Available online at www.sciencedirect.com
ScienceDirect
Journal of Financial Markets 12 (2009) 173–202
Journal of FINANCIAL MARKETS
www.elsevier.com/locate/finmar

Using matched samples to test for differences in trade execution costs ☆

Ryan J. Davies[a,*], Sang Soo Kim[b]

[a]*Finance Division, Babson College, Tomasso Hall, Babson Park, MA 02457, USA*
[b]*Korea Development Bank, 16-3 Yeoido-dong, Yeongdeungpo-gu, Seoul 150-973, Republic of Korea*

Available online 10 June 2008

Abstract

This paper provides guidance on how to use matched samples to test for differences in trade execution costs (e.g., quoted and effective spreads). Based on extensive simulation results, we conclude that the best practice is to match firms one-to-one based on market capitalization and share price, and to test for differences between the matched pairs using a Wilcoxon signed rank test. We demonstrate that pre-sorting by industry groups or discarding apparent poor matches may reduce test power. We show that, in general, tests based on one-to-one nearest-neighbor matching have comparable power and less size distortion than alternatives that place more weight on distant firms. We find that matching without replacement can reduce size distortion when the control sample is relatively small. We highlight conditions under which matched sample estimation may be preferred to the corresponding event study.

JEL classification: G10

Keywords: Matched samples; Market microstructure; Bid-ask spreads

☆The authors are grateful for comments received from an anonymous referee, Vladimir Atanasov, Dan Bernhardt, Chris Brooks, Michael Goldstein, Apostolos Katsaris, Pankaj Jain, Charles Jones, Bruce Lehmann (the editor), Sa Lu, James MacKinnon, Usha Mittoo, Andreas Park, Stephen Sapp, Robert Van Ness, and seminar participants at Babson College, Queen's University, University of Reading, City University (London), the Canadian Economics Association meetings, the Finance and Econometrics Conference at the University of York, the Financial Management Association meetings, the Bachelier Finance Society meetings, the Multinational Finance Society meetings, and the Western Finance Association meetings. Ryan Davies acknowledges financial support from the Babson Faculty Research Fund.
*Corresponding author. Tel.: +1 781 239 5345; fax: +1 781 239 5004.
E-mail address: rdavies@babson.edu (R.J. Davies).

1386-4181/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.finmar.2008.06.001

1. Introduction

Matched sample estimation is widely used in market microstructure research.[1] Typically, this estimation approach takes two groups of stocks that differ in their listing status and matches them in pairs according to various firm characteristics (e.g., market capitalization). A statistical test of the difference in the bid-ask spreads (or another measure of trade execution costs) between the matched pairs is calculated and the test result is used to infer the impact of listing status on the spreads of otherwise similar stocks. Despite its widespread usage, there is little agreement in the existing literature about the appropriate methodological approach for using matched samples. To fill this troubling knowledge gap, this paper uses Monte Carlo simulation to provide the first comprehensive study of the size and power properties of tests for differences in trade execution costs based on matched sample estimation.

Table 1 provides an overview of some of the diverse combinations of firm characteristics used to construct matched samples in the existing microstructure literature. This paper identifies which of these combinations provide the highest testing power, with the least size distortion (probability of a type I error). We find that, in most situations, the best practice is to match firms one-to-one according to market capitalization and share price. We show that test power may be *reduced* when the sample is pre-sorted according to industry groups or when apparent poor matches are eliminated from the sample. We demonstrate that better test properties result from using a non-parametric Wilcoxon signed rank test than a parametric dependent Student *t*-test. We also show that matching without replacement reduces test size distortion when there are relatively few control firms available.

We compare the size and power properties of tests based on a one-to-one nearest neighbor matching estimate with those of tests based on matching estimates that place some weight on distant neighbors, such as a one-to-two nearest neighbor, a one-to-three nearest neighbor, or kernel-based matching estimates. These alternative estimators are potentially less sensitive to a mismatch along unmeasured dimensions; the cost is that they introduce an added mismatch along measured dimensions. While these alternatives are rarely used in finance, many papers contain an apology for using the simpler one-to-one matching approach. We show that no apology is necessary: one-to-one matching of firms typically performs the best in this context. Tests based on this approach have less bias and a lower probability of type I error (rejecting the null hypothesis when the null hypothesis is true) than tests based on the alternative estimators that we consider. For instance, we provide an example in which one-to-one matching has virtually no size distortion, while one-to-three matching has empirical size of 14.4% for a test with nominal size of 5%.

Finally, we provide evidence that matched sample estimation (which selects matches from all available firms) can produce more powerful statistical tests than the corresponding event study (which uses the smaller set of firms that change their status during the sample period) if the number of potential matches is much larger than the number of firms in the event study.

We are not aware of any previous research that examines the size and power of tests for differences in transaction costs based on matching estimates. The closest previous research has been in the context of detecting abnormal operating performance (Barber and Lyon,

[1]For example, Huang and Stoll (1996) compare execution costs on the NYSE and the Nasdaq by constructing matched samples of NYSE-listed firms and comparable Nasdaq-listed firms.

Table 1
Matching characteristics used in selected past studies

Study	Matching characteristics
Affeck-Graves, Hedge, and Miller (1994)	Market capitalization, share price, trade volume, volatility
Aitken, Cook, Harris, and McInish (2006)	Market capitalization, share price, total value of trades
Bacidore and Sofianos (2002)	Market capitalization, share price, volatility (intraday and overnight)
Battalio, Ellul, and Jennings (2007)	Market capitalization, share price, trade volume, volatility
Bessembinder (1999, 2003)	Market capitalization
Bessembinder and Kanfman (1997a, 1997b)	Market capitalization
Boehmer (2005)	Market capitalization, share price, adjusted daily dollar volume, daily relative price range
Chung, Van Ness, and Van Ness (2001)	Share price, number of trades, trade size, return volatility
Chung, Van Ness, and Van Ness (2004)	Market capitalization, share price, number of trades, trade size, return volatility
Huang and Stoll (1996)	Long-term debt level, book value, share price, number of shares outstanding, 2-digit SIC code
Jain and Kim (2006)	Market capitalization, share price, dollar trading volume
SEC (2001)	Market capitalization, share price, adjusted volume, volatility
Silva and Chavez (2002)	Market capitalization, trading volume, return volatility, tick size relative to share price
Venkataraman (2001)	(i) Market capitalization, share price; (ii) Trading volume, share price. With/without pre-sort by Datastream industry classification
Weston (2000)	(i) Market capitalization, share price, volatility; (ii) Market capitalization, share price, volatility, volume; (iii) Market capitalization, share price, volatility, trade size, 2-digit SIC code

1996; Kahle and Walkling, 1996; Lie, 2001) and in the context of detecting abnormal long-run stock returns (Barber and Lyon, 1997; Lyon, Barber, and Tsai (1999); Mitchell and Stafford, 2000). The methodology used in our paper is also similar in spirit to that used in Brown and Warner (1980, 1985). In general, all of these papers focus on very different issues from those considered in this paper.

It is well known from statistics that smoothing matching estimates by placing more weight on distant matches leads to tests with size distortions, but often with higher test power. The amount of the size distortion, however, depends crucially on the relation between the matching characteristic(s) and the property of interest. In our case, we are interested in the non-linear relation between spreads and various firm characteristics. The presence of large outliers and the non-equispaced distribution of firm characteristics means that it is impossible to predict the tradeoffs between test size and test power without direct simulation results. This paper provides those simulation results.

While our results are most applicable to measuring differences in trade execution costs for equity markets, our findings and simulation approach have implications for the use of matched samples in many other contexts. For instance, similar test power trade-offs are likely to occur when using matched samples to test for spread differences in equity

derivatives, such as in Mayhew (2002). In other contexts, when in doubt, researchers might consider conducting simulation experiments similar to those described here.

A main focus of past market microstructure research has been detecting differences between the spreads of NYSE- and Nasdaq-listed stocks. Going forward, this is less likely to be a focus as the link between exchange listing and the location of trading becomes weaker. In such an environment, expected trade execution costs may directly influence the decision to send a trade to a particular trading venue and, consequently, using matched sample estimation to compare costs on the trading venues may not always be feasible. Despite that, other important differences in listing status frequently occur and these differences are often best studied using matched samples. For instance, regulators routinely use a small trial group of stocks to investigate the impact of new rules (e.g., changes in tick size; removal of the up-tick short sale constraint; changes in transparency requirements). The impact of a new rule can be determined by comparing the trade execution costs of the trial group with a matched group of non-trial stocks. Other possible applications of matched sample estimation include studies of the impact on bid-ask spreads of: (i) changes in specialist firms; (ii) the introduction of new derivative products, such as single stock futures; and (iii) the introduction of new alternative trading venues.

It is widely assumed that because market microstructure research typically involves huge amounts of data that it cannot possibly suffer from a lack of power. This is not true. This paper demonstrates that the choice of matching approach could have had an impact on the results of important past studies. Furthermore, as overall bid-ask spreads fall and matched sample estimation is used to study more narrowly defined problems, differences in spreads may become harder to measure. The results in this paper provide researchers with guidance on how best to use matched sample estimation to detect these subtle differences.

The remainder of the paper is organized as follows. Section 2 outlines the matched sample estimation approach. Section 3 describes the data and the selection criteria used. Section 4 presents the Monte Carlo simulation results and robustness checks. Section 5 concludes.

2. Matched sample estimation

Heckman, Ichimura, and Todd (1997, 1998) describe the theory underlying matched sample estimation in the context of labor economics.[2] They use matched sample estimation to compare the mean post-program earnings of job-training program participants with the mean earnings of comparable non-participants. In this section, we adopt the notation of Heckman, Ichimura, and Todd to show how matched sample estimation can be used to compare the trade execution costs of firms that differ in their listing status. For example, we might be interested in comparing the quoted spreads of NYSE-listed firms with those of otherwise comparable Nasdaq-listed firms. We attribute differences in the trade execution costs of these two groups to the difference in listing status. Note that we will use listing status to refer to any difference in the trading environment; it need not refer to the formal exchange listing.

At any given time, a security can have either listing status A (e.g., the firm is NYSE-listed) or listing status 0 (e.g., the firm is Nasdaq-listed); it cannot have both listing statuses simultaneously. The trade execution costs that would occur in listing status A is Y_A and in

[2]Rubin (2006) provides an excellent summary of the matched sample estimation literature in statistics.

listing status 0 is Y_0. Let $D = 1$ if the firm has listing status A; $D = 0$ otherwise. The trade execution costs observed for a firm is $Y = DY_A + (1 - D)Y_0$. The difference attributed to having listing status A is denoted Δ, where $\Delta = Y_A - Y_0$.

Each firm has observed characteristics $\mathbf{X}$, which can be partitioned into two not necessarily mutually exclusive sets of variables, $(\mathbf{T}, \mathbf{Z})$, where the $\mathbf{T}$ variables determine the trade execution costs and the $\mathbf{Z}$ variables determine whether the firm decides to change its listing status.[3] In practice, a firm's characteristics often have an impact on both its listing decision and its trade execution costs. For example, a large industrial firm may be more likely to list on the NYSE and may be more likely to have smaller spreads.

The trade execution costs of a firm with listing status $\ell = \{A, 0\}$ can be written as a function of observables $(\mathbf{T})$ and unobservables U_ℓ, where

$$Y_\ell = g_\ell(\mathbf{T}) + U_\ell, \quad E(U_\ell) = 0$$

and g_ℓ is assumed to be a non-stochastic function. Unobservables include firm characteristics, such as its management style, that directly make an impact on its trading properties, but are difficult, or impossible, to quantify.

The mean effect of listing status A on the trade execution costs for a firm with characteristics $\mathbf{X} \in \mathbf{S}$, where $\mathbf{S}$ is a given set, is given by

$$E(\Delta|\mathbf{X}, D = 1) = g_A(\mathbf{X}) - g_0(\mathbf{X}) + E(U_A - U_0|\mathbf{X}, D = 1).$$

The average effect (across firms) of having listing status A is given by

$$M(\mathbf{S}) = \frac{\int_{\mathbf{S}} E(\Delta|\mathbf{X}, D = 1)\mathrm{d}F(\mathbf{X}|D = 1)}{\int_{\mathbf{S}} \mathrm{d}F(\mathbf{X}|D = 1)}$$

where F is the conditional distribution of $\mathbf{X}$; and $\mathbf{S}$ is a subset of the support of $\mathbf{X}$ given $D = 1$. In practice, the choice of $\mathbf{S}$ can be non-trivial if there does not exist a sufficient number of firms with characteristic $\mathbf{X}$ such that either $D = 1$ or $D = 0$. For example, it may be appropriate to exclude small firms from $\mathbf{S}$ that are unable to meet certain listing requirements or are unable to justify paying higher listing fees or both.

Let $\mathbf{I}_A$ denote the set of indices for firms with listing status A and let $\mathbf{I}_0$ denote the set of indices for firms without listing status A. To estimate the effect of having listing status A for each firm $i \in \mathbf{I}_A$, trade execution costs Y_{Ai} is compared to an average of the trade execution costs Y_{0j} for matched firms $j \in \mathbf{I}_0$ without listing status A. Matches are constructed on the basis of observed characteristics $\mathbf{X}$. A firm without listing status A receives a higher weight in constructing a match when its observed characteristics are closer to those of a firm with listing status A, $i \in \mathbf{I}_A$, using a specific distance measure. The estimated change in trade execution costs for each firm i with listing status A is

$$Y_{Ai} - \sum_{j \in \mathbf{I}_0} W(i,j)Y_{0j} \quad \text{such that} \quad \sum_{j \in \mathbf{I}_0} W(i,j) = 1 \quad \forall i \in \mathbf{I}_A,$$

[3]The overlap between the sets $\mathbf{T}$ and $\mathbf{Z}$ is particularly important in the context of comparing trade execution costs of NYSE-listed stocks that trade primarily on the NYSE with those that trade primarily on the Nasdaq (or a similar situation). In this situation, the trade location of all orders, or a subset of orders (difficult/easy), is likely to be influenced by the *expected* trade execution costs on each exchange. Heckman and Smith (1998) demonstrate that if one makes certain (strong) assumptions about the density of Δ across firms (for example, one possibility is to assume that Δ is the same for all firms), then it may still be possible to use matched sample estimation in such a context.

where $W(i,j)$ is a positive valued weight function. The weighting function assigns weights to the trade execution costs of each firm without listing status A based on distances in the space of observed characteristics, $\mathbf{X}$. Different matching estimates can be constructed by using different weighting functions or different distance measures or both.

Let N_A and N_0 denote the number of firms in $\mathbf{I}_A$ and $\mathbf{I}_0$, respectively. An estimate of the average effect of having listing status A is given by

$$\hat{M}(\mathbf{S}) = \frac{1}{N_A}\sum_{i\in\mathbf{I}_A}\left(Y_{Ai} - \sum_{j\in\mathbf{I}_0} W(i,j)\,Y_{0j}\right).$$

Let $\boldsymbol{\phi}_i$ indicate the set of firms matched to firm $i \in \mathbf{I}_A$. Let k-NN denote a **N**earest **N**eighbor matching estimator over the k closest firms. We focus on the following matching estimators:

1-NN matching estimator: For each $i \in \mathbf{I}_A$, we select as a match the firm $j \in \mathbf{I}_0$ which minimizes the norm distance $\|\mathbf{X}_i - \mathbf{X}_j\|$. We use $(x_i - x_j)/(x_i + x_j)$ as the distance measure for the univariate case. Section 4.4 describes the distance measure used for the multivariate case. $\boldsymbol{\phi}_i$ is a singleton set except for ties that are broken by a random draw. The weighting scheme assigns all the weight to the single match: $W(i,j)$ equals 1 if $j \in \boldsymbol{\phi}_i$; 0 otherwise.

2-NN matching estimator with uniform weights: Now, $\boldsymbol{\phi}_i$ is a set of the two closest firms to firm i according to the distance measure employed. The weighting scheme assigns equal weights to the two closest firms: $W(i,j)$ equals $\frac{1}{2}$ if $j \in \boldsymbol{\phi}_i$; 0 otherwise.

3-NN matching estimator with triangular weights: Now, $\boldsymbol{\phi}_i$ is a set of the three closest firms to firm i. We assign weights of 50% to the closest match, $33\frac{1}{3}$% to the second closest match, and $16\frac{2}{3}$% to the third closest match, as proposed in Stone (1977).

Kernel-based matching estimators: As k increases, k-NN matching estimators are, in effect, kernel-based matching estimators. Here, we focus on univariate kernel-based matching estimators based on characteristic $x \in \mathbf{X}$.[4] Now, matching weights can be assigned to any firm in the control set ($\boldsymbol{\phi}_i = \mathbf{I}_0$) as

$$W(i,j) = \frac{K_{ij}}{\sum_{k\in\mathbf{I}_0} K_{ik}},$$

where $K_{ik} = K((x_i - x_k)/h)$ is a kernel function and h is a bandwidth parameter. We focus on Nadaraya-Watson kernel-based matching estimators, which are based on the standard normal density function

$$K_{ik} = \frac{1}{h}\left[\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x_i - x_k}{h}\right)^2\right\}\right].$$

Intuitively, the bandwidth parameter controls the smoothing across firms: a larger value causes the matching estimate to place more weight on firms that are further away, in terms of the matching characteristic. We first consider as a bandwidth parameter $h_1 = 1.059 s_x N_0^{-1/5}$ where $s_x = \sqrt{N_0^{-1}\sum_{j\in\mathbf{I}_0} x_j^2 - (N_0^{-1}\sum_{j\in\mathbf{I}_0} x_j)^2}$. As explained in Pagan and Ullah (1999, p. 24), this bandwidth parameter choice minimizes the approximate

[4]This paper focuses on *univariate* kernel-based estimators. In theory, the results for multivariate kernel-based estimators should resemble those of k-NN matching estimators based on multiple matching characteristics. That said, the problem with multivariate estimation is that very large sample sizes are needed to get accurate estimates of a density.

mean-integrated squared error of the kernel estimate. The order of magnitude of the parameter, $h \propto n^{-1/5}$, equates the order of the bias and variance of the kernel estimate, while the constant multiplier of 1.059 is derived from an assumption of an underlying normal density.

We also consider a variable window-width kernel estimator where $K_{ik} = K((x_i - x_k)/h_i^{\min})$ and $h_i^{\min}$ is distance of firm i from its closest neighbor. A variable bandwidth may be better suited to account for heterogeneity in matching characteristic(s) across firms (e.g., when matching on market capitalization, a large bandwidth is needed to obtain a match for large firms, whereas a small bandwidth is sufficient to provide numerous close matches for small firms).

Our goal in considering kernel-based estimators is to illustrate the potential problems with smoothing estimates across multiple firms. We acknowledge that the literature on kernel-based estimation is vast, and that more sophisticated kernel estimators exist which will provide tests with better properties than those presented here. These choices, however, are beyond the scope of this paper.

3. Data

Our main simulation results are based on a sample of 1,000 NYSE-listed common stocks contained in the Russell 3000 Index in 2004. For each stock, we obtain from Datastream information on market capitalization, number of shares outstanding, CAPM beta, book value, price volatility, and the percentage of shares outstanding in the public float. From CRSP, we obtain the 2-digit Standard Industrial Classification (SIC) code for each stock. From Compustat, we obtain the level of long-term debt for each firm. These potential matching characteristics are as of December 31, 2003. Our sample excludes stocks with incomplete data and stocks that do not trade during each month in 2004. We also exclude non-U.S. based companies, stocks that changed their ticker symbol during 2004, and stocks with initial share prices above \$100. Table 2 provides a summary of the firm characteristics. Our final representative sample contains stocks with market capitalization ranging from \$649 million to \$311,755 million (mean = \$9,160 million), and beginning share prices ranging from \$2.83 to \$98.73 (mean = \$32.94).

We obtain trades and inside quote revisions from the TAQ database. For each stock, we calculate its trading volume, trading dollar volume, and the number of trades during December 2003. For each stock, we calculate a time-weighted percentage bid-ask (quoted) spread on each day, and then we average across trading days. The percentage bid-ask spread is defined as: $200 \times (ask - bid)/(ask + bid)$. Unless otherwise indicated, the average percentage bid-ask spread is our outcome parameter of interest. Table 2 also reports the trade-weighted percentage effective spread, defined as $200 \times |Trade\ Price - 0.5 \times (ask + bid)|/(ask + bid)$, which is used later in our robustness checks. In general, we find very similar results using effective and quoted spreads.

4. Monte Carlo simulation

Throughout this section, we use variations of the following Monte Carlo simulation approach to answer many of the outstanding questions concerning the optimal usage of matched sample estimation in market microstructure research. For each of 20,000

Table 2
Characteristics of initial sample of 1,000 NYSE-listed common stocks

Market capitalization, number of shares outstanding, CAPM beta, book value, price volatility, and the percentage of shares outstanding in the public float are obtained from Datastream as of December 31, 2003. Long-term debt is obtained from Compustat as of December 31, 2003. Trading volume, trading dollar volume, and the number of trades are measured using the TAQ database during December 2003. Using the 2004 TAQ database, for each stock we calculate a time-weighted quoted percentage bid-ask spread on each day, and then we average across trading days. The percentage bid-ask spread is defined as: $200 \times (ask - bid)/(ask + bid)$. The table also reports the trade-weighted effective percentage spread, defined as $200 \times |Trade\ Price - 0.5 \times (ask + bid)|/(ask + bid)$.

Characteristic	Min	Max	Mean	Median
Market capitalization ($ millions)	649	311,755	9,160	2,449
Number of shares outstanding (millions)	19	950,610	24,810	8,770
Number of shares in public float (millions)	6	950,610	20,014	5,809
Share price ($, December 31, 2003)	2.83	98.73	32.94	30.03
CAPM Beta	−0.15	3.15	0.77	0.67
Book value per share ($)	−43.2	1,556.5	16.5	12.5
Price to book ratio	−41.80	546.95	3.82	2.30
Long-term debt ($ millions)	0	486,876	4,356	560
Trades (December 2003)	1,132	272,601	27,930	18,095
Trade volume (December 2003, thousands)	387	756,480	254,240	9,703
Trade dollar volume (December 2003, $ millions)	3.5	14,805.1	8,105.3	2804.2
Percentage bid-ask spread (2004, bp)	2.89	89.28	10.49	8.92
Percentage effective bid-ask spread (2004, bp)	2.91	85.18	9.64	8.05

Monte Carlo replications:

1. We randomly shuffle and divide the sample of $N_T = 1,000$ firms into two groups: N_A stocks are assigned listing status A and $N_0 = N_T - N_A$ stocks are assigned listing status 0 (no distinction is made for their *true* listing status). The percentage bid-ask spread of each of the N_A stocks with listing status A is artificially changed by θs_y: $\hat{Y}_i = Y_i + \theta s_y$, where s_y is the standard deviation of bid-ask spreads in the sample.
2. Each of the N_A stocks, $i \in \mathbf{I}_A$, is matched with a hypothetical firm created using a weighting of the remaining $N_T - N_A$ firms, $j \in \mathbf{I}_0$. The weighting scheme is implied by the particular matching technique. In general, we allow firms to receive positive weights in multiple matches (for example, $W(a,j) > 0$ does not preclude $W(b,j) > 0$, where $a, b \in \mathbf{I}_A$ and $j \in \mathbf{I}_0$).[5]
3. Based on a comparison of the N_A stocks with induced differences and their hypothetical matched pairs, we then construct a two-sided statistical test of size α of the difference in bid-ask spreads between the two groups. We consider using both a parametric dependent Student t-test and a non-parametric Wilcoxon signed rank test.

The empirical power of the test for a given induced difference (θ) is the share of the replications with a significant test statistic. The empirical size of the test is the share of replications with a significant test statistic when there is no induced difference ($\theta = 0$). A well-specified test will have an empirical size equal to its theoretical (nominal) size of α.

[5]Some researchers use 1-NN matching techniques that require each matched stock to be used only once (matching *without* replacement). We explore this restriction in Section 4.7.

The ability to detect a given induced difference depends on the dispersion of bid-ask spreads in the sample. To account for this, in the results that follow, we typically present the induced difference in spreads in standard deviation terms, rather than basis points. This allows our results to be easily interpreted in other contexts (such as different exchanges and different time periods) in which differences in spreads across stocks might be larger or smaller. To give a sense of scale, it is useful to highlight that in our full sample s_y equals 7.51 bp, or put another way, it equals a spread difference of 2.25 cents for the median priced stock.

Before proceeding, we highlight why our simulations are based on firms listed on a single exchange, rather than based on a larger merged sample comprised of stocks from multiple exchanges. While a merged sample might seem appealing, it would lead to erroneous results. To see why, recall that the *listing decision* of firms is driven by many of the same characteristics we use to construct our matched samples. Since market structure contributes to differences in spreads, a simulation approach based on drawing from a sample of firms from different exchanges would place too much importance on firm characteristics that influence the listing decision rather than those that influence differences in bid-ask spreads. For example, if Nasdaq firms tend to have lower book values, and the market microstructure of Nasdaq leads to larger bid-ask spreads, a simulation based on a merged sample of NYSE and Nasdaq stocks would overstate the power of a firm's book value for predicting its bid-ask spread, *independent* of the firm's listing status. In terms of our previous notation, the purpose of our simulation exercise is to identify firm characteristics in the set **T** (characteristics that determine spreads). In the absence of a suitable propensity score adjustment, we want, as much as possible, these identified characteristics to be absent from the set **Z** (characteristics that determine exchange listing status).

We also note that it is not possible in this context to use a firm-year approach similar to that used in Barber and Lyon (1996). Our simulation requires that each randomly drawn sample has, in expectation, the same independence properties as the actual real-life sample. This would not be true in a firm-year approach. To see why, note that if General Electric matches with IBM in 2004, General Electric is also more likely to match with IBM in 2005. Thus, a firm-year approach would result in a simulation sample with almost identical matched pairs, thereby introducing a bias in the simulation results.[6] In lieu of using a merged sample or a firm-year approach, in Section 4.9 we demonstrate that our results are robust to using different exchanges and time periods.

There are many choices to be made when constructing matched samples. For clarity, we now proceed to consider each choice separately. Although one might conjecture that there could be interesting interactions between these choices, we tried various combinations of approaches (in unreported results) and did not find any evidence of noteworthy interactions. We begin by first considering the choice of test statistic and then considering the choice of matching estimator (one-to-one versus one-to-many). These results are presented based on matching firms according to their market capitalization. (Market capitalization is the most commonly used matching criterion, in part because it is readily

[6]The results in Barber and Lyon (1996) also suffer from this independence problem, but to a much lesser extent. For instance, when they match according to the 2-digit SIC code, 98.1% of their firm-years are matched with a set of five or more firms (Table 4, p. 371). It is likely that this set of firms changes on a yearly basis, thereby largely alleviating independence concerns.

available in databases such as CRSP, Datastream, and Compustat, and in part because it is highly correlated with other factors that influence the magnitude of spreads, such as liquidity and information asymmetry.) Later, we consider matching with various combinations of firm characteristics; these combinations may include or exclude market capitalization. After that, we consider: pre-sorting by industry; excluding apparent poor matches; and matching with and without replacement. We conclude with a comparison of a matched sample approach with an event study approach, and with various robustness checks.

4.1. Which statistical test should we use?

When using matched sample estimation, it is common for researchers to test the difference between the matched pairs using either a dependent (pairwise) Student t-test or a non-parametric Wilcoxon signed rank test. These tests are constructed as follows. If δ_i denotes the difference between the bid-ask spreads of a matched pair $(Y_{Ai} - Y_{0j})$ and s_δ denotes the standard deviation of the differences, then the dependent Student t-test statistic is $(\overline{\delta}\sqrt{N_A})/s_\delta$. The Wilcoxon test orders the absolute values of $\delta_1, \delta_2, \ldots, \delta_{N_A}$, and assigns a rank of R_i to each accordingly. Let $\phi_i = 1$ if $\delta_i > 0$; 0 otherwise. The sum of the rank of positive differences is then calculated: $W^+ = \sum_{i=1}^{N_A} \phi_i R_i$. For $N_A > 25$, the Wilcoxon test statistic is

$$\frac{W^+ - N_A(N_A+1)/4}{\sqrt{N_A(N_A+1)(2N_A+1)/24}} \sim_a N(0,1).$$

For smaller sample sizes, the distribution of W^+ under the null can be calculated exactly using a small simulation (in essence, a coin-flipping exercise) or obtained from a published table. Note that the Wilcoxon test only requires that the distribution of the differences be symmetrical (which always occurs under the null), whereas the dependent Student t-test requires that the distribution of the *differences* between the matched pairs, δ_i, be normally distributed.

In addition to these two test approaches based on matched samples, we also consider using a regression-based approach. In this approach, *all* stocks are included in a regression of spreads on firm characteristics and a dummy variable for listing status; then a t-statistic is used to test the significance of the coefficient on the dummy variable. The potential advantage of a regression approach is that it uses information from more firms than a matched sample approach (since the matched pairs include only a subset of firms). Unfortunately, the response surface between spreads and firm characteristics is highly non-linear and therefore difficult to capture in a linear regression model.

Fig. 1 compares the three test approaches using our Monte Carlo simulation approach. These results are based on constructing matched samples using 1-NN matching with market capitalization. The empirical power of the regression-based approach is the share of Monte Carlo replications with a significant t-statistic on the coefficient β_2 in the OLS regression:

$$(\text{bid} - \text{ask spread})_i = \beta_0 + \beta_1(\text{market capitalization})_i + \beta_2 D_{\text{listing},i} + \varepsilon, \quad (1)$$

where the dummy variable ($D_{\text{listing},i}$) indicates stocks with an induced difference in their spreads. We observe in Fig. 1 that while all three tests have empirical size close to their nominal size of 5%, the non-parametric Wilcoxon signed rank test has much higher power

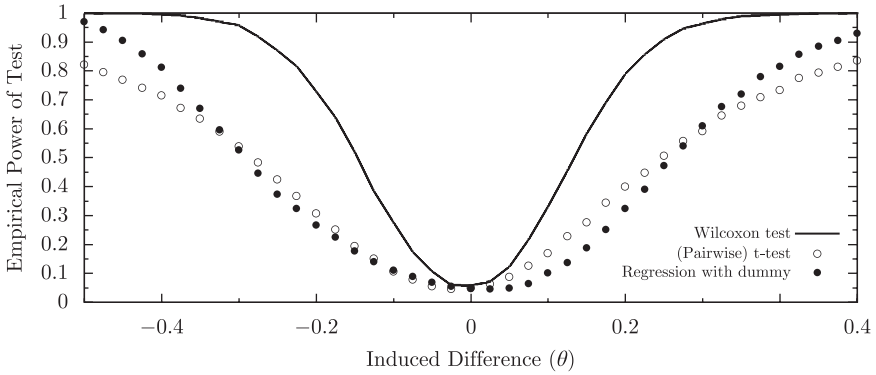

Fig. 1. Comparison of different testing approaches. This graph compares the power of using: (i) matched samples and a Wilcoxon signed rank test; (ii) matched samples and a dependent (pairwise) Student t-test; and (iii) a t-test of the coefficient on the dummy variable for listing status ($D_{\text{listing},i}$) in the regression: $(bid-ask\ spread)_i = \beta_0 + \beta_1(market\ capitalization)_i + \beta_2 D_{\text{listing},i} + \varepsilon$. The reported empirical power of the test is the share of 20,000 Monte Carlo replications with a significant test statistic. For each Monte Carlo replication, we randomly select 50 unique stocks out of the total sample of 1,000 firms and artificially change their observed average percentage bid-ask spread by θs_y: $\hat{Y}_i = Y_i + \theta s_y$. For the matched sample based tests: first each of these 50 stocks is matched one-to-one (1-NN matching) based on market capitalization with one of the remaining 950 firms; then we conduct the specified two-sided test of size $\alpha = 0.05$ of the difference in bid-ask spreads between the 50 stocks with induced differences and their matched pairs. For the regression based test, all 1,000 stocks are included in the regression; the t-test ($\alpha = 0.05$) is for the coefficient of a 0-1 dummy variable that indicates the 50 stocks with induced differences in their spreads.

than the other two approaches. This is strong evidence in support of using the Wilcoxon test. Unless otherwise stated, all of the remaining results in this paper are based on the Wilcoxon test. For comparison with past research, we note that using matched samples with a Student t-test produces similar results to those reported here, except with lower test power.

We note that the regression specification considered in (1) is deliberately simple, so that the three test approaches are compared on the basis of using the same firm information (market capitalization). In unreported results, we find that using additional explanatory variables (e.g., share price) improves test power, but not to the point of exceeding that of the Wilcoxon test. Similar results were also found using logarithmic transformations of the variables.[7] The low empirical power of the regression-based approach confirms why matched sample estimation is generally used in this context.

4.2. Is it best practice to construct one-to-one matches?

To reduce the risk of a bad match, researchers might be tempted to match each stock with a weighted average of two or more stocks, rather than match each stock with another single stock.[8] We now investigate whether matching estimates that smooth over several

[7]We also considered constructing a matched sample first and then running the regression using only stocks included in the matched pairs (as in Harris, 1989).

[8]See, for example, Liu (2006).

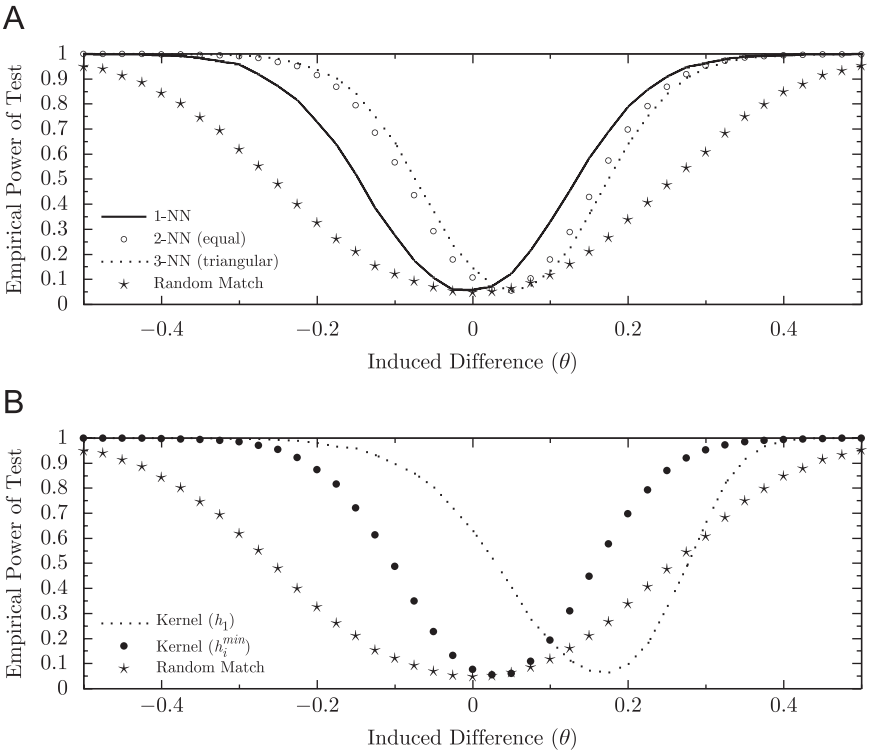

Fig. 2. Comparison of matching estimates. The nearest neighbor matching estimates considered are: 1-NN (one-to-one match), 2-NN (match over two closest firms, equal weights) and 3-NN (match over three closest firms, triangular weights). The kernel-based matching estimates have bandwidth parameters: $h_1 = 1.059 s_x N_0^{-1/5}$ and $h_i^{\min}$ (variable bandwidth). Random match indicates a test based on matched pairs randomly selected without reference to their characteristics. For each of 20,000 Monte Carlo replications: (i) we randomly select 50 unique stocks out of the total sample of 1,000 firms and artificially change their observed average percentage bid-ask spread by θs_y: $\hat{Y}_i = Y_i + \theta s_y$; (ii) each of the 50 stocks is then matched based on market capitalization with a hypothetical firm created using a weighting of the remaining 950 firms, where the weighting scheme is implied by the particular matching estimation technique; (iii) we then conduct a two-sided Wilcoxon signed rank test of size $\alpha = 0.05$ of the difference in bid-ask spreads between the 50 stocks with induced differences and their hypothetical matched pairs. The empirical power of the test is the share of replications with a significant test statistic.

firms are better or worse than estimates based on one-to-one matches. To answer this, Panel A of Fig. 2 compares the power of tests based on 1-NN, 2-NN, and 3-NN estimates for different levels of induced differences, θ, in the percentage bid-ask spread.

We also compare our matching estimates test results with those obtained using an approach that randomly matches pairs of firms independently of any firm characteristics. The random match is the most general alternative against which specific alternatives, such as those based on using market capitalization as a matching criteria, can be evaluated. We are not suggesting that this random match approach should be used in practice; rather we will use this approach to provide a type of lower bound for the power of the matching estimates. In comparison with this lower bound, matching based on market capitalization provides a significant increase in power. That said, even a random match approach can detect large differences. For example, we observe that when spreads differ by $0.4 s_y$

($\approx$ 3.0 bp in our sample), tests based on a random match detect the difference about 84.8% of the time; whereas tests based on 1-NN matching estimates detect the difference 99.5% of the time.

A clear pattern emerges: as k-NN estimates are constructed over more closest firms, the power curve becomes biased away from zero and shifted to the right. Thus, while the 1-NN estimate has virtually no size distortion, the 3-NN estimate has an empirical rejection rate of 14.4% in the presence of no induced difference for a test statistic with a nominal size of 5%. To begin to understand the bias in the power curve, recall that bid-ask spreads (Y) are a function of observed characteristics (X) and unobservables (U), $Y_i = g(X_i) + U_i$. In our simulation, the observed characteristic is market capitalization and the estimated difference between the two samples for a firm with market capitalization $\tilde{X}$ is

$$\hat{\Delta}(\tilde{X}) = g(\tilde{X}) + \theta - \sum_i W_i g(X_i) + \tilde{U} - \sum_i W_i U_i \tag{2}$$

A bias is introduced because $\sum_i W_i g(X_i)$ does not equal $g(\sum_i W_i X_i)$ except in the case of 1-NN matching. This bias increases when more firms are given positive weights (W_i) or when the dispersion of weights in the matching estimate is increased (either by increasing k in k-NN estimates or by increasing h in kernel estimates). In the case of k-NN estimates, the bias is made worse because the X_i values are not equispaced and do not follow a uniform distribution. Intuitively, the distribution of firm sizes is such that it is more likely that positive weights are given to smaller firms than to larger firms, thereby creating an upward bias in the estimated spread of the control group.

While spreading out positive matching weights over more firms increases the bias of our matching estimates, it also helps reduce the variance associated with unobservable firm characteristics, since $\text{plim}_{n\to\infty} n^{-1/2}(\tilde{U} - \sum_{i=1}^{n} W_i U_i)$ equals a constant. Thus, there is a trade-off between bias and variance. The optimal matching technique depends on their relative importance. For example, if there are few matched pairs available, then we should use an estimation technique that minimizes estimation variance. In contrast, if we are attempting to detect a large difference, then we should use an estimation technique that minimizes bias.

Panel B of Fig. 2 illustrates the power curves for tests based on kernel-based matching estimates with the fixed (h_1) and variable ($h_i^{\min}$) bandwidth parameters. The power curves of the kernel-based estimates are biased to the right, much in the same way as increasing k did for the k-NN estimates. For the fixed bandwidth parameter estimator, the size distortion is very severe: tests based on this estimator have empirical size over 55%. In fact, for small positive induced differences, there is a region in which the simplistic random match approach performs better than this kernel-based matching estimate. The problem is that the fixed bandwidth is undersmoothing for the largest firms, but oversmoothing for the smallest firms. In unreported results, we tried various choices of fixed bandwidth parameters and all of them resulted in severely biased test results. In contrast, tests based on the variable bandwidth kernel estimates have much less size distortion. Thus, we conclude that the variable bandwidth approach is more appropriate when there is significant heterogeneity across firms (as is the case in most market microstructure applications).

Fig. 3 considers how test power increases with the number of matched pairs (N_A) for large ($\theta = \pm 0.5$) and small ($\theta = \pm 0.2$) induced differences. Here, θ evenly alternates between positive and negative values across Monte Carlo replications. When the induced

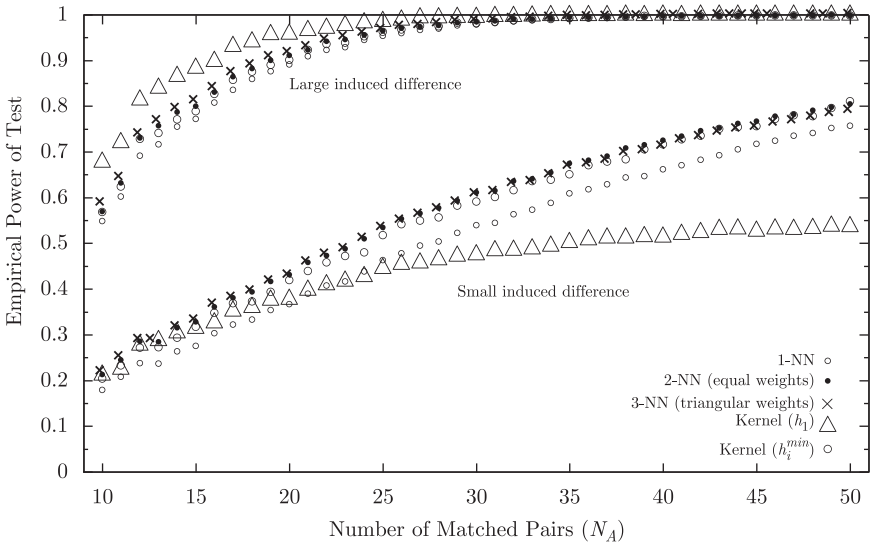

Fig. 3. Number of matched pairs. The upper set of points correspond to a *large* induced difference ($\theta = 0.5$) and the lower set of points correspond to a *small* induced difference ($\theta = 0.2$). Results based on nearest neighbor matching estimates: 1-NN (one-to-one match), 2-NN (match over closest two firms, each assigned equal weight), and 3-NN (match over three closest firms, triangular weights); and kernel-based matching estimates with bandwidth parameters: $h_1 = 1.059 s_x N_0^{-1/5}$ and $h_i^{\min}$ (variable bandwidth). For each of 20,000 Monte Carlo replications: (i) we randomly select N_A unique stocks out of the total sample of 1,000 firms and artificially change their observed average percentage bid-ask spread: $\hat{Y}_i = Y_i + \theta s_y$ for 50% of replications and $\hat{Y}_i = Y_i - \theta s_y$ for 50% of replications; (ii) each of the N_A stocks is then matched based on market capitalization with a hypothetical firm created using a weighting of the remaining $1000 - N_A$ firms, where the weighting scheme is implied by the particular matching technique; (iii) we then conduct a two-sided Wilcoxon signed rank test of size $\alpha = 0.05$ of the difference in bid-ask spreads between the N_A stocks with induced differences and their hypothetical matched pairs. The empirical power of the test is the share of replications with a significant test statistic.

difference is large, each of the four alternative estimates provides *higher* power than the standard 1-NN estimate. Overall, we conclude that when the expected difference is large and the number of matched pairs is small (i.e., in the situation in which minimizing the variance is relatively more important than the estimation bias), a matching estimate that places more weight on distant firms may provide increased testing power. We caution, however, that this gain in test power is small in comparison with the dramatic increase in size distortion.

All of the matching estimation techniques have much less power to detect smaller induced differences. For example, with a small induced difference ($\theta = \pm 0.2$) and 25 matched pairs, the empirical rejection rate is just 40–50%. In this context, the $h_i^{\min}$ kernel-based, the 2-NN, and the 3-NN matching estimates provide higher power than the 1-NN estimate. But, additional smoothing does not guarantee higher power, as evidenced by the poor test power of the h_1 estimates. These results can be understood by referring back to Fig. 2, where we observe that, in comparison with the unbiased 1-NN estimates, the biased smoothing estimates have lower power for $+0.2 s_y$ (≈ 1.5 bp in our sample) and higher power for $-0.2 s_y$. Thus, the overall power depends on the extent to which these differences in power are offsetting.

4.3. Is market capitalization the best matching characteristic?

When firms are matched based on a *single* characteristic, standard practice in the market microstructure literature is to match firms according to their market capitalization. We now consider how using market capitalization compares with using one of 11 alternative matching characteristics: share price, book value, price to book ratio, long-term debt, CAPM beta, shares outstanding, shares in the public float, share price volatility, trade volume, trade dollar volume, and the number of trades. (All matching characteristics are measured prior to the start of the sample period.) As a side benefit, we note that, to the extent that our method establishes the relative importance of these characteristics, our results may have implications for the development of theoretical models of the components of bid-ask spreads.

Panel A of Table 3 shows that market capitalization has better testing power properties than each of the other characteristics, confirming its usage in the literature. The next best test performance comes from matching on trading dollar volume and then from matching on share price. The number of shares outstanding and the number of shares in the public float provide almost identical results. Trading volume, price to book ratio, and book value have relatively poor testing power when used on their own. The empirical size of the tests is close to, but slightly above, its nominal size of 0.05 in all cases. The greatest probability of making a type I error occurs when using book value (empirical size of 0.064).

4.4. Is it better to match using multiple firm characteristics?

We now consider matching over multiple firm characteristics in addition to, or instead of, market capitalization. To do so, we use a weighting scheme similar to that used in Huang and Stoll (1996). Specifically, for each firm $i \in I_A$, we select firm $j \in I_0$ that solves

$$\arg\min_{j \in I_0} \sum_k ((2(x_{Ai}^k - x_{0j}^k))/(x_{Ai}^k + x_{0j}^k))^2, \tag{3}$$

where x_{Ai}^k is firm characteristic k for firm i with listing status A and x_{0j}^k is firm characteristic k for a firm j without listing status A. Panel B of Table 3 reports the empirical test power corresponding to matches based on twelve commonly-used combinations of firm characteristics.

We observe that using characteristics in addition to market capitalization can improve test power, particularly when detecting small differences in spreads. For instance, using market capitalization and share price jointly increases the probability of correctly detecting a difference in spreads of $+0.1s_y$ (≈ 0.75 bp in our sample) by almost twofold compared with using market capitalization alone. We note that our simulation results are based on a post-decimalization sample, in which the minimum tick-size requirement is unlikely to be binding for most stocks. This suggests that price is an important matching characteristic *independent* of tick size differences. We caution that Angel (1997) provides evidence that firms adjust their share prices to reflect the tick size environment. Thus, share prices should only be used as an additional matching characteristic when the tick size environment under the two listing regimes is sufficiently similar.

Of the cases considered, the worst test power is obtained when matching with trading volume, beta, and volatility; clearly, it is important to include at least market capitalization or share price as one of the matching characteristics. We also note that some matching

Table 3
Using different combinations of matching characteristics

When using multiple matching characteristics, for each firm $i \in I_A$, we select firm $j \in I_0$ that solves arg $\min_{j \in I_0} \sum_k ((2(x_{Ai}^k - x_{0j}^k))/(x_{Ai}^k + x_{0j}^k))^2$, where x_{0j}^k is firm characteristic k for a firm j without listing status A and x_{Ai}^k is firm characteristic k for firm i with listing status A. For each of 20,000 Monte Carlo replications: (i) we randomly select 50 unique stocks out of the total sample of 1,000 firms and artificially change their observed average percentage bid-ask spread by θs_y: $\hat{Y}_i = Y_i + \theta s_y$; (ii) each of the 50 stocks is then matched with one of the remaining 950 firms based on the specified matching characteristic; (iii) we then conduct a two-sided Wilcoxon signed rank test of size $\alpha = 0.05$ of the difference in bid-ask spreads between the 50 stocks with induced differences and their matched pairs. The table reports the share of replications with a significant test statistic (empirical power of test).

Matching characteristic(s)	Induced difference (θ)				
	−0.2	−0.1	0	+0.1	+0.2
Panel A. Single matching characteristic					
Market capitalization	0.720	0.252	0.059	0.371	0.816
Trade dollar volume	0.687	0.257	0.057	0.221	0.654
Share price	0.562	0.181	0.052	0.214	0.616
Number of trades	0.526	0.165	0.050	0.212	0.584
Shares outstanding	0.476	0.165	0.052	0.173	0.462
Public float	0.466	0.156	0.056	0.169	0.496
Trade volume	0.425	0.132	0.053	0.214	0.555
Long-term debt	0.412	0.142	0.055	0.139	0.392
CAPM beta	0.368	0.123	0.050	0.148	0.392
Volatility	0.337	0.127	0.053	0.125	0.347
Book value	0.303	0.093	0.064	0.202	0.502
Price to book ratio	0.285	0.092	0.061	0.177	0.415
No characteristic (random match)	0.327	0.123	0.049	0.118	0.340
Panel B. Multiple matching characteristics					
Market capitalization, price	0.926	0.464	0.056	0.548	0.953
Market capitalization, price, number of shares	0.933	0.448	0.052	0.604	0.965
Market capitalization, price, beta	0.931	0.448	0.059	0.619	0.969
Market capitalization, price, volatility	0.932	0.461	0.061	0.638	0.976
Market capitalization, price, volume	0.955	0.488	0.049	0.524	0.963
Market capitalization, price, dollar volume	0.948	0.447	0.057	0.603	0.980
Market capitalization, price, volume, volatility	0.920	0.389	0.069	0.689	0.988
Market cap., price, trades, trade size, volatility	0.945	0.459	0.070	0.753	0.996
Price, trades, trade size, volatility	0.848	0.331	0.050	0.451	0.919
Price, book value, number of shares, long-term debt	0.702	0.259	0.049	0.379	0.830
Volume, beta, volatility	0.444	0.155	0.052	0.159	0.432

combinations cause a larger bias in the power curve (it is shifted slightly to the left), which results in higher relative test power for detecting small positive differences, and lower relative test power for detecting small negative differences. As before, the amount of bias depends on the distribution of the matching characteristic values across firms.

In some instances, using matching characteristics in addition to market capitalization and share price can provide slight incremental test power gains. That said, we caution that these power gains are likely to be offset by two problems. First, in several cases, using additional matching characteristics can lead to a higher likelihood of a type I error (for example, the empirical size of matching with market capitalization, share price, number of trades, trade size, and volatility is 0.070). Second, many of these characteristics are *not* independent of the stock's listing status, as required for unbiased estimates. For instance, it seems reasonable to suppose that the NYSE specialist system may help reduce short-term price volatility; thus, volatility might not be a good matching characteristic for comparisons across trading venues. As well, some characteristics are not directly comparable across exchanges (e.g., Nasdaq trade volume must be adjusted for the possible double-counting of inter-dealer trade). In contrast, share price and market capitalization are largely independent of listing status.[9] Given these advantages and given the wide accessibility of reliable market capitalization data, we recommend 1-NN matching based on market capitalization and share price, unless the specific research context suggests there are benefits to matching over additional characteristics.

Before proceeding, we note that, while the Huang and Stoll (1996) weighting scheme has become the *de facto* standard in the microstructure literature, the Mahalanobis distance metric is more commonly used in the statistics literature. This distance metric is given by

$$\|\mathbf{X}_i - \mathbf{X}_j\| = (\mathbf{X}_i - \mathbf{X}_j)^\top \boldsymbol{\Lambda}^{-1}(\mathbf{X}_i - \mathbf{X}_j), \quad i \in \mathbf{I}_A,\ j \in \mathbf{I}_0,$$

where $\boldsymbol{\Lambda} = [(N_A - 1)\boldsymbol{\Lambda}_A + (N_0 - 1)\boldsymbol{\Lambda}_0]/(N_A + N_0 - 2)$ and where $\boldsymbol{\Lambda}_A$ and $\boldsymbol{\Lambda}_0$ are the sample covariances of the matching characteristics for the control and treated sample, respectively. Rubin (1976) shows that this metric has the property that the percent reduction in bias due to matching is the same for each of the matching variables. This property is desirable if one does not have specific knowledge about the relative importance of firm characteristics in explaining spreads. In unreported results, we find that the Mahalanobis metric produces results similar to those reported here (in the context of 1-NN matching with only two or three matching characteristics).

4.5. Should we restrict matches to firms within the same industry?

A common practice is to sort on industry code first and then construct 1-NN matching estimates. Intuitively, pre-sorting on industry code results in less precise matches on the other matching characteristics, such as market capitalization, for any industry in which the distribution of these characteristics differs between the industry and the full population. To investigate further, we consider restricting matches to within the same 2-digit SIC industry

[9]Biased estimates can result when the sample of stocks with the listing status of interest is not representative of stocks that have a propensity to obtain this listing status. A propensity score adjustment can be used to correct for this deficiency provided that: (i) enough firms change their listing status during the sample period; and (ii) there is an *identified* model of the listing decision. Unfortunately, researchers rarely have either. For example, while Doidge, Karolyi, and Stutz (2004) model the decision to cross-list on a U.S. exchange, the factors included in their model are not independent (in general) of the factors that affect spreads.

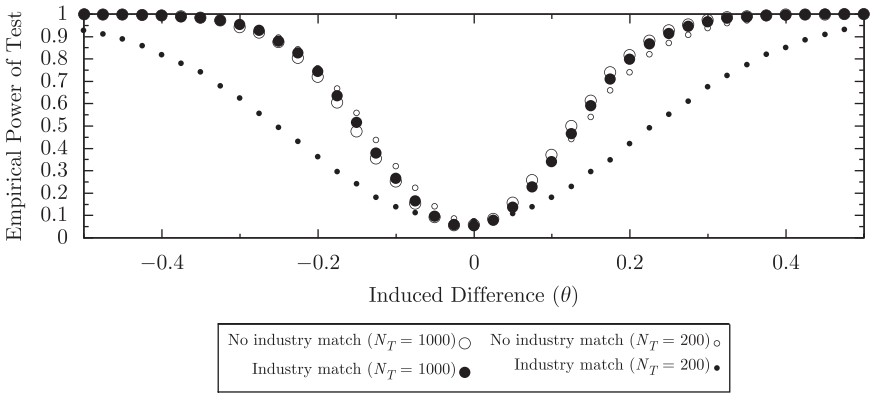

Fig. 4. Pre-sorting by industry. Results based on 1-NN (one-to-one) matching estimates that are based on market capitalization alone and 1-NN matching estimates that also require the matched firm to have the same 2-digit SIC industry code. Based on 50 matched pairs and $N_T = 200$ and 1000 as indicated. When $N_T = 200$, each replication 200 stocks are randomly selected out of the original sample of 1,000 stocks. For each of 20,000 Monte Carlo replications: (i) we randomly select 50 unique stocks out of the total sample of N_T firms and artificially change their observed average percentage bid-ask spread by θs_y: $\hat{Y}_i = Y_i + \theta s_y$; (ii) each of the 50 stocks is then matched with one of the remaining $N_T - 50$ firms; (iii) we then conduct a two-sided Wilcoxon signed rank test of size $\alpha = 0.05$ of the difference in bid-ask spreads between the 50 stocks with induced differences and their matched pairs. The empirical power of the test is the share of replications with a significant test statistic.

code. In our original sample the fewest firms within the same SIC category is 2 (agricultural production crops) and the most is 74 (electric, gas, and sanitary services).

Fig. 4 shows that this industry-sorting technique has virtually no effect on our results when firms are drawn from an original sample of 1,000. This sample size is sufficiently large that pre-sorting by industry has little effect on the distribution of potential matches. In contrast, when there are fewer potential firms to match, pre-sorting by industry can dramatically reduce test power. Specifically, we observe that when $N_T = 200$, pre-sorting by industry reduces test power by almost half for some levels of induced differences. Thus, at best, pre-sorting by industry has little effect; at worst, it dramatically decreases test power. Our recommendation is that in most situations researchers should not restrict matches to firms within the same industry.

4.6. Should we drop poor matches?

Is it better to keep a crudely matched distant neighbor or is it better to drop the firm from the sample for lack of a suitable match? Formally, the exclusion of apparent poor matches is a variant of nearest neighbor matching known as *caliper matching* in which matches are made to firm $i \in \mathbf{I}_A$ only if there exists a firm $j \in \mathbf{I}_0$ such that $\|\mathbf{X}_i - \mathbf{X}_j\| < \varepsilon$, $j \in \mathbf{I}_0$ where ε is a pre-specified tolerance. Otherwise, no match is undertaken, and firm i is omitted. This procedure is designed to circumvent the problem of a substantial gap between i and j.

A common approach in the microstructure literature is to require matches to occur between firms that satisfy $|2(x_{Ai}^k - x_{0j}^k)/(x_{Ai}^k + x_{0j}^k)| < \varepsilon$, for each matching characteristic k

Table 4
Comparison of matching with different caliper tolerance levels

Results based on 1-NN (one-to-one nearest neighbor) matching estimates with caliper tolerance levels of $\varepsilon \in \{\infty, 1.00, 0.75, 0.50, 0.10, 0.05\}$. Based on (up to) 50 matched pairs. For each of 20,000 Monte Carlo replications: (i) we randomly select 50 unique stocks out of the total sample of 1,000 firms and artificially change their observed average percentage bid-ask spread by θs_y: $\hat{Y}_i = Y_i + \theta s_y$; (ii) each of the 50 stocks is then matched based on market capitalization with a stock selected from the remaining 950 firms; (iii) if the match does not satisfy the caliper tolerance criteria, $|2(x^k_{Ai} - x^k_{0j})/(x^k_{Ai} + x^k_{0j})| < \varepsilon$, for any of the matching characteristics, the match is removed from consideration; (iv) We then conduct a two-sided Wilcoxon signed rank test of size $\alpha = 0.05$ of the difference in bid-ask spreads between the remaining matched pairs. The table reports the share of replications with a significant test statistic (empirical power of test).

Matching characteristic(s)	Tolerance	Induced difference (θ)				
	level (ε)	−0.2	−0.1	0	+0.1	+0.2
Market capitalization	∞	0.720	0.252	0.059	0.371	0.816
	1.00	0.710	0.247	0.056	0.362	0.820
	0.75	0.713	0.252	0.056	0.380	0.822
	0.50	0.713	0.252	0.056	0.380	0.822
	0.10	0.712	0.252	0.056	0.377	0.821
	0.05	0.705	0.246	0.056	0.372	0.814
Market capitalization, price	∞	0.926	0.464	0.056	0.548	0.953
	1.00	0.924	0.444	0.052	0.526	0.949
	0.75	0.924	0.445	0.052	0.526	0.949
	0.50	0.923	0.449	0.052	0.539	0.949
	0.10	0.905	0.429	0.051	0.499	0.930
	0.05	0.805	0.343	0.052	0.293	0.767
Market capitalization, price, number of shares	∞	0.933	0.448	0.052	0.604	0.965
	1.00	0.931	0.428	0.054	0.601	0.965
	0.75	0.930	0.425	0.054	0.605	0.965
	0.50	0.932	0.430	0.054	0.601	0.965
	0.10	0.855	0.342	0.053	0.483	0.901
	0.05	0.683	0.265	0.048	0.252	0.655

considered. The existing literature has used, without any theoretical justification, tolerance levels (ε) of 0.5, 0.75 and 1. If no match can be found that satisfies the tolerance level for each matching characteristic considered, then the number of matched pairs is reduced accordingly. In Table 4, we provide illustrative results for (i) matching with market capitalization; (ii) matching with market capitalization and share price; and (iii) matching with market capitalization, share price, and trading volume. The results suggest that using caliper matching with the commonly-used tolerance levels ($\varepsilon \in \{0.5, 0.75, 1\}$) is not sufficiently binding to make a difference to overall test power. In some cases, test power is reduced slightly by using caliper matching; in other cases, it is increased slightly. We also report the results of using more restrictive tolerance levels of $\varepsilon = 0.1$ and $\varepsilon = 0.05$: these tolerance levels reduce test power, especially when multiple matching characteristics are used.

These results highlight the main problem with caliper matching: a priori, it is difficult to know which tolerance level to use. There is a trade-off between reducing the number of matched pairs and improving the closeness of matches. In the absence of a known optimal

tolerance level, we believe that it is better to keep poor matches than to reduce the sample size. In unreported results, we find this holds true for both large and small sample sizes. On the one hand, when there are a large number of available matched pairs, the impact of a poor match is relatively small and the likelihood of a poor match is also normally smaller—thus, discarding a poor match has little impact on a non-parametric test. On the other hand, when there are few available matched pairs, the likelihood of a poor match is higher, but the elimination of a matched pair has a larger negative impact on the power of the statistical test. Thus, since one cannot be sure that an apparent poor match is *really* a poor match (due to unobserved firm characteristics), it is generally better to use the entire available sample.

4.7. How does the size of the control sample affect test properties?

Until now, most of our simulation results have been based on a control sample of $N_0 = 950$ stocks and a treatment sample of $N_A = 50$ stocks. This is similar to the environment found in many applications of matched sample estimation, in which the control sample is much larger than the treated sample; however, this is not always so. For example, the pilot program of Regulation SHO (which eases restrictions for short sales) involved 1,000 stocks; since this is a large share of stocks in the Russell 3000 Index, the available sample of similar sized firms not in the pilot program is small in *relative* terms. In unreported results, we found that the empirical power of tests is largely unaffected by the number of firms in the control sample (when there are more firms in the control sample than the treated sample). Thus, our focus here is how the number of firms in the control sample affects the empirical size of the test.

With a smaller control sample (in relative terms), it is more important to focus on the distinction between 1-NN matching *with* replacement and 1-NN matching *without* replacement. All of our reported results so far have been based on matching with replacement, in which a firm in the control sample can be the nearest neighbor for multiple firms. This contrasts with matching without replacement, in which each firm can be matched only once. While this approach forces some matches to occur between firms that are not closest neighbors, it avoids having one firm being matched multiple times and thereby having disproportionate influence on the matching estimate.

Fig. 5 plots the empirical size of the test against the number of control firms (N_0), for $N_A \in \{25, 50, 100, 250\}$ with replacement and for $N_A = 50$ without replacement. For matching with replacement, empirical size increases as the number of firms in the control sample decreases, reaching about 13% when $N_0 = N_A$. The empirical size is even higher when there are fewer control firms than treatment firms, but clearly this is not a good test design. In contrast, matching without replacement provides a test with (close to) correct empirical size, except when the number of control firms is very close to the number of treatment firms.[10] This holds true for the other values of N_A (not shown for clarity). In unreported results, we found that the cost of improving the empirical size using matching

[10]When matching without replacement, there is a non-trivial problem of deciding the optimal order of the matches (each match reduces the set of firms available to match with the other firms). Rubin (1973) shows that a good approach is to always select, from the remaining unmatched treatment firms, the firm with the *highest* distance score to its closest neighbor as the next match (i.e., match difficult firms first).

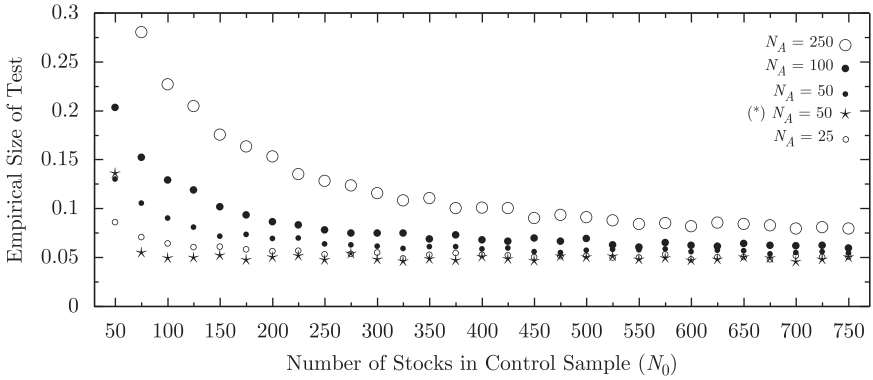

Fig. 5. Empirical size and the number of control stocks. The star (*) indicates matching *without* replacement (each stock can be used in a match only once). All other results are *with* replacement. For each of 20,000 Monte Carlo replications: (i) we randomly select N_A unique stocks for the treated sample and N_0 unique stocks for the control sample from the original sample of 1,000 firms; (ii) each of the N_A stocks is then matched one-to-one (with or without replacement) based on market capitalization with one of the N_0 firms; (iii) we then conduct a two-sided Wilcoxon signed rank test of size $\alpha = 0.05$ of the difference in bid-ask spreads between the matched pairs. The empirical size of the test is the share of replications with a significant test statistic (since there is no induced difference in spreads).

without replacement is a slight reduction in the power of tests to detect small induced differences.

Finally, we note that, in unreported results, when the entire original sample ($N_T = 1,000, N_0 = 950, N_A = 50$) is used, the test properties (empirical power and size) of matching with and without replacement are virtually identical. Essentially, the number of potential matches is sufficiently large that few firms are matched twice. When a firm is matched twice, the sample is large enough that the next closest firm is a close substitute to the nearest neighbor. Importantly, this confirms that all of the other findings reported in this paper would be the same if conducted under a matching without replacement regime.

Thus, our main take-away here is that matching without replacement is a good idea when the ratio of control firms to treated firms is small (say, less than two-to-one). Otherwise, matching with and without replacement provide tests with very similar characteristics.

4.8. Should we use matched sample estimation or an event study?

An alternative to using matched sample estimation is to use an event study approach. This approach restricts attention to the subset of firms that change their listing status over the sample period. An estimate of the impact of the change in listing status is then constructed by comparing trading properties over a period prior to the event date with those over a period after the event date.[11] Ideally, event studies provide a natural experiment that allows for *direct* measurement of the impact of changing listing status on trade execution costs.

[11]For example, see Christie and Huang (1994), Barclay (1997), and Clyde, Schultz, and Zaman (1997).

Despite the obvious advantages of its direct approach, there are some potential problems with using an event study in this context.[12] For instance, an event study with a short time window may not provide a clear indication of the long-term impact of the change in listing status. Also, an event study may provide biased results if listing decisions cluster in time by industry. We acknowledge that there are sophisticated techniques which may help overcome some of these problems and that similar problems can also inflict matched sample estimation. These issues are not the focus of this paper. Here, we are interested in another potential serious limitation of an event study: often few stocks change their listing status in any given timeframe. For instance, Kedia and Panchapagesan (2003) report that over the period 1986–1998, on average, fewer than 40 firms per year switch their listing status from Nasdaq to the NYSE and there are virtually no *voluntary* switches from NYSE to Nasdaq.[13] The number of firms switching exchanges is much fewer than the 600+ Nasdaq companies that currently satisfy NYSE listing standards and would be available for a matched sample approach.

Thus, researchers often face a trade-off: using a more direct event study with a small sample of stocks or using a matched sample approach with a much larger sample of stocks. To explore this issue, we conduct another simulation to illustrate the relative power of the event study approach. Specifically, for each of 20,000 Monte Carlo replications:

1. We randomly select N unique stocks from the sample of 1,000 firms.
2. There are $t = 1, \ldots, 251$ trading days in 2004. For each randomly selected stock i, we choose a random event date τ_i, where $\tau_i \in [31, 221]$. Different stocks may or may not have the same event date.
3. For each stock i, we calculate the average percentage bid-ask spread during the 30 trading days prior to the event date (denoted by Y_i^b) and during the 30 trading days after the event date (denoted by Y_i^a).
4. A difference of $\pm 0.2 s_y$ is induced in the post event date spread: $\hat{Y}_i^a = Y_i^a + 0.2 s_y$ for one half of the replications and $\hat{Y}_i^a = Y_i^a - 0.2 s_y$ for the other half of the replications.
5. We conduct a two-sided Wilcoxon signed rank test of size α of the differences between the post- and pre-event spreads: $[(\hat{Y}_1^a - Y_1^b), (\hat{Y}_2^a - Y_2^b), \ldots, (\hat{Y}_N^a - Y_N^b)]$.

Fig. 6 provides the results. As expected, the event study approach has more power than the matched sample approach *for a given sample size*. However, Fig. 6 also demonstrates that an event study can have less test power than a matched sample approach if the number of firms available for an event study is much fewer than the number of matched pairs available.

[12] Many of the potential problems with an event study in this context are analogous to those discussed in Fama (1998) in the context of detecting long-run abnormal returns.

[13] Other event studies of listing status changes have also used small numbers of firms. Over a 4-year period, Jain and Kim (2006) study 174 stocks that switched their US exchange listing prior to decimalization and the 93 stocks that switched after decimalization. Bennett and Wei (2006) study 39 stocks that switched from Nasdaq to NYSE between January 2002 and March 2003. The small sample problem is worse when studying the impact of cross-listing. For example, Foerster and Karolyi (1998) study just 5 NYSE-, 7 AMEX- and 40 Nasdaq-interlisted stocks and Mittoo (2003) studies just 30 NYSE-, 11 AMEX- and 67 Nasdaq-interlisted stocks. Both of these studies needed to use long time periods (10 and 9 years, respectively) to have a sufficient sample of firms changing their listing status.

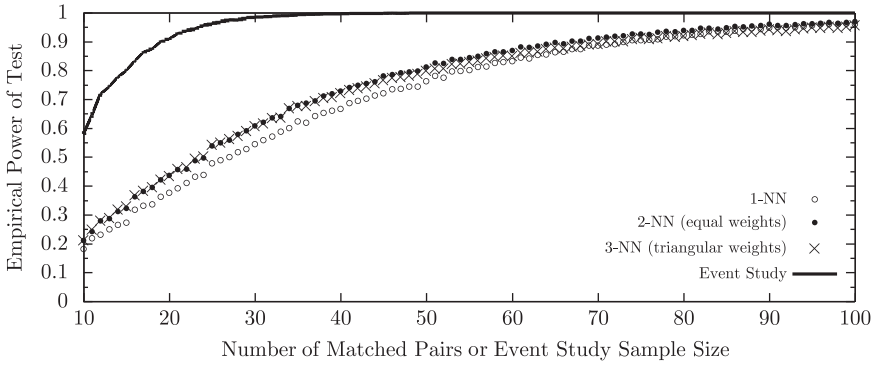

Fig. 6. Comparison of event study approach and matched sample approach. Power graph of two-sided Wilcoxon signed rank test based on: (i) nearest neighbor matching estimates: 1-NN, 2-NN, and 3-NN; and (ii) the corresponding event study. The empirical power of the matching estimates is obtained in the same manner as Fig. 3 and is based on matching according to market capitalization. To construct the empirical power of tests based on the event study, for each of 20,000 Monte Carlo replications: (i) we randomly select N unique stocks from the sample of 1,000 firms; (ii) for each of the randomly selected stocks, we choose a random event date $\tau \in [31, 221]$ (different stocks can have the same event date); (iii) denote Y_i^b and Y_i^a as the average percentage bid-ask spread of stock i during the 30 trading days prior to the event date and during the 30 trading days after the event date, respectively; (iv) the post event date spread has an induced difference of: $\hat{Y}_i^a = Y_i^a + 0.2s_y$ for one half of the replications and $\hat{Y}_i^a = Y_i^a - 0.2s_y$ for the other half of the replications; (v) we conduct a two-sided Wilcoxon signed rank test of size $\alpha = 0.05$ of the difference between $[Y_1^b, Y_2^b, \ldots, Y_N^b]$ and $[\hat{Y}_1^a, \hat{Y}_2^a, \ldots, \hat{Y}_N^a]$. The empirical power of the test is the share of replications with a significant test statistic.

It is important to recognize that our simulation cannot fully capture the complex underlying order flow dynamics around the listing event date. That said, we believe that our model is more likely than not to over-estimate the relative power of an event study approach given the large number of unmodelled factors that could have an impact on spreads in the period prior to and subsequent to the change in listing status. Thus, we conclude that if the number of firms available for an event study is small (e.g., fewer than 20 firms), it may be preferable to use a broad matched sample estimation approach instead.

4.9. Robustness of results

4.9.1. How do our findings change when different sub-samples are used?

One might conjecture that our findings are driven by stocks with very low share prices or very low market capitalizations, since these stocks tend to have much wider bid-ask spreads. To confirm that our findings are robust to excluding these stocks, we evenly split our original sample of 1,000 NYSE stocks into high/low sub-samples according to: (i) market capitalization; and (ii) share price. Table 5 confirms that, for all four sub-samples, matching with market capitalization and share price has greater overall test power than matching with market capitalization alone. As expected, higher test power occurs in the samples of large capitalization stocks and high share price stocks. These stocks tend to have less noise in their spread estimates, thereby making a given induced difference in spreads easier to detect.

Table 5
Sub-samples based on market capitalization and share price

Matching estimates are based on the nearest neighbor (1-NN) approach using as matching characteristics: (i) market capitalization; and (ii) market capitalization and share price. When using market capitalization and share price, for each firm $i \in I_A$, we select firm $j \in I_0$ that solves arg $\min_{j \in I_0} \sum_k^2 ((2(x_{Ai}^k - x_{0j}^k))/(x_{Ai}^k + x_{0j}^k))^2$, where x_{0j}^k is firm characteristic k for a firm j without listing status A and x_{Ai}^k is firm characteristic k for firm i with listing status A. s_y is the standard deviation of the bid-ask spreads of the original sample of 1,000 stocks. For each of 20,000 Monte Carlo replications: (i) we randomly select 50 unique stocks out of the total sample of 500 firms and artificially change their observed average percentage bid-ask spread by θs_y: $\hat{Y}_i = Y_i + \theta s_y$; (ii) each of the 50 stocks is then matched with one of the remaining 450 firms; (iii) we then conduct a two-sided Wilcoxon signed rank test of size $\alpha = 0.05$ of the difference in bid-ask spreads between the 50 stocks with induced differences and their matched pairs. The table reports the share of replications with a significant test statistic (empirical power of test).

Sub-sample	Matching characteristic(s)	Induced difference (θ)				
		−0.2	−0.1	0	+0.1	+0.2
Large stocks	Market capitalization	0.955	0.441	0.073	0.743	0.992
	Market capitalization, price	1.000	0.853	0.055	0.923	1.000
Small stocks	Market capitalization	0.456	0.156	0.059	0.174	0.490
	Market capitalization, price	0.613	0.193	0.053	0.261	0.672
High price stocks	Market capitalization	0.996	0.699	0.058	0.717	0.996
	Market capitalization, price	0.995	0.723	0.054	0.816	0.999
Low price stocks	Market capitalization	0.554	0.192	0.056	0.201	0.553
	Market capitalization, price	0.731	0.251	0.061	0.373	0.834

4.9.2. How does test power differ across different time periods?

Table 6 considers matching based on market capitalization and share price for a sample of 1,000 NYSE-listed U.S.-based common stocks obtained annually from 1993 to 2006. Data on market capitalization and share prices at the beginning of each year are obtained from CRSP. As before, we express induced differences in fractional terms of the standard deviation of the bid-ask spreads in the sample (s_y), rather than in basis point or price terms. This is intended to ensure comparability across time periods, since a given basis point difference is easier to detect when the dispersion of spreads is smaller. In this case, it adjusts the induced difference to account for the dramatic fall in s_y over time: from 48.4 bp in 1993 to 5.3 bp in 2006. Although a single number cannot fully capture all features of the dispersion of spreads across firms in the sample, our adjustment factor does a good job overall, producing remarkably similar magnitudes of the test power across years. This is important, since it ensures that the main findings of this paper are applicable across different time periods.

It is worth highlighting that there is a drop in test power during the stock market boom period of 1998–2000. This suggests that market capitalization and share price became less closely related to a firm's bid-ask spread during this period. One possible explanation is that the inflated firm values and share prices of a subset of stocks caused these firm characteristics to be less reflective of some of the theoretical components of the bid-ask spread, such as informational asymmetry across investors.

Table 6
Annual results, 1993–2006

Each year, we select a sample of 1,000 NYSE U.S.-based common stocks. Matching estimates are based on the nearest neighbor (1-NN) approach using market capitalization and share price as matching characteristics. For each firm $i \in I_A$, we select firm $j \in I_0$ that solves $\arg\min_{j \in I_0} \sum_k^2 ((2(x_{Ai}^k - x_{0j}^k))/(x_{Ai}^k + x_{0j}^k))^2$, where x_{0j}^k is firm characteristic k for a firm j without listing status A and x_{Ai}^k is firm characteristic k for firm i with listing status A. s_y is the standard deviation of the bid-ask spreads of the selected annual sample. $\overline{P}$ is the mean share price of firms in the sample at the beginning of the year. Each year, for each of 20,000 Monte Carlo replications: (i) we randomly select 50 unique stocks out of the total sample of 1,000 firms and artificially change their observed average percentage bid-ask spread by θs_y: $\hat{Y}_i = Y_i + \theta s_y$; (ii) each of the 50 stocks is then matched with one of the remaining 950 firms; (iii) we then conduct a two-sided Wilcoxon signed rank test of size $\alpha = 0.05$ of the difference in bid-ask spreads between the 50 stocks with induced differences and their matched pairs. The table reports the share of replications with a significant test statistic (empirical power of test).

Year	s_y (in bp)	$s_y \times \overline{P}$ (in cents)	Induced difference (θ)				
			−0.2	−0.1	0	+0.1	+0.2
1993	48.4	16.6	0.880	0.439	0.056	0.297	0.811
1994	45.4	16.3	0.837	0.351	0.050	0.333	0.815
1995	39.0	12.5	0.777	0.315	0.053	0.363	0.838
1996	32.5	12.4	0.851	0.382	0.053	0.320	0.805
1997	27.5	11.0	0.828	0.356	0.056	0.442	0.889
1998	26.0	11.7	0.473	0.155	0.056	0.232	0.611
1999	31.8	13.0	0.639	0.242	0.050	0.232	0.655
2000	36.2	13.6	0.601	0.206	0.052	0.244	0.655
2001	24.1	8.9	0.758	0.303	0.053	0.332	0.791
2002	22.1	7.9	0.952	0.532	0.051	0.617	0.973
2003	10.3	3.0	0.921	0.470	0.059	0.663	0.973
2004	7.5	2.5	0.926	0.464	0.056	0.548	0.953
2005	6.4	2.6	0.658	0.209	0.059	0.365	0.798
2006	5.3	2.1	0.787	0.276	0.055	0.413	0.872

Examined from a different perspective, our annual results suggest that the power of tests used in the existing literature may not be as high as often assumed. The large dispersion in spreads in earlier years of the sample made it much harder to detect a given difference *expressed in basis point or price terms*. For instance, in 1993, a difference in percentage bid-ask spreads of $+0.1s_y$ corresponded to about 1.66 cents for the mean priced stock. In today's highly competitive environment, this would be considered a large difference in spreads. Yet, because of the large dispersion in spreads in 1993, the matched sample approach would have detected a difference of this magnitude only 29.7% of the time. In other words, the probability of a type II error would be over 70% in this case.

To put the magnitudes of these induced differences into perspective, we refer to the spread differences found in three highly-cited articles: Venkataraman (2001), Bessembinder (1999) and Weston (2000). Venkataraman (2001) uses 40 matched pairs and finds spread differences of 1.59–2.65 bp in the 1997–1998 period. Our simulation results in Table 6 suggest that test power to detect a similar spread difference during this time period could be as low as 15%. Bessembinder (1999) uses between 136 and 539 matched pairs and finds spread differences of about 25 bp during July–December 1997. In unreported results, we ran a simulation with a similar number of matched pairs during the same time period, and

found that tests would almost always detect a similar large spread difference. Weston (2000) uses as few as 22 matched pairs (in the dollar volume quartiles) and finds spread differences as low as 3 cents during 1996–1997. Again, in unreported results, we ran a simulation with similar conditions and found test power of about 20–40%.

All three of these papers employed extensive robustness checks to verify their results. Our simulation results similarly suggest that the one-to-one matching techniques used in these papers have a low probability of a type I error. Thus, our intention is not to cast doubt on the results of previous research, but rather to highlight that test power can be surprisingly low and that the choice of matching technique might influence test outcomes in commonly-studied contexts. The potential for type II errors is much higher than many researchers in market microstructure typically assume. It is likely that past (unpublished) research may have failed to find important spread differences that actually existed.

4.9.3. Are our findings sensitive to the database used?

We repeat our analysis using a sample of 1,000 stocks listed on the Nasdaq stock market contained in the Russell 3000 index, using the same selection criteria as used for our sample of NYSE stocks. On average, the Nasdaq stocks in our sample are much smaller, with an average market capitalization of $2,450 million (median of $579 million), with a range of $171 million to $295,937 million. Table 7 presents results for Nasdaq stocks that are analogous to those presented in Table 3 for NYSE stocks. The magnitudes of the empirical power and size reported in Table 3 are not directly comparable because of differences in the dispersion of firm characteristics in the two samples. Thus, our primary interest is whether the qualitative ordering of the matching techniques is similar for the two samples.

Many of the same patterns emerge. As before, matching on market capitalization and share price jointly provides superior test power to matching on market capitalization alone. And, again, there is little or no benefit from adding the number of shares outstanding, CAPM beta, or volatility as matching criteria to a matching regime that already includes market capitalization and share price. Again, we also find that jointly matching with trade volume, beta and volatility provides low testing power. Similar to our NYSE results, we find that matching with market capitalization, price, and dollar volume and matching with market capitalization, share price, number of trades, trade size and volatility provides higher test power than matching with market capitalization and share price, but at the cost of greater probability of a type I error. Finally, we observe that matching on market capitalization, share price, and trade volume provides considerable test power (and low size distortion) relative to other methods—this method has higher relative power for Nasdaq stocks than for NYSE stocks, suggesting that trade volume may be a useful matching characteristic for an *intra*-exchange study using only Nasdaq stocks.

Overall, our results using Nasdaq stocks support our earlier findings. In most situations, it is optimal to match on market capitalization and share price. Using additional matching characteristics might be useful in specific research contexts. When comparing across exchanges, adjustments must be made to control for differences in the accounting of order flow volume and differences in trade volume resulting from market structure features, rather than firm-specific features that are independent of exchange listing.

In a previous version of this paper, we conducted many of the same simulations using a sample of common stocks listed on the Toronto Stock Exchange (TSX) in 1998. These results are available from the authors upon request. The TSX sample contained fewer firms than our NYSE and Nasdaq samples, and contained several large firms that did not have

Table 7
Sample of Nasdaq stocks

When using multiple matching characteristics, for each firm $i \in I_A$, we select firm $j \in I_0$ that solves $\arg\min_{j \in I_0} \sum_k ((2(x^k_{Ai} - x^k_{0j}))/(x^k_{Ai} + x^k_{0j}))^2$, where x^k_{0j} is firm characteristic k for a firm j without listing status A and x^k_{Ai} is firm characteristic k for firm i with listing status A. For each of 20,000 Monte Carlo replications: (i) we randomly select 50 unique stocks out of the total sample of 1,000 Nasdaq firms and artificially change their observed average percentage bid-ask spread by θs_y: $\hat{Y}_i = Y_i + \theta s_y$; (ii) each of the 50 stocks is then matched using 1-NN matching with one of the remaining 950 firms; (iii) we then conduct a two-sided Wilcoxon signed rank test of size $\alpha = 0.05$ of the difference in bid-ask spreads between the 50 stocks with induced differences and their matched pairs. The table reports the share of replications with a significant test statistic (empirical power of test).

Matching characteristic(s)	Induced difference (θ)				
	−0.2	−0.1	0	+0.1	+0.2
Market capitalization	0.601	0.217	0.052	0.250	0.607
Market capitalization, price	0.673	0.282	0.050	0.270	0.658
Market capitalization, price, number of shares	0.660	0.272	0.054	0.245	0.623
Market capitalization, price, beta	0.663	0.260	0.055	0.382	0.791
Market capitalization, price, volatility	0.604	0.243	0.052	0.293	0.701
Market capitalization, price, volume	0.844	0.399	0.052	0.383	0.842
Market capitalization, price, dollar volume	0.790	0.297	0.070	0.585	0.926
Market capitalization, price, volume, volatility	0.700	0.263	0.062	0.540	0.912
Market cap., price, trades, trade size, volatility	0.709	0.256	0.068	0.585	0.926
Price, trades, trade size, volatility	0.670	0.216	0.053	0.349	0.798
Price, book value, number of shares, long-term debt	0.321	0.120	0.050	0.148	0.375
Volume, beta, volatility	0.375	0.138	0.051	0.101	0.287

apparent suitable matches. While the heterogeneity of the TSX sample resulted in slightly less test power, our general conclusions about the potential drawbacks of using additional matches, using caliper matching, and matching by industry still held.

4.9.4. *Are our findings similar for other spread measures?*

To investigate, Table 8 considers 1-NN matching using (i) a time-weighted percentage quoted spread: $2 \times (bid - ask)/(bid + ask)$; (ii) a trade-weighted effective percentage spread: $200 \times |Trade\ Price - 0.5 \times (ask + bid)|/(ask + bid)$; (iii) a time-weighted (dollar) quoted spread: $(bid - ask)$; and (iv) a trade-weighted effective (dollar) spread: $|Trade\ Price - 0.5 \times (ask + bid)|$. As before, for all four spread measures, matching based on market capitalization and share price jointly provides more test power than matching based on market capitalization alone. Our results show that differences in effective spreads are harder to detect than differences in quoted spreads; in part, this may be caused by the existence of more (unobservable) factors that influence effective spread measures. As well, we find it much better to compare percentage spreads than dollar spreads; perhaps surprisingly, this finding persists even after matching over share price. Overall, in unreported results, we ran all of our main simulations using effective spreads and found that using effective spreads resulted in lower power, but the same test power trade-offs (across different combinations of matching characteristics), as using quoted spreads. In particular, we did not find a particular combination of matching characteristics that worked much better for effective spreads relative to quoted spreads.

Table 8
Alternative spread measures

The four spreads measures are: (i) a time-weighted percentage quoted spread: $2 \times (bid - ask)/(bid + ask)$; (ii) a trade-weighted effective percentage spread: $200 \times |Trade\ Price - 0.5 \times (ask + bid)|/(ask + bid)$; (iii) a time-weighted (dollar) quoted spread: $(bid - ask)$; and (iv) a trade-weighted effective (dollar) spread: $|Trade\ Price - 0.5 \times (ask + bid)|$. Results based on the 1-NN (one-to-one) matching estimates using market capitalization and using market capitalization and share price. With multiple matching characteristics, for each firm $i \in I_A$, we select firm $j \in I_0$ that solves $\arg\min_{j \in I_0} \sum_{k=1}^{2}((2(x_{Ai}^k - x_{0j}^k))/(x_{Ai}^k + x_{0j}^k))^2$, where x_{0j}^k is firm characteristic k for a firm j without listing status A and x_{Ai}^k is firm characteristic k for firm i with listing status A. For each of 20,000 Monte Carlo replications: (i) we randomly select without replacement 50 stocks out of the total sample of 1,000 firms and artificially change their observed average bid-ask spread by θs_y: $\hat{Y}_i = Y_i + \theta s_y$; (ii) each of the 50 stocks is then matched with one of the remaining 950 firms; (iii) we then conduct a two-sided Wilcoxon signed rank test of size $\alpha = 0.05$ of the difference in bid-ask spreads between the 50 stocks with induced differences and their matched pairs. The table reports the share of replications with a significant test statistic (empirical power of test).

Spread	Matching characteristic(s)	Induced difference (θ)				
		−0.2	−0.1	0	+0.1	+0.2
% Quoted spread	Market capitalization	0.720	0.252	0.059	0.371	0.816
	Market capitalization, price	0.926	0.464	0.056	0.548	0.953
% Effective spread	Market capitalization	0.535	0.172	0.054	0.223	0.601
	Market capitalization, price	0.797	0.287	0.057	0.382	0.859
$ Quoted spread	Market capitalization	0.413	0.176	0.062	0.077	0.233
	Market capitalization, price	0.747	0.294	0.050	0.309	0.750
$ Effective spread	Market capitalization	0.324	0.144	0.061	0.077	0.193
	Market capitalization, price	0.473	0.157	0.051	0.163	0.467

5. Conclusion

Matched sample estimation is an effective and widely-used tool for testing for differences in trade execution costs. In this context, this paper has shown that the *best practice* for constructing matched samples is to match firms one-to-one based on market capitalization and share price. We argue that the slight gain in test power which may occur from using additional matching characteristics over and above these two firm characteristics is likely to be offset by problems that can arise when matching characteristics are not independent of the firm's listing decision and the exchange's market structure.

We provide evidence that, in the absence of a procedure to select a suitable tolerance level, it is better to include an apparent poor match than drop the firm for lack of a suitable match. We demonstrate that restricting matches to the same industry group may lead to a substantial loss of power, particularly when the distribution of firm market capitalizations differs between the industry and the full population. We find that when the size of the control sample is relatively small, matching without replacement reduces size distortion. We also show that, in this context, a non-parametric Wilcoxon signed rank test provides better test properties than a dependent Student t-test. We also provide evidence that matched sample estimation can provide greater test power than the corresponding event

study if the number of matched pairs is sufficiently greater than the number of firms available for the event study.

We show that tests for differences in bid-ask spreads based on one-to-one nearest neighbor matching estimates typically have less size distortion than, and comparable power to, matching estimates that place more weight on distant matches. Any potential power gains that arise from using matching estimates that smooth the measured effect by matching over multiple firms are often more than offset by significant size distortion.

Our results demonstrate that a lack of test power can be a concern in commonly studied contexts, and that the choice of matching technique can have an impact on the test outcome. Finally, robustness checks have shown that our main findings are widely applicable and relevant to studies using different exchanges, sample sizes, spread measures, and time periods.

References

Affleck-Graves, J., Hedge, S.P., Miller, R.E., 1994. Trading mechanisms and the components of the bid-ask spread. Journal of Finance 49 (4), 1471–1488.

Angel, J.J., 1997. Tick size, share prices, and stock splits. Journal of Finance 52 (2), 655–681.

Aitken, M.J., Cook, R.M., Harris, F.H.deB., McInish, T.H., 2006. Market design and execution cost: implications for thinly-traded stocks. Working paper, University of New South Wales.

Bacidore, J.M., Sofianos, G., 2002. Liquidity provision and specialist trading in NYSE-listed non-U.S. stocks. Journal of Financial Economics 63 (1), 133–158.

Barber, B.M., Lyon, J.D., 1996. Detecting abnormal operating performance: the empirical power and specification of test statistics. Journal of Financial Economics 41 (3), 359–399.

Barber, B.M., Lyon, J.D., 1997. Detecting long-run abnormal stock returns: the empirical power and specification of test statistics. Journal of Financial Economics 43 (3), 341–372.

Barclay, M., 1997. Bid-ask spreads and the avoidance of odd-eighth quotes on Nasdaq: an examination of exchange listings. Journal of Financial Economics 45 (1), 35–58.

Battalio, R., Ellul, A., Jennings, R., 2007. Reputation effects in trading on the New York Stock Exchange. Journal of Finance 62 (3), 1243–1271.

Bennett, P., Wei, L., 2006. Market structure, fragmentation and market quality. Journal of Financial Markets 9 (1), 49–78.

Bessembinder, H., 1999. Trade execution costs on Nasdaq and the NYSE: a post-reform comparison. Journal of Financial and Quantitative Analysis 34 (3), 387–407.

Bessembinder, H., 2003. Trade execution costs and market quality after decimalization. Journal of Financial and Quantitative Analysis 38 (4), 747–777.

Bessembinder, H., Kaufman, H., 1997a. A comparison of trade execution costs for NYSE and NASDAQ-listed stocks. Journal of Financial and Quantitative Analysis 32 (3), 287–310.

Bessembinder, H., Kaufman, H., 1997b. A cross-exchange comparison of execution costs and information flow for the NYSE-listed stocks. Journal of Financial Economics 46 (3), 293–319.

Boehmer, E., 2005. Dimensions of execution quality: recent evidence for U.S. equity markets. Journal of Financial Economics 78 (3), 553–582.

Brown, S.J., Warner, J.B., 1980. Measuring security price performance. Journal of Financial Economics 8 (3), 205–258.

Brown, S.J., Warner, J.B., 1985. Using daily stock returns: the case of event studies. Journal of Financial Economics 14 (1), 3–31.

Christie, W.G., Huang, R.D., 1994. Market structures and liquidity: a transactions data study of exchange listings. Journal of Financial Intermediation 3 (3), 300–326.

Chung, K.H., Van Ness, R.F., Van Ness, R.A., 2001. Can the treatment of limit orders reconcile the differences in trading costs between NYSE and Nasdaq issues? Journal of Financial and Quantitative Analysis 36 (2), 267–286.

Chung, K.H., Van Ness, R.F., Van Ness, R.A., 2004. Trading costs and quote clustering on the NYSE and Nasdaq after decimilization. Journal of Financial Research 27 (3), 309–328.

Clyde, P., Schultz, P., Zaman, M., 1997. Trading costs and exchange delisting: the case of firms that voluntarily move from the American Stock Exchange to the Nasdaq. Journal of Finance 52 (5), 2103–2112.

Doidge, C., Karolyi, G.A., Stulz, R.M., 2004. Why are foreign firms listed in the U.S. worth more? Journal of Financial Economics 71 (2), 205–238.

Fama, E.F., 1998. Market efficiency, long-term returns, and behavioral finance. Journal of Financial Economics 49 (3), 283–306.

Foerster, S.R., Karolyi, G.A., 1998. Multimarket trading and liquidity: a transaction data analysis of Canada–US interlistings. Journal of International Financial Markets, Institutions and Money 8 (3–4), 393–412.

Harris, L., 1989. S&P 500 cash stock price volatilities. Journal of Finance 44 (5), 1155–1175.

Heckman, J.J., Ichimura, H., Todd, P.E., 1997. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Review of Economic Studies 64 (4), 605–654.

Heckman, J.J., Ichimura, H., Todd, P.E., 1998. Matching as an econometric evaluation estimator. Review of Economic Studies 65 (2), 261–294.

Heckman, J.J., Smith, J., 1998. Evaluating the welfare state. In: Steinar, Strøm (Ed.), Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium. Econometric Society Monograph Series, pp. 241–318.

Huang, R.D., Stoll, H.R., 1996. Dealer versus auction markets: a paired comparison of execution costs on NASDAQ and the NYSE. Journal of Financial Economics 41 (3), 313–357.

Jain, P.K., Kim, J.-C., 2006. Investor recognition, liquidity, and exchange listings in the reformed markets. Financial Management 35 (2), 21–42.

Kahle, K.M., Walkling, R.A., 1996. The impact of industry classifications on financial research. Journal of Financial and Quantitative Analysis 31 (3), 309–335.

Kedia, S., Panchapagesan, V., 2003. Why do only some Nasdaq firms switch to the NYSE? Evidence from corporate transactions. Working paper, Harvard University.

Lie, E., 2001. Detecting abnormal operating performance: revisited. Financial Management 30 (2), 77–91.

Liu, J.W., 2006. Electronic versus specialist trading: a comparison of trading quality at the Archipelago exchange and the NYSE. Working paper, California State University.

Lyon, J.D., Barber, B.M., Tsai, C.-L., 1999. Improved methods for tests of long-run abnormal stock returns. Journal of Finance 54 (1), 165–201.

Mayhew, S., 2002. Competition, market structure, and bid-ask spreads in stock option markets. Journal of Finance 57 (2), 931–958.

Mitchell, M.L., Stafford, E., 2000. Managerial decisions and long-term stock price performance. Journal of Business 73 (3), 287–329.

Mittoo, U.R., 2003. Globalization and the value of U.S. listing: revisiting Canadian evidence. Journal of Banking and Finance 27 (9), 1629–1661.

Pagan, A., Ullah, A., 1999. Nonparametric Econometrics. Cambridge University Press, Cambridge.

Rubin, D.B., 1973. Matching to remove bias in observational studies. Biometrics 29 (1), 159–183.

Rubin, D.B., 1976. Multivariate matching methods that are equal percent bias reducing, I: some examples. Biometrics 32 (1), 109–120.

Rubin, D.B., 2006. Matched Sampling for Causal Effects. Cambridge University Press, Cambridge.

Silva, A.C., Chavez, G., 2002. Comparison of execution costs: evidence of asymmetric information at the Mexican Stock Exchange. Journal of International Financial Markets, Institutions and Money 12 (3), 253–278.

Stone, C.J., 1977. Consistent non-parametric regression. Annals of Statistics 5 (4), 595–620.

United States Securities and Exchange Commission (SEC), 2001. Report on the comparison of order executions across equity market structures.

Venkataraman, K., 2001. Automated versus floor trading: an analysis of execution costs on the Paris and New York exchanges. Journal of Finance 56 (4), 1445–1485.

Weston, J.P., 2000. Competition on the Nasdaq and the impact of recent market reforms. Journal of Finance 55 (6), 2565–2598.