Interacting Biases, Non-Normal Return Distributions and the Performance of Parametric and Bootstrap 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 6060
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 6060
anne_sgt@iastate.edu

First Draft: October 1996
This Draft (2nd): January 1997

We report simulations of abnormal buy-and-hold stock return tests. Using benchmark portfolios purged of new-listings and rebalancing biases, we find severe misspecification of parametric tests, due in part to skewness. 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. Bootstrap tests detect more negative than positive abnormal returns, but are well-specified and reasonably powerful.

Keywords: Event studies, long-horizon performance, abnormal returns, nonparametric tests, bootstrap.

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 (1996, 1997) report simulation results suggesting that many commonly used methods tend to find positive or negative abnormal performance when none is present. This paper provides a more detailed investigation of some reasons for the poor specification of long-horizon event study tests and examines alternative approaches to improving test specification.

Tests of long-horizon returns can suffer from biases due to skewness, partially contemporaneous holding periods ("overlapping horizons"), benchmark mismatching, new listings, and benchmark rebalancing. In contrast with previous work, the benchmark portfolios in this paper are purged of new listing and rebalancing biases. We argue that insufficient attention has been devoted to the effect of sample size on the skewness bias. This is important because researchers have used a variety of sample sizes in long-horizon event studies. Table 1 lists some empirical studies that examine samples or subsets of samples as small as 25-50 stocks (e.g., Clark and Ofek, 1994 look at 38 distressed firms) to as large as 5000 stocks (e.g., Loughran and Ritter, 1995 look at 4,753 IPOs and 3,702 SEOs.)

Another neglected area is the potential for interactions among skewness, contemporaneous holding periods, and benchmark mismatching to either exacerbate or mitigate biases in test statistics. These interactions can lead to a test that appears to be correctly specified becoming biased in response to one seemingly innocuous change in test condi-

tions. We study these issues by jointly varying the sample size and the holding period in simulations using actual CRSP data. Previous simulation research is limited to samples of 200 or fewer stocks; this paper is the first to examine portfolios sizes as large as 1000 stocks. Our simulation analysis is the first to examine the following three approaches to improving test properties. Winsorizing long-horizon abnormal returns can reduce skewness; using two-groups tests instead of the usual paired-difference tests can mitigate the overlapping horizons bias; and bootstrapping can avoid the distributional assumptions, which are so egregiously violated, of parametric tests.[1]

We report results for simulations of paired-difference, two-groups, and bootstrap tests for buy-and-hold abnormal returns cumulated over one-, three-, and five- year periods for samples of 50, 200 and 1000 stocks. The tests use equal- and value-weighted market indices and benchmark portfolios matched on size and book-to-market ratio as well as control stocks. Consistent with previous research (Kothari and Warner 1997; Barber and Lyon 1997), we find misspecification in parametric tests even with benchmark portfolios void of new-listings and rebalancing biases. 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. Our results indicate that a two-groups tests using value-weighted size and book-to-market benchmark portfolios of winsorized returns is powerful and nearly always exhibits appropriate specifi-

[1] After circulating the first draft, we learned of a contemporaneous paper by Barber, Lyon and Tsai (1996)

cation at all reported portfolios sizes and holding periods. The bootstrap procedure is always well specified and about as powerful as the winsorized test. Both the winsorized and bootstrap tests are more powerful and better specified at large sample sizes than the control-firm approach offered by Barber and Lyon (1997.)

2. Data and simulation approach

2.1 Population sampled

For annual holding periods, we include in the simulation samples NYSE, AMEX and Nasdaq stocks listed on the 1995 CRSP daily file at any time from 1965 through 1994. When the holding period is three or five years, only stocks listed by the end of 1992 or 1990, respectively, are considered. We exclude 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.

To be eligible for the simulation sample for a given holding period, we require a stock to appear on the CRSP daily return file for the first day of the holding period. We do not impose any requirement of prior or subsequent return data on the CRSP file. However, since we consider size- and book-to-market matched benchmark portfolios and control stocks, 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 the Compustat files.)

---

that also provides simulation evidence on bootstrapping.

Benchmark portfolios, described in more detail below, consist of stocks eligible for the simulation samples. For example, when we require a market index return, we construct our own index portfolio according to the above criteria.

2.2 Computation of returns

For each n-year holding period, we construct a replica of the 1995 daily CRSP returns file where we replace each daily return with the buy-and-hold return for the *n*-year period (252*n* trading days) beginning on the same day. The buy-and-hold return of stock *j* for the 252*n*-day holding period beginning at the close of day *t-1* is

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

where r_{jt+l} is the return on day $t+l$ from the CRSP daily returns file.

If a stock stops trading before the end of 1995, we fill in its daily return series for the remainder of the outstanding *n*-year holding periods. We check for a delisting return in the CRSP delisting data structure. (See Shumway, 1997 for a discussion of the potential importance of the delisting return.) When one is present, we compound the delisting return with the final trading return. Sometimes the date that CRSP reports for the delisting return is later than the last trading date. In such a case, we assume that an investor cannot liquidate the stock after the last trading day until the date of the delisting return. To reflect this assumption, the l index in the *HPR* computation advances, but all the implied daily returns through the delisting return date are zero.

After the delisting return date, we assume that the proceeds are invested in a value-weighted market index portfolio, until the end of the *n*-year holding period. The in-

dex portfolio contains all NYSE, AMEX and Nasdaq ordinary domestic common stocks but excludes stocks of firms incorporated outside the U.S., Americus Trust Components, closed-end funds and real estate investment trusts.[2]

2.3 Construction of benchmark portfolio returns

We consider three types of buy-and-hold benchmark portfolios. The first is a market index consisting of all NYSE, AMEX and Nasdaq ordinary domestic common stocks. The second is a size decile portfolio. We use the Fama and French (1993) procedure to construct market-capitalization decile portfolios. We use NYSE-based decile boundaries to assign each NYSE, AMEX, and Nasdaq stock to a portfolio for the year beginning July 1st on the basis of its June 30th size. The third type of benchmark is a size and book-to-market portfolio. Again following Fama and French, we form fifty size and book-to-market portfolios by annually ranking eligible stocks into quintiles by their book-to-market ratio at the end of the most recent fiscal year.[3]

For each type of benchmark portfolio, we compute an equal weighted return and a value weighted return for each holding period. The benchmark return simply is the arith-

---

[2] The market value of a stock at the end of the previous trading day determines its weight. We use a value-weighted index for two reasons. First, we want to simulate a feasible strategy for using the proceeds of a liquidated investment. An institutional investor, and more recently an individual, could invest the proceeds in a passively managed fund that mimics a broad market index. We would have preferred to use the Standard and Poor's 500 stock index, but lacking data on the return with dividends of the S&P, we thought the value weighted index to be a reasonable approximation. Second, because the weights adjust themselves to the return experience, the value-weighted mean buy-and-hold return of a portfolio can be obtained by compounding the daily value-weighted portfolio returns.

[3] We use the end of February as the ranking date instead of the end of June to allow for up to a four month lag in the publication of annual financial statements. Thus, a firm with a calendar fiscal year has a book-to-market ranking for the year starting July 1, 1980 based on its balance sheet and market price at the end of December 1979. A firm with an April fiscal year end has a ranking that uses its April 1979 data, because its statements for the year ending April 1980 may not have been public by July 1980.

metic or weighted mean *n*-year holding period returns of the stocks in the benchmark portfolio. Averaging the buy-and-hold stock returns avoids the need to cumulate or compound benchmark subperiod returns. Thus, we eliminate the rebalancing bias that Barber and Lyon (1997) identify. Also, since we compute the benchmark portfolio return for a given holding period from our stock buy-and-hold return database, we do not include stocks that newly enter the CRSP file during the holding period. For example, a benchmark return for the period starting July 25th would not incorporate data for a stock that went public on July 26th. Including only stocks listed at the start of the holding period eliminates the new-listing bias that Barber and Lyon also observe.

Our benchmark return procedure differs from that used by Barber and Lyon (1997.) They study what they consider to be the typical method, which treats stock returns and benchmark portfolio returns differently. Many existing studies do use benchmark designs that resemble the one in Barber and Lyon, perhaps by extrapolation from short-horizon event studies. In contrast, we adopt a natural procedure for buy-and-hold returns. Specifically, we define the return of a stock across a given *n*-year holding period the same way regardless of whether it is considered as a sample stock or as a component of a benchmark portfolio.

2.4 Simulation sample selection

We simulate long-run event studies by randomly selecting stocks and event dates. We sort stocks into deciles on the basis of the number of holding period start days avail-

able and construct a database of stock PERMNOs eligible for sample selection.[4] Stocks appear in the database in proportion to the decile ranks. The stock with the most return data has ten times the chance of random selection as the stock with the least data. The procedure yields samples that should resemble those of actual event studies, in that stocks that have been on the CRSP file longer are more likely to be present.

For each firm, we select an event date (the beginning date of holding period returns) within its range of available dates. Thus, we do not alter the probability of a firm remaining in the sample by selecting an event date on which it is not listed on CRSP. Barber and Lyon (1996) argue that Kothari and Warner's (1996) procedure, which samples all stocks with equal probability, over-samples stocks with a short return history. This would be true if Kothari and Warner restricted the choice of an event date to a stock's range of available data. However, our reading of Kothari and Warner leads us to believe that they select an event date without regard to data availability. Thus, while a stock with a short return history and one with a long return history are equally likely to come up in the initial selection, the short-history stock is more likely to drop out because it has no data on the chosen date. This is essentially the same procedure as Brown and Warner (1980, 1985) use. We are convinced that our sampling procedure and those of Barber and Lyon (1997) and Kothari and Warner produce similar mixtures of return histories, and that the mixture is reasonably representative of what would be found in many studies of corporate events.

---

[4] The number of holding periods is equal to CRSP variables ENDRET–BEGRET+1, except that a BEGRET less than 631 is set to 631 (the beginning of 1965) and an ENDRET beyond the last starting date of the simula-

We draw 1000 random samples each of sizes 50, 200 and 1000. The sampling is without replacement at the date-selection stage; thus, the same combination of a stock and an 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.

3. Long-horizon event study methods

3.1 Cumulative versus compounded returns

Event studies conventionally test the null hypothesis that the equally-weighted mean abnormal return is equal to zero (or a one-tail-test variant of the same.) Presumably, the use of a hypothesis about the equally-weighted mean stems from an implicit assumption that all stocks experiencing the same type of event are equally interesting, regardless of characteristics such as size or return volatility. This can be appropriate if the investigator expects the event to affect the abnormal returns of all stocks in the same direction. Our impression of the literature over the last 15 years is that researchers typically do make such an assumption in formulating tests of the mean. (Many studies use cross-sectional regressions, rather than adjustments to the mean test, to examine characteristics that cause individuals to deviate from the mean.)

Given that the hypothesis addresses the equally-weighted mean, in a long-term event study the question arises whether the hypothesis applies to the holding period mean, or to an embedded series of subperiod means. For example, in a study of annual abnormal returns following an event, the researcher can test the mean annual holding period return,

---

tion is set to the last starting date. The last starting date is 8180, the end of 1994, for an annual holding

or the mean or sum of twelve monthly means. As Ritter (1991) and Barber and Lyon (1997) observe, a test of the subperiod means requires that the equal weighting of the sample be re-applied each subperiod. The equal weighting of the embedded subperiod means is inherent in the *cumulative abnormal returns* (CAR) measure, where the researcher adds the subperiod abnormal (or actual) returns across the holding period instead of compounding them. In investment parlance, the mean CAR mimics the experience of an investor who holds an equally weighted portfolio and rebalances it every subperiod, selling some shares of stocks that have appreciated and buying more shares of stocks that have declined in value (Roll, 1983.)

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. The AHPAR applies equal weight to the holding period return of each stock, not to subperiod returns, so no portfolio rebalancing is implied. Thus, the AHPAR mimics the experience of an investor following a buy-and-hold strategy for the specified holding period, and is said to be a buy-and-hold abnormal return.

When the holding period is a few days long and the subperiods are single days, the difference between CAR and AHPAR is trivial. However, Roll (1983), Blume and Stambaugh (1983), Conrad and Kaul (1993), and Barber and Lyon (1997) criticize the use of

---

period, 7675 (1992) for the three-year holding period, and 7168 (1990) for the five-year simulations.

CARs to study longer holding periods. Roll argues that actual investment experience more nearly resembles a buy-and-hold practice than a policy of frequent rebalancing. He shows that rebalanced measures are poor estimators for buy-and-hold returns because of serial dependence in subperiod returns. Blume and Stambaugh and Conrad and Kaul emphasize that rebalanced measures exacerbate the bid-ask bias in returns computed from closing prices. Barber and Lyon point out that differences in volatility between individual stock returns and benchmark portfolio returns can induce differences between CAR and AH-PAR. Moreover, the empirical evidence as to the specification of CAR-based tests in Barber and Lyon and Kothari and Warner (1997) does not lead one to conclude that CARs are more trustworthy than AHPARs. Thus, we focus on buy-and-hold return tests.

3.2 Return prediction models

3.2.1 Simple differences versus factor models

Two common forms of long-term abnormal returns are simple differences between the return on a stock and the return on a benchmark, and the prediction error from a factor model. Examples of benchmarks are market indices, portfolios of all stocks with a common characteristic, such as market capitalization, and matched control-firm stocks. Examples of factor models are the ubiquitous single-index market model and the Fama and French (1993) three-factor model. Factor models are problematic for long-term buy-and-hold returns. Presumably one would want to estimate the model parameters using returns of the same holding-period length as one plans to test. Using a three-year holding period, however, few data points will be available for the estimation since it would be unreasonable to assume that model parameters are stable over a period of many years before the

event. The alternative of using a shorter return interval for parameter estimation is unappealing because of the well-known intervaling effect (Levhari and Lev, 1977.) Also, there is no apparent way to translate the time-series standard deviation of short-period factor model residuals into the standard deviation of a buy-and-hold portfolio return. We conclude that factor models are inappropriate for buy-and-hold return tests, and conduct simulations only of simple-difference abnormal returns.

3.2.2 Choice of benchmarks

The benchmarks that we consider are a market index, a portfolio of all stocks that match the test stock on size or size and book-to-market equity, and a control stock that matches the test stock on size and book-to-market equity. These are the main benchmarks that long-run event studies use. The market index makes an interesting benchmark because of its use with considerable success in short-horizon event studies. Portfolios matched on size and book-to-market are particularly appropriate because Fama and French (1992, 1993) report that these characteristics explain much of the cross-sectional and time-series variation in stock returns. 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.

We consider both equal-weighted and value-weighted market indices and benchmark portfolios. Each has commendable properties for purposes of long-term 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 better represent actual investment opportunities.

Some empirical studies use matched control-firm stocks as benchmarks instead of portfolios. Barber and Lyon (1997) find tests based on single control stocks to be well specified. Therefore, in addition to benchmark portfolio tests, we study tests using control stocks matched on the basis of size and book-to-market ratio. Each sample stock is assigned to a control stock by random selection from the same size decile and book-to-market quintile.

3.3 Test statistics

We consider two parametric tests and one non-parametric test, the bootstrap. We also consider a variant of each of the parametric tests using winsorized abnormal returns.

3.3.1 Parametric tests using holding period abnormal returns

As we discuss above, our measure of the abnormal return of a stock is the difference between the holding period compound return on the stock and the corresponding return on a benchmark. A natural way to test the null hypothesis is to use a simple paired-difference Z statistic (t test for smaller samples),

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

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

Abnormal returns can manifest another form of dependence, however. If the holding periods of stocks in the sample sometimes overlap, there is a positive correlation of stock returns across the sample. If the correlation persists in the abnormal returns, then the sample variance of paired differences underestimates the true variance. This can inflate paired difference test statistics and lead to a finding of either positive or negative abnormal performance where none is truly present. We call this potential bias of paired difference statistics the *overlapping horizons bias*. The 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-group difference of means test,

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

The two groups are the stocks and the benchmarks. In a two-group 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. A specific drawback of the two-groups test is that in a random sample with no cross-security dependence, the test will be less powerful than a paired-difference test. More generally, the two-group 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-group test has the virtue of being simple and easy to calculate. Whether it is too simple is an empirical question.

3.3.2 Parametric tests using winsorized holding period abnormal returns

Researchers who use long-run buy-and-hold stock returns often mention the extreme positive sample skewness. The skewness is not particularly surprising, because the largest possible negative return is –100%, while positive returns are unlimited. Extreme skewness, and other forms of non-normality that stem from extreme positive observations, can cause statistical tests to be badly specified.

It is not possible to avoid extreme returns; the best that researchers can do is to use tests that are among the least likely to suffer ill effects from the extreme observations. One way to limit the effect of extreme observations is to set an arbitrary limit on how far away from the rest of the sample an observation is allowed to be, then "pull in" more distant observations by setting them at the limit. The usual term for this procedure is *winsorization*. Some obvious objections to winsorization are possible. The arbitrary nature of the limit affords the temptation for manipulation in search of a desired result, and pulling in an observation throws away part of the information in the data. While we recognize the potential drawbacks, we think that there are good reasons to consider using winsorized data for tests of the mean holding period abnormal return.

The simulation results of Kothari and Warner (1997) and Barber and Lyon (1996, 1997), some of which parallel our own, show that sometimes it is easy to find positive or negative mean long-run abnormal performance when none truly exists. Extreme observations and positive skewness appear to account for a substantial part of the problem. Winsorizing the data allows the investigator to explore the sensitivity of the inference to extreme returns. A plethora of studies finds abnormal performance following various corporate events. With absolutely no derogation of these studies intended, we think it apparent that an unscrupulous or inattentive researcher would not need winsorization to generate a finding of long-term abnormal performance. On the contrary, reporting the results of a procedure that limits extreme observations should help readers to judge the robustness of the conclusions. Additional safeguards, for example stating whether the author tried but did not report alternative limits, can help to control potential data-mining biases.

There are two common approaches to winsorization. The analyst can set the limits at percentiles of the sample; for example, pull in the largest five percent of observations to the 95th percentile, and pull in the smallest five percent to the fifth percentile. Alternatively, the limit can be a specified number of standard deviations from the sample mean, which is the approach that we adopt. The percentile approach treats high and low values symmetrically, but the problems caused by high and low stock returns are not symmetric. The standard deviation approach can pull-in only large positive observations, only large negative ones, or both, depending on the data. In long-run stock return data, the procedure is more likely to affect positive returns than negative returns.

We winsorize holding period abnormal returns 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. We found that 3.0 standard deviation limits produced better parametric test specification in preliminary experiments than 2.5, 3.5 or 4.0 standard deviations.

3.3.3 Bootstrap method

Another approach to the detection of long-run abnormal stock returns is to employ a nonparametric test. Ikenberry, Lakonishok and Vermaelen (1995) use a bootstrap test for mean long-run abnormal stock returns. Their test differs from the usual bootstrap procedure because they use the returns of comparison stocks instead of resampling the original data. This paper provides the first empirical evidence of the specification and power of the Ikenberry *et al.* bootstrap test.

Following Ikenberry, Lakonishok and Vermaelen (1995), for each stock in a sample, we randomly draw 1000 stocks, which we assign to 1000 matching pseudo-samples. The matching samples must have data on the CRSP tape for the same holding-period start date and be in the same size decile and book-to-market quintile as the sample stock. The drawing is conducted with replacement. The sample stock never is drawn as one of the random stocks corresponding to itself, but could be drawn as a random match for another sample stock. Each pseudo-sample parallels the true sample by having the identical set of event dates and the same size and book-to-market classification of each stock with a given event date.

We rank the mean holding period returns of the sample and pseudo-samples from 1 (low) to 1001 and divide the rank of the true sample by 1001 to obtain its fractional rank. The *p*-value of a lower- (upper-; two-) tail test is the fractional rank (one minus the fractional rank; two times the lesser of the fractional rank or one minus the fractional rank.) The null hypothesis is rejected when the performance of the sample is extreme relative to the range of results that randomly drawn samples exhibit under similar conditions.

We carry out the bootstrap using raw holding-period returns only. The ranks of mean abnormal returns are identical to the ranks of mean raw returns because the size and book-to-market benchmark returns are identical across the 1001 samples for each event. Other applications of the bootstrap could necessitate the use of abnormal returns. For example, if the benchmark is a prediction error from a factor model, a matched control-firm stock, or any other benchmark that is specific to the stock, not common to the sample stock and the pseudo-sample stocks, abnormal returns must be used.

One also could use the bootstrap to test hypotheses about the sample median. In this case, the sample medians would be ranked. The median depends on the ranking of abnormal returns *within* a sample. The median stock in an arbitrary pseudo-sample is not necessarily the one that corresponds to the median stock in the true sample. Therefore, all bootstrap tests of the median would need to use abnormal returns.

4. Biases in long-run event study tests

Barber and Lyon (1997) describe three sources of bias in long-run event study tests. The first two, the *new listing bias* and the *rebalancing bias,* relate to asymmetric criteria for sample selection and inclusion in a benchmark universe. The new listing bias occurs when the composition of the benchmark universe changes as new stocks enter during the holding period. New stocks apparently underperform other stocks, creating a downward bias in the benchmark relative to the population that sample stocks come from. The new listing bias also can result from requiring sample stocks, but not the stocks that make up the benchmark, to have a return history over a specified pre-event period.

Barber and Lyon (1997) argue that the new listing bias is common in practice. We do not disagree, but we argue that it is practical to minimize the new listing bias, and that it is desirable to study the performance of tests under conditions where the bias is absent. Therefore, we construct our sample and benchmark universes so as to largely eliminate the new listing bias. We calculate market index and other benchmark returns daily for the one-, three- or five-year holding period beginning at the close of the previous trading day. By using the holding period returns only of stocks listed at the start of the holding period, we exclude the returns of future new stocks. We do not specifically require that sample

stocks have a return history, but the availability of size and book-to-market ranking data does effectively require a history of up to 28 months. Thus, some new listing bias can remain in the market index, and to a lesser extent the size decile, benchmarks. Poorly performing new stocks tend to be small issues, so the value-weighted benchmarks we consider should suffer less from the new listing bias than the equal-weighted benchmarks. The size- and book-to-market benchmarks require the same pre-event history as sample stocks, so tests using them are free of new listing bias.

The rebalancing bias occurs when the benchmark return is computed as a rebalanced portfolio return, for example by summing 36 monthly returns on an equal-weighted market index to arrive at a three-year return. Barber and Lyon (1997) show that monthly benchmark rebalancing leads to a negative bias in buy-and-hold stock returns. However, we do not believe that the use of monthly rebalanced portfolios is necessary or justified. Benchmark returns should be computed by the same buy-and-hold method used for sample stocks. Our benchmark returns incorporate buy-and-hold returns on the underlying stocks, purging our experiments of the rebalancing bias.

The third bias that Barber and Lyon (1997) identify is the skewness bias. Long-run individual stock returns are exceedingly positively skewed. When forming benchmark portfolios, the individual returns are averaged, which reduces the skewness of the portfolio return. Thus, the skewness of the sample stocks is greater than the skewness of the

benchmark portfolio. Therefore the abnormal return is positively skewed too, leading to a negative bias in tests of the mean abnormal return.[5]

However, the skewness bias is more complex than that discussed by Barber and Lyon (1997.) From a hypothesis testing perspective, the key skewness coefficient is that of the sampling distribution of the mean abnormal return. The usual Central Limit Theorem states that the distribution of the sample mean converges toward the normal distribution as the sample size grows, regardless of the underlying population. Thus, given a large enough sample, the abnormal-return skewness converges to zero. However, given the profoundly non-normal population from which we are sampling, the sample size needed to eliminate skewness is an empirical issue. *Ex ante*, we expect the skewness bias to be less pronounced in the largest samples than in the smallest samples. Moreover, because return skewness increases for longer holding periods, we expect the skewness bias to increase as well, all else held constant.

There are two additional biases that can degrade the specification of holding period abnormal returns tests. The first is the *overlapping horizons* bias, which arises from potential cross-sectional dependence among the returns of sampled stocks due to partially contemporaneous holding periods. In samples of 200 or 1000, the number of nontrivially overlapping horizons will be considerable.[6]

---

[5] Positive skewness occurs when the mean exceeds the median. Uniform random samples from a positively skewed population therefore contain more observations below the true mean than above, inducing a negative bias in tests of the mean.

[6] For example, for our three-year periods from 1965–1992, there can be only nine completely distinct holding periods in a sample. In a sample as small as 28 stocks, the best we could hope for is to have each holding period overlapping one or two others by two years and one or two others by one year.

Assuming that the cross-sectional dependence between stocks with overlapping holding period returns is positive, 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.) Unfortunately, the task of designing a corrected test is complicated by the varying amounts of overlap and by the extreme deviations from the normal distribution. We do not attempt an explicit correction, but we do examine the two groups test statistic, which can compensate for cross-sectional dependence by not adjusting for pairwise dependence.

The second additional bias is the *benchmark matching* bias. Kothari and Warner (1997) emphasize this source of bias, although they do not give it a name. Apart from the effects of skewness and non-normality, the expected return of a sample may differ from a given benchmark because the sample stocks and the benchmark have different return-determining characteristics. In other words, a test may be biased not only because of problems in aggregating returns over time or non-normality, but because the abnormal return estimator is inherently biased. This is a fairly obvious point, but we think it important to call attention to it because there is a danger that the complex econometric problems in the analysis of long-run returns can overshadow the need to carefully examine the characteristics of the sample stocks and select an appropriate benchmark.

5. Results

5.1 Descriptive statistics

Table 2 reports descriptive statistics of sample mean three-year holding period raw returns, and abnormal returns using size and book-to-market matched benchmark portfolios and control stocks. One- and five-year holding periods and other benchmarks yield similar insights and so are not reported. However, in order to highlight the changes across sample sizes, we do include samples of 25, 100 and 500 stocks in addition to the sizes used for the main simulations. The table shows that the mean of 1000 sample mean raw returns is around 58% per three years and does not vary much across sample sizes. The mean raw return is extremely skewed and leptokurtic in small samples. In samples of size 25, the average centered skewness and kurtosis coefficients are 2.430 and 15.363, respectively. Samples from a normal distribution have centered coefficients of zero. The variances of the sample coefficients of skewness and kurtosis, assuming large samples from a normal population, are $6/n$ and $24/n$, respectively. The coefficients decline substantially as the sample size increases from 25 to 1000, but even at 1000 the sampling distribution of the sample mean is significantly skewed and leptokurtic relative to the normal distribution.

The effects of winsorizing the raw holding period returns at three standard deviations are seen in a slight decrease in the mean, and large decreases in the skewness and kurtosis from the raw returns. The winsorized raw holding period returns have means of around 53% across all sample sizes. The coefficient of skewness and kurtosis for all sample sizes is positive. Smaller samples (25, 50 and 100) show positive skewness and lep-

tokurtosis that is greater than the 95% confidence limit, whereas the larger samples (200, 500 and 1000) are within the confidence interval.

Like the raw returns, the abnormal returns using size and book-to-market matched portfolios and control stocks show a positive mean, positive skewness, and leptokurtosis. The equal-weighted benchmark portfolio has the smallest mean for each sample size, around 1% triennially, followed by the control stocks at 4%, and value-weighted benchmark portfolios at 6%. The centered coefficients of skewness is positive and significant for each matching procedure and for all sample sizes. The abnormal returns using the value-weighted portfolio have the smallest skewness of all the benchmarks in samples of size 25, 2.516, but the abnormal returns using control stocks are less skewed than those using value-weighted portfolios in samples of 50 or more stocks. Table 2 reports positive skewness of the control-stock mean abnormal returns, contrary to Barber and Lyon (1997), who report a small negative skewness in 200-stock samples. The difference may be due to random differences in the simulation samples. Either positive or negative skewness in the absence of an event-related abnormal return is equally likely, since either a control stock or a sample stock can have an extreme positive abnormal return by chance.

For all three abnormal return methods, the second and higher moments decline almost monotonically as the sample size increases. For example, the average centered skewness 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. Figure 1 reinforces the impression that there is a considerable change in the skewness across the sample sizes. Decreases in the skewness and kurtosis of the sample mean are consistent with the usual

central limit theorem, which predicts that the distribution of the sample mean converges to the normal as the sample size increases. However, the variation in the sampling distribution, especially given the extreme non-normality in samples of modest size, suggests that we cannot have a high degree of confidence that parametric test statistics will exhibit similar properties in much larger samples.

For the abnormal return based on value-weighted benchmark portfolios, table 2 also reports statistics after winsorizing at three standard deviations within each sample. Winsorizing reduces the mean, to less than a tenth of a percent triennially in most sample sizes. The reduction in the mean is not surprising, since more extreme positive outliers than extreme negative outliers are pulled in. The effects 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 at most sample sizes, it remains non-normal.

Table 3 reports descriptive statistics of the test statistics computed from the abnormal returns in table 2. The test statistics should be distributed standard normal, so they should have a mean and centered coefficients of skewness and kurtosis equal to zero, and a standard deviation equal to one. The paired-difference test statistic using the value-weighted benchmark portfolio has a mean of –0.047 in samples of 25 stocks. The mean increases with the sample size, to 1.138 in samples of 1000, suggesting that the test is

likely to become positively biased at larger sample sizes. The standard deviation decreases from 1.117 to 0.941 as the sample size rises from 25 to 1000. The test statistic using equal-weighted benchmark portfolios has a negative mean and a standard deviation greater than one at all sample sizes, although the values are close to zero and one in samples of 1000. Thus, the bias in this test should diminish at larger sample sizes.

The paired-difference test statistic using control stocks has a standard deviation close to one at all sample sizes, but the mean increases with the sample size, from 0.072 in samples of 25 stocks, to 0.191 in samples of 200, to 0.529 in samples of 1000. Thus, this test may tend to become more positively biased at larger sample sizes.

The means of all the paired-difference test statistics increase with the sample size, and all are negatively skewed. Both results are consistent with the effect of positive skewness of the abnormal returns as predicted by Barber and Lyon (1997), and with our further observation that the importance of skewness diminishes as the sample size increases. Winsorizing the abnormal return based on value-weighted benchmarks reduces the mean paired-difference statistic, to a range of –0.100 to 0.129 from a range of –0.047 to 1.138 without winsorization. Thus, the test based on winsorized abnormal returns should be less likely to produce a false positive result and not much more likely to produce a false negative. With winsorization, the standard deviation, skewness and kurtosis move closer to the standard normal values.

In general, the two-groups test statistics have means closer to zero than the paired-difference tests. The standard deviations are mostly less than one, but not much closer to one than those of the paired-difference statistics. For example, the two-groups statistic for

the abnormal returns using the equal-weighted benchmark in samples of 200 has a mean of –0.138 and a standard deviation of 0.932, whereas the paired difference statistic has a mean of –0.191 and standard deviation of 1.091. Patterns across sample sizes are similar to the paired-difference test, so that the value-weighted test and the control-stock test tend to become more positively biased, and the equal-weighted test tends to become less negatively biased, as the sample size increases. Negative skewness is present in all two-groups tests. However, only the value-weighted benchmark (no winsorization) and larger sample sizes have means that are significantly different from zero.

The test statistic based on winsorized abnormal returns from value-weighted benchmarks has a mean that is no larger than 0.120 in absolute value, which is the smallest of the four two-groups tests and smaller than any of the paired-difference tests. Other moments of the sampling distribution also are closer to standard normal for the winsorized two-groups test than for any other except the control-stock test, which is closer to standard normal for some sample sizes. The general impression from table 3 is that two-groups tests are closer to the expected distribution than paired-difference tests, and that the winsorized value-weighted test is less biased than other tests.

5.2 Specification of parametric tests

5.2.1 Paired-difference tests

Table 4 reports on the specification of paired-difference tests. 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. When the benchmark is the

equal-weighted market index or either equal-weighted (size or size and book-to-market) matched portfolio, the test is negatively biased. It rejects the null hypothesis against a lower-tail alternative significantly more often than five percent, but the rejection rate against an upper-tail alternative is significantly less than five percent. For example, using the market index, a five-year holding period, and a sample of size of 50, the null is rejected in 14.3% of lower-tail tests, but only 1.2% of the time in upper-tail tests. In most cases, the rejection rate against a two-tail alternative significantly exceeds five percent. 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.

The pattern of rejection rates across equal-weighted benchmark portfolios and sample sizes 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 all three equal-weighted benchmarks. For example, using the size-and-book-to-market matched portfolios, 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 shorter holding periods, the rejection rate appears to vary randomly across sample sizes for the market index and size-decile benchmarks, but decreases with sample size for the size- and book-to-market benchmark.

Tests using value-weighted benchmarks are positively biased in samples of size 1000, but not necessarily in smaller samples. The upper-tail rejection rates sharply increase

with the sample size for a given holding period, while the lower-tail rejection rates decrease. This is consistent with a decrease in the skewness bias as the sample size increases.

Holding constant the sample size for value-weighted benchmarks, the upper-tail rejection rates tend to increase, and the lower-tail rates to decrease, with the holding period. However, in 50-stock samples and a one-year holding period, tests using value-weighted benchmarks are properly specified or mildly negatively biased. As the holding period increases, tests on samples of 50 stocks using the value-weighted market index become positively biased, while test using matched size- or size-and-book-to-market benchmarks become more negatively biased. The difference in the behavior of the market index and the matched benchmarks likely comes from the larger positive mean of the index-based abnormal return. The skewness bias would normally have a larger negative effect in longer periods, but the overlapping horizons bias also has a larger effect, and it tends to inflate the test statistic in the same direction as the sample mean. The mean abnormal returns using matched benchmarks also are positive, but smaller, allowing the skewness bias to take over.

Interestingly, the tests using the value-weighted market index or value-weighted size-matched portfolios are correctly specified in nearly all one- and two-tailed tests when the sample size is 50 and the holding period is one or three years. The skewness bias and the overlapping horizons bias (combined with the positive mean abnormal return) offset each other to produce an apparently unbiased test. 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.

Table 4 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. The pattern is reversed in the five-year holding period, where 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 Barber and Lyon's are roughly consistent for samples of 200 stocks or fewer, where 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.

5.2.2 Two-groups tests

Table 5 shows that two-groups tests based on market indices and size-matched benchmark portfolios are severely biased. Tests based on equal-weighted portfolios are negatively biased and those based on value-weighted portfolios are positively biased. As our analysis of the overlapping horizons bias leads us to expect, the biases are less than the corresponding biases in the paired-difference tests. A similar result holds for tests using size- and book-to-market matched benchmark portfolios. 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 rejection rates range from 5.1% to 6.5%, all within the 95% binomial confidence interval for a 5% significance level. The size-and-book-to-market test using equal-weighted benchmarks rejects too often only in samples of size 50, but is conservative in upper-tailed tests. Its rejection rates range from 0.3% to 1.6%, all of which are below the lower 95% binomial confidence limit for a 5% significance level. The tests using value-weighted benchmarks and control stocks reject too often only in 1000-stock samples.

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.

5.2.3 Parametric tests using winsorized abnormal returns

Table 6 shows that paired-difference and two-groups tests using winsorized abnormal returns based on market indices tend to be positively biased. The bias increases with the holding period and the sample size. The two-groups test using size matching rejects the null hypothesis 8.7% of the time in the upper tail with 1000 stock samples and 3-year holding periods, and 7.9% of the time in the lower tail with 50-stock samples with a five-year holding period. Otherwise, it is conservative to correctly specified in one-tailed tests, and in two-tailed tests except in the same case where a one-tail rejection rate is 8.7%. The corresponding paired-difference test is misspecified in all sample sizes and all holding periods.

The two-groups test using size- and book-to-market matched benchmark portfolios is nearly always correctly specified in two-tail tests. Its two-tail rejection rates range from 3.2% (just below the lower 95% confidence limit) to 6.1%. However, the test is conservative in most upper-tail tests (except in 1000-stock samples with three- and five-year holding periods.) It is correctly specified or slightly conservative in lower-tail tests except in 1000-stock samples with a one-year holding period, where it rejects the null 11.7% of the time, and in 50-stock samples with a five-year holding period, where the rejection rate is 7.6%. The corresponding paired difference tests, and the two-groups and paired difference tests using control stocks, are often misspecified.

The overall impressions from table 6 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 or size and book-to-market are the least likely to produce excessive rejections of the null hypothesis, though the tests are not correctly specified in all situations.

5.3 Power of parametric tests

Table 7 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 do not report power for tests using market indices or size-matched benchmarks as they are no better specified, and often worse, than the tests we do report. 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 7 shows that 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.

5.4 Summary of parametric test results

So far we have shown that parametric tests are highly sensitive to the choice of a benchmark portfolio, and that most such tests are misspecified, even when all the benchmarks considered are purged of new listings bias and rebalancing bias. The result is consistent with the findings of Barber and Lyon (1996, 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 parametric 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 approach to parametric testing 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 was observed, at least for the purpose of computing a test statistic. In response, we would point out that the contorted distributions of long-term 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.

5.5 Bootstrap test results

We limit our investigation of the specification and power of the bootstrap procedure that Ikenberry, Lakonishok and Vermaelen (1995) use to the three-year holding pe-

riod due to the computer resources needed.[7] Table 8 shows that the bootstrap test is well specified. When no abnormal return is induced by the simulation program, the bootstrap test rejects the null hypothesis from 4.8% to 6.8% of the time. The type I error rates are within or at the limit of the 95% binomial confidence interval around 5% in all cases. Surprisingly, the power function has a negative tilt. In samples of 50, 200, and 1000 stocks, with a –10% induced triennial abnormal return, the one-tail test rejection rates are 17.6%, 31.4%, and 71.3% respectively. With a positive 10% induced abnormal return, the rates are 8.7%, 17.4% and 55.4% respectively. It is not clear why it should be easier for the bootstrap test to detect a low than a high return. Possibly the right skewness of returns helps ensure that an abnormally low sample mean return falls in the tail of the bootstrap distribution, but allows a sample mean return that is high *for that sample* to be swamped by a handful of large non-event related returns in the pseudo-samples.

The power is sensitive to the sample size. At an absolute value of 20% induced abnormal return, the range of rejection frequencies across one- and two-tailed tests is 7.9% to 37.3% in 50-stock samples, 27.4% to 69.6% in 200-stock samples, and 96.9% to 98.9% in 1000-stock samples. In general, the bootstrap test is less powerful in upper-tail tests and more powerful in lower-tail tests than the control-stock tests. However, it is not usually more powerful than the two-groups, winsorized test even in the lower tail. In samples of 200 stocks, the bootstrap detects nearly all abnormal returns with an absolute value of 40% or more, and in samples of 1000 stocks, it detects nearly all abnormal returns of

[7] Simulating the bootstrap test for 1000 samples of size 1000 involves assembling 1.001×10^9 holding pe-

20% or more. Overall, the Ikenberry *et al.* bootstrap procedure appears to be well specified and reasonably powerful.

6. Concluding remarks

Consistent with previous research (Kothari and Warner 1996; 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 con-

---

riod stock returns and calculating 1.001×10^6 sample means for each alternative hypothesis.

temporaneous 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 suggest two positive approaches 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. Second, we analyze a non-parametric test, an out-of-sample bootstrap developed by Ikenberry, Lakonishok and Vermaelen (1995.) The bootstrap is well specified and powerful. However, its power function is asymmetric in that it tends to detect negative abnormal performance better than positive performance. Future research may be able to provide more insight into the reasons for the asymmetry.

References

Agrawal, Anup, Jeffrey F. Jaffe, and Gershon N. Mandelker, 1992, The post-merger performance of acquiring firms: A re-examination of an anomaly, Journal of Finance 48, 1605–1621.

Asquith, Paul and David W. Mullins, Jr., 1986, Equity issues and offering dilution, Journal of Financial Economics 15, 61–89.

Barber, Brad M. and Lyon D. Lyon, 1997, Detecting long-run abnormal stock returns: The empirical power and specification of test-statistics, forthcoming, Journal of Financial Economics.

Barber, Brad M. and Lyon D. Lyon, 1996, How can long-run abnormal stock returns be both positively and negatively biased?, working paper (Graduate School of Management, University of California — Davis, Davis, California.)

Barber, Brad M., Lyon D. Lyon, and Chih-Ling Tsai, 1996, Improved methods for tests of long-run abnormal stock returns, working paper (Graduate School of Management, University of California — Davis, Davis, California.)

Blume, Marshall E. and Robert F. Stambaugh, 1983, Biases in computed returns: An application to the size effect, Journal of Financial Economics 12(3), 387–404.

Brav, Alon, Christopher Geczy, and Paul A. Gompers, 1995, The long-run underperformance of seasoned equity offerings revisited, Working paper (Harvard University, Graduate School of Business Administration, Boston, MA.)

Brown, Stephen J. and Jerold B. Warner, 1980, Measuring security price information, Journal of Financial Economics 8(3), 205–258.

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

Clark, Kent and Eli Ofek, 1994, Mergers as a means of restructuring distressed firms: An empirical investigation, Journal of Financial and Quantitative Analysis 29, 541–565.

Conrad, Jennifer and Gautam Kaul, 1993, Long-term market overreaction or biases in computed returns?, Journal of Finance 48(1), 39–64.

Cusatis, Patrick J., James A. Miles, and J. Randall Woolridge, 1993, Restructuring through spinoffs, Journal of Financial Economics 33, 293–311.

Desai, Hemang and Prem C. Jain, 1997, Long-run common stock returns following stock splits and reverse splits, forthcoming, Journal of Business.

Dharan, Bala G. and David L. Ikenberry, 1995, The long-run negative drift of post-listing stock returns, Journal of Finance 50, 1547–1574.

Fama, Eugene F. and Kenneth R. French, 1992, The cross-section of expected stock returns, Journal of Finance 47(2), 427–466.

Fama, Eugene F. and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, Journal of Financial Economics 33(1), 3–56.

Ikenberry, David and Josef Lakonishok, 1993, Corporate governance through the proxy contest: Evidence and implications, Journal of Business 66, 405–435.

Ikenberry, David, Josef Lakonishok, and Theo Vermaelen, 1995, Market underreaction to open market share repurchases, Journal of Financial Economics 39, 181–208.

Kothari, S.P. and Jerold B. Warner, 1997, Measuring long-horizon security price performance, forthcoming, Journal of Financial Economics.

Lakonishok, Josef and Theo Vermaelen, 1990, Anomalous price behavior around repurchase tender offers, Journal of Finance 45, 455–477.

Levhari, David and Haim Levy, 1977, The capital asset pricing model and the investment horizon, Review of Economics and Statistics 59(1), 92–104.

Loughran, Tim, 1993, Market microstructure or the poor performance of initial public offerings?, Journal of Financial Economics 33, 241–260.

Loughran, Tim and Jay R. Ritter, 1995, The new issues puzzle, Journal of Finance 50, 23–51.

Michaely, Roni, Richard H. Thaler, and Kent L. Womack, 1995, Price reactions to dividend initiations and omissions: Overreaction or drift?, Journal of Finance 50, 573–608.

Ritter, Jay R., 1991, The long-run performance of initial public offerings, The Journal of Finance 46, 3–27.

Roll, Richard, 1983, On computing mean returns and the small firm premium, Journal of Financial Economics 12(3), 371–386.

Shumway, Tyler, 1997, The delisting bias in CRSP data, Journal of Finance 52(1), forthcoming.

Spiess, D. Katherine and John Affleck-Graves, 1995, Underperformance in long-run stock returns following seasoned equity offerings, Journal of Financial Economics 38, 243–267.

Teoh, Siew Hong, Ivo Welch, and T. J. Wong, 1995, Earnings management and the post-issue underperformance of seasoned equity offerings, Working paper (Mitsui Life Financial Research Center, School of Business Administration, The University of Michigan, Ann Arbor, Michigan.)

Teoh, Siew Hong, Ivo Welch, and T. J. Wong, 1996, Earnings management and the long-run market performance of initial public offerings, Working paper (Mitsui Life Financial Research Center, School of Business Administration, The University of Michigan, Ann Arbor, Michigan.)

Wahal, Sunil, 1996, Pension fund activism and firm performance, Journal of Financial and Quantitative Analysis 31, 1–23.

Table 1
Studies examining long-horizon security performance: corporate event, event period, and sample size

For each of eight corporate events, the table lists empirical studies examining long-horizon security performance. The event is trading-day, month, or year zero. Sample sizes used are given. When long-horizon security returns are presented for subsamples, these sizes are also listed.

Corporate event Author(s)	Event period	Sample size
Initial public offerings (IPO) and seasoned equity offerings (SEO)		
Asquith and Mullins (1986)	Day -480 to +480	SEO: industrial n=189; utilities n=211
Ritter (1991)	Month 1 to 36	IPO: n=1,526
Loughran (1993)	Month 1 to 72	IPO: n=3,656
Loughran and Ritter (1995)	Year 1 to 3; 1 to 5	IPO: n=4,753; SEO: n=3,702
Spiess and Affleck-Graves (1995)	Month 1 to 60	SEO: n=1,247
Brav, Geczy, and Gompers (1995)	Month 1 to 60	SEO: n=3,931
Teoh, Welch, and Wong (1995)	Month -11 to +60	SEO: n=1,265; quartiles n=318
Teoh, Welch, and Wong (1996)	Month 1 to 36	IPO: n=1,526
Open market repurchases (OMR)		
Ikenberry, Lakonishok, and Vermaelen (1995)	Month 1 to 48	OMR: n=1,239
Dividend initiation (Init) and omission (Omit)		
Michaely, Thaler, and Womack (1995)	Month -12 to +36	Omit: n=887; Init: n=561
Stock split (SS) and reverse splits (RS)		
Desai and Jain (1997)	Month -6 to +36	SS: n=2,740; RS: n=15
Spin-off (SO)		
Cusatis, Miles, and Woolridge (1993)	Month 1 to 36	SO: n=146; Parent: n=131
Merger activity (Merg) and tender offers (TO)		
Lakonishok, and Vermaelen (1990)	Month 1 to 24	TO: n=221
Agarwal, Jaffe, Mandeker (1992)	Month 1 to 60	Merg: n=937; TO: n=227
Clark and Ofek (1994)	Year -3 to +3	Merg (distressed firms): n=38
Proxy contests (PC) and pension fund activism (Act)		
Ikenberry and Lakonishok (1993)	Month -60 to +60	PC: n=97 (subsets n=50; n=20; n=45; n=11; n=34; n=29; n=21)
Wahal (1996)	Year -2 to +1	Act: n=135 (subsets n=53; n=82)
Exchange changes		
Dharan and Ikenberry (1995)	Month 1 to 36	total: n=2,889; subsets n=1,146 n=393, n=349, n=254, n=470, n=388, n=284, n=283, n=271, n=197, n=44, n=139, n=82, n=186, n=12, n=86

Table 2
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
Raw returns:						
No winsorizing						
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
Winsorizing at three standard deviations						
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
Abnormal returns using size and book-to-market matching:						
Value-weighted portfolios						
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
Equal-weighted portfolios						
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
Control stock						
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
Winsorized abnormal returns using size and book-to-market matching:						
Value-weighted winsorized portfolios						
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 3
Descriptive statistics of the paired difference and 2-groups test statistics

This table contains descriptive information for paired difference and two-groups test statistics. The tests examine three-year holding-period 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 test statistics using abnormal returns based on value-weighted benchmark portfolios, statistics are shown with and without winsorizing at three standard deviations.

	Sample size					
	25	50	100	200	500	1000
Paired difference test statistics:						
Value-weighted benchmark portfolio with no winsorizing						
Mean	-0.047	-0.011	0.155	0.384	0.790	1.138
Std. Dev.	1.117	1.106	1.060	0.984	0.934	0.941
Skewness	-0.865	-0.972	-0.823	-0.541	-0.551	-0.398
Kurtosis	1.115	1.443	0.901	0.232	0.702	0.263
Equal-weighted benchmark portfolio with no winsorizing						
Mean	-0.288	-0.208	-0.221	-0.191	-0.117	-0.038
Std. Dev.	1.129	1.125	1.140	1.091	1.075	1.049
Skewness	-0.644	-0.749	-0.580	-0.373	-0.492	-0.449
Kurtosis	0.337	0.673	-0.003	-0.273	0.388	0.163
Control stocks with no winsorizing						
Mean	0.072	0.157	0.144	0.191	0.333	0.529
Std. Dev.	0.993	1.004	1.026	1.021	0.992	0.997
Skewness	-0.163	-0.205	-0.184	-0.025	-0.166	-0.178
Kurtosis	0.102	0.136	-0.094	-0.163	-0.180	-0.113
Value-weighted benchmark portfolio winsorizing at three standard deviations						
Mean	-0.100	-0.167	-0.136	-0.067	0.060	0.129
Std. Dev.	1.126	1.145	1.152	1.119	1.140	1.178
Skewness	-0.768	-0.740	-0.588	-0.276	-0.375	-0.173
Kurtosis	0.961	0.928	0.493	-0.169	0.457	0.024
Two-groups test statistics:						
Value-weighted benchmark portfolio with no winsorizing						
Mean	-0.001	0.027	0.162	0.363	0.727	1.044
Std. Dev.	0.919	0.929	0.921	0.877	0.845	0.860
Skewness	-0.644	-0.757	-0.696	-0.439	-0.479	-0.345
Kurtosis	0.470	0.800	0.583	0.037	0.516	0.146
Equal-weighted benchmark portfolio with no winsorizing						
Mean	-0.215	-0.140	-0.156	-0.138	-0.082	-0.016
Std. Dev.	0.933	0.929	0.961	0.932	0.932	0.920
Skewness	-0.628	-0.581	-0.469	-0.257	-0.371	-0.358
Kurtosis	0.505	0.323	-0.101	-0.367	0.141	-0.002
Control stocks with no winsorizing						
Mean	0.070	0.156	0.145	0.189	0.322	0.509
Std. Dev.	0.903	0.931	0.968	0.966	0.946	0.952
Skewness	-0.156	-0.139	-0.155	-0.009	-0.141	-0.158
Kurtosis	-0.039	0.011	-0.058	-0.132	-0.171	-0.136
Value-weighted benchmark portfolio winsorizing at three standard deviations						
Mean	-0.050	-0.113	-0.097	-0.042	0.062	0.120
Std. Dev.	0.923	0.950	0.976	0.959	0.979	1.015
Skewness	-0.549	-0.561	-0.486	-0.202	-0.318	-0.127
Kurtosis	0.379	0.456	0.333	-0.194	0.388	0.029

Table 4
Specification of paired difference tests at 5% significance level

This table reports percentage rejection frequencies in 1000 samples for paired difference 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 benchmark. The benchmarks include equal-weighted and value-weighted versions of three portfolios: the NYSE-AMEX-Nasdaq index, size matched, and size and book-to-market matched. The final benchmark reported is a randomly selected control stock matched to the sample stock using size and book-to-market ratio. 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		
	Alternative	Sample size			Sample size			Sample size		
Weighting and benchmark	hypotheses	50	200	1000	50	200	1000	50	200	1000
Equal-weighted market index	upper-tail	*1.3%*	*1.6%*	*1.1%*	*1.0%*	*1.9%*	*1.1%*	*1.2%*	*1.6%*	*1.8%*
	lower-tail	**13.4**	**12.5**	**13.2**	**12.3**	**14.0**	**12.7**	**14.3**	**11.2**	**9.0**
	2-tail	**9.4**	**8.1**	**7.9**	**8.6**	**10.1**	**8.4**	**10.0**	**8.3**	6.7
Equal-weighted size matched	upper-tail	*1.6*	*1.5*	*1.5*	*1.1*	*1.8*	*1.0*	*1.2*	*1.6*	*1.6*
	lower-tail	**12.9**	**11.4**	**12.6**	**12.5**	**13.9**	**12.9**	**13.7**	**11.7**	**8.7**
	2-tail	**9.3**	**8.3**	6.9	**8.3**	**10.4**	**9.0**	**10.8**	**8.2**	6.8
Equal-weighted size/book-to-market matched	upper-tail	*1.8*	*2.2*	*2.5*	*1.2*	*2.9*	3.6	*1.3*	*2.1*	*3.1*
	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 market index	upper-tail	5.0	**8.2**	**35.9**	5.7	**24.0**	**83.5**	**9.7**	**31.1**	**93.3**
	lower-tail	4.5	*1.8*	*0.3*	3.8	*0.4*	*0.0*	*3.1*	*0.0*	*0.0*
	2-tail	4.2	4.6	**23.2**	4.9	**12.7**	**73.3**	5.1	**16.9**	**86.8**
Value-weighted size matched	upper-tail	3.5	4.6	**17.7**	*3.0*	**8.6**	**32.8**	3.5	**8.8**	**36.9**
	lower-tail	5.9	3.6	*1.0*	6.9	*3.0*	*0.4*	**9.0**	*2.9*	*0.1*
	2-tail	4.8	3.7	**9.6**	5.7	4.2	**21.3**	6.8	5.6	**23.8**
Value-weighted size/book-to-market matched	upper-tail	*2.8*	*3.1*	**9.6**	*2.4*	**7.9**	**30.3**	*2.9*	**8.0**	**32.7**
	lower-tail	**8.0**	5.5	*2.4*	**7.7**	3.8	*0.5*	**9.3**	*3.0*	*0.2*
	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	upper-tail	**7.0**	**9.1**	**13.3**	5.8	**7.6**	**13.5**	*3.0*	*2.7*	*1.2*
	lower-tail	3.5	*2.6*	*2.0*	4.4	3.3	*1.6*	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

Table 5
Specification of two-groups tests at 5% significance level

This table reports percentage rejection frequencies in 1000 samples for 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 benchmark. The benchmarks include equal-weighted and value-weighted versions of three portfolios: the NYSE-AMEX-Nasdaq index, size matched, and size and book-to-market matched. The final benchmark reported is a randomly selected control stock matched to the sample stock using size and book-to-market ratio. 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		
	Alternative	Sample size			Sample size			Sample size		
Weighting and benchmark	hypotheses	50	200	1000	50	200	1000	50	200	1000
Equal-weighted market index	upper-tail	*0.7%*	*0.9%*	*0.8%*	*0.5%*	*1.2%*	*0.6%*	*0.6%*	*1.4%*	*1.0%*
	lower-tail	**9.2**	**8.6**	**8.5**	**8.4**	**10.3**	**10.0**	**9.4**	**8.0**	6.6
	2-tail	6.0	5.1	4.9	5.4	**7.2**	6.3	5.7	5.4	4.2
Equal-weighted size matched	upper-tail	*0.6*	*0.9*	*0.5*	*0.4*	*1.3*	*0.6*	*0.7*	*1.2*	*0.8*
	lower-tail	**8.6**	**7.9**	**7.0**	**8.1**	**10.7**	**9.9**	**9.3**	**7.7**	6.8
	2-tail	5.5	4.5	4.6	4.9	6.4	6.0	5.8	4.8	3.5
Equal-weighted size/book-to-market matched	upper-tail	*0.8*	*1.2*	*1.1*	*0.3*	*1.6*	*1.9*	*0.6*	*1.2*	*1.4*
	lower-tail	**7.4**	5.1	4.3	6.4	6.5	5.0	**7.5**	5.6	4.1
	2-tail	3.5	*3.1*	*2.8*	3.6	3.8	*2.7*	4.7	*2.8*	*2.1*
Value-weighted market index	upper-tail	3.4	6.5	**30.2**	4.7	**22.2**	**81.9**	**7.9**	**27.8**	**92.3**
	lower-tail	*3.0*	*1.2*	*0.1*	3.4	*0.3*	*0.0*	*2.5*	*0.0*	*0.0*
	2-tail	*2.5*	*3.1*	**18.9**	3.8	**11.1**	**71.9**	3.8	**14.1**	**84.9**
Value-weighted size matched	upper-tail	*1.1*	*3.0*	**12.8**	*1.2*	6.1	**28.3**	*1.3*	6.9	**31.0**
	lower-tail	3.5	*2.1*	*0.6*	5.0	*1.6*	*0.4*	5.7	*1.7*	*0.0*
	2-tail	*2.3*	*1.9*	6.0	3.7	*2.2*	**16.3**	*3.1*	*2.3*	**16.8**
Value-weighted size/book-to-market matched	upper-tail	*1.0*	*2.2*	6.4	*1.7*	4.5	**24.8**	*1.6*	5.8	**25.2**
	lower-tail	4.2	*2.9*	*1.6*	4.8	*1.6*	*0.3*	5.8	*1.3*	*0.1*
	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	upper-tail	4.7	6.9	**10.8**	5.2	6.4	**12.0**	*2.2*	*2.0*	*0.9*
	lower-tail	*2.4*	*2.3*	*1.5*	*3.2*	*2.5*	*1.3*	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 6
Specification of value-weighted paired difference tests at 5% significance level using returns winsorized at three standard deviations

This table reports percentage rejection frequencies in 1000 samples for value-weighted paired difference 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 benchmark. The benchmarks include equal-weighted and value-weighted versions of three portfolios: the NYSE-AMEX-Nasdaq index, size matched, and size and book-to-market matched. The final benchmark reported is a randomly selected control stock matched to the sample stock using size and book-to-market ratio. 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.

		Rejection rates at 5% significance								
		1 year holding period			3 year holding period			5 year holding period		
	Alternative	Sample size			Sample size			Sample size		
Weighting, benchmark, and test	hypotheses	50	200	1000	50	200	1000	50	200	1000
Value-weighted market index,	upper-tail	3.5%	4.6%	**11.0%**	5.1%	**17.3%**	**60.2%**	3.7	**20.9**	**71.8**
two-groups test	lower-tail	4.1	*2.7*	*1.9*	4.2	*1.6*	*0.2*	5.2	*0.7*	*0.0*
	2-tail	*2.9*	*3.2*	6.8	4.3	**11.2**	**46.9**	**8.9**	**12.0**	**60.2**
Value-weighted market index,	upper-tail	5.2	6.0	**15.0**	6.2	**19.5**	**62.4**	**10.4**	**24.5**	**75.2**
paired difference test	lower-tail	6.1	4.3	*2.8*	4.8	*1.9*	*0.2*	4.6	*0.9*	*0.0*
	2-tail	5.4	4.8	**9.9**	5.6	**12.5**	**49.4**	6.8	**14.6**	**63.8**
Value-weighted size matched,	upper-tail	*1.2*	*1.9*	*2.9*	*1.3*	3.3	**8.7**	*1.4*	4.3	6.9
two-groups test	lower-tail	4.3	4.6	5.8	6.1	5.1	4.2	**7.9**	5.4	*2.5*
	2-tail	*3.0*	*2.3*	4.3	4.0	3.7	**7.1**	4.4	3.7	5.1
Value-weighted size matched,	upper-tail	4.1	3.3	5.4	*2.9*	5.9	**11.3**	4.2	**7.0**	**10.5**
paired difference test	lower-tail	**8.5**	**8.2**	**10.1**	**9.4**	**7.3**	6.2	**11.2**	**8.3**	4.5
	2-tail	5.7	6.4	**8.4**	6.5	6.5	**10.8**	**9.5**	**8.2**	**8.2**
Value-weighted size/book-to-market	upper-tail	*1.1*	*1.4*	*1.2*	*1.6*	*2.1*	5.8	*1.5*	*3.0*	5.2
matched, two-groups test	lower-tail	5.6	6.8	**11.7**	5.9	5.5	5.4	**7.6**	5.6	*3.0*
	2-tail	3.6	3.3	6.1	3.5	3.5	5.8	4.7	*3.2*	4.0
Value-weighted size/book-to-market	upper-tail	*3.2*	*2.2*	*2.3*	*2.5*	4.5	**9.4**	3.5	6.1	**8.9**
matched, paired difference test	lower-tail	**10.2**	**11.8**	**19.0**	**9.8**	**8.3**	**7.4**	**11.9**	**9.9**	6.8
	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	upper-tail	5.4	**7.0**	**9.6**	5.3	5.6	**7.8**	*2.5*	*1.6*	*0.1*
control stocks, two-groups test	lower-tail	*2.1*	*1.1*	*0.5*	*2.7*	*2.6*	*2.1*	4.5	5.4	**11.9**
	2-tail	*3.1*	4.4	5.1	4.0	3.5	4.1	3.4	3.3	**7.4**
Size/book-to-market matched	upper-tail	**14.2**	**18.5**	**31.3**	**13.0**	**14.9**	**19.0**	**7.7**	6.7	*1.7*
control stocks, paired difference test	lower-tail	**12.7**	**8.6**	5.4	**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**

Table 7
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 amount, ranging from 10% to 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				
			Induced abnormal returns (IAR) for three year holding period				
Test Weighting and benchmark	Alternative hypotheses	Sign of IAR	10.0%	20.0%	30.0%	40.0%	50.0%
Paired difference tests							
Equal-weighted size/book-to-market matched	upper-tail	+	7.8%	23.5%	49.7%	77.9%	93.5%
	lower-tail	-	24.8	39.6	56.1	66.3	76.8
	2-tail	+	4.2	13.5	34.3	63.4	86.1
	2-tail	-	17.9	32.6	48.6	60.7	70.4
Value-weighted size/book-to-market matched	upper-tail	+	11.6	31.8	61.0	86.5	97.1
	lower-tail	-	19.5	33.6	49.4	65.0	76.2
	2-tail	+	6.2	19.1	44.5	74.8	92.8
	2-tail	-	13.9	27.3	42.1	57.2	70.3
Size/book-to-market matched control stocks	upper-tail	+	13.2	25.3	41.2	58.4	71.0
	lower-tail	-	8.5	18.0	29.9	45.3	59.6
	2-tail	+	8.0	15.7	28.8	45.5	59.6
	2-tail	-	6.7	10.6	21.1	33.4	49.0
Two-groups test							
Equal-weighted size/book-to-market matched	upper-tail	+	3.7	14.8	36.0	66.0	88.1
	lower-tail	-	17.4	31.9	48.0	61.4	72.1
	2-tail	+	1.4	5.2	19.1	46.0	74.8
	2-tail	-	11.8	24.1	37.9	54.2	64.9
Value-weighted size/book-to-market matched	upper-tail	+	6.3	22.4	49.8	78.1	95.2
	lower-tail	-	13.4	27.6	43.3	60.1	73.0
	2-tail	+	3.1	11.1	30.2	60.3	86.0
	2-tail	-	8.1	19.8	34.4	50.1	65.4
Size/book-to-market matched, control stocks	upper-tail	+	10.7	21.0	36.8	54.1	67.8
	lower-tail	-	7.0	14.6	26.1	40.2	56.0
	2-tail	+	6.3	12.2	23.1	39.9	56.5
	2-tail	-	4.3	8.6	16.8	29.2	44.1
Paired difference test with winsorized data							
Value-weighted size/book-to-market matched	upper-tail	+	12.5	33.6	62.3	88.7	97.8
	lower-tail	-	25.7	42.1	60.8	75.6	86.8
	2-tail	+	7.0	21.3	48.9	78.9	95.8
	2-tail	-	18.0	34.9	52.4	68.0	82.3
Two-groups test with winsorized data							
Value-weighted size/book-to-market matched	upper-tail	+	7.0	22.9	50.9	80.0	95.9
	lower-tail	-	17.1	34.3	52.8	69.9	83.3
	2-tail	+	3.3	11.8	32.5	64.2	89.3
	2-tail	-	10.9	27.9	42.7	60.6	75.2

Table 7- Continued

Panel B: 200 stock sample size			Rejection frequencies at 5% significance				
			Induced abnormal returns (IAR) for three year holding period				
Test Weighting and benchmark	Alternative hypotheses	Sign of IAR	10.0%	20.0%	30.0%	40.0%	50.0%
Paired difference tests							
Equal-weighted size/book-to-market matched	upper-tail	+	20.4	66.4	96.4	99.5	99.8
	lower-tail	-	37.0	64.5	84.6	93.4	96.8
	2-tail	+	11.3	50.7	91.3	99.2	99.5
	2-tail	-	29.5	57.3	77.8	90.5	95.3
Value-weighted size/book-to-market matched	upper-tail	+	43.5	87.4	99.2	99.7	99.9
	lower-tail	-	21.3	50.1	74.2	87.4	94.0
	2-tail	+	26.6	76.6	98.2	99.6	99.7
	2-tail	-	14.6	40.8	65.9	83.3	92.2
Size/book-to-market matched control stocks	upper-tail	+	24.6	52.6	79.0	93.8	98.3
	lower-tail	-	16.6	39.1	63.9	82.3	92.6
	2-tail	+	16.0	40.1	66.7	89.2	96.6
	2-tail	-	10.5	30.1	52.8	76.1	88.3
Two-groups test							
Equal-weighted size/book-to-market matched	upper-tail	+	13.3%	54.6%	92.7%	99.5%	99.8%
	lower-tail	-	29.3	58.9	80.8	91.9	96.1
	2-tail	+	7.3	37.8	85.0	98.9	99.5
	2-tail	-	20.9	50.1	73.3	88.2	94.5
Value-weighted size/book-to-market matched	upper-tail	+	34.5	81.6	98.7	99.7	99.9
	lower-tail	-	16.0	44.0	69.4	86.0	93.4
	2-tail	+	18.3	65.9	96.9	99.6	99.7
	2-tail	-	9.3	35.4	61.4	81.0	90.5
Size/book-to-market matched control stocks	upper-tail	+	21.8	48.4	75.7	92.2	98.0
	lower-tail	-	13.6	35.3	61.6	80.8	91.8
	2-tail	+	13.8	35.5	63.6	87.0	96.1
	2-tail	-	8.0	26.3	49.3	73.8	87.4
Paired difference test with winsorized data							
Value-weighted size/book-to-market matched	upper-tail	+	38.2	84.7	99.3	100.0	100.0
	lower-tail	-	41.4	74.9	94.4	98.6	99.5
	2-tail	+	23.8	74.4	99.0	100.0	100.0
	2-tail	-	32.6	68.1	91.6	98.1	99.3
Two-groups test with winsorized data							
Value-weighted size/book-to-market matched	upper-tail	+	26.7	76.0	99.0	100.0	100.0
	lower-tail	-	32.9	69.2	92.5	98.2	99.4
	2-tail	+	15.8	62.2	97.2	100.0	100.0
	2-tail	-	23.1	59.2	86.9	97.4	99.3

Table 7- Continued

Panel C: 1000 stock sample size			Rejection frequencies at 5% significance				
			Induced abnormal returns (IAR) for three year holding period				
Test Weighting and benchmark	Alternative hypotheses	Sign of IAR	10.0%	20.0%	30.0%	40.0%	50.0%
Paired difference tests							
Equal-weighted size/book-to-market matched	upper-tail	+	72.1	99.9	100.0	100.0	100.0
	lower-tail	-	63.9	95.5	98.1	99.2	99.9
	2-tail	+	59.5	99.4	100.0	100.0	100.0
	2-tail	-	55.4	93.9	97.8	98.5	99.9
Value-weighted size/book-to-market matched	upper-tail	+	97.2	100.0	100.0	100.0	100.0
	lower-tail	-	29.4	83.1	97.5	99.5	99.8
	2-tail	+	93.9	100.0	100.0	100.0	100.0
	2-tail	-	21.7	76.9	96.4	99.2	99.8
Size/book-to-market matched control stocks	upper-tail	+	70.7	98.5	100.0	100.0	100.0
	lower-tail	-	29.2	79.8	96.1	98.6	99.7
	2-tail	+	58.1	96.0	100.0	100.0	100.0
	2-tail	-	20.1	72.4	95.6	98.1	99.2
Two-groups test							
Equal-weighted size/book-to-market matched	upper-tail	+	63.5%	99.9%	100.0%	100.0%	100.0%
	lower-tail	-	58.5	95.2	98.1	99.2	99.9
	2-tail	+	47.8	99.1	100.0	100.0	100.0
	2-tail	-	48.5	92.6	97.7	98.5	99.8
Value-weighted size/book-to-market matched	upper-tail	+	95.1	100.0	100.0	100.0	100.0
	lower-tail	-	24.3	80.4	97.4	99.5	99.8
	2-tail	+	92.0	100.0	100.0	100.0	100.0
	2-tail	-	16.7	73.4	96.1	99.2	99.7
Size/book-to-market matched, control stocks	upper-tail	+	67.7	98.0	100.0	100.0	100.0
	lower-tail	-	26.5	78.5	96.1	98.6	99.7
	2-tail	+	53.8	94.9	100.0	100.0	100.0
	2-tail	-	17.9	70.4	95.0	98.0	99.2
Paired difference test with winsorized data							
Value-weighted size/book-to-market matched	upper-tail	+	90.7	100.0	100.0	100.0	100.0
	lower-tail	-	80.3	99.6	100.0	100.0	100.0
	2-tail	+	85.3	100.0	100.0	100.0	100.0
	2-tail	-	72.7	99.3	100.0	100.0	100.0
Two-groups test with winsorized data							
Value-weighted size/book-to-market matched	upper-tail	+	85.8	100.0	100.0	100.0	100.0
	lower-tail	-	74.4	99.4	100.0	100.0	100.0
	2-tail	+	77.1	100.0	100.0	100.0	100.0
	2-tail	-	63.8	98.8	100.0	100.0	100.0

Table 8
Power of bootstrap tests for three year holding period returns

This table reports percentage rejection frequencies in 1000 samples for the bootstrap test using three year compounded holding-period returns. Abnormal performance is induced by adding a positive or negative amount, ranging from 0% to 50% over the three years, to the actual return. P-values are computed by ranking the sample mean return and the mean returns of 1000 pseudo-samples samples. The pseudo-samples are generated by randomly selecting stocks with the same event dates from the same size and book-to-market categories as the sample firms. 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.

			Rejection frequencies at 5% significance					
			Induced abnormal returns (IAR) for three year holding period					
Sample size	Alternative hypotheses	Sign of IAR	0.0%	10.0%	20.0%	30.0%	40.0%	50.0%
50 stocks	upper-tail	+	4.8%	8.7%	16.8%	29.5%	49.1%	70.4%
	lower-tail	-	6.7	17.6	37.3	60.0	76.8	86.7
	2-tail	+	5.5	4.9	7.9	14.5	25.4	44.1
	2-tail	-	5.5	13.5	29.0	50.9	70.5	83.7
200 stocks	upper-tail	+	4.8	17.4	48.4	83.4	97.8	99.8
	lower-tail	-	5.8	31.4	69.6	90.0	97.4	99.1
	2-tail	+	5.3	9.0	27.4	63.9	92.2	99.2
	2-tail	-	5.3	22.9	60.8	86.9	96.0	98.8
1000 stocks	upper-tail	+	5.6	55.4	98.9	100.0	100.0	100.0
	lower-tail	-	6.5	71.3	97.8	100.0	100.0	100.0
	2-tail	+	6.8	40.1	97.0	100.0	100.0	100.0
	2-tail	-	6.8	62.5	96.9	99.9	100.0	100.0

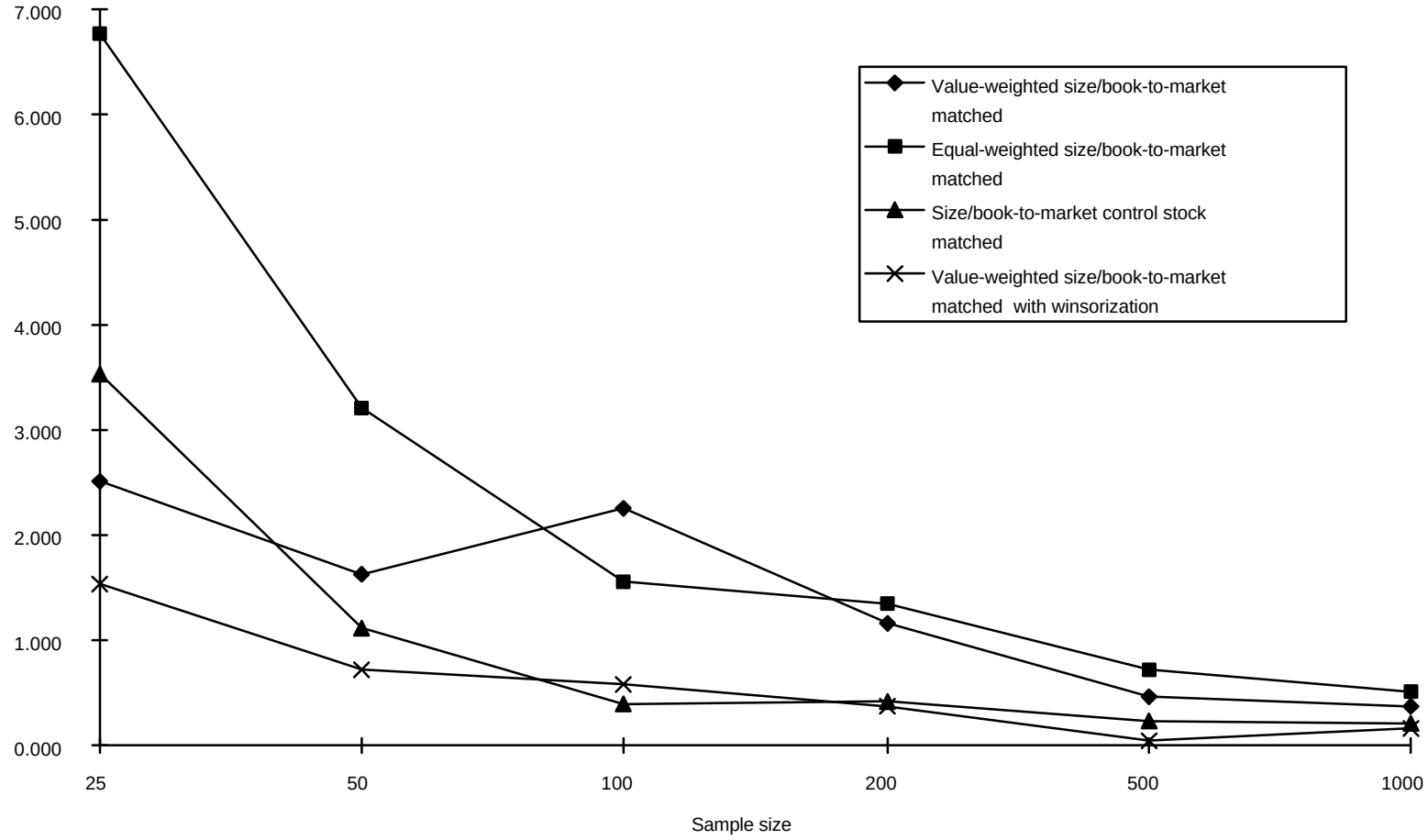

Fig. 1. Skewness for sample sizes ranging from 25 to 1000 stocks for three year holding period returns matching on size and book-to-market using value-weighted (with and without winsorization at three standard deviations) and equal-weighted portfolios and control stock procedures.