

Interacting Biases, Non-Normal Return Distributions and the Performance of Tests for Long-Horizon Event Studies

Arnold R. Cowan
Department of Finance
College of Business
Iowa State University
Ames, Iowa 50011–2063 USA
Voice: +1 515 294 9439
Fax: +1 515 294 3525
arnie@iastate.edu

Anne M.A. Sergeant
Department of Accounting
College of Business
Iowa State University
Ames, Iowa 50011–2063 USA
Voice: +1 515 294 2204
Fax: +1 515 294 3525
anne_sgt@iastate.edu

First Draft: October 1996
This Draft (5th): October 1999

We report simulations of one-, three-, and five-year abnormal buy-and-hold stock return tests. Using benchmark portfolios purged of new-listings and rebalancing biases, we find severe misspecification of most tests, due in part to skewness. Control-firm matching also results in misspecification, particularly in large samples. We document a negative relation between skewness bias and sample size, and an overlapping-horizons bias. Both biases become more severe as the holding period lengthens. The biases interact such that tests can be well-specified in one situation but not another. A two-groups test using winsorized abnormal returns yields correct specification and considerable power in many situations.

Keywords: Event studies, long-horizon performance, abnormal returns, winsorization.

JEL Classification: G12, G14, G30, M40

Comments are welcome.

1. Introduction

Many studies investigate long-horizon stock-price performance. Several report large abnormal returns following corporate events. The interpretation of such findings is controversial. Significant long-horizon abnormal returns are inconsistent with market efficiency. Kothari and Warner (1997) and Barber and Lyon (1997) report simulation results suggesting that many commonly used methods tend to find positive or negative abnormal performance when none is present.

This paper evaluates two new approaches to improving test specification and compares their performance to earlier test designs. First, we propose replacing the commonly used, simple paired difference test statistic with an equally simple two groups test statistic. The paired difference test implicitly controls for pairwise dependence between the return on a stock and the return on a size- and book-to-market-matched benchmark portfolio or control stock. Absent any other specification problem, ignoring positive pairwise dependence would lead to a weak test, that is, one that does not reject a false null hypothesis often enough. However, a typical long-horizon event study sample manifests another type of dependence that the paired difference test ignores. Specifically, overlapping time periods create positive dependence between the returns of different stocks in the sample (see Fama, 1998). By ignoring positive cross-sectional dependence while controlling for positive pairwise dependence, the paired difference test can be expected to produce too many rejections of a true null hypothesis. We refer to this effect as the *overlapping horizons bias*. The two groups test omits any adjustment for either type of dependence, allowing their interaction to potentially produce a better specified test. Our simulation results show that two groups tests indeed do generally exhibit better specification than paired difference tests.

The second approach to improving test specification is *winsorization* of abnormal returns, which we use to address the *skewness bias* described by Barber and Lyon (1997). Winsorization is a well-known procedure that sets a limit on how far away from the rest of the sample an extreme observation is allowed to be. More extreme observations are set equal to the limit, in effect giving the most extreme observations a lower weight but not removing them from the sample. Individual stock buy-and-hold long-horizon returns experience much larger extreme positive observations than extreme negative observations, resulting in substantial positive skewness. Benchmark portfolio returns, being broad averages, are less skewed than the individual stock returns from which they are subtracted to determine abnormal returns. As a result, long horizon abnormal returns are highly positively skewed like raw returns. The positive skewness leads to negatively biased test statistics (that is, excessive rejections of the null in favor of a lower-tail alternative; see Barber and Lyon for a detailed explanation). For example, in 1000 samples of 50 stocks with random event dates, tests of non-winsorized, buy-and-hold abnormal returns relative to an equally-weighted, firm size- and book-to-market matched benchmark, the paired difference test rejects the lower-tail null hypothesis about ten times as often as the upper-tail null.

Winsorization allows the investigator to explore the sensitivity of the inference to extreme returns. A plethora of studies finds abnormal performance following various corporate events. Extreme observations and the resulting positive skewness could account for these findings. Reporting the results of tests on both winsorized and non-winsorized abnormal returns should help readers judge the robustness of the conclusions.

Barber and Lyon (1997) identify two additional biases in previously studied tests. First, the *new listing* bias occurs when the long-horizon return of a benchmark portfolio reflects new listings that occur after the event date. Sample stocks, by definition, include only those already

trading as of the event date. Empirically, new listings tend to have lower long-horizon returns than other stocks, thereby biasing abnormal returns of sample stocks upward. We eliminate this bias by matching each sample stock with a benchmark portfolio of stocks already trading on the event date, and excluding subsequent new listings from the benchmark portfolio.

The *rebalancing bias* stems from the practice of using an equally weighted market index or benchmark portfolio, which is rebalanced to equal weights each month or even daily. This rebalancing leads to an inflated long-horizon benchmark return and thus a downward-biased abnormal return. (See Barber and Lyon, 1997 and Canina *et al.*, 1998 for more detailed discussion.) The tests in this paper avoid the rebalancing bias with equal-weighted benchmarks by not rebalancing after the event date.¹

Another contribution of this paper is the consideration of sample size as a factor in the specification of long-horizon event study tests. Sample sizes vary widely across existing studies, but the effect of sample size on specification has not been studied.² As the sample size increases sufficiently, the distribution of the sample mean should resemble the normal and the skewness bias should be small. However, whether the skewness bias is too small to detect in relatively large samples in practice is an empirical question. The question is also complicated by the potential for different biases to interact. For example, suppose that a sample of 200 stocks has a larger expected return than the matched benchmarks because of a sample selection bias of which the researcher is unaware. This would tend to bias tests toward finding positive abnormal performance and away from finding negative abnormal performance. However, the skewness bias may

¹ A contemporaneous paper by Lyon, Barber and Tsai (1999) uses similar approaches to eliminating the new listing and rebalancing biases.

² Sample sizes range from less than 50 stocks (Clark and Ofek, 1994; Ikenberry and Lakonishok, 1993) to more than 3,500 stocks (Loughran, 1993; Loughran and Ritter, 1995; Brav, Geczy, and Gompers, 1995).

offset the effect of sample selection bias, leading to an apparently well specified test. If the sample size is increased to 1000, the skewness bias diminishes but the sample selection bias need not. Thus, test specification can vary with a change in the sample size.

Finally, we study the effect of holding period length, which also varies across studies, on test performance.³ The overlapping horizons bias is expected to increase with the holding period as the number of contemporaneous horizons increases. Tests may appear well specified in shorter holding periods because the overlapping horizons problem is small, but specification can deteriorate as the holding period is increased. Moreover, because return skewness increases for longer holding periods, we expect the skewness bias to increase as well, all else held constant. Thus, tests that are misspecified in shorter holding periods are expected to become even worse in longer holding periods due to the interaction of the two biases.

This paper uses the simulation method introduced by Brown and Warner (1980) and adopted by many event methodology studies since: randomly select firms and event dates, and compute test statistics based upon actual return data. This type of simulation cannot cleanly isolate each potential bias, but has the advantage of revealing the combined effects of biases present in actual data. Additionally, we gain insights into the relative effects of individual biases by jointly varying the sample size and the holding period, and by comparing tests that should be differently sensitive to particular sources of bias.

We report results for simulations of paired difference and two groups tests for buy-and-hold abnormal returns cumulated over one-, three-, and five- year periods for samples of 50, 200

³ Commonly used longer periods are three years (Cusatis, Miles, and Woolridge, 1993; Loughran and Ritter, 1995; Desai and Jain, 1997; Teoh, Welch, and Wong, 1998) and five years (Spiess and Affleck-Graves, 1995; Loughran and Ritter, 1995; Brav, Geczy, and Gompers, 1995; Agarwal, Jaffe, and Mandelker, 1992).

and 1000 stocks. Previous long-horizon simulation research is limited to samples of 200 or fewer stocks; this paper is the first to examine 1000 stock samples. The tests use equal- and value-weighted benchmark portfolios, matched on size and book-to-market ratio, as well as control stocks. We find test misspecification even with benchmark portfolios void of new-listings and rebalancing biases, consistent with previous research (Kothari and Warner, 1997; Barber and Lyon, 1997). We document a negative relation between the magnitude of skewness bias and sample size and find significant skewness bias even in samples as large as 1000 stocks. There also is evidence of overlapping horizons bias, which like the skewness bias becomes serious with longer holding periods. However, a two groups test of winsorized returns using value-weighted, size and book-to-market matched benchmark portfolios is successful in reducing these biases. The test is powerful and nearly always exhibits appropriate specification at the reported portfolio sizes and holding periods. It also is more powerful and better specified for large sample sizes than the control-firm approach offered by Barber and Lyon.

2. Methods and data

2.1. Tests statistics studied

When the abnormal return of a stock is defined as the difference between the holding-period compound return and the corresponding benchmark return, a natural way to test the null hypothesis is to use a simple paired difference Z statistic (t test for smaller samples),

$$Z = \frac{AHPAR}{\sqrt{\hat{\sigma}_{HPAR}^2/N}},$$

where $AHPAR$ is the average holding period abnormal return (defined more precisely below) and

$\hat{\sigma}_{HPAR}^2$ is the cross-sectional sample variance of the holding-period abnormal returns. The null

hypothesis of the simple paired difference Z statistic is that there is no difference between the test stock and the benchmark. The procedure is common in the literature, and both Barber and Lyon (1997) and Kothari and Warner (1997) apply such a test to buy-and-hold abnormal returns. The rationale for using the variance of differences, as opposed to the variances of the stock and benchmark holding-period returns, is that the observations are not independent, but pairwise dependent. In the case of abnormal returns, the stock return and benchmark return are expected to be positively dependent. Ignoring the dependence would cause the sample standard deviation to be overestimated. The paired difference test implicitly controls for the pairwise dependence.

If the holding periods of stocks in the sample overlap, however, there is a potential cross-sectional dependence among the returns of sampled stocks. In typical event study samples, the number of nontrivially overlapping horizons will be considerable.⁴ Assuming that the cross-sectional dependence between stocks with overlapping holding-period returns is positive and that the correlation persists in the abnormal returns, the effect will be to increase the variance of the sample mean return. Conventional test statistics do not adjust for the dependence, so they can underestimate the variance, which inflates the absolute value of test statistics and causes the tests to reject the null hypothesis too often (see Brown and Warner, 1980). This *overlapping horizons bias* should not be large when the holding period is one year, but as the holding period lengthens, the bias is expected to grow.

To compensate for cross-sectional dependence, we consider a two groups difference of means test,

⁴ For example, there can be only nine completely distinct three-year holding periods between 1965 and 1992.

$$Z_{2G} = \frac{AHPAR}{\sqrt{\frac{\hat{\sigma}_{HPR}^2}{N} + \frac{\hat{\sigma}_{HPR(benchmark)}^2}{N}}}.$$

The null hypothesis is that the means of the two populations, the stocks and the benchmarks, are equal. In a two groups test, any pairing of the data is disregarded. The null hypothesis for which the statistic is derived is that the means of the two populations are equal. The two populations are assumed independent; there is no correction for either pairwise dependence or cross-sectional dependence. Both kinds of dependence are expected in long-run returns. However, since the two kinds affect the variance in opposite directions, the lack of a correction for either can potentially improve test specification relative to a test that corrects for only one form. In a random sample with no cross-sectional dependence, the test will be less powerful than a paired difference test. More generally, the two groups test admittedly is a crude approach to the issue of dependence. Arguably, it would be better to try and model the data more thoroughly. However, long-run returns exhibit such dramatic departures from the usual distributions as to render such modeling excessively complex. The two groups test has the virtue of being simple and easy to calculate. Whether it is too simple is an empirical question.⁵

2.2. *Population sampled*

We develop simulation samples from NYSE, AMEX and Nasdaq stocks listed on the 1995 CRSP daily file from 1965 through 1995 minus n where n is the year(s) in the holding period. Stocks of firms incorporated outside the U.S., American Depository Receipts, Americus Trust components, closed-end funds, unit investment trusts and real estate investment trusts are excluded. To be eligible for the test and benchmark simulation samples, a stock must appear on

⁵ Lyon, Barber and Tsai (1999) consider a test statistic based on more detailed modeling; however, it addresses only one source of bias, skewness.

the CRSP daily return file for the first day of the holding period. No requirements for prior or subsequent return data are imposed. However, because we consider size- and book-to-market matched benchmarks, some stocks drop out of the simulation samples for lack of market price and share data (from the CRSP file) and book-to-market data (from Compustat files).

2.3. *Simulation sample selection*

To simulate long-run event studies, we randomly select stocks and event dates. For the one-year holding period tests, stocks are sorted into deciles based on the number of trading days from 1965 through 1994 that the stock is listed on the CRSP daily file. The process is the same for three- and five-year holding periods, except the ending date becomes 1992 or 1990 respectively. Stocks are randomly selected in proportion to the number of available trading days. Thus, a stock with twice as many CRSP trading days as another would be twice as likely to be selected. For each stock selected, an event date (the holding-period start date) is randomly selected from within the available dates for that stock. We believe this sampling procedure is representative of samples found in many studies of corporate events and produces results that can be compared to Barber and Lyon (1997) and Kothari and Warner (1997).

We examine portfolios sizes of 50, 200 and 1000 stocks. For each portfolio size, we draw 1000 random stock-event date combinations. Stocks are replaced after being selected, but event dates are not replaced. Thus, the same combination of a stock and event date does not appear twice in a set of 1000 samples. Separate samples are drawn for the one-, three-, and five-year holding periods.

2.4. *Computation of returns*

For one-, three-, and five-year holding periods, we construct a n -year holding-period database which replaces each 1995 CRSP daily return with a buy-and-hold return that starts on that

day being replaced (t) and ends n years ($252n$ trading days) later. The n -year buy-and-hold return of stock j beginning at the close of day $t-1$ is

$$HPR_{jt} = \left[\prod_{\ell=0}^{252n-1} (1 + r_{jt+\ell}) \right] - 1,$$

where $r_{jt+\ell}$ is the return on day $t+\ell$ from the CRSP daily returns file. In other words, for each CRSP trading day, we compute an n year buy-and-hold return for each stock. Test stocks are drawn from this database, as are the benchmark portfolios.

If a stock stops trading before the end of the holding period, its daily return series is filled in with the delisting return when available, followed by a market return (see Shumway, 1997 for a discussion of the potential importance of the delisting return). We assume that the proceeds from the delisting are invested in a value-weighted market index portfolio, from the delisting date through the end of the n -year holding period. Daily value-weighted market index portfolio returns are constructed from the n -year holding-period return database weighted by the market value at the end of the previous trading day. An index return is used because it simulates a feasible strategy for an individual trader who could invest the residual proceeds in a passively managed fund that mimics a broad market index.

2.5. *Choice of benchmarks*

The benchmarks that we consider are a portfolio of all stocks that match the test stock on size and book-to-market equity, and a control stock that also matches the test stock on size and book-to-market equity. These are the main benchmarks that long-run event studies use. Portfolios matched on size and book-to-market are popular because Fama and French (1992, 1993) report that these characteristics explain much of the cross-sectional and time-series variation in stock returns.

We consider both equal-weighted and value-weighted benchmark portfolios. Each has commendable properties for purposes of long-horizon event studies. Event study means and test statistics conventionally are equal-weighted. Assume that identical past and future survival criteria are applied to event study samples and indices or benchmark portfolios. On average across many studies, the composition of equal-weighted benchmarks will perfectly match the composition of test samples, theoretically ensuring a zero mean. A significant source of misspecification is thereby eliminated. However, except in simulation studies like this one, neither identical survival criteria nor averaging across many samples is the norm. For example, takeover bidders tend to be firms that have survived. Value-weighted benchmarks also better represent actual investment opportunities.

Some studies use single control-firm stocks as benchmarks. Barber and Lyon (1997) report that the substitution of a control stock for a control portfolio improves the specification of statistical tests. The improvement comes from greater similarity of return skewness between sample stocks and individual stocks. Portfolio returns manifest less skewness than individual stocks. Therefore, in addition to benchmark portfolio tests, we study tests using control stocks matched on the basis of size and book-to-market ratio.

2.6. *Construction of benchmark portfolio returns*

We consider size and book-to-market buy-and-hold benchmark portfolios constructed from our n -year holding-period return database. Again following Fama and French (1993), 50 size and book-to-market portfolios are formed by annually ranking eligible stocks within the size deciles into quintiles based on their most recent fiscal year-end book-to-market ratio available at the end of February. We use the end of February as the book-to-market ranking date to allow up to a four month lag in the publication of annual financial statements.

For each benchmark portfolio, we compute an equal-weighted return and a value-weighted return. The equal (or value) weighted benchmark return is the arithmetic (or weighted) mean of all n -year buy-and-hold stock returns available starting on the event date for a particular benchmark portfolio. Averaging the full holding-period returns available on the specific trading date avoids the need to cumulate or compound benchmark subperiod returns, thus eliminating the rebalancing bias that Barber and Lyon (1997) identify. Moreover, the new listing bias is avoided because only the stocks appearing in the benchmark at the start of the holding period are included.⁶

In addition to portfolio benchmarks, we examine control-firm benchmarks. One stock is randomly selected from the appropriate portfolio and matched to the test stock.

2.7. *Benchmark matching and abnormal return calculation*

The average compounded, or *holding-period abnormal return*,

$$AHPAR = \frac{1}{N} \sum_{j=1}^N \left(HPR_j - HPR_{benchmark_j} \right),$$

incorporates the actual value change of each stock during the holding period, with any dividends reinvested.

When our simulations call for winsorized abnormal returns, we winsorize at plus or minus three standard deviations from the sample mean. The rationale for three standard deviations is that in a sample from a normal distribution, nearly all observations would be expected to fall within three standard deviations of the mean. By using standard deviation as the limiting criterion

⁶ Our method of constructing benchmark returns differs from that used by Barber and Lyon (1997). They use a method commonly found in existing studies which treats stock returns and benchmark portfolio returns differently. Stock returns are developed using buy-and-hold methods, similar to our study. However, their benchmark portfolios returns are computed by compounding cross-sectional average returns, thus introducing new-listing and rebalancing biases.

for heavily skewed distributions, the upper and lower tails of the distribution are differentially truncated. Specifically, positive abnormal returns, which are the source of the observed skewness, are much more likely to be truncated than negative abnormal returns. Winsorizing at three standard deviations from the sample mean affected only the upper tail of our distributions. In preliminary experiments, we found that 3.0 standard deviation limits produced better test specification than 2.5, 3.5 or 4.0 standard deviations.

When matching the benchmark portfolios or control firms to stocks, it is critical for the researcher to consider the effect of the event on size and book-to-market ratio. For example, Mitchell and Stafford (1998) and Rau and Vermaelen (1998) show that many acquiring firms change their size and book-to-market classifications as a result of the acquisition. Simulation studies like this one use random event dates, and thus they do not account for an event-induced shift in classification. This does not necessarily limit the generality of the conclusions, because studies investigating actual events can compensate for such a shift. For example, if a merger causes an acquirer to change in size and book-to-market ratio, the benchmark portfolio with the same size and book-to-market ranking as the *merged* firm, not the pre-merger acquiring firm, should be used.

Events do not necessarily induce only discrete shifts in size and book-to-market, however. It is also possible that classifications drift gradually over a period of years. Such a drift may be unrelated to the event, but occur for non-event firms in the same size- and book-to-market category. In this case, matching on size and book-to-market ratio at the event date, and maintaining the match for several years, can make the benchmark return be a noisier measure of expected return. When the drift is not event-related, simulations using random event dates will suffer as much or as little as actual event studies, so the applicability of our findings should be maintained.

On the other hand, size and book-to-market may drift in a systematic way for a long time because of an event. For example, suppose that mergers systematically result in plant closings or asset write-downs over a period of years, but the implementation of these decisions contains no news about firm value and the market is efficient. The market-value consequences of the decisions would be fully anticipated at the event date, but the book value of the firm would decline if not offset by increases in other assets. Thus, the book-to-market ratio could gradually drift downward. Such a systematic event-induced drift would cause event firms and benchmarks to be mismatched in the actual event study but not in a simulation with random event dates.

Researchers potentially can compensate for event-related ranking drift by appropriately selecting a sequence of benchmarks. For example, Brav and Gompers (1997) allow the size and book-to-market benchmark portfolio to change quarterly. Such periodic re-matching of event firms and benchmarks implies that a buy-and-hold investment strategy for the event firms is being compared to a rebalanced strategy for the benchmark. Suppose that a sequence of equal-weighted size and book-to-market benchmark portfolios is used. The wealth resulting from buying and holding one benchmark portfolio for a period is rolled over into a different portfolio, which starts the new period with equal weights. Thus, the attempt to compensate for benchmark-matching bias due to event-induced drift re-introduces the rebalancing bias that a buy-and-hold benchmark strategy avoids. We do not claim that the trade-off between the two biases should always be resolved in favor of a buy-and-hold benchmark strategy. As Fama (1998) points out, neither buy-and-hold nor rebalanced procedures are immune to “bad models” problems. However, simulation analysis of methods other than buy-and-hold is beyond the scope of this study.

2.8. *Method of artificially inducing abnormal returns*

To evaluate the power of the tests, we artificially induce various levels of positive and negative abnormal return. For example, to induce a 10% abnormal return, we simply add 10% to each stock's HPR before performing the remaining calculations to arrive at a test statistic. This could be interpreted as simulating either a 10% abnormal return on a single day or equivalent smaller daily abnormal returns compounded over the holding period.

3. **Results**

3.1. *Descriptive statistics of returns*

To understand the problem of non-normal return distributions and its relation to sample sizes, we provide descriptive statistics for raw and abnormal returns in Table 1. We report only the three-year holding period for size and book-to-market matching procedures. Evidence from one- and five-year holding periods yields similar insights. In order to highlight the changes across sample sizes, we include sample sizes of 25, 100 and 500 stocks in addition to the 50, 200 and 1000 used for the main simulations. We show the mean, standard deviation, centered coefficient of skewness and centered coefficient of kurtosis of the 1000 simulation samples for each sample size. Large samples (size=N) from a normal distribution have expected centered coefficients of skewness and kurtosis of zero and variances of the sample coefficients equal to $6/N$ and $24/N$, respectively. Given our 1000 simulation trials, 95% confidence limits for skewness and kurtosis are plus or minus 0.152 and 0.304, respectively.

Raw returns Table 1 shows that the mean of 1000 sample mean raw returns is around 58% triennially and does not vary much across sample sizes. The mean raw return is positively skewed and leptokurtic, especially in small samples. For example, in samples of size 25, the average centered skewness and kurtosis coefficients are 2.430 and 15.363, respectively. The coeffi-

cients decline substantially as the sample size increases from 25 to 1000, but even at sample sizes of 1000 stocks the sampling distribution of the sample mean is significantly skewed (0.337) and leptokurtic (0.386) relative to the normal distribution.

Abnormal returns Using size and book-to-market matching procedures, abnormal returns for equal-weighted, value-weighted and control stock matching are reported in Table 1. The abnormal returns have a positive mean, skewness, and kurtosis. These abnormal return statistics are expected to be zero under the standard event study null hypothesis. The buy-and-hold abnormal return (BHAR) based on the equal-weighted benchmark portfolio has the mean nearest zero for each sample size, around 1% triennially, followed by the control stock-based BHARs at 4%, and value-weighted benchmark portfolio-based BHARs at 6%. The centered coefficient of skewness is positive and significant for each matching procedure and for all sample sizes. The smallest skewness for samples of 25 stocks comes from BHARs using the value-weighted portfolios and is 2.516. In samples of 50 or larger, BHARs using control stocks are less skewed than the other matching portfolio abnormal returns. Consistent with the central limit theorem, skewness decreases monotonically with increasing sample size. For example, the average centered skewness coefficient of the mean abnormal return using control stocks decreases from 3.530 in samples of 25, to 0.407 in samples of 200, to 0.200 in samples of 1000. Like skewness, the centered coefficient of kurtosis reported in Table 1 is positive and significant for each matching procedure and for all sample sizes. A monotonic decline in kurtosis is present as sample size increases, which again is consistent with the central limit theorem.

Winsorized abnormal returns Table 1 reports the descriptive statistics for BHARs based on value-weighted benchmark portfolios after winsorizing at three standard deviations. Winsorizing reduces the mean BHARs, to about one-half of a percent triennially in most sample sizes.

The reduction in the mean is not surprising given positive skewness and truncation based on standard deviation. The effects of winsorizing on the second and higher moments are substantial as well. The standard deviation of the sample mean declines by about 25%. The skewness coefficient decreases from 2.516 to 1.545 in samples of size 25, and from 0.376 to 0.158 in samples of 1000. The kurtosis coefficient experiences similar reductions. While the sampling distribution of the winsorized mean abnormal return is closer to the normal distribution than that of other abnormal returns examined, it remains non-normal, particularly at smaller sample sizes.

3.2. *Specification of tests*

Although it is important to understand the distribution of abnormal returns underlying the test statistic, ultimately it is the performance of the test statistic itself that is critical. Table 2 reports the specification of paired difference and two groups tests for one-, three-, and five-year holding periods using equal- and value-weighted size/book-to-market benchmark portfolios and size/book-to-market matched control stocks.⁷ We report only five percent significance level tests; the one percent significance level tests yield similar conclusions. The numbers in the table are percentages of 1000 samples in which a test rejects the null hypothesis when no abnormal return is artificially induced.

3.2.1. *Paired difference tests*

Equal-weighted benchmark Table 2 shows that when the benchmark is the equal-weighted size and book-to-market matched portfolio, the paired difference test is negatively biased. It rejects the null hypothesis against a lower-tail alternative significantly more often than

⁷ The paired difference and two groups tests are biased for all holding periods and sample sizes when the benchmark is the market index alone or size alone. This holds for both equal- and value-weighted benchmarks. Thus, we do not present market index and size matching procedures nor further analyze them.

five percent, but the rejection rate against an upper-tail alternative is significantly less than five percent. For example, using a one-year holding period, and a sample of size of 50, the null is rejected in 12.3% of lower-tail tests, but only 1.8% of the time in upper-tail tests. The rejection rate against a two-tail alternative significantly exceeds five percent for all but the 1000 stock portfolios. The pattern of rejection rates is generally consistent with the skewness bias decreasing as the sample size increases. When the holding period is five years, the lower-tail and two-tail rejection rates decrease, and the upper-tail rate increases monotonically, across sample sizes. For example, the lower-tail rejection rate drops from 12.7% in samples of 50 stocks to 6.9%, just barely indistinguishable from 5%, in samples of 1000 stocks. In general, the equal-weighted benchmark results are consistent with Barber and Lyon (1997). Thus, while our benchmarks are purged of the new listing and rebalancing biases, the skewness bias still renders the test significantly misspecified.

Value-weighted benchmark Tests using value-weighted size and book-to-market benchmarks are positively biased for both one- and two-tail tests in samples of size 1000. This bias increases as the holding period is extended. For example, the upper-tail rejection rates for 1000 stock portfolios is 9.6% for one-year holding periods and 32.7% for five-year holding periods. As the sample size decreases, the positive bias is reduced so that at 50 stocks there is a negative bias for one-tail tests. Again, the bias is larger for longer holding periods.

Recall that the skewness bias, given positive long-horizon return skewness, causes tests to be negatively biased. The observed increase in positive bias as sample size increases in Table 2 is consistent with a decrease in the skewness bias. Moreover, the skewness bias has a larger negative effect in longer holding periods. However, the overlapping horizons bias also has a larger effect in longer holding periods, and the net of these effects tends to inflate the test statistic in the

same direction as the sample mean. The mean abnormal return using benchmark portfolios are positive, so tests based upon them are positively biased in most situations.

The skewness bias and the overlapping horizons bias (combined with the positive mean abnormal return) offset each other to produce an apparently unbiased two-tailed test, for example at the 200 stock sample size. The potential for offsetting biases to produce correct specification that is not stable across sample sizes or holding periods suggests the need for caution in interpreting the results of simulation studies.

Control stocks Table 2 also shows that paired difference tests using a size- and book-to-market-matched control stock are usually correctly specified for samples of 50 stocks. However, as the sample size increases, the tests become positively biased in one- and three-year holding periods, rejecting the null in 13.3–13.5% of upper-tail tests in samples of 1000. The increasing positive bias is consistent with the skewness bias diminishing in larger samples. It is not clear what other effects may be influencing the results, because the pattern is reversed in the five-year holding period. In samples of 1000 stocks, the null hypothesis is rejected in 1.2% of upper-tail tests but 10.4% of lower-tail tests.

Holding the sample size constant, the upper-tail rejection rate decreases as the holding period increases, from 7.0% to 3.0% in samples of 50 stocks, from 9.1% to 2.7% in samples of 200, and from 13.3% to 1.2% in samples of 1000. As the upper-tail rejections decrease, lower-tail rejections increase in a similar fashion. The increasing negative bias is consistent with the skewness bias increasing as the holding period increases, with an insufficiently large positive mean abnormal return, or an insufficient increase in the overlapping horizons bias, to offset it.

The results for the control-stock tests differ from those of Barber and Lyon (1997), who report correct, and reasonably symmetric, upper- and lower-tail rejection rates. The difference in

results could be the result of slightly different matching procedures between their study and ours. However, our results and theirs are roughly consistent for samples of 200 stocks or fewer, where similar to Barber and Lyon (1997) we observe much smaller biases than in equal-weighted benchmark tests, correctly specified two-tail tests, and only mild to moderate biases in the upper-tail tests. The more dramatic difference occurs in samples of 1000 stocks, which Barber and Lyon do not investigate. The somewhat severe biases of the tests in large samples casts doubt on the claim that replacing index or portfolio benchmarks with control stocks is a sufficient safeguard against erroneous conclusions caused by extreme buy-and-hold return distributions.

3.2.2. Two groups tests

Table 2 shows that two groups tests based on equal-weighted portfolios matched on size and book-to-market tend to be negatively biased to conservative for upper-tail tests. The negative bias is more strongly observed for the 50 stock sample size and does not systematically change over the different holding-periods. Upper-tail rejection rates for all holding periods and sample sizes range from 0.3% to 1.9%, all below the 95% binomial confidence limit for a 5% significance level. Lower-tail rejection rates are within the confidence limit for all but the 50 stock sample size, which is overstated. The two-tail tests tend to be understated, ranging from 2.1% to 3.8%.

In contrast, the two groups tests using value-weighted benchmarks are positively biased for the 1000 stock sample size particularly for the longer holding-periods. The upper-tail (two-tail) rejection rate for 1000 stocks for five-year holding period is 25.2% (12.8%). The smaller-sample-size rejection rates are conservative for the upper-tail and two-tail tests but tend to be appropriately specified for lower-tail tests for all holding periods. The range of upper-tail and two-tail rejection rates for 50 stocks is 1.0% to 3.2%.

The control-stock two groups tests are generally well specified for the 50 and 200 stock sample sizes, with some tendency toward conservative rejection rates for lower-tail tests at one- and three- year holding periods and for upper-tail tests at the five-year holding period. However, when sample sizes increase to 1000 stocks, the test results are less well-behaved. The upper tail tests for 1000 stocks are 10.8% and 12.0% for one- and three-year holding periods and 0.9% for five-year holding period.

The control-stock two groups test is equivalent to the common procedure of collecting a matching, non-event sample to see if it experiences significantly different post-event performance from the event sample. The results suggest that while the test often works well, non-event-related extreme returns on event stocks or control stocks are likely to create unpredictable results in large samples. Thus, researchers should not assume that they can rely on a control sample to verify the robustness of long-horizon event study results.

As our analysis of the overlapping horizons bias leads us to expect, the biases with two groups tests are less than the corresponding biases in the paired difference tests. For example, the equal-weighted benchmark, paired difference test has a lower-tail rejection rate in 200-stock samples of 9.3% to 11.4%, depending on the holding period. The corresponding two-groups test rejection rates range from 5.1% to 6.5%, all within the 95% binomial confidence interval for a 5% significance level.

3.2.3. *Tests using winsorized abnormal returns*

Table 3 shows the paired-difference and two-groups test rejection rates using winsorized abnormal returns based on size and book-to-market matching for equal- and value-weighted portfolios and control stock portfolios. The results show profound misspecification for all the paired difference tests. Tests with equal- and value-weighted benchmark portfolios exhibit a negative

bias whereas the rejection rates with control stocks are too high for upper-, lower-, and two-tail tests. Consistent with overlapping horizons bias, these misspecifications tend to increase as the holding period increases. For example, the lower-tail rejection rates for 200 stocks using equal-weighted control stock portfolios increase from 8.6% for one-year to 13.3% for three-year to 24.6% for five-year holding periods. The biases tend to increase as the portfolio size increases. The two-tail rejection rates for three-year holding period control stock portfolios increase from 17.4% for 50 stocks to 21.3% for 1000 stocks. An exception occurs with the tests using value-weighted benchmarks for three- and five-year holding periods, for which the biases decrease as sample size increases.

The two-groups test results tend to be correctly specified with the exception of the tests using equal-weighted portfolios, which demonstrate a negative bias. The two-tail rejection rates for value-weighted and control stock portfolios range from 3.1 to 7.4% (both just outside the 95% confidence limit). However, these tests are conservative in most upper-tail value-weighted portfolios (except in 1000-stock samples with three- and five-year holding periods). The upper-tail rejection rates with one- and three-year holding periods and control stocks are misspecified in the larger samples.

The overall impressions from Table 3 are that winsorizing abnormal returns at three standard deviations tends to reduce upper-tail rejection frequencies and increase lower-tail rejection frequencies. Winsorized two groups tests using value-weighted benchmark portfolios matched on size and book-to-market is the least likely to produce excessive rejections of the null hypothesis, though the tests are not correctly specified in all situations.⁸

⁸ When event-date clustering (perfectly overlapping horizons) occurs, none of the tests analyzed in this paper is appropriately specified. Simulations with common event dates for all stocks in a sample (not reported) indicate that

3.3. *Power of tests*

Table 4 reports the power of two groups tests based on size-and book-to-market benchmark portfolios, and the paired difference test using control stocks. We report the power of value-weighted size and book-to-market matched paired difference and two groups tests using winsorized data. We did not find the control-stock test or the unwinsorized, value-weighted benchmark test reported in the table to be close to well-specified, but we report their power for comparison. We report only the three-year holding period as the insights gained from other holding periods are similar.

Table 4 shows that the power of the unwinsorized, equal-weighted test and the winsorized, value-weighted test are asymmetric, but the nature of the asymmetry varies with the abnormal return and the sample size. Both tests reject more often in the lower tail than the upper tail when the absolute value of induced abnormal return is small and the sample size is small. As the sample size and absolute value of abnormal return increase, the pattern reverses. The pattern is consistent with the reduction in the skewness bias as the sample size increases. The winsorized value-weighted test is more powerful in most situations than the control-stock test and the equal-weighted test. In 50-stock samples, the control-stock test is more powerful when we induce a positive 10% abnormal return, but as previously discussed, the control-stock test is positively biased in this situation. In 1000-stock samples, the two-tailed winsorized test detects abnormal returns of positive and negative 10% in 77.1% and 63.8% of cases, respectively, versus 47.8% and 48.5% for the equal-weighted test and 58.1% and 20.1% for the control-stock test.

winsorized two-groups tests that use size- or size and book-to-market benchmark portfolios generally are better specified than most other tests, particularly with smaller samples. However, none of the tests can be recommended for this situation.

3.4. *Summary of test results*

We have shown that long horizon tests are highly sensitive to the choice of testing procedures, and that many tests are misspecified, even when the benchmarks are purged of new listings bias and rebalancing bias. The result is consistent with the findings of Barber and Lyon (1997) for benchmarks that do suffer from the two biases. The results also show that the third bias documented by Barber and Lyon (1997), the skewness bias, is necessary but not sufficient to explain the behavior of the tests. Specifically, as the sample size increases, the skewness of the sample mean decreases, making tests less negatively biased, holding other effects constant. The lessening of the skewness bias may allow other biases to emerge. For example, a test that is negatively biased in small samples, like the paired difference test based on value-weighted matching portfolios, can switch to being positively biased. The positive bias occurs because the abnormal return has a positive mean either due to benchmark mismatching or by chance, combined with an underestimated standard error. A downward-biased standard error can result from the overlapping horizons effect, which increases with the length of the holding period.

Even the control-stock test advocated by Barber and Lyon (1997) can suffer from skewness bias, though to a considerably smaller degree than most benchmark portfolio tests. Moreover, the use of matched control stocks instead of benchmark portfolios involves a different source of instability. Randomly occurring extreme returns on either a sample stock or a control stock are equally probable. This helps reduce the average skewness bias, but can increase the frequency of unpredictable results.

The most promising testing approach that we have uncovered is to match value-weighted benchmark portfolios to sample firms on size and book-to-market, winsorize the abnormal returns at three standard deviations to reduce skewness, and conduct a two groups test instead of a

paired difference test. The resulting statistic comes closer to the standard normal distribution and produces lower type I error rates and better power than the control-stock test. However, the specification of the test is not perfect, and winsorization may disturb some researchers since it involves changing some data from what is observed, at least for the purpose of computing a test statistic. In response, we would point out that the contorted distributions of long-horizon stock returns should give pause to anyone considering the use of any parametric test, as Kothari and Warner (1997) emphasize. The combination of winsorization, value-weighted benchmark portfolios, and the two groups test, cautiously applied, seems to us to be a reasonable approach at least for examining the robustness of results and perhaps as the main test.

4. Concluding remarks

Consistent with previous research (Kothari and Warner, 1997; Barber and Lyon, 1997), our simulations provide further evidence that most paired difference tests of long-horizon abnormal returns based on benchmark portfolios tend to find abnormal performance more often than the nominal significance level when none is induced. The poor specification occurs even though our benchmark portfolios are purged of the new listing bias and rebalancing bias and our procedures attempt to mimic feasible investment strategies.

No benchmark portfolio construction method overcomes the skewness bias discussed in Barber and Lyon (1997). We find that sample size is an important determinant of the magnitude of skewness bias. The larger the sample size, the smaller is the skewness bias. Further, we find evidence that although a control stock approach reduces skewness bias, it still can produce incorrectly specified results, probably because random extreme returns in the sample stocks and the control stock do not always offset each other, even in large samples.

We report evidence of an overlapping horizons bias that previously has not been analyzed in detail. As the length of the time horizon increases, the potential for cross-sectional dependence among the returns of sample stocks increases due to partially contemporaneous holding periods. If the cross-sectional dependence in returns is positive, conventional test statistics will underestimate the variance, causing the tests to reject the null hypothesis too often. However, the skewness and overlapping horizons biases interact with each other and with any benchmark matching bias to produce unpredictable results in conventional tests.

Finally, we offer two suggestions to improving the quality of inference in long-horizon studies. First, we report that winsorizing abnormal returns at three standard deviations, using value-weighted benchmark portfolios matched to sample stocks on size and book-to-market, and computing a two groups statistic, provides a parametric procedure that is better specified, and often more powerful, than previously proposed tests. The procedure is not correctly specified in every instance, but we believe it to be potentially useful to researchers who want a parametric test.⁹ Second, we recommend researchers carefully consider their sample size when selecting test procedures and evaluating the results. Our simulations are limited to 1000 and smaller stock portfolios whereas researchers have used samples three to four times larger. While increases in sample size beyond 1000 should produce smaller changes in return distribution than those under 1000, the magnitude of the effect of such large sample sizes remains undocumented.

⁹ Ikenberry, Lakonishok and Vermaelen (1995) and Rau and Vermaelen (1998) apply a nonparametric bootstrapping method to long-horizon event studies. Lyon, Barber and Tsai (1999) present simulation evidence that suggests that bootstrapping tests can be well-specified and powerful, while Brav (1999) cautions that these tests may be sensitive to post-event residual variation. We leave further analysis of bootstrapping methods to future research.

References

- Agrawal, A.; J. F. Jaffe; and G. N. Mandelker, 1992, The post-merger performance of acquiring firms: A re-examination of an anomaly. *Journal of Finance* 48, 1605–1621.
- 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., 1999, Inference in long-horizon event studies: A Bayesian approach with application to initial public offerings, *Journal of Finance*, forthcoming.
- Brav, A.; C. Geczy; and P. A. Gompers, 1995, The long-run underperformance of seasoned equity offerings revisited, working paper, Harvard University, Graduate School of Business Administration, Boston, MA.
- 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.
- Brown, S. J., and J. B. Warner, 1980, Measuring security price performance, *Journal of Financial Economics* 8, 205–258.
- Canina, L.; R. Michaely; R. Thaler; and K. Womack, 1998, Caveat compounder: A warning about using the daily CRSP equal-weighted index to compute long-run excess returns, *Journal of Finance* 53, 403–416.
- Clark, K., and E. Ofek, 1994, Mergers as a means of restructuring distressed firms: An empirical investigation, *Journal of Financial and Quantitative Analysis* 29, 541–565.
- Cusatis, P. J.; J. A. Miles; and J. R. Woolridge, 1993, Restructuring through spinoffs, *Journal of Financial Economics* 33, 293–311.
- Desai, H., and P. C. Jain, 1997, Long-run common stock returns following stock splits and reverse splits, *Journal of Business* 70, 409–433.
- 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, 427–466.
- 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.
- Ikenberry, D., and J. Lakonishok, 1993, Corporate governance through the proxy contest: Evidence and implications, *Journal of Business* 66, 405–435.
- Ikenberry, D.; J. Lakonishok; and T. Vermaelen, 1995, Market underreaction to open market share repurchases, *Journal of Financial Economics* 39, 181–208.

- Kothari, S.P., and J.B. Warner, 1997, Measuring long-horizon security price performance, *Journal of Financial Economics* 43, 301–339.
- Loughran, T., 1993, Market microstructure or the poor performance of initial public offerings?, *Journal of Financial Economics* 33, 241–260.
- Loughran, T., and J. R. Ritter, 1995, The new issues puzzle, *Journal of Finance* 50, 23–51.
- Lyon, J.D., B.M. Barber, and C. Tsai, 1999, Improved methods for tests of long-run abnormal stock returns, *Journal of Finance* 54, 165–201.
- Mitchell, M.L., and E. Stafford, 1998, Managerial decisions and long-term stock price performance, working paper, The University of Chicago.
- Rau, P.R., and T. Vermaelen, 1998, Glamour, value and the post-acquisition performance of acquiring firms, *Journal of Financial Economics* 49, 223–253.
- Shumway, T., 1997, The delisting bias in CRSP data, *Journal of Finance* 52, 327–340.
- Spiess, D. K., and J. Affleck-Graves, 1995, Underperformance in long-run stock returns following seasoned equity offerings, *Journal of Financial Economics* 38, 243–267.
- Teoh, S. H.; I. Welch; and T. J. Wong, 1998, Earnings management and the long-run market performance of initial public offerings, *Journal of Financial Economics* 53, 1935–1974.

Table 1
Descriptive Statistics of Three-Year Holding-Period Raw and Abnormal Returns

This table contains descriptive statistics for three-year holding-period returns, and abnormal returns using size and book-to-market matched benchmark portfolios and control stocks. Sample size varies from 25 to 1000 stocks. Mean, standard deviation, and skewness and kurtosis coefficients are given for 1000 samples of each size. For raw returns and abnormal returns using value-weighted benchmark portfolios, statistics are shown with and without winsorizing at three standard deviations.

	Sample size					
	25	50	100	200	500	1000
<i>Raw returns:</i>						
<u>No winsorizing</u>						
Mean	0.593	0.576	0.580	0.584	0.588	0.586
Std. Dev.	0.318	0.224	0.171	0.120	0.076	0.055
Skewness	2.430	1.492	2.089	1.070	0.442	0.337
Kurtosis	15.363	6.525	14.709	3.456	0.644	0.386
<u>Winsorizing at three standard deviations</u>						
Mean	0.563	0.531	0.527	0.528	0.532	0.532
Std. Dev.	0.273	0.182	0.131	0.092	0.061	0.045
Skewness	1.474	0.600	0.514	0.302	0.102	0.130
Kurtosis	6.926	1.917	1.639	0.260	0.298	0.139
<i>Abnormal returns using size and book-to-market matching:</i>						
<u>Value-weighted portfolios</u>						
Mean	0.069	0.052	0.056	0.060	0.062	0.060
Std. Dev.	0.310	0.218	0.168	0.117	0.072	0.052
Skewness	2.516	1.631	2.247	1.169	0.468	0.376
Kurtosis	15.672	7.222	16.710	4.072	0.667	0.444
<u>Equal-weighted portfolios</u>						
Mean	0.016	0.019	0.009	0.004	0.004	0.005
Std. Dev.	0.366	0.234	0.166	0.117	0.072	0.052
Skewness	6.772	3.198	1.567	1.360	0.725	0.517
Kurtosis	97.390	30.929	7.907	5.549	1.963	0.911
<u>Control stock</u>						
Mean	0.048	0.057	0.040	0.035	0.034	0.037
Std. Dev.	0.440	0.294	0.211	0.147	0.092	0.067
Skewness	3.530	1.108	0.397	0.407	0.233	0.200
Kurtosis	44.418	8.721	1.573	0.897	0.458	0.130
<i>Winsorized abnormal returns using size and book-to-market matching:</i>						
<u>Value-weighted winsorized portfolios</u>						
Mean	0.039	0.007	0.003	0.004	0.007	0.007
Std. Dev.	0.263	0.174	0.126	0.088	0.056	0.041
Skewness	1.545	0.720	0.571	0.365	0.044	0.158
Kurtosis	7.225	2.364	2.179	0.349	0.130	0.134

Table 2
Specification of Paired Difference and Two-Groups Tests at 5% Significance Level

This table reports percentage rejection frequencies in 1000 samples for paired difference and two-groups tests using one, three, and five year compounded holding-period returns with zero induced abnormal return. The tests report abnormal returns computed as the difference between the return of a sample stock and the return on a size and book-to-market matched benchmark. The benchmarks include equal-weighted and value-weighted portfolios and a randomly selected control stock matched to the sample stock. Three sample sizes for each of the holding periods are reported: 50, 200 and 1000 stocks. Upper and lower one-tail results and two-tail results are presented. A 5% significance level is used. Bolded rejection frequencies exceed the binomial upper 95% confidence limit and italicized fall below the lower 95% limit.

		Rejection rates at 5% significance								
		1 year holding period			3 year holding period			5 year holding period		
Weighting and benchmark	Alternative hypotheses	Sample size			Sample size			Sample size		
		50	200	1000	50	200	1000	50	200	1000
Equal-weighted size/book-to-market matched, paired difference test	upper-tail	<i>1.8</i>	<i>2.2</i>	<i>2.5</i>	<i>1.2</i>	<i>2.9</i>	<i>3.6</i>	<i>1.3</i>	<i>2.1</i>	<i>3.1</i>
	lower-tail	12.3	9.3	8.9	11.3	11.4	7.6	12.7	10.0	6.9
	2-tail	9.4	6.9	5.3	7.3	8.4	5.7	9.6	7.2	5.2
Value-weighted size/book-to-market matched, paired difference test	upper-tail	<i>2.8</i>	<i>3.1</i>	9.6	<i>2.4</i>	7.9	30.3	<i>2.9</i>	8.0	32.7
	lower-tail	8.0	5.5	<i>2.4</i>	7.7	3.8	<i>0.5</i>	9.3	<i>3.0</i>	<i>0.2</i>
	2-tail	5.6	4.3	5.6	6.3	4.1	18.8	7.5	4.7	20.2
Size/book-to-market matched control stocks, paired difference test	upper-tail	7.0	9.1	13.3	5.8	7.6	13.5	<i>3.0</i>	<i>2.7</i>	<i>1.2</i>
	lower-tail	3.5	2.6	<i>2.0</i>	4.4	3.3	<i>1.6</i>	5.9	7.6	10.4
	2-tail	5.4	6.0	8.2	5.5	5.0	8.0	5.0	5.3	6.7
Equal-weighted size/book-to-market matched, two-groups test	upper-tail	<i>0.8</i>	<i>1.2</i>	<i>1.1</i>	<i>0.3</i>	<i>1.6</i>	<i>1.9</i>	<i>0.6</i>	<i>1.2</i>	<i>1.4</i>
	lower-tail	7.4	5.1	4.3	6.4	6.5	5.0	7.5	5.6	4.1
	2-tail	3.5	<i>3.1</i>	<i>2.8</i>	3.6	3.8	<i>2.7</i>	4.7	2.8	<i>2.1</i>
Value-weighted size/book-to-market matched, two-groups test	upper-tail	<i>1.0</i>	<i>2.2</i>	6.4	<i>1.7</i>	4.5	24.8	<i>1.6</i>	5.8	25.2
	lower-tail	4.2	2.9	<i>1.6</i>	4.8	<i>1.6</i>	<i>0.3</i>	5.8	<i>1.3</i>	<i>0.1</i>
	2-tail	2.9	2.2	2.9	3.2	2.0	13.1	3.1	2.0	12.8
Size/book-to-market matched control stocks, two-groups test	upper-tail	4.7	6.9	10.8	5.2	6.4	12.0	2.2	2.0	<i>0.9</i>
	lower-tail	2.4	2.3	<i>1.5</i>	3.2	2.5	<i>1.3</i>	4.4	5.5	8.4
	2-tail	3.3	4.1	5.0	3.7	3.9	6.8	2.9	3.6	4.6

Table 3
Specification of Paired Difference and Two-Groups Tests at 5% Significance Level Using Returns Winsorized at Three Standard Deviations

This table reports percentage rejection frequencies in 1000 samples for paired difference and two-groups tests using one, three, and five year compounded holding-period returns with zero induced abnormal return. The tests report abnormal returns computed as the difference between the return of a sample stock and the return on a size and book-to-market matched benchmark. The benchmarks include equal-weighted and value-weighted portfolio and a randomly selected control stock matched. The returns for the sample and benchmark have been winsorized at three standard deviations. Three sample sizes for each of the holding periods are reported: 50, 200 and 1000 stocks. Upper and lower one-tail results and two-tail results are presented. A 5% significance level is used. Bolded rejection frequencies exceed the binomial upper 95% confidence limit and italicized fall below the lower 95% limit.

Weighting, benchmark, and test hypotheses		Rejection rates at 5% significance								
		1 year holding period			3 year holding period			5 year holding period		
		Sample size			Sample size			Sample size		
		50	200	1000	50	200	1000	50	200	1000
Equal-weighted size/book-to-market matched, paired difference test	upper-tail	<i>2.0</i>	<i>1.7</i>	<i>0.3</i>	<i>1.5</i>	<i>1.4</i>	<i>0.4</i>	<i>1.6</i>	<i>1.4</i>	<i>0.0</i>
	lower-tail	15.0	19.8	40.9	13.9	23.3	43.1	16.1	23.3	42.8
	2-tail	10.9	19.9	30.6	9.4	17.3	33.5	12.1	16.5	32.1
Value-weighted size/book-to-market matched, paired difference test	upper-tail	<i>3.2</i>	<i>2.2</i>	<i>2.3</i>	<i>2.5</i>	<i>4.5</i>	9.4	<i>3.5</i>	<i>6.1</i>	8.9
	lower-tail	10.2	11.8	19.0	9.8	8.3	7.4	11.9	9.9	<i>6.8</i>
	2-tail	7.4	8.2	13.3	7.8	7.2	9.9	9.4	8.5	8.7
Size/book-to-market matched control stocks, paired difference test	upper-tail	14.2	18.5	31.3	13.0	14.9	19.0	<i>7.7</i>	<i>6.7</i>	<i>1.7</i>
	lower-tail	12.7	8.6	<i>5.4</i>	12.0	13.3	10.5	22.2	24.6	38.4
	2-tail	18.9	19.9	27.9	17.4	19.4	21.3	20.9	21.7	31.4
Equal-weighted size/book-to-market matched, two-groups test	upper-tail	<i>0.7</i>	<i>0.9</i>	<i>0.1</i>	<i>0.4</i>	<i>0.7</i>	<i>0.3</i>	<i>0.6</i>	<i>0.8</i>	<i>0.0</i>
	lower-tail	8.7	11.0	27.8	8.0	15.3	33.4	8.6	13.6	28.7
	2-tail	<i>4.3</i>	<i>5.4</i>	17.7	<i>4.7</i>	10.1	23.0	<i>5.9</i>	7.9	19.2
Value-weighted size/book-to-market matched, two-groups test	upper-tail	<i>1.1</i>	<i>1.4</i>	<i>1.2</i>	<i>1.6</i>	<i>2.1</i>	<i>5.8</i>	<i>1.5</i>	<i>3.0</i>	<i>5.2</i>
	lower-tail	<i>5.6</i>	<i>6.8</i>	11.7	<i>5.9</i>	<i>5.5</i>	<i>5.4</i>	7.6	<i>5.6</i>	<i>3.0</i>
	2-tail	<i>3.6</i>	<i>3.3</i>	<i>6.1</i>	<i>3.5</i>	<i>3.5</i>	<i>5.8</i>	<i>4.7</i>	<i>3.2</i>	<i>4.0</i>
Size/book-to-market matched control stocks, two-groups test	upper-tail	<i>5.4</i>	7.0	9.6	<i>5.3</i>	<i>5.6</i>	7.8	<i>2.5</i>	<i>1.6</i>	<i>0.1</i>
	lower-tail	<i>2.1</i>	<i>1.1</i>	<i>0.5</i>	<i>2.7</i>	<i>2.6</i>	<i>2.1</i>	<i>4.5</i>	<i>5.4</i>	11.9
	2-tail	<i>3.1</i>	<i>4.4</i>	<i>5.1</i>	<i>4.0</i>	<i>3.5</i>	<i>4.1</i>	<i>3.4</i>	<i>3.3</i>	7.4

Table 4
Power of Parametric Tests for Three-Year Holding-Period Returns

This table reports percentage rejection frequencies in 1000 samples for paired difference and two-groups tests using three year compounded holding-period returns. Abnormal performance is induced by adding a positive or negative 0%, 10%, 30% or 50% over the three years, to the actual return. The tests report abnormal returns computed as the difference between the return of a sample stock and the return on a benchmark. The benchmarks include equal-weighted and value-weighted portfolios matched on size and book-to-market ratio and a randomly selected control stock matched to the sample stock using size and book-to-market ratio. Three sample sizes are reported: 50, 200 and 1000 stocks. Upper and lower one-tail results and two-tail results are presented. A 5% significance level is used.

Panel A: 50 stock sample size			Rejection frequencies at 5% significance			
<i>Test</i>	Alternative hypotheses	Sign of IAR	Induced abnormal returns (IAR) for three year holding period			
<u>Weighting and benchmark</u>			0.0%	10%	30.0%	50.0%
<i>Paired difference tests</i>						
Equal-weighted size/book-to-market matched	upper-tail	+	1.2%	7.8%	49.7%	93.5%
	lower-tail	-	11.3	24.8	56.1	76.8
	2-tail	+	7.3	4.2	34.3	86.1
	2-tail	-	7.3	17.9	48.6	70.4
Value-weighted size/book-to-market matched	upper-tail	+	2.4	11.6	61.0	97.1
	lower-tail	-	7.7	19.5	49.4	76.2
	2-tail	+	6.3	6.2	44.5	92.8
	2-tail	-	6.3	13.9	42.1	70.3
Size/book-to-market matched control stocks	upper-tail	+	5.8	13.2	41.2	71.0
	lower-tail	-	4.4	8.5	29.9	59.6
	2-tail	+	5.5	8.0	28.8	59.6
	2-tail	-	5.5	6.7	21.1	49.0
<i>Two-groups test</i>						
Equal-weighted size/book-to-market matched	upper-tail	+	0.3	3.7	36.0	88.1
	lower-tail	-	6.4	17.4	48.0	72.1
	2-tail	+	3.6	1.4	19.1	74.8
	2-tail	-	3.6	11.8	37.9	64.9
Value-weighted size/book-to-market matched	upper-tail	+	1.7	6.3	49.8	95.2
	lower-tail	-	4.8	13.4	43.3	73.0
	2-tail	+	3.2	3.1	30.2	86.0
	2-tail	-	3.2	8.1	34.4	65.4
Size/book-to-market matched, control stocks	upper-tail	+	5.2	10.7	36.8	67.8
	lower-tail	-	3.2	7.0	26.1	56.0
	2-tail	+	3.7	6.3	23.1	56.5
	2-tail	-	3.7	4.3	16.8	44.1
<i>Paired difference test with winsorized data</i>						
Value-weighted size/book-to-market matched	upper-tail	+	2.5	12.5	62.3	97.8
	lower-tail	-	9.8	25.7	60.8	86.8
	2-tail	+	7.8	7.0	48.9	95.8
	2-tail	-	7.8	18.0	52.4	82.3
<i>Two-groups test with winsorized data</i>						
Value-weighted size/book-to-market matched	upper-tail	+	1.6	7.0	50.9	95.9
	lower-tail	-	5.9	17.1	52.8	83.3
	2-tail	+	3.5	3.3	32.5	89.3
	2-tail	-	3.5	10.9	42.7	75.2

Table 4 Continued

Panel B: 200 stock sample size			Rejection frequencies at 5% significance			
<i>Test</i>	Alternative hypotheses	Sign of IAR	Induced abnormal returns (IAR) for three year holding period			
			0.0%	10.0%	30.0%	50.0%
<i>Weighting and benchmark</i>						
<i>Paired difference tests</i> ^{4%}						
Equal-weighted size/book-to-market matched	upper-tail	+	2.9%	20.4%	96.4%	99.8%
	lower-tail	-	11.4	37.0	84.6	96.8
	2-tail	+	8.4	11.3	91.3	99.5
	2-tail	-	8.4	29.5	77.8	95.3
Value-weighted size/book-to-market matched	upper-tail	+	7.9	43.5	99.2	99.9
	lower-tail	-	3.8	21.3	74.2	94.0
	2-tail	+	4.1	26.6	98.2	99.7
	2-tail	-	4.1	14.6	65.9	92.2
Size/book-to-market matched control stocks	upper-tail	+	7.6	24.6	79.0	98.3
	lower-tail	-	3.3	16.6	63.9	92.6
	2-tail	+	5.0	16.0	66.7	96.6
	2-tail	-	5.0	10.5	52.8	88.3
<i>Two-groups test</i>						
Equal-weighted size/book-to-market matched	upper-tail	+	1.6	13.3	92.7	99.8
	lower-tail	-	6.5	29.3	80.8	96.1
	2-tail	+	3.8	7.3	85.0	99.5
	2-tail	-	3.8	20.9	73.3	94.5
Value-weighted size/book-to-market matched	upper-tail	+	4.5	34.5	98.7	99.9
	lower-tail	-	1.6	16.0	69.4	93.4
	2-tail	+	2.0	18.3	96.9	99.7
	2-tail	-	2.0	9.3	61.4	90.5
Size/book-to-market matched control stocks	upper-tail	+	6.4	21.8	75.7	98.0
	lower-tail	-	2.5	13.6	61.6	91.8
	2-tail	+	3.9	13.8	63.6	96.1
	2-tail	-	3.9	8.0	49.3	87.4
<i>Paired difference test with winsorized data</i>						
Value-weighted size/book-to-market matched	upper-tail	+	4.5	38.2	99.3	100.0
	lower-tail	-	8.3	41.4	94.4	99.5
	2-tail	+	7.2	23.8	99.0	100.0
	2-tail	-	7.2	32.6	91.6	99.3
<i>Two-groups test with winsorized data</i>						
Value-weighted size/book-to-market matched	upper-tail	+	2.1	26.7	99.0	100.0
	lower-tail	-	5.5	32.9	92.5	99.4
	2-tail	+	3.5	15.8	97.2	100.0
	2-tail	-	3.5	23.1	86.9	99.3

Table 4 Continued

Panel C: 1000 stock sample size			Rejection frequencies at 5% significance			
<i>Test</i>	Alternative hypotheses	Sign of IAR	Induced abnormal returns (IAR) for three year holding period			
			0.0%	10.0%	30.0%	50.0%
<i>Paired difference tests</i>						
Equal-weighted size/book-to-market matched	upper-tail	+	3.6%	72.1%	100.0%	100.0%
	lower-tail	-	7.6	63.9	98.1	99.9
	2-tail	+	5.7	59.5	100.0	100.0
	2-tail	-	5.7	55.4	97.8	99.9
Value-weighted size/book-to-market matched	upper-tail	+	30.3	97.2	100.0	100.0
	lower-tail	-	0.5	29.4	97.5	99.8
	2-tail	+	18.8	93.9	100.0	100.0
	2-tail	-	18.8	21.7	96.4	99.8
Size/book-to-market matched control stocks	upper-tail	+	13.5	70.7	100.0	100.0
	lower-tail	-	1.6	29.2	96.1	99.7
	2-tail	+	8.0	58.1	100.0	100.0
	2-tail	-	8.0	20.1	95.6	99.2
<i>Two-groups test</i>						
Equal-weighted size/book-to-market matched	upper-tail	+	1.9	63.5	100.0	100.0
	lower-tail	-	5.0	58.5	98.1	99.9
	2-tail	+	2.7	47.8	100.0	100.0
	2-tail	-	2.7	48.5	97.7	99.8
Value-weighted size/book-to-market matched	upper-tail	+	24.8	95.1	100.0	100.0
	lower-tail	-	0.3	24.3	97.4	99.8
	2-tail	+	13.1	92.0	100.0	100.0
	2-tail	-	13.1	16.7	96.1	99.7
Size/book-to-market matched, control stocks	upper-tail	+	12.0	67.7	100.0	100.0
	lower-tail	-	1.3	26.5	96.1	99.7
	2-tail	+	6.8	53.8	100.0	100.0
	2-tail	-	6.8	17.9	95.0	99.2
<i>Paired difference test with winsorized data</i>						
Value-weighted size/book-to-market matched	upper-tail	+	9.4	90.7	100.0	100.0
	lower-tail	-	7.4	80.3	100.0	100.0
	2-tail	+	9.9	85.3	100.0	100.0
	2-tail	-	9.9	72.7	100.0	100.0
<i>Two-groups test with winsorized data</i>						
Value-weighted size/book-to-market matched	upper-tail	+	5.8	85.8	100.0	100.0
	lower-tail	-	5.4	74.4	100.0	100.0
	2-tail	+	5.8	77.1	100.0	100.0
	2-tail	-	5.8	63.8	100.0	100.0