Using Matched Samples to Test for Differences in Trade Execution Costs*

Ryan J. Davies†
Finance Division
Babson College
Tomasso Hall
Babson Park MA 02457

Sang Soo Kim
Korea Development Bank
16-3 Yeoido-dong
Yeongdeungpo-gu
Seoul, Korea 150-973

May 12, 2007

Abstract

We consider the properties of tests for differences in trade execution costs (e.g. bid-ask spreads) based on matched samples. We consider using various weighting schemes and various combinations of twelve commonly-used matching characteristics. Based on extensive simulation results, we conclude that the best practice for constructing matched samples is to match firms one-to-one based on market capitalization and share price. We demonstrate that pre-sorting by industry groups or eliminating 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 provide guidance on optimal estimation when there are few available matched pairs or large data measurement errors or both. We demonstrate 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 current version of this paper can be found at: http://faculty.babson.edu/rdavies/. Ryan Davies acknowledges financial support from the Babson Faculty Research Fund.

†Tel: +1–781–239–5345. Fax: +1–781–239–5004. email: rdavies@babson.edu

Using Matched Samples to Test for Differences in Trade Execution Costs

We consider the properties of tests for differences in trade execution costs (e.g. bid-ask spreads) based on matched samples. We consider using various weighting schemes and various combinations of twelve commonly-used matching characteristics. Based on extensive simulation results, we conclude that the best practice for constructing matched samples is to match firms one-to-one based on market capitalization and share price. We demonstrate that pre-sorting by industry groups or eliminating 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 provide guidance on optimal estimation when there are few available matched pairs or large data measurement errors or both. We demonstrate 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.

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 bid-ask 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.

The market microstructure literature has matched firms using characteristics such as market capitalization, industry, trading volume, trading dollar volume, number of trades, price volatility, share price, book value, price-to-book ratio, beta, and the number of shares outstanding. Table 1 provides an overview of some of the combinations of matching characteristics used in past research. In this paper, we identify which combination of matching characteristics provides 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 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 also show that better test properties result from using a nonparametric Wilcoxon signed rank test than a parametric Student t-test.

We also investigate whether using different nearest-neighbor weighting schemes or using kernel-based matching estimation can improve the size-power properties of tests for dif-

[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. Similar matched sample approaches are used by Keim and Madhavan (1997), Bessembinder and Kaufamn (1997a, 1997b), LaPlante and Muscarella (1997), Bessembinder (1999, 2003), Weston (2000), Venkataraman (2001), SEC (2001), Bacidore and Sofianos (2002), Boehmer (2003), Chung and Chuwonganant (2004), Chung et al. (2001, 2004), Aitken et al. (2006), Battalio et al. (2006), and many others, to study the effect of changes in exchange rules and to compare execution costs across trading platforms.

ferences in trade execution costs. Essentially, these alternative estimators weight closest neighbors by more, but still place some weight on more distant neighbors. The potential benefit is that these estimators are less sensitive to a mis-match along un-measured dimensions, but the cost is that they introduce an added mis-match along measured dimensions. While these more complicated 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. The nearest neighbor approach has less bias and less probability of type I error (rejecting the null hypothesis when the null hypothesis is true) than the alternative estimators that we consider. For instance, we find that a one-to-one matching approach has virtually no size distortion; whereas we find that a three-to-one matching approach has empirical size of 14.4% for a test with nominal size of 5%. When the number of matched pairs is small, however, additional test power may be obtained by using a matching technique that places additional weight on more distant firms (e.g. kernel-based matching estimates).

Finally, we provide evidence that matched sample estimation (which uses all available firms) can produce tests with more statistical power 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, 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).[2] The methodology used in our paper is also similar in spirit to that used in Brown and Warner (1980, 1985). In large part, these studies were concerned with potentially severe cross-sectional dependence across firms and with controlling for differences

[2]Kahle and Walkling (1996) explore how the ability to detect abnormal performance varies between tests using matched samples based on firm size only and those based on firm size and industry classification. They show that industry-matching increases test power and that the actual database source of the industry classifications matters. Lie (2001) shows that test statistics of abnormal performance can be severely biased if control firms are drawn from firms with pre-defined characteristics.

in historical performance and returns. These issues are not of particular relevance to most market microstructure applications. For instance, bid-ask spreads do not exhibit significant cross-sectional dependence (except during major market crashes) and spreads do not have sufficiently long memories to introduce biases from historical spreads. Thus, the focus of our paper is much different from this past work.

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 nonlinear relation between bid-ask 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 test size-power tradeoffs without direct simulation results. Our results thereby provide insights into which matching techniques are appropriate in this important, widely-used context.

A main focus of past market microstructure research has been detecting differences between the bid-ask 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. But, other important differences in listing status will always occur and these differences are often best studied using the matched sample approach. For instance, exchange regulators frequently 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). 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 in measuring differences in trade execution costs include studies on the impact of changes in specialist firms, the impact of the introduction of new derivative products (e.g. single stock futures), and the impact of the introduction of new trading venues (e.g. cross-listing on an international exchange or a new alternative trading system).

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, the differences in bid-ask spreads caused by differences in listing status will 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.[3] They use matched sample estimation to compare the mean post-programme earnings of job-training programme participants with the mean earnings of comparable non-participants. In this section, we adopt the notation of Heckman et al. to show how matched sample estimation can be used to compare the trade execution costs (e.g. percentage bid-ask spread) of firms whose listing status differs. For example, we might be interested in comparing the bid-ask spreads of NYSE-listed firms with the spreads of otherwise comparable Nasdaq-listed firms. Alternatively, we might be interested in comparing the bid-ask spreads of firms cross-listed on the Toronto Stock Exchange and a US exchange with the spreads of firms listed in Toronto only. We attribute differences in the trading properties of these two groups to the difference in listing status.

Let Y_A denote the trading property outcome (e.g. trade execution costs) that would occur if the security has listing status A (e.g. the firm is listed on NYSE) and let Y_0 denote the trading property outcome that would occur if the security does not have listing status A (e.g. the firm is listed on Nasdaq). Let $D = 1$ if the firm has listing status A; $D = 0$ otherwise. The trading property outcome observed for a firm is $Y = DY_A + (1 - D)Y_0$. The difference

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

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 trading property outcome and the $\mathbf{Z}$ variables determine whether the firm decides to change its listing status. In practice, a firm's characteristics often have an impact on both its listing decision and its trading properties. For example, a firm that conducts a lot of business outside of its domestic market may be more likely to interlist to increase product awareness in its foreign markets and may also be more likely to have larger bid-ask spreads, reflecting the added uncertainty of competing in multiple jurisdictions. The trading property of a firm with listing status A can be written as a function of observables $(\mathbf{T})$ and unobservables U_A, where

$$Y_A = g_A(\mathbf{T}) + U_A, \qquad E(U_A) = 0$$

and g_A is assumed to be a non-stochastic function. Similarly, the trading property of a firm without listing status A can be written as

$$Y_0 = g_0(\mathbf{T}) + U_0, \qquad E(U_0) = 0.$$

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 trading property 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 of having listing status A is given by:

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

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 $i \in \mathbf{I}_A$, trading property Y_{Ai} is compared to an average of the outcomes 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 a trading property for each firm i with listing status A is

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

where $W(i,j)$ is a positive valued weight function. The weighting function assigns weights to the trading properties 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 and/or different distance measures.

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

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

Let ϕ_i indicate the set of firms matched to firm $i \in \mathbf{I}_A$. We consider the following alternative matching estimators:

Nearest neighbor (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 4.3 describes the distance measure used for the multivariate case. ϕ_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 $j \in \phi_i$ and equals 0 otherwise.

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

3-NN matching estimator with triangular weights: Now, ϕ_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.

Kernel-based matching estimators: As k increases, k-NN matching estimators become effectively Nadaraya-Watson kernel-based matching estimators. Univariate kernel-based matching estimates based on characteristic $x \in \mathbf{X}$ are constructed as follows. Kernel matching sets $\phi_i = \mathbf{I}_0$ and defines

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

where $K_{ik} = K((x_i - x_k)/h)$ is a kernel function and h is a bandwidth parameter. We use a kernel based on the standard normal density function,

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

To investigate the sensitivity of predictions to the bandwidth parameter, we first consider two bandwidth parameters: $h_1 = 1.059 s_x N_0^{-1/5}$ and $h_2 = 1.059 s_x N_0^{-1/3}$; where

$$s_x = \sqrt{\frac{1}{N_0}\sum_{j \in \mathbf{I}_0} x_j^2 - \left(\frac{1}{N_0}\sum_{j \in \mathbf{I}_0} x_j\right)^2}.$$

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. The choice of h_1 is motivated by the desire to minimize the approximate mean-integrated squared error. This essentially equates the trade-off between the bias and the variance of the kernel estimate (see Pagan and Ullah, p. 24, 1999). The choice of h_2 is motivated by the observation that when constructing bootstrap confidence intervals, as the number of bootstrap replications increases, the variance of the kernel estimate from a single replication becomes less important. Thus, it may be optimal to minimize the potential bias of the estimate, rather than jointly minimize the bias and variance of the estimate. The bias of a kernel estimate is minimized by a bandwidth parameter that is proportional to $n^{-1/3}$, instead of the usual $n^{-1/5}$ (see Davison and Hinkley, p. 228, 1997).

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 nearest neighbor. The variable bandwidth parameter may be better suited to account for the significant heterogeneity in the 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).

3 Data

Our main Monte Carlo 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. These potential firm matching characteristics are as of December 31, 2003. Our sample excludes stocks with incomplete matching characteristics 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 sample provides a representative sample of stocks with market capitalization ranging from $649 million to $311,755 million (average = $9,160 million), and beginning share prices ranging from $2.83 to $98.73 (average = $32.94).

We obtain trades and inside quote revisions from the 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: $200 \times (ask - bid)/(ask + bid)$. Unless otherwise indicated, the average percentage bid-ask spread is our outcome parameter of interest. Table 2 also reports the trade-weighted effective percentage spread, defined as $200 \times |TradePrice - 0.5 \times (ask + bid)|/(ask + bid)$, which is used later in our robustness checks. Using TAQ, for each stock we also calculate trading volume, trading dollar volume, and the number of trades during December 2003. These are used as potential additional matching characteristics.

4 Monte Carlo Simulation

In this section, we use Monte Carlo simulation to determine the size and power properties of tests for differences in trade execution costs based on different matched sample estimation approaches. We proceed as follows. For each of 20,000 Monte Carlo replications:

1. We randomly shuffle and divide the sample of $N_T = 1000$ 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, $i \in \mathbf{I}_A$, are 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 estimation 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 $j \in \mathbf{I}_0$.).[4]

3. Based on a comparison of the N_A stocks with induced differences and their hypothetical matched pairs, we then construct a two-sided nonparametric Wilcoxon signed rank test of size α of the difference in bid-ask spreads between the two groups.

The empirical test power for a given induced difference (θ) is the share of the Monte Carlo replications with a significant test statistic. The empirical test size is the share of Monte Carlo replications with a significant test statistic when there is no induced difference ($\theta = 0$). A well-specified test will have an empirical test size equal its theoretical (nominal) size of α. We use this Monte Carlo approach to answer many of the outstanding questions concerning the optimal usage of matched sample estimation in market microstructure research.

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

[4]Some researchers use 1-NN matching techniques that require each matched stock to be used only once. Clearly, this restriction is not desirable in the context of k-NN or kernel-based matching estimates. We explore this restriction in the context of 1-NN matching estimates in section 4.7.

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 7.51bp, or put another way, it equals a spread difference of 2.25 cents for the median priced stock.

Before proceeding to the results, it is important to highlight why our simulations are based on firms listed on a single exchange. One might be tempted to conduct a simulation based on a larger merged sample comprising of stocks from multiple exchanges. Unfortunately, this approach would lead to erroneous results. To see why, recall that the *listing decision* of firms is also 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 influencing 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 approach 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 $\mathbf{T}$ (characteristics that determine bid-ask 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 $\mathbf{Z}$ (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 Monte Carlo 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.[5] 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 different time periods.

We now proceed to our results. We begin by matching firms based on their market capitalization. Market capitalization is the most commonly used matching criterion, in part because it is readily available in databases such as CRSP, Datastream, and Compustat; and in part because it is highly correlated with other factors which contribute to bid-ask 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.

4.1 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.[6] We now investigate whether matching estimates that smooth over several firms are better or worse than estimates based on one-to-one matches. To answer this, figure 1 compares the size and power of tests based on 1-NN, 2-NN (equal weights) and 3-NN (triangular weights) estimates for different levels of induced differences, θ, in the percentage bid-ask spread. A clear pattern emerges: As nearest neighbor 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),

$$Y_i = g(X_i) + U_i.$$

[5]Barber and Lyon (1996) results 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, page 371). It is likely that this set of firms changes on a yearly basis, thereby largely alleviating independence concerns.

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

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

$$\hat{\Delta}(\tilde{X}) = g(\tilde{X}) + \theta - \sum_i W_i g(X_i) + \tilde{U} - \sum_i W_i U_i \tag{1}$$

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 one-to-one 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 nearest neighbor 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.

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.

Before proceeding to explore the bias-variance trade-off further, it is useful to 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 the random match 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, figure 1 illustrates that if spreads differ by $0.4 s_y$ (≈3.0bp in our sample), then tests based on a random match will detect the difference about 84.8% of the time; whereas tests based on 1-NN matching estimates will detect the difference 99.5% of the time.

Figure 2 illustrates the size and power of tests based on kernel-based matching estimates with bandwidth parameters h_1, h_2, and h_i^{min} (variable bandwidth) for different levels of induced changes, θ, in the percentage bid-ask spread. For comparison purposes, we also include the previous 1-NN and random match results. We have two main observations. Our first observation is that the theoretical bandwidth parameters (h_1 and h_2) appear much too wide – a bandwidth that is sufficiently large to obtain matches for the largest firms (which have large differences in market capitalization) causes over-smoothing in matching estimates of the smaller firms. As a result, the power curves of the kernel estimates are biased to the right, much in the same way as increasing k did for the k-NN estimates. The size distortion, however, is much more severe: tests based on the kernel-based estimates using the h_1 and h_2 bandwidth parameters have empirical size of over 40%. In fact, for small positive induced differences, there is a region in which the simplistic random match approach actually performs better than the kernel-based matching estimate with bandwidth parameters h_1 and h_2. Our second observation is that tests based on the variable bandwidth kernel estimates have much less size distortion than tests based on fixed bandwidth kernel estimates. 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).

Figure 3 considers how test power changes with the number of matched pairs (N_A). We consider two levels of induced differences: high ($\theta = 0.5$) and low ($\theta = 0.2$). The results are based on 20,000 Monte Carlo replications in which the induced difference is θs_y for 50% of the replications and $-\theta s_y$ for 50% of the replications. As expected, the power of all of the tests increases with the number of matched pairs.

When the induced difference is large, each of the five alternative estimates provide *higher* power than the standard nearest neighbor estimate. The potential importance of the choice of bandwidth parameter is highlighted by the observation that the $h_1 = 1.059 s_x N_0^{-1/3}$ kernel-based matching estimate has the highest power, while the $h_2 = 1.059 s_x N_0^{-1/5}$ kernel-based matching estimate has the second worst power. Overall, we conclude that in the situation in which 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), it may be desirable to obtain more testing power by using a matching estimate that places more weight on more distant firms. Note, however, that in this case the gain in test power is relatively small in comparison with the dramatic increase in size distortion.

The figure also illustrates the power graphs for a smaller induced difference of $\theta = 0.2$. All of the matching estimation techniques have much less power than they did in the case of a large induced difference. For example, with a small induced difference and 25 matched pairs, the empirical rejection rate is just 40–50%. In this context, the 2-NN and 3-NN nearest neighbor and the h_i^{min} kernel-based matching estimates provide higher power than the standard 1-NN nearest neighbor approach. But, additional smoothing does not guarantee higher power, as evident by the poor test power of the h_1 and h_2 estimates. These results can be understood by referring back to figures 1 and 2, where we observe that, in comparison with the unbiased 1-NN nearest neighbor estimates, the biased smoothing estimates have lower power for $+0.2s_y$ ($\approx$1.5bp in our sample) and higher power for $-0.2s_y$. Thus, the overall power depends on the extent to which these differences in power are offsetting.

4.2 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 the following ten commonly suggested alternative matching characteristics: share price, book value, price to book ratio, 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.

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).

4.3 Is it better to match using multiple firm characteristics?

We now consider whether test properties can be improved by matching over multiple firm characteristics in addition to, or instead of, market capitalization. A common approach is to 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:

$$\text{argmin}_{j \in I_0} \ \sum_k \left((2(x_{Ai}^k - x_{0j}^k))/(x_{Ai}^k + x_{0j}^k) \right)^2,$$

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. As we previously reported in table 1, a vast variety of combinations of firm characteristics have been used in past research.

In table 4, we consider the test power of employing eleven of the most commonly-used combinations of matching characteristics. Specifically, we study matches based on: (i) market capitalization alone; (ii) market capitalization and share price; (iii) market capitalization, share price, and number of shares outstanding; (iv) market capitalization, share price, and beta; (v) market capitalization, share price and volatility; (vi) market capitalization, share price, and trading volume; (vii) market capitalization, share price and dollar trade volume; (viii) market capitalization, share price, trading volume, and volatility; (ix) market capitalization, share price, number of trades, average trade size, and volatility; (x) share price, number of trades, average trade size, and volatility; and (xi) trading volume, beta, and volatility.

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 $+0.1s_y$ ($\approx$0.75bp in our sample) by almost two-fold 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 combinations provide higher test power for detecting differences of $+0.1s_y$. This higher test power is caused by a bias in the power curve (the curve is shifted slightly to the left in these cases); thus, most of these methods also provide lower power test power to detect small negative differences in spreads.

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. For instance, 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.[7]

[7]Matched sample estimation may produce biased results when the available sample of stocks with the listing status of interest is not representative of stocks that have a propensity to obtain this listing status. Ideally, we could use a propensity score adjustment to correct for this deficiency if enough firms changed their listing status during our sample period and if there was an accurate model of the listing decision. Unfortunately, researchers rarely have either. In particular, a model of the listing decision would need to be

Given these advantages and given the wide accessibility of reliable market capitalization data, we recommend matching based on market capitalization and share price, unless the specific research context suggests there are benefits to matching over additional characteristics.

4.4 Should we restrict matches to firms within the same industry?

A common practice is to sort on industry-code first and then construct nearest-neighbor matching estimates. Intuitively, a matching technique that sorts on industry-code first will have less-precise matches on the market capitalization characteristic for any industry in which the distribution of firm market capitalizations differs between the industry and the full population. To investigate the effect of industry-based matching, we consider restricting matches to within the same 2-digit SIC code. From our original sample, the smallest number of firms within the same SIC category is 2 (agricultural production crops) and the largest number is 74 (electric, gas, and sanitary services).

Figure 4 shows that this simple industry-sorting technique has virtually no effect on our results when firms are drawn from an original sample of 1,000. In other words, our full sample contains enough firms that pre-sorting by industry has little effect on the distribution of potential matches. That said, if there are a smaller number of potential firms to match, pre-sorting by industry can have a dramatic negative impact on test power. To illustrate this, we consider the case in which $N_T = 200$. In this case, 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.

4.5 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

sufficiently well identified such that it incorporates firm characteristics that have implications for the listing decision but do not affect trade execution costs. This is not possible with the data available in most studies. For example, while Doidge, et al. (2004) provide a model of the decision to cross-list on a US exchange, the factors included in their model are not independent (in general) of the factors that affect bid-ask spreads.

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 market microstructure literature is to require matches to occur between firms that satisfy the restriction

$$\left|2(x_{Ai}^k - x_{0j}^k)/(x_{Ai}^k + x_{0j}^k)\right| < \varepsilon,$$

for each matching characteristic k 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 5, 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 these tolerance levels makes little 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 consider an alternative caliper matching approach in which the tolerance level varies across firms depending on the closeness of potential matches. Specifically, the tolerance level for a potential match for firm i is defined as a share of the variance of $(x_{Ai} - x_{0j})/(x_{Ai} + x_{0j})$ across all potential matches $j \in \mathbf{I}_0$. As figure 5 illustrates, this caliper matching approach reduces test power at these tolerance levels.

In summary, the main problem with caliper matching is that it is difficult, a priori, 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. 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, eliminating the poor match

has little impact on a nonparametric 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 nonparametric test. Thus, since we 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.

4.6 Should we use a nonparametric test or parametric test?

All of our reported results so far have been based on a nonparametric Wilcoxon signed rank test. Figure 6 verifies that the Wilcoxon test has much higher power than the most commonly-used parametric alternative – a (dependent) Student t-test of the difference between the matched pairs. Thus, we conclude that in this context the Wilcoxon signed rank test should be used instead of a parametric test. Intuitively, the distribution of bid-ask spreads is non-symmetric, bounded by zero, and has large positive outliers; thus it is not well-fitted by the normal distribution implied by the Student t-test.

4.7 Should we allow the same firm to be matched multiple times?

When using a one-to-one nearest neighbor approach, some researchers use a matching *without replacement* approach 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. In contrast, all of our reported results so far have been based on matching *with replacement* (a firm can be the nearest neighbor for multiple firms). We now investigate, in the context of 1-NN matching based on market capitalization, the test properties of matching with replacement compared with those of matching without replacement. Table 6 compares the two approaches when $N_T = 1000$, $N_T = 200$, and $N_T = 100$. First, we observe that when we use our entire original sample ($N_T = 1000$), the results 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 findings reported

in this paper would be the same if conducted under a matching without replacement regime.

We also find almost identical results for the two methods when $N_T = 200$. With this sample size, the matching with replacement method has slightly more power for small induced differences, and slightly less power for large induced differences. Finally, we consider the extreme case in which $N_T = 100$, such that $N_A = N_0 = 50$. In this case, when matching without replacement, every firm in the sample must be in a matched pair. Here, we observe that matching with replacement provides better test power than matching without replacement. In this case, the reduction of power that arises when matches are forced to occur between non-closest firms dominates any possible increase in power from reducing the potential influence (through multiple matches) that one or more outlier firms can have on the estimates. That said, neither method works well in this extreme case: both methods have high size distortion.

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.[8] Ideally, event studies provide a natural experiment that allows for *direct* measurement of the impact of changing listing status on trade execution costs.

Despite the obvious advantages of its direct approach, there are some potential problems with using an event study in this context.[9] 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. And, 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

[8]For example, Christie and Huang (1994), Barclay (1997), and Clyde, et al. (1997) use an event study approach to examine the impact on spreads of stocks that switch their US exchange listing.

[9]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.

issues are not the focus of this paper. Here, we are interested in another potential serious limitation of an event study: the sample of stocks that change their listing status in any given year may be small. For instance, Kedia and Panchapagesan (2003) report that 460 firms switched their listing status from the Nasdaq to the NYSE over the period 1986 to 1998. Thus, in any given year, on average fewer than 40 firms switch their listing status from Nasdaq to the NYSE and there are virtually no *voluntary* switches from NYSE to Nasdaq.[10] The number of firms switching exchanges is much smaller 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 Monte Carlo 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, the average percentage bid-ask spread during the 30 trading days prior to the event date is denoted by Y_i^b and the average percentage bid-ask spread during the 30 trading days after the event date is denoted by Y_i^a.

4. An artificial difference of θ is induced in the post event date spread: $\hat{Y}_i^a = Y_i^a + \theta$ for one half of the replications and $\hat{Y}_i^a = Y_i^a - \theta$ for the other half of the replications.

5. We conduct a two-sided nonparametric Wilcoxon signed rank test of size α of the difference between the sample $\left[Y_1^b, Y_2^b, \ldots, Y_N^b\right]$ and the corresponding sample $\left[\hat{Y}_1^a, \hat{Y}_2^a, \ldots, \hat{Y}_N^a\right]$.

[10] Other event studies of listing status changes have also used small numbers of firms. Over a four year period, Jain and Kim (2006) study 174 US-listed stocks that switched their listing prior to decimalization of NYSE, AMEX, and Nasdaq and the 93 stocks that switched after decimalization. Bennett and Wei (2006) study 39 stocks that switch 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.

Figure 7 provides the Monte Carlo results. As expected, the event study approach has more power than the matched sample approach *for a given sample size*. But, the figure 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 smaller than the number of matched pairs available.

It is important to recognize that our Monte Carlo simulation cannot fully capture the complex underlying order flow dynamics around the listing event date.[11] 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 conduct a broad matched sample estimation approach instead.

4.9 Robustness of results

How do our findings change when different sub-samples are used? To gain further insights, we run our Monte Carlo simulations using four sub-samples of the original sample of 1,000 NYSE stocks: (i) the largest 500 stocks by market capitalization; (ii) the smallest 500 stocks by market capitalization; (iii) the 500 stocks with the highest share price; and (iv) the 500 stocks with the lowest share price. Table 7 reports the corresponding test power based on matching with market capitalization and matching with market capitalization and share price. 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. The results for all four sub-samples confirm our earlier finding that matching with market capitalization and share price has greater overall test power than matching with market capitalization alone.

How does test power differ across different time periods? Table 8 considers

[11] In the context of detecting abnormal returns, Boehmer, Musumeci, and Poulsen (1991) show that event studies may be severely affected by event-induced variance in returns. It seems reasonable, therefore, to suppose that a change in listing status may cause transitory changes in the bid-ask spread not related to the long-term impact of the change in listing status.

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 2005. 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 48.4bp in 1993 to 6.4bp in 2005. 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.

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.

How do these induced differences compare to the differences found in past studies? To illustrate, we consider three highly-cited articles by Venkataraman (2001),

Bessembinder (1999) and Weston (2000).

Venkataraman (2001) constructs a matched sample of 40 stocks listed on the NYSE and Paris Bourse over the period April 1997 to March 1998. When the NYSE tick-size is a sixteenth, he finds that the difference between the time-weighted percentage quoted spreads of the two samples is between 1.59bp and 2.65bp. Referring to our simulation results in table 8, this corresponds to an induced difference of less than $\theta = 0.1$ in the 1997–1998 period, suggesting that the test power in this case might have been as low as 15% to 45%. This highlights the importance of the other steps taken in Venkataraman (2001) to verify significance, such as conducting a bootstrap simulation and using different combinations of matching characteristics.

Bessembinder (1999) uses matched samples to find a statistically significant difference of 24.8bp between NYSE and Nasdaq percentage bid-ask spreads during the post-reform period of July to December 1997. Bessembinder (1999) uses a t-test (rather than a more powerful non-parametric test) and uses market capitalization as the sole matching characteristic (rather than matching with market capitalization and share price). But, our previous results suggest that the impact of these choices is likely to be swamped by the large number of matched pairs (between 136 and 539 pairs) and the very large difference in spreads (corresponding to about $\theta = 1$ in table 8). To verify, we calculate time-weighted average percentage bid-ask spreads for the largest 1,000 NYSE stocks during July-December 1997 and repeat our simulation analysis for an induced difference of -24.8bp and 150 matched pairs. As expected, we find that test power is close to 100% for both a t-test based on matching with market capitalization and a Wilcoxon test based on matching with market capitalization and share price.

Weston (2000) explores the phased introduction of new Nasdaq order-handling rules, by constructing 88 matched pairs of Nasdaq and NYSE stocks, during the period October 1996 to June 1997. He further divides these matched pairs into dollar volume quartiles of 22 matched pairs each. In the 90 days after the implementation of the new rules, he finds that the difference in quoted spreads is about 2.8 cents for the entire sample, and that Nasdaq spreads are larger than NYSE spreads for the smallest two dollar volume quartiles, but NYSE spreads are larger than Nasdaq spreads for the largest dollar volume quartile. In this time pe-

riod, a difference of 2.8 cents roughly corresponds to the small induced difference in figure 3. While Weston (2000) does not report a test statistic for the difference between NYSE and Nasdaq, our results in figure 3 suggest that a Wilcoxon test of the difference would have test power of about 40% for the quartile-based results. Weston (2000) is careful to confirm the robustness of his results by using three different combinations of matching characteristics, and by running regressions to investigate the sensitivity of his results to the matching procedure.

Our intention is not to cast doubt on the results in Venkataraman (2001), Bessembinder (1999), and Weston (2000), but rather to highlight when the choice of matching technique might influence the test outcome in commonly-studied contexts. Overall, our results suggest that test power and potential type II errors are more important than many researchers in market microstructure typically assume.

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 9 presents results for Nasdaq stocks that are analogous to those presented in table 4 for NYSE stocks. The magnitudes of the empirical power and size reported in tables 4 and 9 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, we observe that matching on market capitalization and share price jointly provides superior results to matching on market capitalization alone. And, again we find that adding the number of shares outstanding, CAPM beta, or volatility as a matching criteria to a matching regime that already includes market capitalization and price provides little or no benefit. Again, we also find that 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 (TSE) in 1998. These results are available from the authors upon request. The TSE sample contained fewer firms than our NYSE and Nasdaq samples, and contained several large firms that did not have apparent suitable matches. While the heterogeneity of the TSE 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 followed.

Are our findings similar for other spread measures? To investigate, we calculate a trade-weighted effective percentage spread, defined as $200 \times |TradePrice - 0.5 \times (ask + bid)|/(ask + bid)$, for each stock. We then re-run our simulations with effective spreads and find test power patterns similar to those previously reported. In general, matched sample tests for differences in effective spreads have slightly less power than the counterpart using percentage bid-ask spreads. Figure 8 provides an illustrative example based on one-to-one matching. As the figure reveals, the power of tests for differences in effective spreads is somewhat lower than those of percentage spreads, but the same pattern emerges: matching based on market capitalization and share price jointly provides additional testing power over matching based on market capitalization alone.

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. And, 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 also show that in the context of comparing bid-ask spreads, a non-parametric Wilcoxon signed rank test provides better test properties than a parametric Student t-test. As well, we 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.

Our simulation results also provide important insights about the impact of smoothing estimates over multiple firms. 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 important impact on the test outcome. Finally, numerous robustness checks have shown that our main findings are widely applicable and relevant to studies using different exchanges, different sized samples, different spread measures, and different time periods.

References

Affleck-Graves, J., S.P. Hedge, R.E. Miller, 1994. Trading Mechanisms and the Components of the Bid-Ask Spread. Journal of Finance 49(4), 1471–1488.

Ahn, H-J., C.Q. Cao, H. Choe, 1998. Decimalization and competition among stock markets: Evidence from the Toronto Stock Exchange cross-listed securities. Journal of Financial Markets 1(1), 51–87.

Ang, J.S., J.C. Brau, 2002. Firm transparency and the costs of going public. Journal of Financial Research 25(1), 1–17.

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

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

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

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

Barber, B.M., J.D. Lyon, 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.

Baruch, S., G.A. Karolyi, M.L. Lemmon, 2004. Multi-Market Trading and Liquidity: Theory and Evidence. Journal of Finance, forthcoming.

Battalio, R., A. Ellul, R. Jennings, 2006. Reputation effects in trading on the New York Stock Exchange. Journal of Finance, forthcoming.

Bennett, P., L. Wei, 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., H. Kaufman, 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., H. Kaufman, 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., J. Musumeci, A.B. Poulsen, 1991. Event-study methodology under conditions of event-induced variance. Journal of Financial Economics 30(2), 253–272.

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

Brown, S.J., J.B. Warner, 1985. Using daily stock returns: The Case of Event Studies. Journal of Financial Economics 14(1), 3–31.

Chowhdry, B., V. Nanda, 1991. Multimarket trading and market liquidity. Review of Financial Studies 4(3), 623–656.

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

Chung, K.H., C. Chuwonganant, 2004. Tick size, order handling rules, and trading costs, Financial Management 33(1), 47–62.

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

Clyde, P., P. Schultz, M. Zaman, 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.

Davison, A.C., D.V. Hinkley, 1997. Bootstrap Methods and their Application. Cambridge University Press.

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

Domowitz, I., J. Glen, A. Madhavan, 1998. International cross-listing and order flow migration: Evidence from an Emerging Market. Journal of Finance 53(6), 2001–2028.

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

Foerster, S.R., G.A. Karolyi, 1993. International listings of stocks: The case of Canada and the US. Journal of International Business Studies 24(4), 763–784.

Foerster, S.R., G.A. Karolyi, 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.

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

Huang, R.D., H.R. Stoll, 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. J-C. Kim, 2006. Investor Recognition, Liquidity, and Exchange Listings in the

Reformed Markets, Financial Management 35(2), 21-42.

Jiang, C.X., J.-C. Kim, 2005. Trading costs of non-U.S. stocks on the New York Stock Exchange: The effect of institutional ownership, analyst following, and market regulation. Journal of Financial Research 28(3), 439–459.

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

Karolyi, A.G., 1998. Why do companies list shares abroad?: A survey of the evidence and its managerial implications. Financial Markets, Institutions, and Instruments 7(1), 1–60.

Kedia, S., V. Panchapagesan, 2003. Why do only some Nasdaq firms switch to the NYSE? Evidence from Corporate Transactions, Working Paper, Harvard University.

Keim, D.B., A. Madhavan, 1997. Transactions costs and investment style: An inter-exchange analysis of institutional equity trades, Journal of Financial Economics, 46(3), 265–292.

LaPlante, M., C.J. Muscarella, 1997. Do institutions receive comparable execution in the NYSE and Nasdaq markets? A transaction study of block trades. Journal of Financial Economics, 45(1), 97–134.

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

Mitchell, M.L., E. Stafford, 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.

Neal, R., 1992. A Comparison of Transaction Costs between Competitive Market Maker and Specialist Market Structures, Journal of Business 65(3), 317–334.

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

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

Silva, A.C, G. Chavez, 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.

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.

Table 1: **Matching characteristics used in selected past studies.**

Study	**Matching characteristics**
Affleck-Graves, et al. (1994)	Market capitalization, share price, trade volume, volatility
Aitken, et al. (2006)	Market capitalization, share price, total value of trades
Bacidore and Sofianos (2002)	Market capitalization, share price, volatility (intraday and overnight)
Battalio, et al. (2006)	Market capitalization, share price, trade volume, volatility
Bessembinder (1999, 2003)	Market capitalization
Bessembinder and Kaufman (1997a, 1997b)	Market capitalization
Boehmer (2003)	Market capitalization, share price, adjusted daily dollar volume, daily relative price range
Chung, et al. (2001)	Share price, number of trades, trade size, return volatility
Chung, et al. (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

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. 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 $200 \times |TradePrice - 0.5 \times (ask + bid)|/(ask + bid)$.

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
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
Effective percentage bid-ask spread (2004, bp)	2.91	85.18	9.64	8.05

Table 3: **Empirical power and size of tests based on different matching characteristics using the nearest neighbor (1-NN) approach.** 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 are then matched with one of the remaining 950 firms based on the specified matching characteristic; (iii) We then conduct a two-sided nonparametric 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 test power is the share of Monte Carlo replications with a significant test statistic.

	Induced Difference (θ)				
Matching characteristic	-0.2	-0.1	0	$+0.1$	$+0.2$
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
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

Table 4: **Empirical power and size of tests based on different matching characteristics using the nearest neighbor (1-NN) approach.** When using multiple matching characteristics, for each firm $i \in I_A$, we select firm $j \in I_0$ that solves $\text{argmin}_{j \in I_0} \sum_k \left((2(x_{Ai}^k - x_{0j}^k))/(x_{Ai}^k + x_{0j}^k)\right)^2$, 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 are then matched with one of the remaining 950 firms; (iii) We then conduct a two-sided nonparametric 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 test power is the share of Monte Carlo replications with a significant test statistic.

	Induced Difference (θ)				
Matching characteristic(s)	-0.2	-0.1	0	$+0.1$	$+0.2$
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
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 capitalization, 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
Volume, beta, volatility	0.444	0.155	0.052	0.159	0.432

Table 5: **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 $\varepsilon = \infty$, $\varepsilon = 1$, $\varepsilon = 0.75$, and $\varepsilon = 0.5$. 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 are 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, $\left|2(x_{Ai}^k - x_{0j}^k)/(x_{Ai}^k + x_{0j}^k)\right| < \varepsilon$, for any of the matching characteristics, the match is removed from consideration; (iv) We then conduct a two-sided nonparametric Wilcoxon signed rank test of size $\alpha = 0.05$ of the difference in bid-ask spreads between the remaining matched pairs. The empirical test power is the share of Monte Carlo replications with a significant test statistic.

	Tolerance	Induced Difference (θ)				
Matching characteristic(s)	level	-0.2	-0.1	0	$+0.1$	$+0.2$
Market capitalization	∞	0.720	0.252	0.059	0.371	0.816
	1	0.710	0.247	0.056	0.362	0.820
	0.75	0.713	0.252	0.056	0.380	0.822
	0.5	0.713	0.252	0.056	0.380	0.822
Market capitalization, price	∞	0.926	0.464	0.056	0.548	0.953
	1	0.924	0.444	0.052	0.526	0.949
	0.75	0.924	0.445	0.052	0.526	0.949
	0.5	0.923	0.449	0.052	0.539	0.949
Market capitalization, price, number of shares	∞	0.933	0.448	0.052	0.604	0.965
	1	0.931	0.428	0.054	0.601	0.965
	0.75	0.930	0.425	0.054	0.605	0.965
	0.5	0.932	0.430	0.054	0.601	0.965

Table 6: **Comparison of matching with replacement and matching without replacement.** Based on 50 matched pairs and $N_T \in \{1000, 200, 100\}$ as indicated. When $N_T = 200$ or $N_T = 100$, each replication N_T 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 are then matched based on market capitalization with one of the remaining firms without a modified bid-ask spread (when matching without replacement, once a firm has been used in a match it cannot be used again); (iii) We then conduct a two-sided nonparametric 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 test power is the share of Monte Carlo replications with a significant test statistic.

		Induced Difference (θ)				
N_T	Matching Method	-0.2	-0.1	0	$+0.1$	$+0.2$
1000	With replacement	0.720	0.252	0.059	0.371	0.816
	Without replacement	0.721	0.248	0.053	0.352	0.814
200	With replacement	0.759	0.321	0.075	0.333	0.740
	Without replacement	0.776	0.307	0.052	0.303	0.773
100	With replacement	0.753	0.351	0.128	0.367	0.714
	Without replacement	0.645	0.324	0.148	0.314	0.646

Table 7: **Empirical power and size of tests for different sub-samples.** 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 $\text{argmin}_{j \in I_0} \sum_k^2 \left((2(x_{Ai}^k - x_{0j}^k))/(x_{Ai}^k + x_{0j}^k) \right)^2$, 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 are then matched with one of the remaining 450 firms; (iii) We then conduct a two-sided nonparametric 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 test power is the share of Monte Carlo replications with a significant test statistic.

		Induced Difference (θ)				
Sub-sample	Matching characteristic(s)	-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

Table 8: **Empirical power and size of tests across time.** 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 $\text{argmin}_{j \in I_0} \sum_k^2 \left((2(x_{Ai}^k - x_{0j}^k))/(x_{Ai}^k + x_{0j}^k)\right)^2$, 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 are then matched with one of the remaining 950 firms; (iii) We then conduct a two-sided nonparametric 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 test power is the share of Monte Carlo replications with a significant test statistic.

	s_y	$s_y \times \overline{P}$	Induced Difference (θ)				
Year	(in bp)	(in cents)	-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

Table 9: **Empirical power and size of tests based on different matching characteristics using the nearest neighbor (1-NN) approach with a 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 $\text{argmin}_{j \in I_0} \sum_k \left((2(x_{Ai}^k - x_{0j}^k))/(x_{Ai}^k + x_{0j}^k) \right)^2$, 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 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 are then matched with one of the remaining 950 firms; (iii) We then conduct a two-sided nonparametric 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 test power is the share of Monte Carlo replications with a significant test statistic.

	Induced Difference (θ)				
Matching characteristic(s)	-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 capitalization, 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
Volume, beta, volatility	0.375	0.138	0.051	0.101	0.287

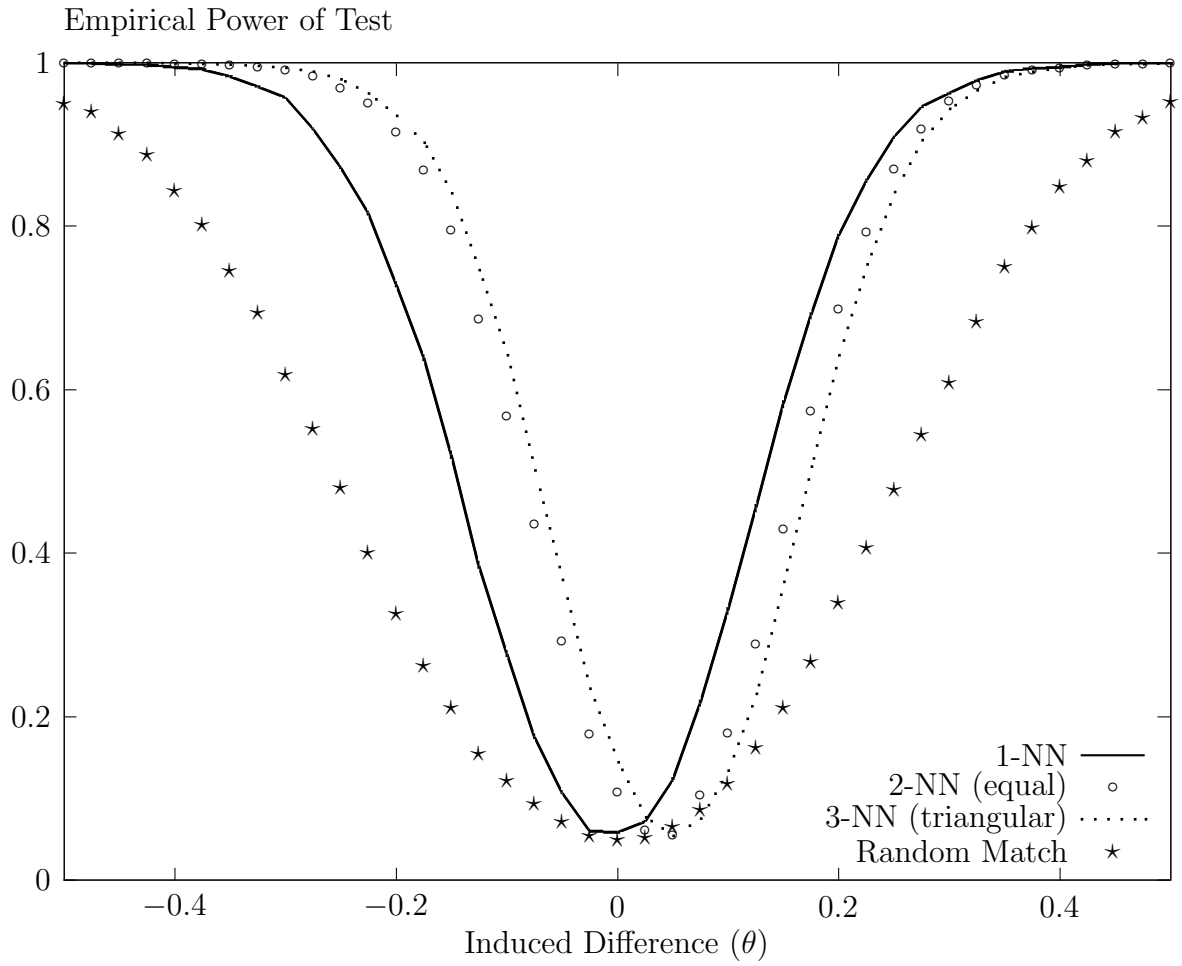

Figure 1: **Nearest neighbor matching estimates.** The matching estimates considered are: 1-NN (one-to-one match), 2-NN (match over two closest firms, each assigned equal weight) and 3-NN (match over three closest firms, triangular weights). Random Match indicates a test based on matched pairs randomly 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 are 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 nonparametric 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 test power is the share of Monte Carlo replications with a significant test statistic.

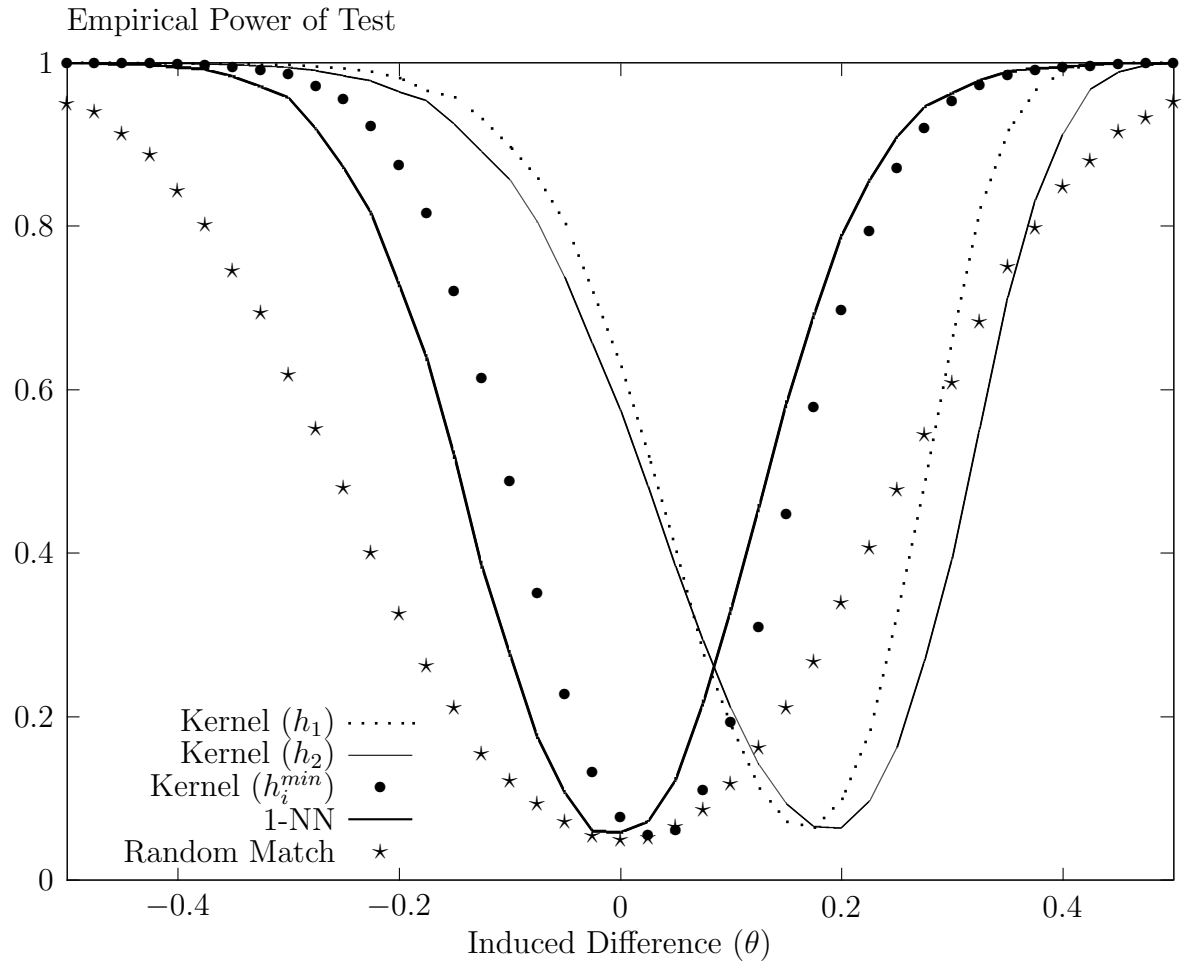

Figure 2: **Kernel-based matching estimates.** Results based on kernel-based matching estimates with bandwidth parameters: $h_1 = 1.059 s_x N_0^{-1/5}$, $h_2 = 1.059 s_x N_0^{-1/3}$, and h_i^{min} (variable bandwidth). Also reported for comparison purposes are the power of 1-NN (one-to-one) matching estimates and the power of a random match (matched pairs 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 are 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 nonparametric 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 test power is the share of Monte Carlo replications with a significant test statistic.

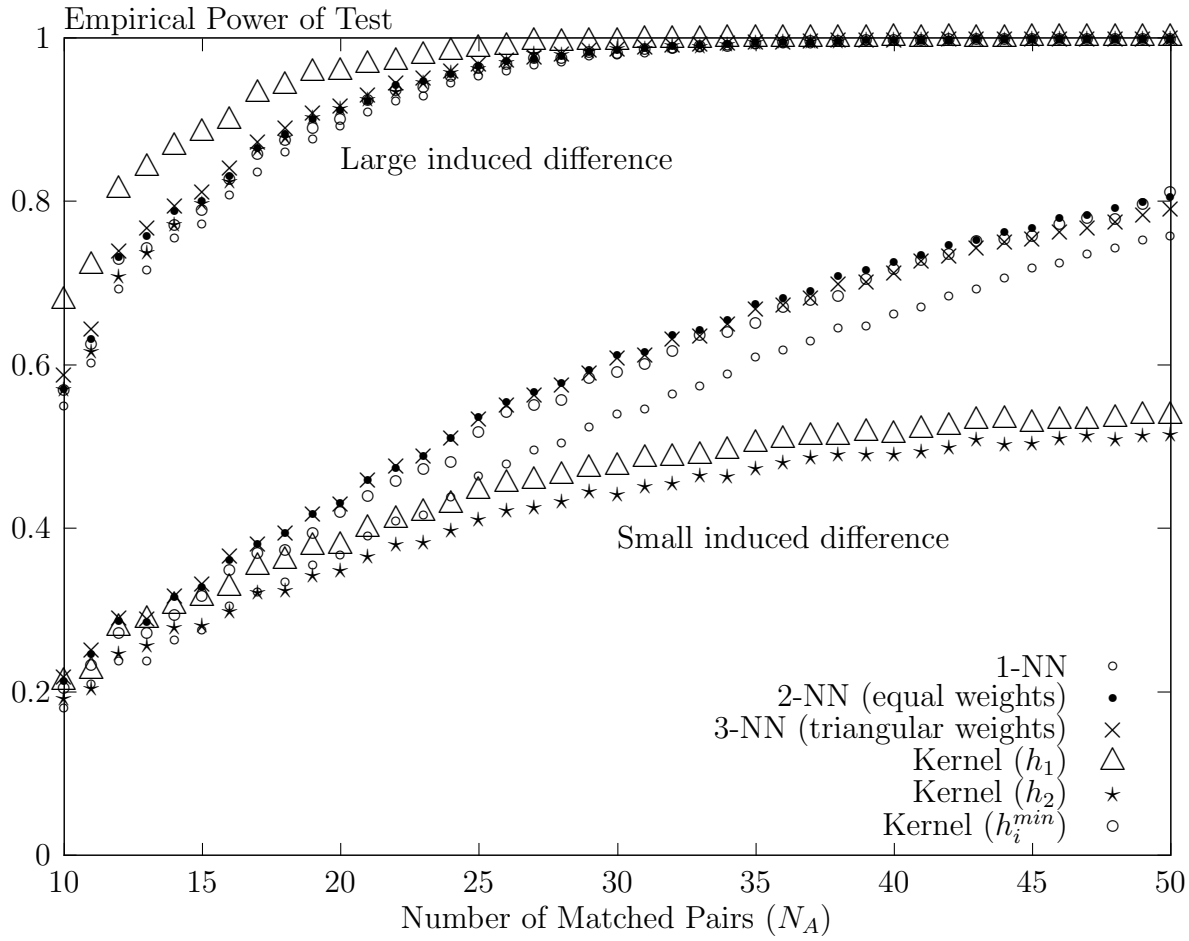

Figure 3: **Number of matched pairs.** The upper set of points correspond to a *large* induced difference ($\theta = 0.5$) and the lower set of points correspond to a *small* induced difference ($\theta = 0.2$). Results based on nearest neighbor matching estimates: 1-NN (one-to-one match), 2-NN (match 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: $h_1 = 1.059 s_x N_0^{-1/5}$, $h_2 = 1.059 s_x N_0^{-1/3}$ 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 are 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 nonparametric 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 test power is the share of Monte Carlo replications with a significant test statistic.

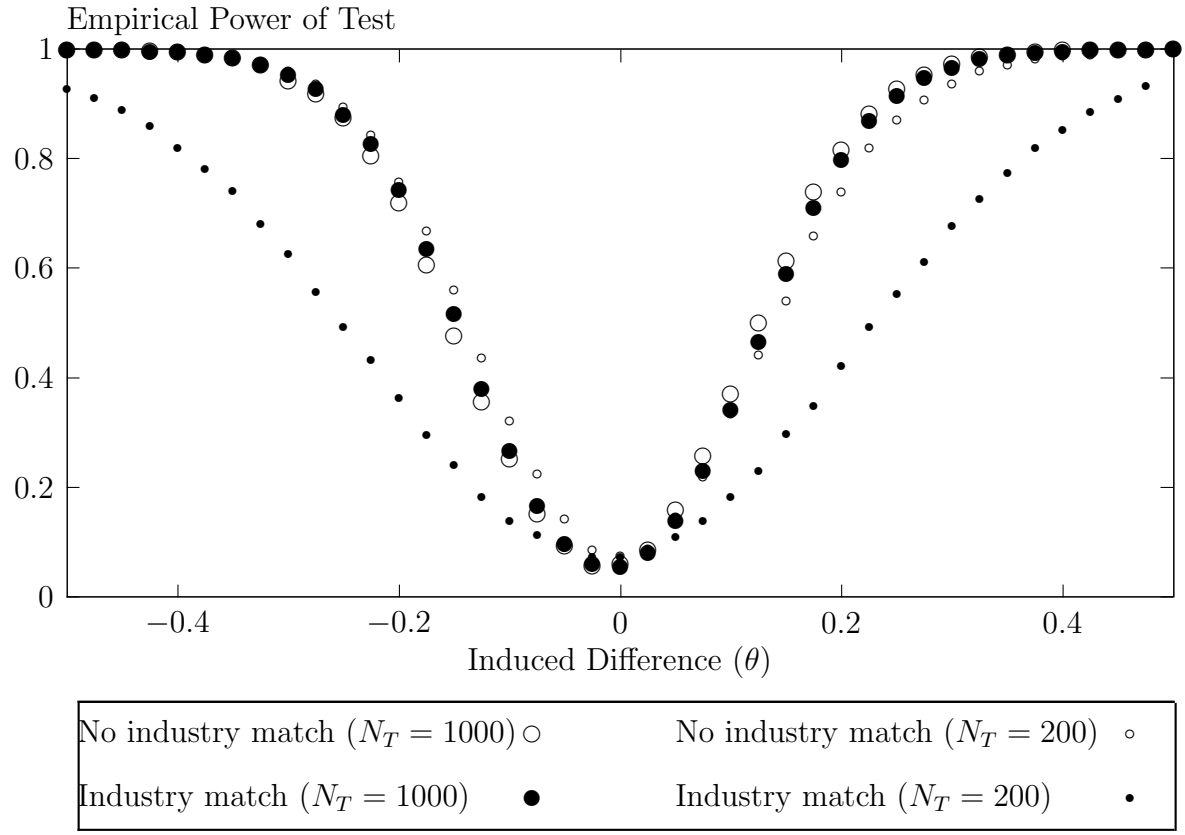

Figure 4: **Pre-sorting by industry.** Results based on 1-NN (one-to-one) matching estimates that are based on market capitalization alone and 1-NN matching estimates that also require the matched firm to have the same 2-digit SIC industry code. Based on 50 matched pairs and $N_T = 200$ and $N_T = 1000$ 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 are then matched with one of the remaining $N_T - 50$ firms; (iii) We then conduct a two-sided nonparametric 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 test power is the share of Monte Carlo replications with a significant test statistic.

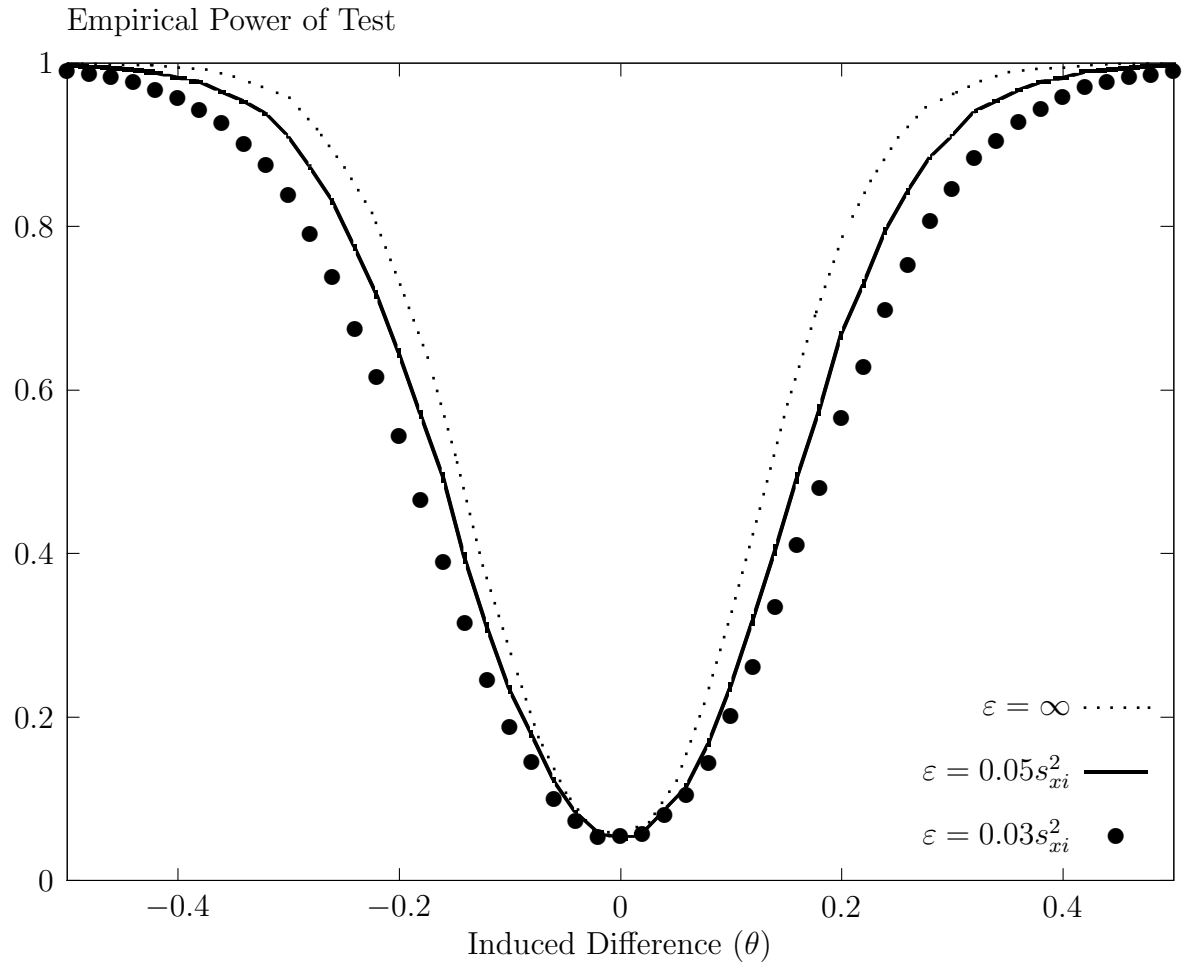

Figure 5: **Eliminating apparent poor matches.** Results based on 1-NN (one-to-one nearest neighbor) matching estimates with caliper tolerance levels of $\varepsilon = \infty$, $\varepsilon = 0.05 s_{xi}^2$, and $\varepsilon = 0.03 s_{xi}^2$, where s_{xi}^2 for a given firm i is defined as the variance of $(x_{Ai} - x_{0j})/(x_{Ai} + x_{0j})$ across all potential matches $j \in \mathbf{I}_0$. The tolerance level varies across firms depending on the closeness of potential matches. 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 are 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, $|(x_{Ai} - x_{0j})/(x_{Ai} + x_{0j})| < \varepsilon$, the match is removed from consideration; (iv) We then conduct a two-sided nonparametric Wilcoxon signed rank test of size $\alpha = 0.05$ of the difference in bid-ask spreads between the remaining matched pairs. The empirical test power is the share of Monte Carlo replications with a significant test statistic.

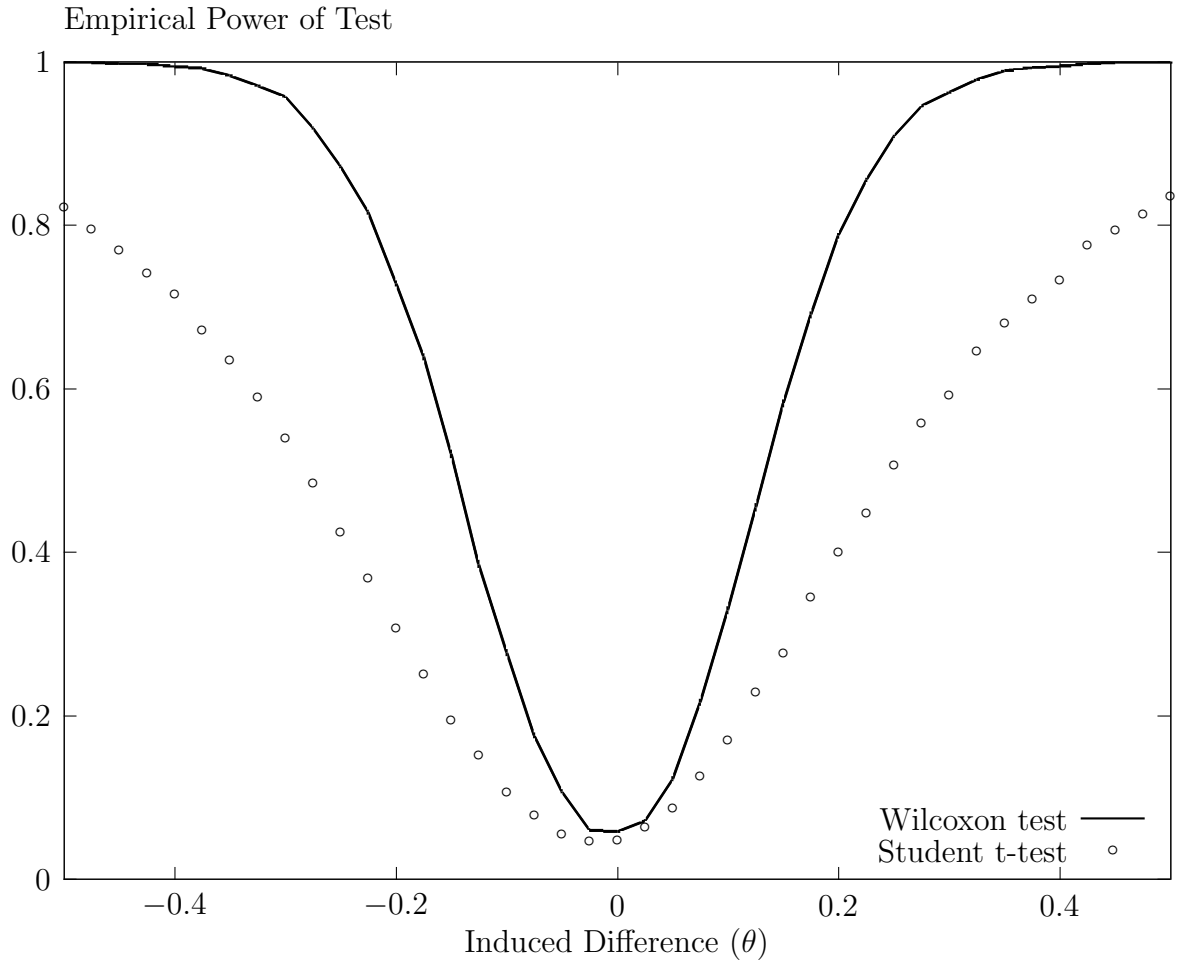

Figure 6: **Comparison of parametric and nonparametric test.** This graph compares the power of using a two-sided Wilcoxon signed rank test versus a dependent Student t-test. Both are based on the 1-NN (one-to-one) matching estimates. For each of 20,000 Monte Carlo replications: (i) We randomly select 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 are then matched based on market capitalization with one of the remaining 950 firms; (iii) We then conduct a 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. The empirical test power is the share of Monte Carlo replications with a significant test statistic.

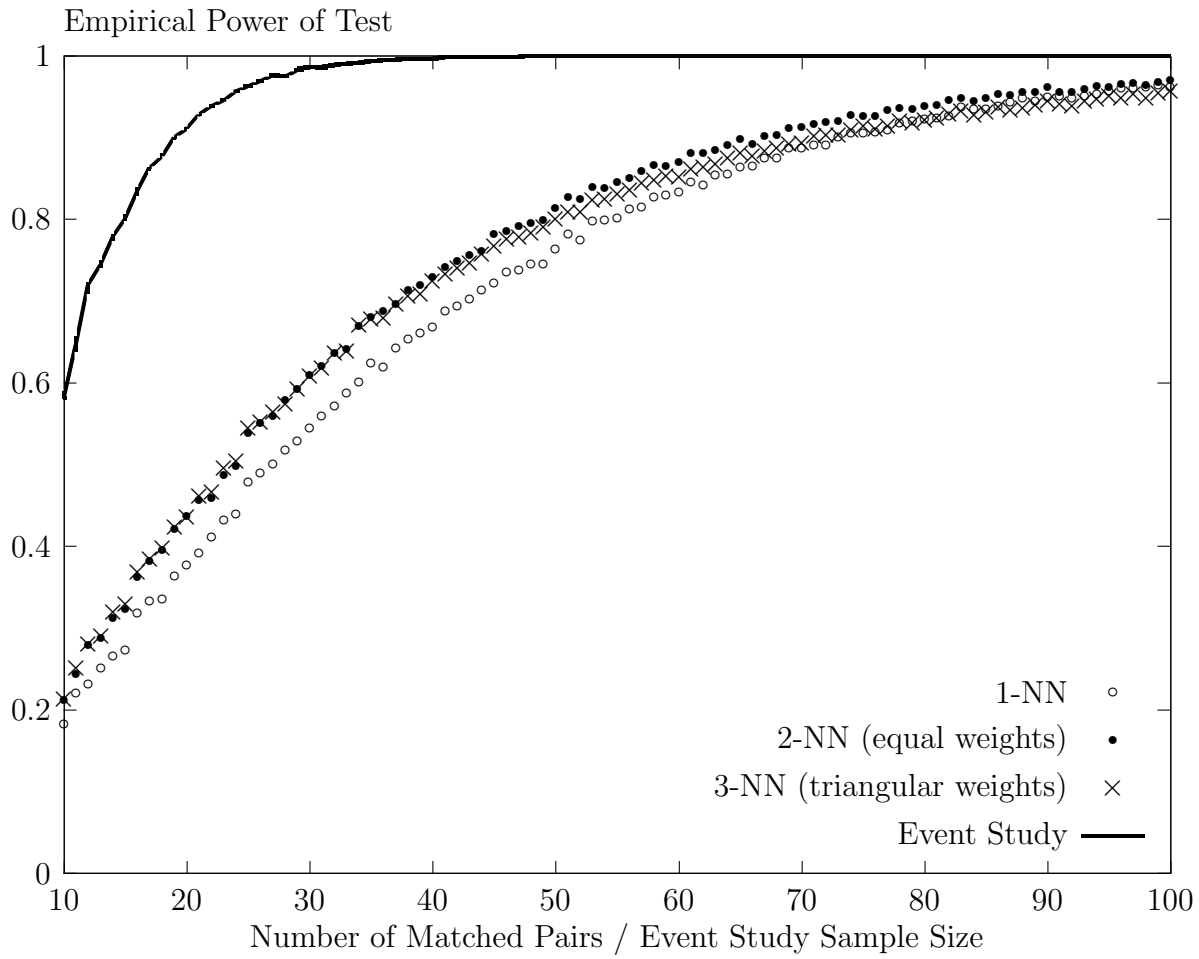

Figure 7: **Comparison of event study approach and matched sample approach.** Power graph of two-sided Wilcoxon signed rank test based on: (i) nearest neighbor matching estimates: 1-NN (one-to-one match), 2-NN (match over closest two firms, equal weights), and 3-NN (match over three closest firms, triangular weights); and (ii) the corresponding event study. The empirical power of the matching estimates is obtained in the same manner as Figure 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 nonparametric Wilcoxon signed rank test of size $\alpha = 0.05$ of the difference between $\left[Y_1^b, Y_2^b, \ldots, Y_N^b\right]$ and $\left[\hat{Y}_1^a, \hat{Y}_2^a, \ldots, \hat{Y}_N^a\right]$. The empirical test power is the share of Monte Carlo replications with a significant test statistic.

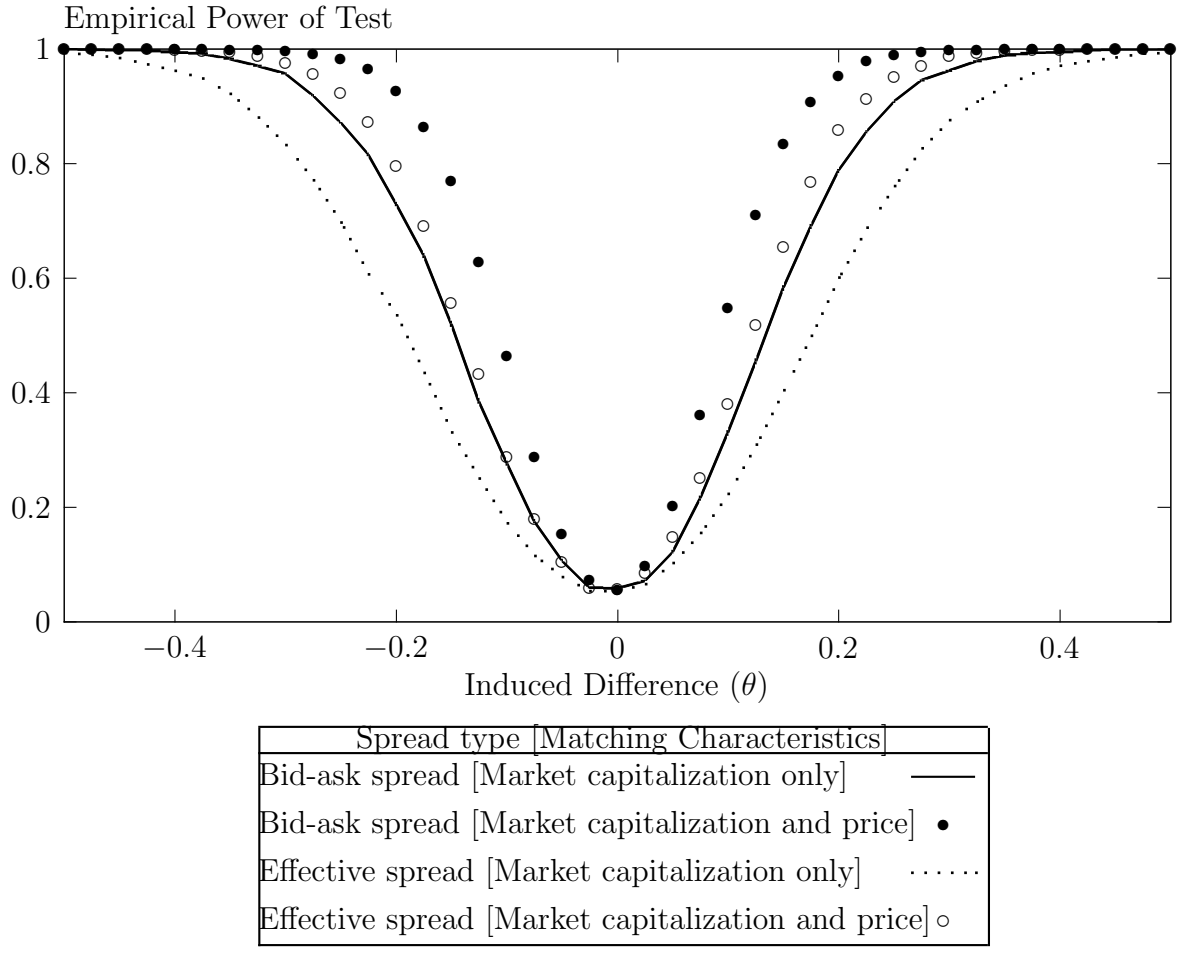

Figure 8: **Percentage bid-ask spreads and effective bid-ask spreads.** Results based on the 1-NN (one-to-one) matching estimates using market capitalization and using market capitalization and share price. With multiple matching characteristics, for each firm $i \in I_A$, we select firm $j \in I_0$ that solves $\text{argmin}_{j \in I_0} \sum_{k=1}^{2} \left((2(x_{Ai}^k - x_{0j}^k))/(x_{Ai}^k + x_{0j}^k) \right)^2$, 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 are then matched with one of the remaining 950 firms; (iii) We then conduct a two-sided nonparametric 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 test power is the share of Monte Carlo replications with a significant test statistic.