Chapter 14
Evaluating Long-Horizon Event Study Methodology

James S. Ang[*]
Florida State University

Shaojun Zhang[†]
Hong Kong Polytechnic University

Abstract

We describe the fundamental issues that long-horizon event studies face in choosing the proper research methodology, and summarize findings from existing simulation studies about the performance of commonly used methods. We document in details how to implement a simulation study and report our own findings on large-size samples. The findings have important implications for future research.

We examine the performance of more than twenty different testing procedures that fall into two categories. First, the buy-and-hold benchmark approach uses a benchmark to measure the abnormal buy-and-hold return for every event firm, and tests the null hypothesis that the average abnormal return is zero. Second, the calendar-time portfolio approach forms a portfolio in each calendar month consisting of firms that have had an event within a certain time period prior to the month, and tests the null hypothesis that the intercept is zero in the regression of monthly portfolio returns against the factors in an asset-pricing model. We find that using the sign test and the single most correlated firm being the benchmark provides the best overall performance for various sample sizes and long horizons. In addition, the Fama-French three-factor model performs better in our simulation study than the four-factor model, as the latter leads to serious overrejection of the null hypothesis.

We evaluate the performance of bootstrapped Johnson's skewness-adjusted t-test. This computation-intensive procedure is considered because the distribution of long-horizon abnormal returns tends to be highly skewed to the right. The bootstrapping method uses repeated random sampling to measure the significance of relevant test statistics. Due to the nature of random sampling, the resultant measurement of significance varies each time such a procedure is used. We also evaluate simple nonparametric tests, such as the Wilcoxon signed-rank test or the Fisher's sign test, which are free from random sampling variation.

Citation:
James S. Ang and Shaojun Zhang, 2015, "Evaluating long-horizon event study methodology", ***Handbook of Financial Econometrics and Statistics***, Chapter 14, 383-411.

[*] Contact author. Department of Finance, College of Business, Florida State University, Tallahassee, FL 32306, USA. Tel.: +1-850-644-8208. Fax: +1-850-644-4225. E-mail: jang@cob.fsu.edu.
[†] School of Accounting and Finance, Faculty of Business, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.

14.1 Introduction

A large number of papers in finance literature have documented evidence that firms earn abnormal returns over a long time period (ranging from one to five years) after certain corporate events. Kothari and Warner (2007) report that a total of 565 papers reporting event study results were published between 1974 and 2000 in 5 leading journals: the Journal of Business (JB), Journal of Finance (JF), Journal of Financial Economics (JFE), Journal of Financial and Quantitative Analysis (JFQA), and the Review of Financial Studies (RFS). Approximately 200 of the 565 event studies use a maximum window length of 12 months or more.

The evidence of long-horizon abnormal returns contradicts the Efficient Market Hypothesis that stock prices adjust to information fully within a narrow time window (a few days). To reconcile the contradiction, Fama (1998) argues that, "Most important, consistent with the market efficiency prediction that apparent anomalies can be due to methodology, most long-term return anomalies tend to disappear with reasonable changes in technique." Several simulation studies such as Kothari and Warner (1997) and Barber and Lyon (1997) document evidence that statistical inference in long horizon event studies is sensitive to the choice of methodology. Therefore, it is crucial to gain an understanding of the properties and limitations of the available approaches before choosing a methodology for a long-horizon event study.

At the core of a long-horizon event study lie two tasks: the first is to measure the event-related long horizon abnormal returns; and the second is to test the null hypothesis that the distribution of these long horizon abnormal returns concentrates around zero. A proper testing procedure for long-horizon event studies has to do both tasks well. Otherwise, two types of error could arise and lead to incorrect inference. The first error occurs when the null hypothesis is rejected, not because the event has generated true abnormal returns, but because a biased benchmark has been used to measure abnormal returns. A biased benchmark shifts the concentration of abnormal returns away from zero and leads to too many false rejections of the null hypothesis. The second error occurs when the null hypothesis is accepted, not because the event has no impact, but because the test itself does not have enough power to statistically discriminate the mean abnormal return from zero. A test with low power is undesirable, as it will lead researchers to reach

the incorrect inference that long term effect is statistically insignificant. Thus the researchers would want a procedure that minimizes both sources of error, or at least, choose a balance between them.

Two approaches have been followed in recent finance literature to measure and test long-term abnormal returns. The first approach uses a benchmark to measure the abnormal buy-and-hold return for every event firm in a sample, and tests whether the abnormal returns have a zero mean. The second approach forms a portfolio in each calendar month consisting of firms that have had an event within a certain time period prior to the month, and tests the null hypothesis that the intercept is zero in the regression of monthly calendar-time portfolio returns against the factors in an asset-pricing model. To follow either approach, researchers need to make a few choices as illustrated in Figure 14.1. For the calendar-time portfolio approach, researchers choose an asset-pricing model and an estimation technique to fit the model. Among the most popular asset-pricing models are Fama and French's (1993) three-factor model and its four-factor extension proposed by Carhart (1997) that includes an additional momentum-related factor. Two techniques are commonly used to fit the pricing model: the ordinary least squares (OLS) technique and the weighted least squares (WLS) technique. On the other hand, if adopting the buy-and-hold benchmark approach, researchers choose either a reference portfolio or a single control firm as the benchmark for measuring abnormal returns and select either parametric or nonparametric statistic for testing the null hypothesis of zero abnormal return.

Permutations of these choices under both approaches generate a large number of possible testing procedures that can be used in a long-horizon event study. It is neither practical nor sensible to implement all the testing procedures in an empirical study of a financial event. Therefore, it would be very useful to provide guidance on the strength and weakness of the procedures based on simulation results. Simulation study generates large number of repetitions under various circumstances for each testing procedure, which allows the tabulations of these two types of error for comparison.

We organize this chapter as follows. Section 14.2 discusses the fundamental issues in long-horizon event studies that have been documented in the literature. Section

14.3 reviews existing simulation studies. Section 14.4 reports results from a simulation study of large-size samples. Section 14.5 contains some suggestions for future research.

14.2 Fundamental issues in long-horizon event studies

14.2.1 The buy-and-hold benchmark approach

The long-term buy-and-hold abnormal return of firm i, denoted as AR_i, is calculated as

$$AR_i = R_i - BR_i \ , \tag{14.1}$$

where R_i is the long-term buy-and-hold return of firm i, and BR_i is the long-term return on a particular benchmark of firm i. The buy-and-hold return of firm i over τ months is obtained by compounding monthly returns, that is,

$$R_i = \prod_{t=1}^{\tau}(1+r_{it}) - 1 \ , \tag{14.2}$$

where r_{it} is firm i's return in month t. Calculation of the benchmark return BR_i is given below. The benchmark return, BR_i, estimates the return that an event firm would have had if the event had not happened.

Several articles clearly show that long-term abnormal returns are very sensitive to choice of benchmarks, see, e.g. Ikenberry Lakonishok and Vermaelen (1995), Kothari and Warner (1997), Barber and Lyon (1997), and Lyon, Barber and Tsai (1999). If wrong benchmarks were used in measuring long-term abnormal returns, inference on the significance of a certain event would be erroneous. Most existing studies use either a single matched firm or a matched reference portfolio as the benchmark. Barber and Lyon (1997) point out that the control firm approach eliminates the new listing bias, the rebalancing bias, and the skewness problem. It also yields well-specified test statistics in virtually all the situations they consider. Further, Lyon, Barber and Tsai (1999) advocate a reference portfolio of firms that match on size and BE/ME. The issue on choice of the benchmark is practically unresolved. Ang and Zhang (2004) additionally argue that the control firm method overcomes another important problem that is associated with the event firm not being representative in important aspects of the respective matched portfolio in the reference portfolio approach. This leads to the matched portfolio return

generating a biased estimate of expected firm return. This problem is particularly severe with small firms.

A common practice in computing an event firm's long-term abnormal return is to utilize a benchmark that matches the event firm on size and BE/ME. The practice is often justified by quoting the findings in Fama and French (1992) that size and BE/ME combine to capture the cross-sectional variation in average monthly stock returns and that market beta has no additional power in explaining cross-sectional return differences. However, in a separate paper, Fama and French (1993) demonstrate that expected monthly stock returns are related to three factors: a market factor, a size related factor and a book-to-market equity ratio (BE/ME) related factor. To resolve this issue, Ang and Zhang (2004) show that matching based on beta in addition to size and BE/ME does not improve the performance of the approach.

A recent trend is to use computation-intensive bootstrapping-based tests, such as the bootstrapped Johnson's skewness-adjusted t-statistic (e.g., Sutton (1993), and Lyon, Barber and Tsai (1999)) and the simulated empirical p-values (e.g., Brock, Lakonishok, and LeBaron (1992), and Ikenberry, Lakonishok, and Vermaelen (1995)). These procedures rely on repeated random sampling to measure the significance of relevant test statistics. Due to the nature of random sampling, the resultant measurement of significance varies every time such a procedure is used. As a consequence, different researchers could reach contradictory conclusions using the same procedure on the same sample of event firms. In contrast, simple nonparametric tests, such as the Wilcoxon signed-rank test or the Fisher's sign test, are free from random sampling variation. Barber and Lyon (1997) examined the performance of the Wilcoxon signed-rank test in a large-scale simulation study. They show that the performance depends on choice of the benchmark. The signed-rank test is well specified when the benchmark is a single size and BE/ME matched firm, and misspecified when the benchmark is a size and BE/ME matched reference portfolio. However, Barber and Lyon (1997) present only simulation results for one-year horizon. *No simulation study in the finance literature has examined*

the performance of these simple nonparametric tests for three- or five-year horizons, which are the common holding periods in long-horizon event studies.[1]

Power is an important consideration in statistical hypothesis testing. Lyon, Barber and Tsai (1999) report that bootstrapping-based tests are more powerful than Student's t-test in testing one-year abnormal returns in a large-scale simulation study. However, they do not report evidence on the power of these tests for the longer three- or five-year horizon. In statistics literature, bootstrapping is primarily for challenging situations when the sampling distribution of the test statistic is either indeterminate or difficult to obtain, and that bootstrapping is less powerful in hypothesis testing than other parametric or simple non-parametric methods when both bootstrapping and other methods are applicable (See, e.g., Efron and Tibshirani (1993, Chapter 16) and Davison and Hinkley (1997, Chapter 4)). In a recent study on five-year buy-and-hold abnormal returns to holders of the seasoned equity offerings, Eckbo, Masulis, and Norli (2000) note that bootstrapping gives lower significance level relative to the Student's t-test.

Ang and Zhang (2004) find that most testing procedures have very low power for samples of medium size over long event horizons (three or five years). This raises concern about how to interpret long-horizon event studies that fail to reject the null hypothesis. Failure to reject is often interpreted as evidence that supports the null hypothesis. However, when power of the test is low, such interpretation may no longer be warranted. This problem gets even worse when event firms are primarily small firms. They observe that all tests, except the sign test, have much lower power for samples of small firms.

More recently, Schultz (2003) argue via simulation that the long-run IPO underperformance could be related to the endogeneity of the number of new issues. Firms choose to go IPO at the time when they expect to obtain high valuation in the stock market. Therefore, IPOs cluster after periods of high abnormal returns on new issues. In such a case, even if the ex ante returns on IPO are normal, the ex post measures of abnormal returns may be negative on average. Schultz suggests using calendar-time returns to overcome the bias. However, Dahlquist and de Jong (2008) find that it is

[1] The sign test has an advantage over the signed-rank test in that it does not require a symmetric underlying distribution while the signed-rank test does.

unlikely that the endogeneity of the number of new issues explains the long-run underperformance of IPOs. Viswanathan and Wei (2008) present a theoretical analysis on event abnormal returns when returns predict events. They show that, when the sample size is fixed, the expected abnormal return is negative and becomes more negative as the holding period increases. This implies that there is a small-sample bias in the use of long-run event returns. Asymptotically, abnormal returns converge to zero provided that the process of the number of events is stationary. Nonstationarity in the process of the number of events is needed to generate a large negative bias.

The issues discussed above are associated with the buy-and-hold approach to testing long-term abnormal returns.[2] In addition, this approach suffers from the cross-correlation problem and the bad model problem (Fama (1998), Brav (1999), and Mitchell and Stafford (2000)). The cross-correlation problem arises because matching on firm-specific characteristics fails to completely remove the correlation between event firms' returns. The bad model problem arises because no benchmark gives perfect estimate of the *counter factual* (i.e., what if there was no event) return of an event firm and benchmark errors are multiplied in computing long-term buy-and-hold returns. Therefore, Fama (1998) advocates a calendar-time portfolio approach.[3]

14.2.2 The calendar-time portfolio approach

In the calendar-time portfolio approach, for each calendar month, an event portfolio is formed, consisting of all firms that have experienced the same event within the τ months prior to the given month. Monthly return of the event portfolio is computed as the equally weighted average of monthly returns of all firms in the portfolio. Excess returns of the event portfolio are regressed on the Fama-French three factors as in the following model:

$$R_{pt} - R_{ft} = \alpha + \beta\left(R_{mt} - R_{ft}\right) + sSMB_t + hHML_t + \varepsilon_t \ , \qquad (14.3)$$

[2] Variations of this approach have been used extensively, see, e.g. Ritter (1991), Ikenberry, Lakonishok, and Vermaelen (1995), Ikenberry, Rankine, and Stice (1996), and Desai and Jain (1997), among many others.

[3] Loughran and Ritter (1995), Brav and Gompers (1996), and Brav, Geczy and Gompers (2000), among others, have used the calendar-time portfolio approach.

where R_{pt} is the event portfolio's return in month t, R_{ft} is the one-month Treasury bill rate, observed at the beginning of the month, R_{mt} is the monthly market return, SMB_t is the monthly return on the zero investment portfolio for the common size factor in stock returns, and HML_t is the monthly return on the zero investment portfolio for the common book-to-market equity factor in stock returns.[4] Under the assumption that the Fama-French three-factor model provides a complete description of expected stock returns, the intercept, α, measures the average monthly abnormal return on the portfolio of event firms and should be equal to zero under the null hypothesis of no abnormal performance.

A later modification that has gained popularity is the four-factor model that added a momentum-related factor to the Fama-French three factors:

$$R_{pt} - R_{ft} = \alpha + b\left(R_{mt} - R_{ft}\right) + sSMB_t + hHML_t + pPR12_t + \varepsilon_t, \qquad (14.4)$$

where $PR12_t$ is the momentum-related factor advocated by Carhart (1997). Typically, we compute $PR12_t$ by first ranking all firms by their previous 11-month stock return lagged one month and then taking the average return of the top one third (i.e. high past return) stocks minus the average return of the bottom one third (i.e. low past return) stocks.

Under the assumption that the asset pricing model adequately explains variation in expected stock returns, the intercept, α, measures the average monthly abnormal return of the calendar-time portfolio of event firms and should be equal to zero under the null hypothesis of no abnormal performance. If the test concludes that the time series conforms to the asset pricing model, the event is said to have had no significant long-term effect; otherwise, the event has produced significant long-term abnormal returns. Lyon, Barber and Tsai (1999) report that the calendar-time portfolio approach together with the Fama-French three-factor model, which shall be referred to as the Fama-French calendar-time approach later, is well specified for random samples in their simulation study.

However, we do not know how much power the Fama-French calendar-time approach has. Loughran and Ritter (1999) criticize the approach as having very low power. They argue that reduction in power is caused by using returns on contaminated

[4] See Fama and French (1993) for details on construction of the mimicking portfolios for the common size and book-to-market equity factors. We thank Eugene Fama for providing us with returns on R_{ft}, R_{mt}, SMB_t and HML_t.

portfolios as factors in the regression, by weighting each month equally and by using value-weighted returns of the calendar-time portfolios. However, their empirical evidence is based only on one carefully constructed sample of firms and is hardly conclusive. No large-scale simulation study has been done to examine power of the Fama-French calendar-time approach, which we will remedy in this paper.

The Fama-French calendar time approach, estimated with the ordinary least squares (OLS) technique, could suffer from a potential heteroskedasticity problem due to unequal and changing number of firms in the calendar-time portfolios. The weighted least square (WLS) technique, which is helpful in addressing the heteroskedasticity problem, has been suggested as a way to deal with the changing size of calendar-time portfolios. When applying WLS, we use the monthly number of firms in the event portfolio as weights.

14.3 A review of simulation studies on long-horizon event study methodology

Several papers have documented performance of testing procedures in large-scale simulations. Table 14.1 surveys these papers with reference to testing procedures under their investigation and their simulation settings. The simulation technique was pioneered by Brown and Warner (1980, 1985) to evaluate size and power of testing procedures. In this section, we review these simulation studies.

[Table 14.1 is about here]

As shown in Figure 14.1, there are two approaches for a long-term event study: the calendar-time portfolio approach versus the buy-and-hold benchmark approach. There has been a debate on which approach prescribes the best procedure for long-term event studies. Both approaches have been under criticisms. The buy-and-hold benchmark approach is susceptible to biases associated with cross-sectional correlation, insufficient matching criteria, new equity issues, periodic balancing, and skewed distribution of long-term abnormal returns, while the calendar-time portfolio approach may suffer from an improper asset pricing model and heteroskedasticity in portfolio returns. See Kothari and Warner (1997), Barber and Lyon (1997), Fama (1998), Loughran and Ritter (1999), Lyon, Barber, and Tsai (1999) and others for more detailed

discussions. Kothari and Warner (1997) argue that the combined effect of these issues is difficult to specify a priori and, thus, "a simulation study with actual security return data is a direct way to study the joint impact, and is helpful in identifying the potential problems that are empirically most relevant."

In their simulation study, Kothari and Warner (1997) measure the long-term (up to 3 years) impact of an event by cumulative monthly abnormal returns, where monthly abnormal returns are computed against four common models: the market-adjusted model, the market model, the capital asset pricing model, and the Fama-French three-factor model. They find that tests for cumulative abnormal returns are severely misspecified. They identify sample selection, survival bias, and bias in variance estimation as potential sources of the misspecification and suggest that nonparametric and bootstrap tests are likely to reduce misspecification.

Barber and Lyon (1997) address two main issues in their simulation study. First, they argue that buy-and-hold return is a better measure of investors' actual experience over a long horizon and should be used in long-term event study (up to 5 years). They show simulation evidence that approaches using cumulative abnormal returns cause severe misspecification, which is consistent with the observation in Kothari and Warner (1997). Second, they use simulations to measure both size and power of testing procedures that follow the buy-and-hold benchmark approach. An important finding is that using a single control firm as benchmark yields well-specified tests, whereas using reference portfolio causes substantial over-rejection.

In a later paper, Lyon, Barber and Tsai (1999) report another simulation study (for up to the 5-year horizon) that investigates the performance of both buy-and-hold benchmark approach and calendar-time portfolio approach. They find that using the Fama-French three-factor model yields a well-specified test. However, they advocate a test that uses carefully constructed reference portfolio as benchmark and the bootstrapped Johnson's statistic for testing abnormal returns. They present evidence that this test is well specified and has high power at the one-year horizon.

Two questions remain unanswered in Lyon, Barber and Tsai (1999). First, how much power does the bootstrap test have for event horizons longer than 1 year (e.g. 3 or 5 years that is common in long-horizon studies)? It is known in statistics literature that a

bootstrap test is not as powerful as simple non-parametric tests in many occasions (See Efron and Tibshirani (1993, Chapter 16) and Davison and Hinckley (1997, Chapter 4)). It is necessary to know the actual power of such test for event horizons beyond one year. Second, is the calendar-time portfolio approach as powerful as the buy-and-hold benchmark approach? Loughran and Ritter (2000) argue that the calendar-time portfolio approach has low power, using simulations and empirical evidence from a sample of new equity issuers. However, they do not measure how much power the approach actually has, which makes it impossible to compare the two approaches directly in more general settings.

Mitchell and Stafford (2000) is the only study that empirically measures power of the calendar-time portfolio approach using simulations. Their main focus is to assess performance of several testing procedures in three large samples of major managerial decisions, i.e., mergers, seasoned equity offerings, and share repurchases (up to 3 years). They find that different procedures lead to contradicting conclusions and argue that the calendar-time portfolio approach is preferred. To resolve Loughran and Ritter (2000)'s critique that the calendar-time portfolio approach has low power, they conduct simulations to measure the empirical power and find that the power is actually very high with an empirical rejection rate of 99% for induced abnormal returns of $\pm 15\%$ over a three-year horizon. Since they have a large sample size, this finding is actually consistent with what we document in Table 14.5. However, their simulations focus on only samples of 2,000 firms. Many event studies have much smaller sample sizes, especially after researchers slice and dice a whole sample into sub-samples. More evidence is needed in order to have great confidence in applying the calendar-time portfolio approach in such studies.

Cowan and Sergeant (2001) focus on the buy-and-hold benchmark approach in their simulations. They find that using the reference portfolio approach cannot overcome the skewness bias discussed in Barber and Lyon (1997), and that the larger the sample size, the smaller the magnitude of the skewness bias. They also argue that cross-sectional dependence among event firms' abnormal returns increases in event horizon due to partially contemporaneous holding periods, which may cause the overlapping horizon bias. They propose a two-group test using abnormal returns winsorized at three standard

deviations to deal with these two biases, and report evidence that this test yields correct specifications and considerable power in many situations.

All previous simulation studies use only size and BE/ME to construct benchmarks, which is often justified by the findings in Fama and French (1992) that size and BE/ME together adequately capture the cross-sectional variations in average monthly stock returns. Ang and Zhang (2004) use two other matching criteria to explore whether better benchmarks could be used for future studies. The two criteria are market beta and pre-event correlation coefficient. Using market beta is motivated by the fact that Fama and French's (1993) three-factor model has a market factor, a size-related factor, and a BE/ME related factor. Matching on the basis of size and BE/ME does not account for the influence of the market factor. The rationale for using pre-event correlation coefficient is that matching on size and BE/ME may fail to control for other factors that could influence stock returns, such as industry factor, seasonal factor, momentum factor, and other factors shared by only firms of same characteristics, such as geographical location, ownership and governance structures. Matching on the basis of pre-event correlation coefficient helps remove the effect of these factors on the event firm's long-term return.

The main findings in Ang and Zhang (2004) include the following. First, the four-factor model is inferior to the well-specified three-factor model in the calendar-time portfolio approach in that the former causes too many rejections of the null hypothesis relative to the specified significance level. Second, WLS improves the performance of the calendar-time portfolio approach over OLS, especially for long event horizons. Third, the Fama-French three-factor model has relatively high power in detecting abnormal returns, although power decreases sharply as event horizon increases. Fourth, the simple sign test is well specified when it is applied with a single firm benchmark, but misspecified when used with reference portfolio benchmarks. More importantly, the combination of the sign test and the benchmark with the single most correlated firm consistently has much higher power than any other test in our simulations and is the only testing procedure that performs well in samples of small firms.

Jegadeesh and Karceski (2009) propose a new test of long-run performance that allows for heteroskedasticity and autocorrelation. Previous tests used in Lyon, Barber and Tsai (1999) implicitly assume that the observations are cross-sectionally uncorrelated.

This assumption is frequently violated in nonrandom samples such as samples with industry clustering or with overlapping returns. To overcome the cross-correlation bias in event firms' returns, they recommend a t-statistic that is computed using a generalized version of the Hansen and Hodrick (1980) standard error. Their simulation studies show that the new tests they propose are reasonably well-specified in random samples, in samples that are concentrated in particular industries, and also in samples where event firms enter the sample on multiple occasions within the holding period.

In summary, these simulation studies show that testing procedures differ dramatically in performance. Some procedures reject the null hypothesis at an excessively high rate, while others have very low power. These findings confirm the Fama (1998) statement that evidence for long-term return anomalies is dependent upon methodology, and suggest that caution must be exercised in choosing the proper methodology for a long-term event study.

14.4 A simulation study of large-size samples

A simulation study of large-size samples serves two purposes. First, it is well documented that the distribution of buy-and-hold abnormal returns tends to be skewed to the right. Kothari and Warner (2007) mentions that the extent of skewness bias is likely to decline with sample size. It is of interest to provide evidence on how much is the level of right-skewness in the average abnormal returns of large-size samples. Second, although it is expected that testing power increases with sample size, it is of practical interest to know more precisely how much power a test can have in a sample of 1,000 observations. Large sample simulation defines the limits of a procedure.

14.4.1 Research design

In this simulation study, we construct 250 samples each consisting of 1,000 event firms. To produce one sample, we randomly select, with replacement, 1,000 event months between January 1980 and December 1992, inclusively.[5] [6] This allows us to

[5] We use a pseudorandom number generator developed by Matsumoto and Nishimura (1998) to ensure high quality of random sampling.

calculate five-year abnormal returns until December 1997. For each selected event month, we randomly select, without replacement, one firm from a list of qualified firms. The qualified firms satisfy the following requirements: (i) they are publicly traded firms, incorporated in U.S., and have ordinary common shares with Center for Research in Security Prices (CRSP) share codes 10 and 11; (ii) they have return data found in the CRSP monthly returns database for the 24-month period prior to the event month; (iii) they have nonnegative book values on COMPUSTAT prior to the event month so that we can calculate their book-to-market equity ratios.

The 250 samples, each of 1,000 randomly selected firms, comprise the simulation setting for comparing the performance of different testing procedures.[7] We apply all testing procedures under our study to the same samples. Such controlled comparison is more informative because it eliminates difference in performance due to variation in the samples.

For the buy-and-hold approach, we compute the long-term buy-and-hold abnormal return of firm i as the difference between the long-term buy-and-hold return of firm i and the long-term return of a benchmark. The buy-and-hold return of firm i over τ months is obtained by compounding monthly returns. In case that firm i does not have return data for all τ months, we replace missing returns by the same-month returns of a size and BE/ME matched reference portfolio.[8] We evaluate a total of five benchmarks and four test statistics in this study. We briefly describe them in the following and give the details in the Appendix.

Three of the benchmarks are reference portfolios. The *first* reference portfolio consists of firms that are similar to the event firm in both size and BE/ME. We follow the same procedure as in Lyon, Barber and Tsai (1999) to construct the two-factor reference portfolio. We use the label "SZBM" for this benchmark. The *second* reference

---

[6] Kothari and Warner (1997) use 250 samples, each of 200 event months between January 1980 and December 1989 inclusively. Barber and Lyon (1997) use 1,000 samples, each of 200 event months in a much longer period from July 1963 through December 1994. The period under our study, between January 1980 and December 1992, is of similar length to Kothari and Warner's.

[7] Ang and Zhang (2004) examine two other simulation settings. Under one setting, they have another 250 samples of 200 event firms, a smaller sample size than the setting in this paper. Under the other setting, they have the sample size of 200 with the requirement that event firms belong to the smallest quintile sorted by NYSE firm size. The second setting is used to examine the effect of small firms.

[8] Filling in missing returns is a common practice in calculating long-term buy-and-hold returns, e.g. see Barber and Lyon (1997), Lyon, Barber and Tsai (1999) and Mitchell and Stafford (2000).

portfolio consists of firms that are similar to the event firm not only in size and BE/ME but also in market beta. We use the label "SZBMBT" for this benchmark. The *third* reference portfolio consists of ten firms that are most correlated with the event firm prior to the event. We use the label "MC10" for this benchmark.

The other two of the five benchmarks consist of a single firm. The *first* single firm benchmark is the firm that matched the event firm in both size and BE/ME. To find the two-factor single firm benchmark, we first identify all firms whose market value is within 70% to 130% of the event firm's market value and then choose the firm that has the BE/ME ratio closest to that of the event firm. We use the label "SZBM1" for this benchmark. The *second* single firm benchmark is the firm that has the highest correlation coefficient with the event firm prior to the event. We use the label "MC1" for this benchmark.

We apply four test statistics to test the null hypothesis that the mean long-term abnormal return is zero. They include Student's t-test, Fisher's sign test, Johnson's skewness-adjusted t-test, and the bootstrapped Johnson's t-test. Fisher's sign test is a nonparametric test and is described in details in Hollander and Wolfe (1999, Chapter 3). Johnson's skewness-adjusted t-statistic was developed by Johnson (1978) to deal with the skewness-related misspecification error in Student's t-test. Sutton (1992) proposes to apply Johnson's t-test with a computationally intensive bootstrap re-sampling technique when the population skewness is severe and the sample size is small. Lyon, Barber and Tsai (1999) advocate use of the bootstrapped Johnson's t-test because long-term buy-and-hold abnormal returns are highly skewed when buy-and-hold reference portfolios are used as benchmarks. We follow Lyon, Barber and Tsai (1999) and set the re-sampling size in the bootstrapped Johnson's t-test to be one quarter of the sample size.

For the Fama-French calendar-time approach, we use both the Fama-French three-factor model and the four-factor model. We apply both ordinary least squares (OLS) and weighted least squares (WLS) techniques to estimate parameters in the pricing model. The WLS is used to correct the heteroskedasticity problem due to the monthly variation in the number of firms in the calendar-time portfolio. When applying WLS, we use the number of event firms in the portfolio as weights.

14.4.2 Simulation results for the buy-and-hold benchmark approach

In this section, we examine the performance of testing procedures that follow the buy-and-hold benchmark approach. Implementation of the buy-and-hold benchmark approach involves choosing both benchmark and test statistic. For this reason, rather than focusing on what is the best among all benchmarks, or focusing on what is the best among all test statistics, we address the more practical question of finding the best combination of benchmark and test statistic. Combination of the five benchmarks and the four test statistics yields 20 testing procedures, out of which we look for the best combination.

For each sample of 1,000 abnormal returns, we compute mean, median, standard deviation, inter-quartile range, skewness coefficient, and kurtosis coefficient. Table 14.2 reports the average of these statistics over 250 samples.

[Table 14.2 is about here]

Since these event firms, being randomly selected, may not experience any event or may experience events that have offsetting effects on averaged stock returns, we expect their abnormal returns to concentrate around zero. In Table 14.2, means are close to zero for all five benchmarks at all three holding periods, but medians differ systematically according to the type of benchmark used. Medians are clearly negative under the three reference portfolio benchmarks (i.e., SZBM, SZBMBT, and MC10), but close to zero under the two single firm benchmarks (i.e., SZBM1 and MC1). The evidence suggests that reference portfolio benchmarks overestimate holding period returns of many event firms, resulting in far too many event firms having negative abnormal returns under the portfolio-based benchmarks. The extent of the overestimation bias by portfolio-based benchmarks is quite severe, and gets worse as the time horizon lengthens. The bias, as measured by the magnitude of median, ranges from around 4% at a one-year horizon, to 12% at a three-year horizon, and to more than 20% at a five-year horizon. Bias of this magnitude could cause too many events to be falsely identified as having significant long-term impact.

Volatility of abnormal returns increases with the length of holding period under all five benchmarks. For the same holding period, volatility is higher under the two single firm benchmarks than under the three reference portfolio benchmarks. This is expected

because reference portfolios have lower volatility due to averaging. As for kurtosis, all five benchmarks produce highly leptokurtic abnormal returns, with kurtosis coefficients ranging from 41.4 to 67.5, which are far greater than three, the kurtosis coefficient of any normal distribution. At last, skewness coefficients for the two single firm benchmarks are close to zero regardless of event horizons, while skewness coefficients for the three portfolio benchmarks are excessively positive.

To sum up, probability distributions of long-term abnormal returns exhibit different properties, depending on whether the benchmark is a reference portfolio or a single firm. Under a reference portfolio benchmark, the distribution is highly leptokurtic and positively skewed, with a close-to-zero mean but a highly negative median. Under a single firm benchmark, the distribution is highly leptokurtic but symmetric, with both mean and median close to zero. Statistical properties of long-term abnormal returns have important bearings on performance of test statistics. Overall, it seems single firm benchmarks have more desirable properties. Between the two single firm benchmarks, MC1 shows better performance than SZBM1, because the abnormal returns based on MC1 have both mean and median being closer to zero and smaller standard deviation.

A superior test should control for the probability of committing two errors. First, it is important to control for the probability of misidentifying an insignificant event as having statistical significance; in other words, the empirical size of the test, which is computed from simulations, is close to the pre-specified significance level at which the test is conducted. When this happens, the test is well specified. Second, power of the test should be large, that is, the probability of finding a statistically significant event if one did exist.

Table 14.3 reports empirical size of all 20 tests for three holding periods. Empirical size is calculated as the proportion of 250 samples that rejects the null hypothesis at the 5% nominal significance level. With only a few exceptions, Student's t-test is well specified against the two-sided alternative hypothesis. Despite excessively high skewness in abnormal returns from reference portfolio benchmarks, Student's t-test is well specified against two-sided alternative hypothesis because the effect of skewness at both tails cancels out (See, e.g., Pearson and Please (1975)). When testing against the two-sided alternative hypothesis, Johnson's skewness-adjusted t-test is in general

misspecified, but its bootstrapped version is well specified in most situations. The sign test is misspecified when applied to abnormal returns from reference portfolio benchmarks, and the extent of misspecification is quite serious and increases in the length of holding period. This is not surprising because abnormal returns from reference portfolio benchmarks have highly negative medians.

[Table 14.3 is about here]

Table 14.4 reports empirical power of testing the null hypothesis of zero abnormal return against the two-sided alternative hypothesis. We follow Brown and Warner (1980, 1985) to measure empirical power by intentionally forcing the mean abnormal return away from zero with induced abnormal returns. We induce nine levels of abnormal returns ranging from –20% to 20% at an increment of 5%. To induce an abnormal return of -20%, for example, we add -20% to the observed holding period return of an event firm. Empirical power is calculated as the proportion of 250 samples that rejects the null hypothesis at 5% significance level.

[Table 14.4 is about here]

With a large sample size of 1,000, the power of these tests remains reasonably high at the longer holding period. Ang and Zhang (2004) report that, with the sample size of 200, the power of all tests deteriorates sharply as holding period lengthens from one- to three- and to five-years and is alarmingly low at the five-year horizon. For example, when the induced abnormal return is –20% over a five-year horizon, the highest power of the bootstrapped Johnson's t-test is 13.6 percent for a sample of 200 firms, whereas the highest power is 62.8 percent for a sample of 1,000 firms.

We compare the power of the three test statistics: Student's t-test, the bootstrapped Johnson's skewness-adjusted t-test, and the sign test. All three test statistics are applied together with the most-correlated single firm benchmark. The evidence shows that all three tests are well specified. However, the sign test clearly has much higher power than the other two tests.

14.4.3 Simulation results for the calendar-time portfolio approach

Table 14.5 reports the rejection frequency of the calendar-time portfolio approach in testing the null hypothesis that the intercept is zero in the regression of monthly

calendar-time portfolio returns, against the two-sided alternative hypothesis. Rejection frequency is measured as the proportion of the total 250 samples that reject the null hypothesis. We compute rejection frequencies at nine *nominal* levels of induced abnormal returns, ranging from –20% to 20% at an increment of 5%. Since monthly returns of the calendar-time portfolio are used in fitting the model, to examine the power of testing the intercept, we need to induce abnormal returns by adding an extra amount to actual monthly returns of every event firm before forming the calendar-time portfolios. For example, in order to induce the –20% nominal level of abnormal holding period return, we add the extra amount of –1.67% (=-20%/12) to an event firm's twelve monthly returns for a one-year horizon, or add the abnormal amount of –0.56% (= -20%/36) to the firm's 24 monthly returns for a three-year horizon, or the abnormal amount of –0.33% (=-20%/60) to the firm's 60 monthly returns for a five-year horizon.

[Table 14.5 is about here]

Note that the *nominal* induced holding period return is different from the *effective* induced abnormal holding period return, because adding the abnormal amount each month does not guarantee that an event firm's holding period return will be increased or decreased by the exact nominal level. We measure the *effective* induced holding period return of an event firm as the difference in the firm's holding period return between before and after adding the monthly abnormal amount. The *average* effective induced holding period return is computed over all event firms in the 250 samples. The average induced holding period return allows us to compare power of the buy-and-hold benchmark approach with that of the calendar-time portfolio approach at the scale of holding period return.

We first examine empirical size of the calendar-time portfolio approach, which is equal to the rejection frequency when no abnormal return is induced. In Table 14.5, the empirical size is in the column with zero induced return. It is very surprising that when the four-factor model is used, the test has excessively high rejection frequency at three-year and five-year horizons. The rejection frequency, for example, is 94.0% at the five-year horizon with the WLS estimation! In contrast, when the Fama-French three-factor model is used, the empirical sizes are not significantly different from the 5% significance level. The evidence strongly suggests that the three-factor model is preferred for the

calendar-time portfolio approach, whereas the four-factor model suffers from overfitting and should not be used.

Table 14.5 shows that, for a sample of 1,000 firms, the power of this approach remains high as event horizon increases. WLS estimation does improve the power of the procedure over the OLS, and the extent of improvement becomes greater as holding period gets longer. By comparing Tables 14.4 and 14.5, we find that the power of the Fama-French calendar-time approach implemented with WLS technique, i.e. (FF, WLS), has almost the same power as the buy-and-hold benchmark approach implemented with the most-correlated single firm and the sign test, i.e. (MC1, sign), at the one-year horizon, but slightly less at the three- and five-year horizons.

14.5 Conclusion

Comparing the simulation results in Section 14.4 with those in Ang and Zhang (2004), we find that sample size has a significant impact on the performance of tests in long-horizon event studies. With a sample size of 1,000, a few tests perform reasonably well, including the Fama-French calendar-time approach implemented with WLS technique and the buy-and-hold benchmark approach implemented with the most-correlated single firm (MC1) and the sign test. In particular, they have reasonably high power even for the long five-year holding period. On the contrary, with a sample size of 200, Ang and Zhang (2004) find that the power of most well-specified tests is very low for the five-year horizon, only in the range of 10% to 20% against a high level of induced abnormal returns, while the combination of the most-correlated single firm and the sign test stands out with a power of 41.2%. Thus, the most correlated single firm benchmark dominates for most practical sample sizes and, in addition, the simplicity of the sign test is appealing.

The findings have important implications for future research. For long-horizon event studies with a large sample, it is likely to be more fruitful to spend efforts on understanding the characteristics of the sample firms, than on implementing various sophisticated testing procedures. The simulation results here show that the commonly used tests following both the Fama-French calendar-time approach and the buy-and-hold benchmark approach perform reasonably well. In a recent paper, Butler and Wan (2010)

reexamine the long-run underperformance of bond-issuing firms and find that straight debt and convertible debt issuers appear to have systematically better liquidity than benchmark firms, and controlling for liquidity by having an additional matching criterion eliminates the underperformance. This resonates well with Barber and Lyon (1997)'s suggestion that "as future research in financial economics discovers additional variables that explain the cross-sectional variation in common stock returns, it will also be important to consider these additional variables when matching sample firms to control firms" (pp. 370–71). One reason why the benchmark with a single most correlated firm performs well in our simulations may be that returns of highly correlated firms are likely to move in tandem in response to changes in risk factors that are well known, such as the market, size, book-to-market ratio, but also changes in other factors, such as industry, liquidity, momentum, and seasonality, etc.

On the other hand, for long-horizon event studies with a small sample, it may be necessary to use a wide range of tests and interpret their outcome with care. This prompts researchers to continue searching for better test statistics. For example, Kolari and Pynnonen (2010) find that even relatively low cross-correlation among abnormal returns in a short event window causes serious over-rejection of the null hypothesis. They propose both cross-correlation and volatility-adjusted as well as cross-correlation-adjusted scaled test statistics and demonstrate that these statistics perform well in samples of 50 firms. It is an open and interesting question whether these statistics have high power in long-horizon event studies with a small sample.

Appendix

This appendix includes the details on the benchmarks and the test statistics that are used in our simulation studies. We use five benchmarks. The *first benchmark* is a reference portfolio constructed on the basis of firm size and BE/ME. We follow Lyon, Barber and Tsai (1999) to form 70 reference portfolios at the end of June in each year from 1979 to 1997. At the end of June of year t, we calculate the size of every qualitified firm as price per share multiplied by shares outstanding. We sort all NYSE firms by firm size into ten portfolios, each having the same number of firms, and then place all AMEX/Nasdaq firms into the ten portfolios based on firm size. Since a majority of Nasdaq firms are small, approximately 50 percent of all firms fall in the smallest size decile. To obtain portfolios with the same number of firms, we further partition the smallest size decile into five subportfolios by firm size without regard to listing exchange. We now have 14 size portfolios. Next, we calculate each qualified firm's BE/ME as the ratio of the book equity value (COMPUSTAT data item 60) of the firm's fiscal year ending in year t-1 to its market equity value at the end of December of year t-1. We then divide each of the 14 portfolios into five subportfolios by BE/ME, and conclude the procedure with 70 reference portfolios on the basis of size and BE/ME.

The size and BE/ME matched reference portfolio of an event firm is taken to be the one of the 70 reference portfolios constructed at the month of June prior to the event month that matches the event firm in size and BE/ME. The return on a size and BE/ME matched reference portfolio over τ months is calculated as:

$$BR_i^{SZBM} = \prod_{t=0}^{\tau-1}\left[1+\frac{\sum_{j=1}^{n_t} r_{jt}}{n_t}\right]-1, \qquad (14A.1)$$

where month $t=0$ is the event month, n_t is the number of firms in month t, and r_{jt} is the monthly return of firm j in month t. We use the label 'SZBM' for the benchmark that is based on firm size and BE/ME.

The *second benchmark* is a reference portfolio constructed on the basis of firm size, BE/ME, and market beta. The Fama-French three factor model suggests that

expected stock returns are related to three factors: a market factor, a size related factor and a BE/ME related factor. Reference portfolios constructed on the basis of size and BE/ME account for the systematic portion of expected stock returns due to the size and BE/ME factors, but not the portion due to the market factor. Our second benchmark is based on firm size, BE/ME, and market beta to take all three factors into account.

To build a three-factor reference portfolio for a given event firm, we first construct the 70 size and BE/ME reference portfolios as above and identify the one that matches the event firm. Next, we pick firms within the matched portfolio that have returns in CRSP monthly returns database for all 24 months prior to the event month and compute their market beta by regressing the 24 monthly returns on the value weighted CRSP return index. Lastly, we divide these firms that have market beta into three portfolios by their rankings in beta and pick the one that matches the event firm in beta as the three-factor reference portfolio. The return on a three-factor portfolio over τ months is calculated as:

$$BR_i^{SZBMBT} = \prod_{t=0}^{\tau-1}\left[1+\frac{\sum_{j=1}^{l_t} r_{jt}}{n_t}\right]-1, \qquad \text{(14A.2)}$$

where month $t=0$ is the event month, n_t is the number of firms in month t, and r_{jt} is the monthly return of firm j in month t. We use the label 'SZBMBT' to indicate that the benchmark is based on firm size, BE/ME, and market beta.

The *third benchmark* is a reference portfolio constructed on the basis of firm size, BE/ME, and pre-event correlation coefficient. The rational for using pre-event correlation coefficient as an additional dimension is that returns of highly correlated firms are likely to move in tandem in response to not only changes in "global" risk factors, such as the market factor, the size factor, and the BE/ME factor in the Fama-French model, but also changes in other "local" factors, such as the industry factor, the seasonal factor, liquidity factor, and the momentum factor. Over a long time period following an event, both global and local factors experience changes that affect stock returns. It is reasonable to expect more correlated stocks would be affected by these factors similarly, and should have resulting stock return patterns that are closer to each other. Therefore, returns of a

reference portfolio on the basis of pre event size, BE/ME, and pre-event correlation coefficient are likely to be better estimate of the *status quo* (i.e., what if there was no event) return of an event firm.

To build a reference portfolio on the basis of size, BE/ME and pre-event correlation coefficient, we first construct the same 70 size and BE/ME reference portfolios as above and identify the combination that matches the event firm. Next, we pick firms within the matched size and BE/ME reference portfolio that have returns in CRSP monthly returns database for all 24 months prior to the event month, and compute their correlation coefficients with the event firm over the pre-event 24 months. Lastly, we choose the ten firms that have the highest pre event correlation coefficient with the event firm to form the reference portfolio. Return of the portfolio over τ months is calculated as:

$$BR_i^{MC10} = \sum_{j=1}^{10} \frac{\prod_{t=0}^{\tau-1}\left(1+r_{jt}\right)-1}{10}, \qquad (14A.3)$$

where month t=0 is the event month, r_{jt} is the monthly return of firm j in month t. We use the label 'MC10' to indicate that the benchmark consists of the most correlated ten firms. The benchmark return is the return of investing equally in the ten most correlated firms over the τ months beginning with the event month. The benchmark is to be considered as a hybrid between the reference portfolio discussed above, and the matching firm approach shown below.

The *fourth benchmark* is a single firm matched to the event firm in size and BE/ME. Barber and Lyon (1997) report that using a size and BE/ME matched firm as benchmark gives measurements of long-term abnormal return that is free of the new listing bias, the rebalancing bias, and the skewness bias documented in Kothari and Warner (1997) and Barber and Lyon (1997). To select the size and BE/ME matched firm, we first identify all firms that have a market equity value between 70% and 130% of that of the event firm, and then choose the firm with BE/ME closest to that of the event firm. The buy-and-hold return of the matched firm is computed as in equation (14.2). We use the label 'SZBM1' to represent the single size and BE/ME matched firm.

The *fifth and last benchmark* is a single firm that has the highest pre-event correlation coefficient with the event firm. Specifically, to select the firm, we first construct the 70 size and BE/ME reference portfolios and identify the one that matches the event firm. Next, we pick firms within the matched size and BE/ME reference portfolio that have returns in CRSP monthly returns database for all 24 months prior to the event month, and compute their correlation coefficients with the event firm over the pre-event 24 months. We choose the firm with the highest pre event correlation coefficient with the event firm as the benchmark. The buy-and-hold return of the most correlated firm is computed as in equation (14.2). We use the label 'MC1' to represent the most correlated single firm.

We apply four test statistics to test the null hypothesis of no abnormal returns: (a) Student's t-test, (b) Fisher's sign test, (c) Johnson's skewness-adjusted t-test, (d) bootstrapped Johnson's t-test.

(a) Student's t-test

Given the long-term buy-and-hold abnormal returns for a sample of n event firms, we compute Student's t-statistic as follows:

$$t = \frac{\overline{AR}}{s(AR)/\sqrt{n}}, \tag{14A.4}$$

where $\overline{AR}$ is the sample mean and $s(AR)$ the sample standard deviation of the given sample of abnormal returns. The Student's t-statistic tests the null hypothesis that the population mean of long-term buy-and-hold abnormal returns is equal to zero. The usual assumption for applying the Student's t-statistic is that abnormal returns are mutually independent and follow the same normal distribution.

(b) Fisher's sign test

To test the null hypothesis that the population median of long-term buy-and-hold abnormal returns is zero, we compute Fisher's sign test statistic as follows:

$$B = \sum_{i=1}^{n} \mathrm{I}(AR_i > 0), \tag{14A.5}$$

where $\mathrm{I}(AR_i > 0)$ equals 1 if the abnormal return on the ith firm is greater than zero, and 0 otherwise. At the chosen significance level of α, the null hypothesis is rejected in

favor of the alternative of non-zero median if $B \geq b(\alpha/2, n, 0.5)$ or $B < [n - b(\alpha/2, n, 0.5)]$, or in favor of positive median if $B \geq b(\alpha, n, 0.5)$, or in favor of negative median if $B < [n - b(\alpha, n, 0.5)]$. The constant $b(\alpha, n, 0.5)$ is the upper α percentile point of the binomial distribution with sample size n and success probability of 0.5. The usual assumption for applying the sign test is that abnormal returns are mutually independent and follow the same continuous distribution. Note that application of the sign test does not require the population distribution to be symmetric. When the population distribution is symmetric, the population mean equals the population median and the sign test then indicates the significance of the population mean (See Hollander and Wolfe (2000, Chapter 3)).

(c) Johnson's skewness-adjusted t-test

Johnson (1978) developed the following skewness-adjusted t-test to correct the misspecification of Student's t-test caused by the skewness of the population distribution. Johnson's test statistic is computed as follows:

$$J = t + \frac{1}{3\sqrt{n}} t^2 \gamma + \frac{1}{6\sqrt{n}} \gamma , \tag{14A.6}$$

where t is Student's t-statistic given in equation (14A.4) and γ is an estimate of the coefficient of skewness given by $\gamma = \frac{\sum_{i=1}^{n} (AR_i - \overline{AR})^3}{s(AR)^3 n}$. Johnson's t-test is applied to test the null hypothesis of zero mean under the assumption that abnormal returns are mutually independent and follow the same continuous distribution. At the chosen significance level of α, the null hypothesis is rejected in favor of the alternative of non-zero mean if $J > t(\alpha/2, \upsilon)$ or $J < -t(\alpha/2, \upsilon)$, or in favor of positive mean if $J > t(\alpha, \upsilon)$, or in favor of negative mean if $J < -t(\alpha, \upsilon)$. The constant $t(\alpha, \upsilon)$ is the upper α percentile point of the Student t distribution with the degrees of freedom $\upsilon = n - 1$.

(d) Bootstrapped Johnson's skewness-adjusted t-test

Sutton (1992) proposes to apply Johnson's t-test with a computer-intensive bootstrap resampling technique when the population skewness is severe and the sample size is small. He demonstrates it by an extensive Monte Carlo study that the bootstrapped Johnson's t-test reduces both type I and type II errors compared to Johnson's t-test.

Lyon, Barber and Tsai (1999) advocate the bootstrapped Johnson's t-test in that long-term buy-and-hold abnormal returns are highly skewed when buy-and-hold reference portfolios are used as benchmarks. They report that the bootstrapped Johnson's t-test is well specified and has considerable power in testing abnormal returns at the one-year horizon. In this paper, we document its power at three- and five-year horizons.

We apply the bootstrapped Johnson's t-test as follows. From the given sample of n event firms, we draw m firms randomly with replacement counted as one resample until we have 250 resamples. We calculate Johnson's test statistic as in equation (14A.6) for each resample and end up with 250 J values, labeled as $J_1, \cdots, J_{250}$. Let J_0 denotes the J value of the original sample. To test the null hypothesis of zero mean at the significance level of α, we first determine two critical values, c_1 and c_2, such that the percentage of J values less than c_1 equals $\alpha/2$ and the percentage of J values greater than c_2 equals $\alpha/2$, and then reject the null hypothesis if $J_0 < c_1$ or $J_0 > c_2$. We follow Lyon, Barber and Tsai (1999) to apply the bootstrapped Johnson's t-test with $m = 50$.[9]

[9] Noreen (1989, Chapter 4) cautions that bootstrap hypothesis tests can be unreliable and that extensive research is necessary to determine which one of many possible specifications can be trusted in a particular hypothesis testing situation. We also apply the bootstrapped Johnson's t-test with $m = 100$, 200. We find no significant difference in the test's performance.

References

Ang, J. S. and S. Zhang. 2004. "An evaluation of testing procedures for long horizon event studies." *Review of Quantitative Finance and Accounting* 23, 251–274.

Jegadeesh, N. and J. Karceski. 2009. "Long-run performance evaluation: correlation and heteroskedasticity-consistent tests." *Journal of Empirical Finance* 16, 101–111.

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

Brav, A. and P. A. Gompers. 1997. "Myth or reality? the long-run underperformance of initial public offerings: evidence from venture and nonventure capital-backed companies." *Journal of Finance* 52, 1791-1821.

Brav, A., C. Geczy and P. A. Gompers. 2000. "Is the abnormal return following equity issuances anomalous?" *Journal of Financial Economics* 56, 209-249.

Brown, S. J. and J. B. Warner. 1980. "Measuring security price performance." *Journal of Financial Economics* 8, 205-258.

Brown, S. J. and J. B. Warner. 1985. "Using daily stock returns: the case of event studies." *Journal of Financial Economics* 14, 3-31.

Butler, A. W. and H. Wan. 2010. "Stock market liquidity and the long-run stock performance of debt issuers." *Review of Financial Studies* 23, 3966-3995.

Carhart, M. M. 1997. "On persistence in mutual fund performance." *Journal of Finance* 52, 57-82.

Cowan, A. R. and A. M. A. Sergeant. 2001. "Interacting biases, non-normal return distributions and the performance of tests for long-horizon event studies." *Journal of Banking and Finance* 25, 741-765.

Dahlquist, M. and F. de Jong. 2008. "Pseudo market timing: a reappraisal." *Journal of Financial and Quantitative Analysis* 43, 547–580.

Davison, A. C. and D. V. Hinkley. 1997. *Bootstrap methods and their application*, Cambridge University Press.

Eckbo, B. E., R. W. Masulis and O. Norli. 2000. "Seasoned public offerings: resolutions of the 'new issues puzzle'." *Journal of Financial Economics* 56, 251-291.

Eckbo, B. E., and O. Norli. 2005. “Liquidity risk, leverage, and long-run IPO returns.” *Journal of Corporate Finance* 11, 1–35.

Efron, B. and R. J. Tibshirani. 1993. *An introduction to the bootstrap*, Chapman & Hall.

Fama, E. F. 1998. “Market efficiency, long-term returns and behavioral finance.” *Journal of Financial Economics* 49, 283-306.

Fama, E. F. and K. R. French. 1992. “The cross-section of expected stock returns.” *Journal of Finance* 47, 427-465.

Fama, E. F. and K. R. French. 1993. “Common risk factors in the returns on stocks and bonds.” *Journal of Financial Economics* 33, 3-56.

Hollander, M. and D. A. Wolfe. 1999. *Nonparametric statistical methods*, John Wiley & Sons.

Jegadeesh, N. 2000. “Long-term performance of seasoned equity offerings: benchmark errors and biases in expectations.” *Financial Management* 29, 5-30.

Jegadeesh, N. and S. Titman. 1993. “Returns to buying winners and selling losers: implications for stock market efficiency.” *Journal of Finance* 48, 65-91.

Jegadeesh, N. and S. Titman. 2001. “Profitability of momentum strategies: an evaluation of alternative explanations.” *Journal of Finance* 56, 699-720.

Johnson, N. J. 1978. “Modified *t* tests and confidence intervals for asymmetrical populations.” *Journal of the American Statistical Association* 73, 536-544.

Kolari, J. W. and S. Pynnonen. 2010. “Event study testing with cross-sectional correlation of abnormal returns.” *Review of Financial Studies* 23, 3996-4025.

Kothari, S. P. and J. B. Warner. 1997. “Measuring long-horizon security price performance.” *Journal of Financial Economics* 43, 301-340.

Kothari, S. P., and J. B. Warner. 2007. “Econometrics of event studies.” in *Handbooks of Corporate Finance: Empirical Corporate Finance*, B. E. Eckbo (Eds.), Amsterdam: Elsevier/North-Holland.

Lyon, J. D., B. M. Barber and C.-L. Tsai. 1999. “Improved methods for tests of long-run abnormal stock returns.” *Journal of Finance* 54, 165-201.

Loughran, T. and J. R. Ritter. 1995. “The new issues puzzle.” *Journal of Finance* 50, 23-52.

Loughran, T. and J. R. Ritter. 2000. “Uniformly least powerful tests of the market efficiency.” *Journal of Financial Economics* 55, 361-389.

Matsumoto, M. and T. Nishimura. 1998. “Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator.” *ACM Transactions on Modeling and Computer Simulation* 8, 3-30.

Mitchell, M. L. and E. Stafford. 2000. “Managerial decisions and long-term stock price performance.” *Journal of Business* 73, 287-329.

Noreen, E. W. 1989. *Computer intensive methods for testing hypotheses: an introduction*, John Wiley & Sons.

Pearson, E. S. and N. W. Please. 1975. “Relation between the shape of population distribution and the robustness of four simple test statistics.” *Biometrika* 62, 223-241.

Schultz, P. 2003. “Pseudo market timing and the long-run underperformance of IPOs.” *Journal of Finance* 58, 483-517.

Sutton, C. D. 1993. “Computer-intensive methods for tests about the mean of an asymmetrical distribution.” *Journal of the American Statistical Association* 88, 802-808.

Viswanathan, S. and B. Wei. 2008. “Endogenous events and long-run returns.” *Review of Financial Studies* 21, 855-888.

Table 14.1 Summary of existing simulation studies

Authors (Year)	Procedures under Investigation				Validation Settings		
	Calendar-Time Portfolio Approach		Buy-and-Hold Benchmark Approach		Simulation Design	Evidence on Specification at Horizon of	Evidence on Power at Horizon of
	Pricing Model	Estimation Method	Matching Criteria	Test Statistics			
Kothari and Warner (1997)	This paper examines procedures that are based on cumulating monthly abnormal returns. Such procedures are severely misspecified in most cases and are not recommended.				250 simulated samples of 200 firms each	1 month, 1, 2, and 3 years	1 year
Barber and Lyon (1997)			Size, BE/ME	t-test, Wilcoxon test	1,000 simulated samples of 200 firms each	1, 3, and 5 years	1 year
Lyon, Barber, and Tsai (1999)	Three-factor model	OLS	Size, BE/ME	t-test, Johnson's test, bootstrapped test	1,000 simulated samples of 200 firms each	1, 3, and 5 years	1 year, only for the buy-and-hold approach
Mitchell and Stafford (2000)	Three-factor model	OLS			1,000 simulated samples of 2,000 firms each	3 years	3 years
Cowan and Sergeant (2001)			Size, BE/ME	t-test, two-group test with winsorized data	1,000 simulated samples with sample size of 50, 200, and 1,000	1, 3, and 5 years	3 year
Ang and Zhang (2004)	Three-factor model, Four-factor model	OLS, WLS	Size, BE/ME, Beta, and correlation coefficient	t-test, Johnson's test, bootstrapped test, sign test	250 simulated samples with sample size of 200 and 1,000	1, 3, and 5 years	1, 3, and 5 years
Jegadeesh and Karceski (2009)			Size, BE/ME	t-test, t-test adjusted for heteroskedasticity and serial correlation	1,000 simulated samples of 200 firms each	1, 3, and 5 years	1, 3, and 5 years

Table 14.2 Descriptive statistics of abnormal returns in samples of *1,000 firms*

Benchmark	Descriptive Statistics					
	Mean	Median	Standard Deviation	Inter-Quartile Range	Skewness Coefficient	Kurtosis Coefficient
Panel A: One-Year Holding Period						
SZBM	0.009	-0.032	0.574	0.453	4.332	60.763
SZBMBT	-0.001	-0.043	0.586	0.462	4.074	58.462
MC10	0.000	-0.040	0.591	0.463	3.853	56.733
SZBM1	0.005	0.005	0.814	0.638	-0.203	53.034
MC1	0.002	-0.003	0.780	0.584	0.229	53.202
Panel B: Three-Year Holding Period						
SZBM	0.034	-0.112	1.240	0.963	4.561	57.644
SZBMBT	-0.001	-0.139	1.264	0.982	4.258	54.616
MC10	0.000	-0.126	1.286	0.982	3.996	53.153
SZBM1	0.023	0.022	1.746	1.305	-0.137	51.176
MC1	0.016	-0.006	1.658	1.200	0.736	43.430
Panel C: Five-Year Holding Period						
SZBM	0.068	-0.209	2.034	1.490	5.287	67.521
SZBMBT	0.002	-0.248	2.073	1.514	4.982	64.364
MC10	0.007	-0.223	2.106	1.516	4.652	61.091
SZBM1	0.054	0.039	2.802	1.979	0.269	41.428
MC1	0.036	0.000	2.745	1.834	0.500	50.365

This table reports descriptive statistics that characterize the probability distribution of long-term abnormal returns, in samples of *1,000 firms*. Abnormal return is calculated as the difference in holding period return between the event firm and its benchmark. We use five benchmarks: a reference portfolio matched by size and BE/ME (SZBM), a reference portfolio matched by size, BE/ME, and beta (SZBMBT), a reference portfolio consisting of ten firms, within the event firm's size and BE/ME matched portfolio, whose returns are most correlated with the event firm's (MC10), a single firm matched by size and BE/ME (SZBM1), and a single firm, from the event firm's size and BE/ME matched portfolio, whose returns have the highest correlation with the event firm's (MC1). We compute mean, median, standard deviation, inter-quartile range, skewness coefficient, and kurtosis coefficient for abnormal returns in every sample. Since there are 250 samples in the simulation, entries in the table are the average of these statistics over the 250 samples.

Table 14.3 Specification of tests in samples of *1,000 firms*

Benchmark	Two-Tail Test				Lower-Tail Test				Upper-Tail Test			
	t	Jt	BJt	sign	t	Jt	BJt	sign	t	Jt	BJt	sign
Panel A: One-Year Holding Period												
SZBM	4.0	7.2	6.4	75.6*	2.8	2.0*	1.6*	85.6*	8.4*	15.2*	13.6*	0.0*
SZBMBT	5.2	6.4	4.0	92.0*	9.6*	9.2*	7.6	96.0*	2.0*	4.8	4.0	0.0*
MC10	5.6	6.4	6.4	85.6*	10.0*	8.0*	6.8	92.8*	2.4*	5.6	4.8	0.0*
SZBM1	4.4	5.6	3.6	4.0	2.8	4.4	2.8	1.6*	6.0	8.0*	6.4	10.8*
MC1	3.6	5.2	3.2	9.6*	6.0	8.0*	4.8	12.8*	6.8	7.6	6.4	2.4
Panel B: Three-Year Holding Period												
SZBM	11.2*	14.4*	12.8*	99.6*	1.2*	1.2*	0.8*	100.0*	17.6*	22.8*	21.6*	0.0*
SZBMBT	5.2	5.2	5.6	100.0*	7.6	7.6	5.6	100.0*	2.8	3.2	3.6	0.0*
MC10	4.8	6.8	5.6	100.0*	6.8	5.6	4.4	100.0*	3.2	6.0	5.6	0.0*
SZBM1	6.0	7.6	5.2	9.2*	2.0*	2.8	1.6*	0.4*	11.2*	13.6*	10.0*	15.6*
MC1	6.8	8.4*	6.4	6.4	2.8	2.8	2.8	7.6	7.6	8.0*	6.8	1.2*
Panel C: Five-Year Holding Period												
SZBM	17.6*	20.4*	19.6*	100.0*	0.4*	0.4*	0.4*	100.0*	22.8*	28.8*	28.0*	0.0*
SZBMBT	3.6	4.4	2.8	100.0*	5.6	4.0	3.2	100.0*	2.8	5.6	3.6	0.0*
MC10	2.0*	4.4	2.8	100.0*	3.6	3.2	2.4	100.0*	3.6	6.4	5.6	0.0*
SZBM1	8.0*	10.4*	6.4	12.0*	1.2*	1.2*	1.2*	0.0*	13.6*	15.2*	11.2*	19.6*
MC1	6.0	8.4*	5.2	2.4	1.6*	2.0*	2.0*	4.0	10.8*	12.4*	9.2*	3.2

This table reports empirical size of testing the null hypothesis of zero abnormal return against two-tailed, lower-tailed, and upper-tailed alternative hypothesis, in samples of *1,000 firms*. Empirical size is calculated as the proportion of 250 samples that reject the null hypothesis at 5% significance level. Abnormal return is calculated as the difference in holding period return between the event firm and its benchmark. We use *five benchmarks*: a reference portfolio matched by size and BE/ME (SZBM), a reference portfolio matched by size, BE/ME, and beta (SZBMBT), a reference portfolio consisting of ten firms, within the event firm's size and BE/ME matched portfolio, whose returns are most correlated with the event firm's (MC10), a single firm matched by size and BE/ME (SZBM1), and a single firm, from the event firm's size and BE/ME matched portfolio, whose returns have the highest correlation with the event firm's (MC1); and *four test statistics*: the conventional t-test (t), Fisher's sign test (sign), Johnson's skewness-adjusted t-test (Jt), and the bootstrapped Johnson's skewness-adjusted t-test (BJt). It is indicated by * that the empirical size is significantly different from the 5% significance level. The significance is judged against the critical values $0.05 \pm 1.96\sqrt{0.05(1-0.05)/250}$, where 0.05 is the theoretical size, 1.96 is the 97.5th percentile of the standard normal distribution, and 250 is the sample size.

Table 14.4 Power of tests in samples of *1,000 firms*

Test	Benchmark	Induced abnormal return over the holding period (%)								
		-20	-15	-10	-5	0	5	10	15	20
Panel A: One-Year Holding Period										
t	SZBM	100.0	99.6	98.4	62.0	4.0	92.8	100.0	100.0	100.0
	SZBMBT	100.0	99.6	98.8	76.8	5.2	79.2	100.0	100.0	100.0
	MC10	100.0	99.6	98.4	73.6	5.6	77.6	100.0	100.0	100.0
	SZBM1	100.0	99.6	93.6	46.8	4.4	58.4	97.6	99.2	100.0
	MC1	100.0	99.6	97.2	50.8	3.6	58.0	96.8	99.6	100.0
Jt	SZBM	89.2	94.4	93.2	55.2	7.2	94.4	100.0	100.0	100.0
	SZBMBT	89.6	94.4	95.2	69.6	6.4	83.2	100.0	100.0	100.0
	MC10	91.2	95.6	95.2	66.4	6.4	80.0	100.0	100.0	100.0
	SZBM1	98.4	97.6	92.0	47.6	5.6	58.8	94.8	98.0	98.0
	MC1	98.0	98.4	95.6	50.0	5.2	59.2	95.6	98.0	98.0
BJt	SZBM	80.8	86.0	85.2	47.6	6.4	93.2	100.0	100.0	100.0
	SZBMBT	79.2	85.2	86.0	57.2	4.0	81.2	100.0	100.0	100.0
	MC10	81.6	86.4	87.2	56.4	6.4	78.8	100.0	100.0	100.0
	SZBM1	96.0	96.0	87.2	40.4	3.6	51.6	90.0	95.2	94.0
	MC1	95.6	95.6	88.8	44.4	3.2	51.6	91.6	95.6	95.6
sign	SZBM	100.0	100.0	100.0	100.0	75.6	28.4	100.0	100.0	100.0
	SZBMBT	100.0	100.0	100.0	100.0	92.0	10.4	99.2	100.0	100.0
	MC10	100.0	100.0	100.0	100.0	85.6	17.2	100.0	100.0	100.0
	SZBM1	100.0	100.0	100.0	72.0	4.0	92.0	100.0	100.0	100.0
	MC1	100.0	100.0	100.0	93.6	9.6	90.4	100.0	100.0	100.0

This table reports empirical power of testing the null hypothesis of zero abnormal return against the two-sided alternative hypothesis, in samples of *1,000 firms*. Empirical power is calculated as the proportion of 250 samples that reject the null hypothesis at 5% significance level. Abnormal return is calculated as the difference in holding period return between the event firm and its benchmark. We use *five benchmarks*: a reference portfolio matched by size and BE/ME (SZBM), a reference portfolio matched by size, BE/ME, and beta (SZBMBT), a reference portfolio consisting of ten firms, within the event firm's size and BE/ME matched portfolio, whose returns are most correlated with the event firm's (MC10), a single firm matched by size and BE/ME (SZBM1), and a single firm, from the event firm's size and BE/ME matched portfolio, whose returns have the highest correlation with the event firm's (MC1); and *four test statistics*: the conventional t-test (t), Johnson's skewness-adjusted t-test (Jt), the bootstrapped Johnson's skewness-adjusted t-test (BJt) and Fisher's sign test (sign). We study power at nine levels of induced abnormal return, ranging from -20% to 20% at an increment of 5%.

Table 14.4 (continued)

Test	Benchmark	Induced abnormal return over the holding period (%)								
		-20	-15	-10	-5	0	5	10	15	20
Panel B: Three-Year Holding Period										
t	SZBM	96.0	80.8	43.2	9.6	11.2	58.0	96.0	100.0	100.0
	SZBMBT	98.4	93.2	70.8	30.8	5.2	19.2	73.2	98.4	100.0
	MC10	98.4	92.4	70.0	26.8	4.8	19.2	72.4	98.8	100.0
	SZBM1	88.4	63.6	30.8	10.0	6.0	27.6	64.4	85.6	96.4
	MC1	92.4	74.0	36.4	10.4	6.8	22.4	64.0	91.2	97.6
Jt	SZBM	91.2	74.8	38.4	9.6	14.4	66.4	96.4	100.0	100.0
	SZBMBT	94.8	88.0	65.6	26.0	5.2	24.4	78.4	98.8	100.0
	MC10	94.0	87.6	62.8	24.4	6.8	23.2	76.8	99.2	100.0
	SZBM1	86.4	63.2	32.4	12.4	7.6	29.2	64.0	84.8	94.4
	MC1	90.4	72.4	36.0	12.0	8.4	24.4	64.8	90.4	97.6
BJt	SZBM	84.8	66.0	32.4	7.6	12.8	62.4	96.0	100.0	100.0
	SZBMBT	88.8	82.0	58.4	21.6	5.6	21.6	74.8	98.4	100.0
	MC10	90.4	79.6	54.8	19.6	5.6	21.6	73.6	98.0	100.0
	SZBM1	81.6	56.4	27.6	8.0	5.2	24.4	56.4	79.6	88.8
	MC1	86.0	65.6	29.6	8.8	6.4	20.4	55.2	86.8	94.4
sign	SZBM	100.0	100.0	100.0	100.0	99.6	63.6	6.0	27.2	88.8
	SZBMBT	100.0	100.0	100.0	100.0	100.0	89.6	27.6	6.8	64.4
	MC10	100.0	100.0	100.0	100.0	100.0	78.4	15.6	12.4	78.8
	SZBM1	100.0	94.8	56.4	14.0	9.2	54.8	94.0	100.0	100.0
	MC1	100.0	100.0	95.2	50.0	6.4	37.2	86.4	100.0	100.0

Table 14.4 (continued)

Test	Benchmark	Induced abnormal return over the holding period (%)								
		-20	-15	-10	-5	0	5	10	15	20
Panel C: Five-Year Holding Period										
t	SZBM	58.0	28.4	9.6	1.6	17.6	40.8	79.2	97.6	99.2
	SZBMBT	84.8	63.2	37.6	14.8	3.6	8.4	32.8	66.0	92.0
	MC10	80.4	61.2	32.8	11.6	2.0	10.4	32.4	69.6	92.0
	SZBM1	38.0	18.4	7.2	4.0	8.0	21.6	41.6	64.8	82.4
	MC1	50.4	23.6	10.4	4.0	6.0	17.2	38.0	61.2	81.2
Jt	SZBM	44.4	23.6	7.2	5.6	20.4	52.0	85.2	98.8	99.6
	SZBMBT	72.0	51.6	27.6	11.2	4.4	14.8	40.0	73.2	94.8
	MC10	71.6	51.2	27.6	7.6	4.4	15.6	38.0	73.6	96.4
	SZBM1	38.0	20.0	8.8	6.0	10.4	23.6	42.4	65.2	82.0
	MC1	48.8	24.4	11.2	6.0	8.4	18.8	38.4	60.4	79.6
BJt	SZBM	35.2	19.6	5.2	2.8	19.6	48.0	82.8	98.0	99.6
	SZBMBT	62.8	43.2	21.6	8.4	2.8	12.8	36.8	70.0	94.0
	MC10	60.0	42.4	20.0	6.4	2.8	14.4	36.0	72.0	94.8
	SZBM1	30.0	16.0	5.6	4.0	6.4	16.4	33.6	54.4	70.4
	MC1	38.4	17.6	9.6	3.2	5.2	15.2	30.8	50.4	70.8
sign	SZBM	100.0	100.0	100.0	100.0	100.0	95.6	72.0	22.8	3.2
	SZBMBT	100.0	100.0	100.0	100.0	100.0	99.6	93.2	61.6	19.2
	MC10	100.0	100.0	100.0	100.0	100.0	98.4	81.6	40.8	7.6
	SZBM1	91.2	63.2	20.8	4.0	12.0	48.8	86.0	97.2	100.0
	MC1	99.2	92.8	59.2	22.8	2.4	20.4	67.2	93.6	99.2

Table 14.5
Rejection frequency of calendar-time portfolio approach in samples of *1,000 firms*

Panel A: One-Year Holding Period

		Average effective induced holding period return (%)								
		-20.4	-15.7	-10.7	-5.5	0	5.7	11.7	17.9	24.4
Three Factors	OLS	100.0	100.0	99.2	53.2	2.4	78.8	100.0	100.0	100.0
	WLS	100.0	100.0	99.6	74.4	2.0*	82.8	100.0	100.0	100.0
Four Factors	OLS	100.0	99.2	90.8	18.0	28.0*	97.6	100.0	100.0	100.0
	WLS	100.0	99.6	93.2	20.8	25.2*	98.8	100.0	100.0	100.0

Panel B: Three-Year Holding Period

		Average effective induced holding period return (%)								
		-25.2	-19.3	-13.2	-6.8	0	7.1	14.5	22.3	30.4
Three Factors	OLS	98.0	86.8	38.0	3.6	2.4	32.0	84.8	99.6	99.6
	WLS	100.0	97.2	65.2	10.0	1.2*	36.0	91.6	100.0	100.0
Four Factors	OLS	69.2	22.0	1.6	6.4	55.2*	94.0	99.6	100.0	100.0
	WLS	92.0	38.0	4.0	10.4	75.6*	99.6	100.0	100.0	100.0

Panel C: Five-Year Holding Period

		Average effective induced holding period return (%)								
		-31.1	-23.9	-16.3	-8.3	0	8.7	17.9	27.4	37.5
Three Factors	OLS	64.8	31.2	10.0	0.8	4.0	27.6	62.4	90.8	99.6
	WLS	94.4	58.4	14.8	0.4	4.0	36.0	81.2	99.2	100.0
Four Factors	OLS	12.4	1.6	5.2	32.8	70.8*	89.2	98.8	100.0	100.0
	WLS	14.0	1.2	14.8	62.4	94.0*	100.0	100.0	100.0	100.0

This table reports rejection frequency in testing the null hypothesis that the intercept in the regression of monthly calendar-time portfolio returns is zero, in samples of *1,000 firms*. Both the Fama-French three-factor model and the four-factor model are used in the regression. Model parameters are estimated with both OLS and WLS estimation technique. Rejection frequency is equal to the proportion of 250 samples that reject the null hypothesis at 5% significance level. We measure rejection frequency at nine levels of induced abnormal returns. We induce abnormal returns by adding an extra amount to monthly returns of every event firm before forming the calendar-time portfolios. The *effective* induced holding period return of an event firm is equal to the difference in the firm's holding period return between before and after adding the monthly extra amount. The *average* effective induced holding period return is computed over all event firms in the 250 samples.

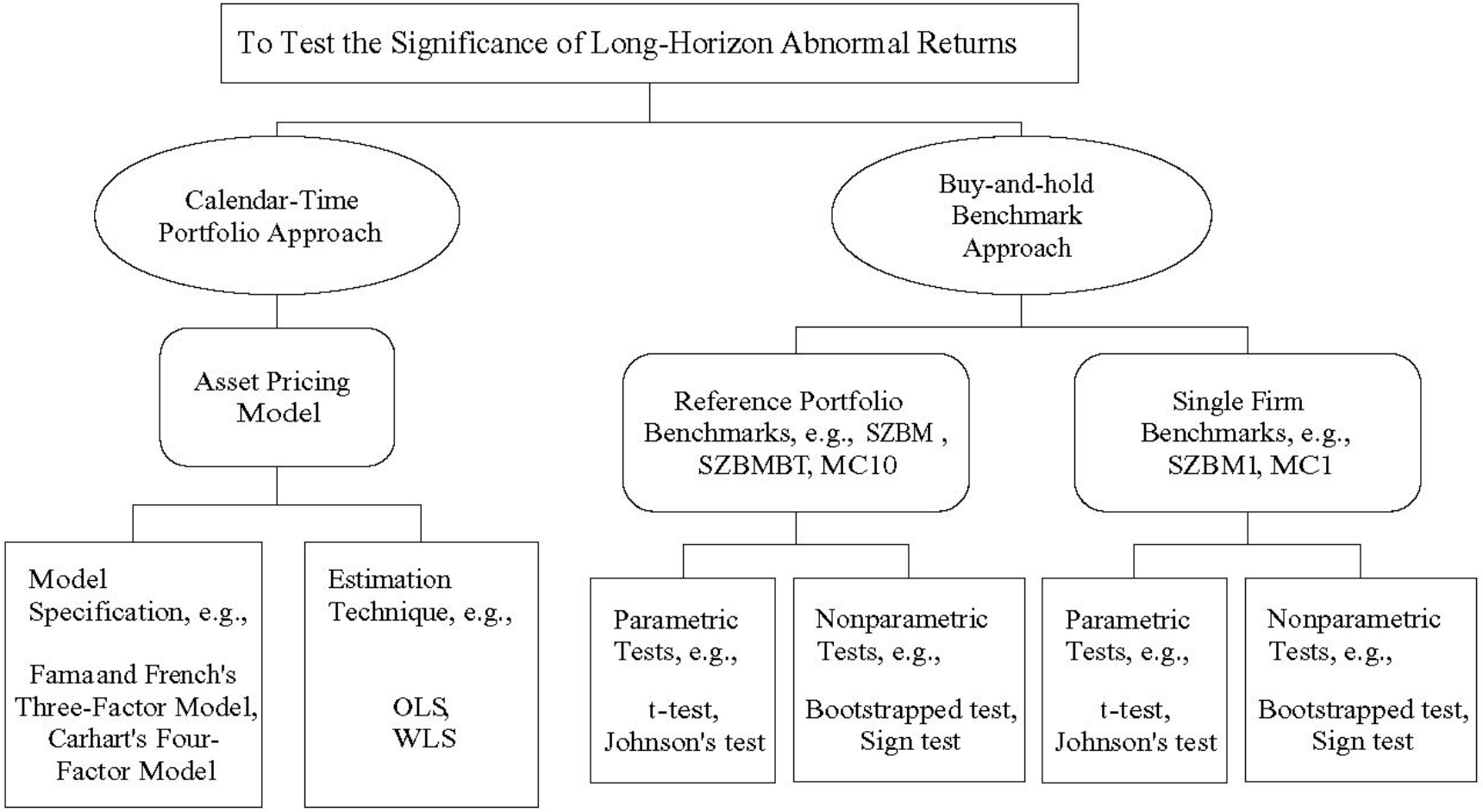

Figure 14.1 Overview of the two approaches to choose a methodology for long-horizon event study.