alpha_2 - \alpha_1$, where $\alpha_2(\alpha_1)$ is trans cost to buy (sell). $\underset{\alpha_1,\alpha_2,\beta,\sigma}{Max} \left\{ \begin{array}{l} \prod_1 \frac{1}{\sigma} n \left[\frac{R_t + \alpha_1 - \beta R_{mt}}{\sigma} \right] \\ \times \prod_0 \left[N\left(\frac{\alpha_2 - \beta R_{mt}}{\sigma} \right) - N\left(\frac{\alpha_1 - \beta R_{mt}}{\sigma} \right) \right] \\ \times \prod_2 \frac{1}{\sigma} n \left[\frac{R_t + \alpha_2 - \beta R_{mt}}{\sigma} \right], \end{array} \right.$ where R_t (R_{mt}) is the own return (market return), σ is the return volatility, and β is the stock's market sensitivity, $S.T. \quad \alpha_1 \leq 0, \alpha_2 \geq 0, \beta \geq 0, \sigma \geq 0$. LOT Mixed is capped at a max value of 1.5. Region 0 is $R_{jt} = 0$, region 1 is $R_{jt} \neq 0$ and $R_{mt} > 0$, and region 2 is $R_{jt} \neq 0$ and $R_{mt} < 0$.
Goyenko et al. (2008)	$LOT\ Y\text{-}split = \alpha_2 - \alpha_1$ and everything is the same as $LOT\ Mixed$, except region 0 is $R_{jt} = 0$, region 1 is $R_{jt} > 0$, and region 2 is $R_{jt} < 0$ and no upper bound cap is imposed.
Pastor and Stambaugh (2003)	*Pastor and Stambaugh* = Γ, from the regression: $r_{t+1}^e = \theta + \phi r_t + \Gamma\ \text{sign}(r_t^e)(Volume_t) + \varepsilon_t$, where r_t^e is the stock's excess return above the CRSP VWMR on day t, θ is the intercept, ϕ and Γ are regression coefficients, and ε_t is the error term.
Roll (1984)	$Roll = \begin{cases} 2\sqrt{-Cov(\Delta P_t, \Delta P_{t-1})}/\bar{P} & \text{When } Cov(\Delta P_t, \Delta P_{t-1}) < 0, \\ 0 & \text{When } Cov(\Delta P_t, \Delta P_{t-1}) \geq 0. \end{cases}$
Lesmond et al. (1999)	$Zeros = \frac{ZRD}{TD + NTD}$, where ZRD is the number of zero returns days, TD the number of trading days, and NTD the number of no-trade days in a given stock-month.

References

Acharya, V., Pedersen, L., 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.

Bekaert, G., Harvey, C., Lundblad, C., 2007. Liquidity and expected returns: lessons from emerging markets. Review of Financial Studies 20, 1783–1831.

Cao, C., Field, L., Hanka, G., 2004. Does insider trading impair market liquidity? Evidence from IPO lockup expirations. Journal of Financial and Quantitative Analysis 39, 25–46.

Chan, K., Chung, P., Johnson, H., 1995. The intraday behavior of bid-ask spreads for NYSE stocks and CBOE options. Journal of Quantitative and Financial Analysis 30, 329–346.

Chan, L.K.C., Jegadeesh, N., Lakonishok, J., 1996. Momentum strategies. Journal of Finance 51, 1681–1713.

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

Christie, W., Schultz, P., 1994. Why do NASDAQ market makers avoid odd-eighth quotes? Journal of Finance 49, 1813–1840.

Chung, K., Van Ness, B., Van Ness, R., 2003. Limit orders and the bid-ask spread. Journal of Financial Economics 53, 255–287.

Chung, K., Zhao, X., 2003. Intraday variation in the bid-ask spread: evidence after the market reform. Journal of Financial Research 26, 191–206.

Corwin, S., Schultz, P., 2009. A simple way to estimate bid–ask spreads from daily high and low prices. University of Notre Dame, Working Paper.

De Bondt, W., Thaler, R., 1985. Does the stock market overreact? Journal of Finance 40, 793–805.

Dennis, P., Strickland, D., 2003. The effect of stock splits on liquidity and excess returns: evidence from shareholder ownership composition. Journal of Financial Research 26, 355–370.

Eckbo, E., Norli O., 2002. Pervasive liquidity risk. Dartmouth College, Working Paper.

Fama, E., French, K., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–56.

Foucault, T., Kadan, O., Kandel, E., 2005. Limit order book as a market for liquidity. Review of Financial Studies 18, 1171–1217.

Fujimoto, A., 2004. Macroeconomic sources of systematic liquidity. Yale University, Working Paper.

George, T., Kaul, G., Nimalendran, M., 1991. Estimation of the bid–ask spreads and its components: a new approach. Review of Financial Studies 4, 623–656.

Glosten, L., Harris, L., 1988. Estimating the components of the bid–ask spread. Journal of Financial Economics 21, 123–142.

Goyenko, R., Holden, C., Trzcinka, C., 2009. Do liquidity measures measure liquidity? Journal of Financial Economics 92, 153–181.

Harris, L., 1991. Stock price clustering and discreteness. Review of Financial Studies 4, 389–415.

Harris, L., 2003. Trading and exchanges: market microstructure for practitioners. Oxford University Press, New York, New York.

Hasbrouck, J., 2004. Liquidity in the futures pits: inferring market dynamics from incomplete data. Journal of Financial and Quantitative Analysis 39, 305–326.

Hasbrouck, J., 2009. Trading costs and returns for US equities: estimating effective costs from daily data. Journal of Finance 46, 1445–1477.

Helfin, F., Shaw, K., 2000. Blockholder ownership and market liquidity. Journal of Financial and Quantitative Analysis 35, 621–633.

Huang, R., Stoll, H., 1997. The components of the bid–ask spread: a general approach. Review of Financial Studies 10, 995–1034.

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling lowers: implications for market efficiency. Journal of Finance 48, 65–92.

Jegadeesh, N., Titman, S., 2001. Profitability of momentum strategies: an evaluation of alternative explanations. Journal of Finance 56, 699–720.

Kalev, P., Pham, P., Steen, A., 2003. Underpricing, stock allocation, ownership structure and post-listing liquidity of newly listed firms. Journal of Banking and Finance 27, 919–947.

Lee, C., Muchlow, B., Ready, M., 2003. Spreads, depths, and the impact of earnings information. Review of Financial Studies 6, 345–374.

Lerner, J., Schoar, A., 2004. The illiquidity puzzle: theory and evidence from private equity. Journal of Financial Economics 72, 3–40.

Lesmond, D., 2005. Liquidity of emerging markets. Journal of Financial Economics 77, 411–452.

Lesmond, D., Ogden, J., Trzcinka, C., 1999. A new estimate of transaction costs. Review of Financial Studies 12, 1113–1141.

Lesmond, D., Senbet, L., O'Connor, P., 2005. Leverage recapitalizations and liquidity. Tulane University, Working Paper.

Lipson, M., Mortal, S., 2004a. Liquidity and firm characteristics: evidence from mergers and acquisitions. Journal of Financial Markets 10, 342–361.

Lipson, M., Mortal, S., 2004b. Capital structure decision and equity market liquidity. Working paper, University of Georgia.

Madhavan, A., Richardson, M., Roomans, M., 1997. Why do security prices change? A transaction-level analysis of NYSE Stocks. Review of Financial Studies 10, 1035–1064.

Pastor, L., Stambaugh, R., 2003. Liquidity risk and expected stock returns. Journal of Political Economy, 642–685.

Roll, R., 1984. A simple implicit measure of the effective bid–ask spread in an efficient market. Journal of Finance 39, 1127–1139.

Rouwenhorst, G., 1998. International momentum strategies. Journal of Finance 53, 267–284.

Sadka, R., 2003. Liquidity risk, and asset pricing. Northwestern University, Working Paper.

Schrand, C., Verrecchia, R., 2004. Disclosure choice and cost of capital: evidence from underpricing in initial public offerings. Working paper, University of Pennsylvania.

Stoll, H., 1989. Inferring the components of the bid–ask spread: theory and empirical tests. Journal of Finance 44, 115–134.