THE JOURNAL OF FINANCE • VOL. LIV, NO. 1 • FEBRUARY 1999

Improved Methods for Tests of Long-Run Abnormal Stock Returns

JOHN D. LYON, BRAD M. BARBER, and CHIH-LING TSAI*

ABSTRACT

We analyze tests for long-run abnormal returns and document that two approaches yield well-specified test statistics in random samples. The first uses a traditional event study framework and buy-and-hold abnormal returns calculated using carefully constructed reference portfolios. Inference is based on either a skewness-adjusted t-statistic or the empirically generated distribution of long-run abnormal returns. The second approach is based on calculation of mean monthly abnormal returns using calendar-time portfolios and a time-series t-statistic. Though both approaches perform well in random samples, misspecification in nonrandom samples is pervasive. Thus, analysis of long-run abnormal returns is treacherous.

COMMONLY USED METHODS TO TEST for long-run abnormal stock returns yield misspecified test statistics, as documented by Barber and Lyon (1997a) and Kothari and Warner (1997).[1] Simulations reveal that empirical rejection levels routinely exceed theoretical rejection levels in these tests. In combination, these papers highlight three causes for this misspecification. First, the *new listing* or *survivor bias* arises because in event studies of long-run abnormal returns, sampled firms are tracked for a long post-event period, but firms that constitute the index (or reference portfolio) typically include firms that begin trading subsequent to the event month. Second, the *rebalancing bias* arises because the compound returns of a reference portfolio, such as an equally weighted market index, are typically calculated assuming periodic (generally monthly) rebalancing, whereas the returns of sample firms are compounded without rebalancing. Third, the *skewness bias* arises because the distribution of long-run abnormal stock returns is positively skewed, which also contributes to the misspecification of test statistics. Generally, the new listing bias creates a positive bias in test statistics, and the rebalancing and skewness biases create a negative bias.

---

* Graduate School of Management, University of California, Davis. This paper was previously entitled "Holding Size while Improving Power in Tests of Long-Run Abnormal Stock Returns." We have benefited from the suggestions of John Affleck-Graves, Peter Bickel, Alon Brav, Sandra Chamberlain, Arnold Cowan, Masako Darrough, Eugene Fama, Ken French, Peter Hall, Inmoo Lee, Tim Loughran, Michael Maher, Terrance Odean, Stephen Penman, N. R. Prabhala, Raghu Rau, Jay Ritter, René Stulz, Brett Trueman, Ralph Walkling, two anonymous reviewers, and seminar participants at the Australian Graduate School of Management, State University of New York–Buffalo, Ohio State, University of California–Berkeley, University of California–Davis, the University of Washington, and 1997 Western Finance Association Meetings. Chih-Ling Tsai was supported by National Science Foundation grant DMS 95-10511. All errors are our own.

[1] Numerous recent studies analyze long-run abnormal stock returns in an event-study context. Citations to many of these studies can be found in Kothari and Warner (1997) and Barber and Lyon (1997a).

In this research, we evaluate two general approaches for tests of long-run abnormal stock returns that control for these three sources of bias. The first approach is based on a traditional event study framework and buy-and-hold abnormal returns. In this approach we first carefully construct reference portfolios that are free of the new listing and rebalancing biases. Consequently, these reference portfolios yield a population mean abnormal return measure that is identically zero and, therefore, reduce the misspecification of test statistics. Then we control for the skewness bias in tests of long-run abnormal returns by applying standard statistical methods recommended for settings when the underlying distribution is positively skewed. Two statistical methods virtually eliminate the skewness bias in random samples: (1) a bootstrapped version of a skewness-adjusted t-statistic, and (2) empirical p values calculated from the simulated distribution of mean long-run abnormal returns estimated from pseudoportfolios. The first method is developed and analyzed based on a rich history of research in statistics that considers the properties of t-statistics in positively skewed distributions, which dates back at least to Neyman and Pearson (1928) and Pearson (1929a, 1929b). In the second method, based on the empirical methods of Brock, Lakonishok, and LeBaron (1992) and Ikenberry, Lakonishok, and Vermaelen (1995), we generate the empirical distribution of mean long-run abnormal stock returns under the null hypothesis. The statistical significance of the sample mean is evaluated based on this empirically generated distribution.[2] Kothari and Warner (1997) note that these methods "seem like a promising framework for alternative tests which can potentially reduce misspecification."

These two statistical methods yield well-specified test statistics in random samples, and in combination with carefully constructed reference portfolios, they control well for the new listing, rebalancing, and skewness biases. However, the methods are unable to control for two additional sources of misspecification: Cross-sectional dependence in sample observations, and a poorly specified asset pricing model. Brav (1997) argues that cross-sectional dependence in sample observations can lead to poorly specified test statistics in some sampling situations and we concur.

The second general approach that we consider is based on calendar-time portfolios, as discussed by Fama (1997) and implemented in recent work by Loughran and Ritter (1995), Brav and Gompers (1997), and Brav, Géczy, and Gompers (1995). Using this approach, calendar-time abnormal returns are calculated for sample firms. Inference is based on a t-statistic derived from the time-series of the monthly calendar-time portfolio abnormal returns. This approach eliminates the problem of cross-sectional dependence among sample firms but, unlike buy-and-hold abnormal returns, the abnormal return measure does not precisely measure investor experience.

---

[2] Research independently undertaken by Cowan and Sergeant (1996) considers the same general issues discussed here.

The most serious problem with inference in studies of long-run abnormal stock returns is the reliance on a model of asset pricing. All tests of the null hypothesis that long-run abnormal stock returns are zero are implicitly a joint test of (i) long-run abnormal returns are zero and (ii) the asset pricing model used to estimate abnormal returns is valid. In this research, we document that controlling for size and book-to-market alone is not sufficient to yield well-specified test statistics when samples are drawn from nonrandom samples, regardless of the approach used. In short, the rejection of the null hypothesis in tests of long-run abnormal returns is not a sufficient condition to reject the theoretical framework of market efficiency.

The remainder of this paper is organized as follows. We discuss the data and the construction of reference portfolios in Section I. We discuss the details of long-run abnormal return calculations in Section II. Our statistical tests and empirical methods are presented in Section III, followed by results in Section IV. We discuss the use of cumulative abnormal returns and calendar-time portfolio methods in Section V. We make concluding remarks in Section VI.

I. Data and the Construction of Reference Portfolios

Our analysis begins with all NYSE/AMEX/Nasdaq firms with available data on the monthly return files created by the Center for Research in Security Prices (CRSP). Our analysis covers the period July 1973 through December 1994 (we begin in 1973 because return data for Nasdaq firms are not available from CRSP prior to 1973). Since event studies of long-run abnormal returns focus on the common stock performance of corporations, we delete the firm-month returns on securities identified by CRSP as other than ordinary common shares (CRSP share codes 10 and 11). Thus, for example, we exclude from our analysis returns on American Depository Receipts, closed-end funds, foreign-domiciled firms, Primes and Scores, and real estate investment trusts.

A key feature of our analysis is the careful construction of reference portfolios, which alleviates the new listing and rebalancing biases (Barber and Lyon (1997a), Kothari and Warner (1997)). Our reference portfolios are formed on the basis of firm size and book-to-market ratios in July of each year, 1973 through 1993. Though we restrict our analysis to portfolios formed on the basis of these two variables, the method we employ to construct our reference portfolios can also be used to construct portfolios on the basis of other firm characteristics (e.g., prior return performance, sales growth, industry, or earnings yields).

We construct 14 size reference portfolios as follows:

1. We calculate firm size (market value of equity calculated as price per share multiplied by shares outstanding) in June of each year for all firms.
2. In June of year *t*, we rank all NYSE firms in our population on the basis of firm size and form size decile portfolios based on these rankings.

3. AMEX and Nasdaq firms are placed in the appropriate NYSE size decile based on their June market value of equity.
4. We further partition the smallest size decile, decile one, into quintiles on the basis of size rankings of all firms (without regard to exchange) in June of each year.
5. The returns of the size portfolios are tracked from July of year t for τ months.

Since Nasdaq is populated predominantly with smaller firms, this ranking procedure leaves many more firms in the smallest decile of firm size than in the other nine deciles (approximately 50 percent of all firms fall in the smallest size decile); therefore we further partition the smallest size decile into quintiles on the basis of size rankings without regard to exchange.

We construct ten book-to-market reference portfolios as follows:

1. We calculate a firm's book-to-market ratio using the book value of common equity (COMPUSTAT data item 60) reported on a firm's balance sheet in year $t - 1$ divided by the market value of common equity in December of year $t - 1$.[3]
2. In December of year $t - 1$, we rank all NYSE firms in our population on the basis of book-to-market ratios and form book-to-market decile portfolios based on these rankings.
3. AMEX and Nasdaq firms are placed in the appropriate NYSE book-to-market decile based on their book-to-market ratio in year $t - 1$.
4. The returns of the book-to-market portfolios are tracked from July of year t for τ months.

The extreme book-to-market deciles contain slightly more firms than deciles two through eight, but no book-to-market decile portfolio contains more than 20 percent of the firms.

We calculate one-, three-, and five-year returns for the size and book-to-market reference portfolios in two ways. First, in each month we calculate the mean return for each portfolio and then compound this mean return over τ months:

$$R_{ps\tau}^{\text{reb}} = \prod_{t=s}^{s+\tau} \left[1 + \frac{\sum_{i=1}^{n_t} R_{it}}{n_t} \right] - 1, \tag{1}$$

where s is the beginning period, τ is the period of investment (in months), R_{it} is the return on security i in month t, and n_t is the number of securities in month t.

---

[3] Firms with negative book value of equity, though this is relatively rare, are excluded from the analysis.

Though research in financial economics commonly uses long-horizon reference portfolio returns calculated in this manner, for two reasons they do not accurately reflect the returns earned on a passive buy-and-hold strategy of investing equally in the securities that constitute the reference portfolio. First, this portfolio return assumes monthly rebalancing to maintain equal weights. This rebalancing leads to an inflated long-horizon return on the reference portfolio, which can likely be attributed to bid-ask bounce and nonsynchronous trading.[4] We refer to this as the *rebalancing bias*. Second, this portfolio return includes firms newly listed subsequent to portfolio formation (period s). Ritter (1991) documents that firms that go public underperform an equally weighted market index, though Brav and Gompers (1997) document that this underperformance is confined to small, high-growth firms. Since it is likely that firms that go public are a significant portion of newly listed firms, the result is a downwardly biased estimate of the long-horizon return from investing in a *passive* (i.e., not rebalanced) reference portfolio in period s. We refer to this as the *new listing* bias. In reference to the rebalanced nature of this return calculation, we denote the return calculated in this manner with the superscript "reb."

Our second method of calculating the long-horizon returns on a reference portfolio involves first compounding the returns on securities constituting the portfolio and then summing across securities:

$$R_{ps\tau}^{\text{bh}} = \sum_{i=1}^{n_s} \frac{\left[\prod_{t=s}^{s+\tau} (1 + R_{it})\right] - 1}{n_s}, \tag{2}$$

where n_s is the number of securities traded in month s, the beginning period for the return calculation. The return on this portfolio represents a passive equally weighted investment in all securities constituting the reference portfolio in period s. There is no investment in firms newly listed subsequent to period s, nor is there monthly rebalancing of the portfolio. Consequently, in reference to the buy-and-hold nature of this return calculation, we denote the return calculated in this manner with the superscript "bh."

An unresolved issue in the calculation of the buy-and-hold portfolio return is where an investor places the proceeds of investments in firms delisted subsequent to period s. In this research, we assume that the proceeds of delisted firms are invested in an equally weighted reference portfolio, which is rebalanced monthly. Thus, missing monthly returns are filled with the mean monthly return of firms comprising the reference portfolio.[5]

---

[4] For a discussion of these issues, see Blume and Stambaugh (1983), Roll (1983), Conrad and Kaul (1993), and Ball, Kothari, and Wasley (1995).

[5] As we document below, this leads to a small overstatement of the returns that can be earned from investing in the portfolio of small firms (portfolio 1A). However, this would only lead to biases in tests of long-run abnormal returns if sample firms are delisted significantly more or less frequently than the firms comprising the benchmark portfolio.

In Table I, we present the annualized one-, three-, and five-year rebalanced and buy-and-hold returns on our fourteen size portfolios (Panel A) and ten book-to-market portfolios (Panel B) assuming investment in July of each year, 1973 through 1993.[6] The last three columns of this table present the difference between the rebalanced portfolio returns and the buy-and-hold portfolio returns. Examination of Panel A reveals the well-documented small firm effect. However, the returns from investing in unusually small firms (portfolio 1A) are overstated by more than 10 percent per year when calculated with monthly rebalancing.

The calculations in Table I, Panel A, are based on reinvestment of delisted firms in an equally weighted, monthly rebalanced portfolio of firms of similar size. If we assume reinvestment of delisted firms in a CRSP value-weighted market index (results not reported in the table), the returns on size portfolio 1A (1B) range from 26.1 percent (18.6 percent) at a one-year holding period to 21.8 percent (19.3 percent) at a five-year holding period. If we assume reinvestment of delisted firms in a firm from the same size class, the returns on size portfolio 1A (1B) range from 27.2 percent (18.4 percent) at a one-year holding period to 23.0 percent (19.3 percent) at a five-year holding period. Returns on the remaining size portfolios are affected by fewer than than 100 basis points, regardless of whether the reinvestment is in a similar size firm or a value-weighted market index.

Examination of Panel B reveals that the returns from investing in high book-to-market firms (portfolio 10) are overstated by 3 to 4 percent per year when calculated with monthly rebalancing. The calculations in Panel B are based on reinvestment of delisted firms in an equally–weighted, monthly rebalanced portfolio of firms of similar book-to-market ratios. If we assume reinvestment of delisted firms in a CRSP value-weighted market index (results not reported in the table), the returns on book-to-market portfolio 10 (9) range from 24.2 percent (23.0 percent) at a one-year holding period to 21.9 percent (22.2 percent) at a five-year holding period. If we assume reinvestment of delisted firms in a firm from the same book-to-market class, the returns on book-to-market portfolio 10 (9) range from 25.3 percent (23.4 percent) at a one-year holding period to 22.8 percent (23.3 percent) at a five-year holding period. The returns on the remaining book-to-market portfolios are affected by fewer than than 100 basis points, regardless of whether the reinvestment is in a firm with similar book-to-market ratio or a value-weighted market index.

Our results indicate that the use of compounded, equally weighted monthly returns also yields inflated returns on reference portfolios, particularly in reference portfolios composed of small firms or firms with high book-to-market ratios. Canina et al. (1998) document a similar bias when the daily returns of an equally weighted market index are used in lieu of monthly

---

[6] The three- and five-year return means include overlapping years, while the one-year return mean does not. The last one-, three-, and five-year return assumes initial investment in July 1993, July 1991, and July 1989, respectively.

Table I

Annualized Returns from Investing in Size or Book-to-Market Decile Portfolios with Monthly Rebalancing or Buy-and-Hold Strategy: July 1973 to June 1994

The rebalanced portfolio returns assume monthly rebalancing to maintain equal weights and include firms newly listed subsequent to July of each year. The buy-and-hold portfolio returns assume equal initial investments in each security traded in July of each year. If a firm is delisted or is missing return data, the return on the equally weighted size decile portfolio for that month is spliced into the return series for that firm.

Panel A reports size deciles created in July of each year based on rankings of NYSE firms with available size measures (price times shares outstanding) in June of that year. AMEX and Nasdaq firms are then place in the corresponding size decile portfolio. Size decile one (small firms) is further partitioned into quintiles based on size rankings for NYSE/AMEX/Nasdaq firms. Portfolio 1A (1E) contains the smallest (largest) 20 percent of firms in size decile one.

Panel B reports book-to-market deciles created in July of each year based on rankings of NYSE firms with available book-to-market data (book value of common equity divided by market value of common equity) in December of the preceding year. AMEX and Nasdaq firms are then placed in the corresponding book-to-market decile portfolio. Decile 1 contains growth firms (low book-to-market); decile 10 contains value firms (high book-to-market).

	Rebalanced Returns (%)			Buy-and-Hold Returns (%)			Difference (%)		
Decile	1 yr.	3 yrs.	5 yrs.	1 yr.	3 yrs.	5 yrs.	1 yr.	3 yrs.	5 yrs.
Panel A: Size Decile Portfolio Returns over Three Investment Periods									
1A	41.4	40.4	36.5	28.9	27.7	26.4	12.6	12.7	10.1
1B	24.1	23.8	23.8	19.3	20.0	20.7	4.8	3.9	3.1
1C	20.0	19.8	19.8	17.8	19.3	19.6	2.2	0.6	0.2
1D	17.8	17.5	18.0	17.1	17.7	18.5	0.7	−0.1	−0.6
1E	17.6	17.1	17.3	17.2	17.9	18.7	0.4	−0.8	−1.4
2	17.6	17.3	17.6	17.7	18.7	19.2	−0.1	−1.4	−1.6
3	18.1	18.0	18.2	18.1	19.1	19.3	0.0	−1.1	−1.0
4	18.2	18.0	18.2	18.1	18.6	19.1	0.1	−0.6	−0.8
5	18.9	18.8	18.7	19.1	18.1	18.3	−0.2	0.6	0.4
6	16.7	16.8	16.6	16.4	17.0	17.3	0.3	−0.2	−0.8
7	16.9	16.9	17.1	16.4	16.4	16.3	0.5	0.5	0.8
8	16.3	16.4	16.3	15.9	16.1	16.1	0.4	0.3	0.2
9	14.7	15.1	15.2	14.3	14.9	15.3	0.4	0.2	−0.1
10	12.8	13.4	14.0	12.4	13.2	13.9	0.4	0.2	0.1
Panel B: Book-to-Market Decile Portfolio Returns over Three Investment Periods									
1	10.6	9.6	9.9	8.3	9.4	11.1	2.2	0.2	−1.2
2	16.8	16.6	16.9	15.7	16.2	17.4	1.1	0.4	−0.5
3	18.2	18.2	18.4	16.8	17.4	17.4	1.4	0.8	1.0
4	19.7	19.0	19.0	18.2	17.9	18.3	1.5	1.1	0.7
5	21.1	20.7	20.8	19.4	19.0	19.3	1.7	1.7	1.5
6	20.6	20.4	20.1	19.7	19.6	19.7	0.9	0.8	0.3
7	21.3	21.6	21.7	21.0	20.9	21.2	0.4	0.6	0.5
8	23.0	23.5	23.2	21.0	22.4	22.3	2.0	1.0	0.9
9	25.2	25.4	25.0	23.4	23.6	23.6	1.9	1.7	1.4
10	29.3	28.4	26.9	25.1	24.5	23.8	4.2	3.9	3.1

returns on the same index. We later document that the use of our buy-and-hold reference portfolios significantly reduces much of the misspecification problems that plague tests of long-run abnormal returns.

II. Calculation of Abnormal Returns

We calculate long-horizon buy-and-hold abnormal returns as:

$$AR_{i\tau} = R_{i\tau} - E(R_{i\tau}), \tag{3}$$

where $AR_{i\tau}$ is the τ period buy-and-hold abnormal return for security i, $R_{i\tau}$ is the τ period buy-and-hold return, and $\mathrm{E}(R_{i\tau})$ is the τ period expected return for security i. In this research, we use either: (i) the rebalanced return on a size/book-to-market reference portfolio ($R_{ps\tau}^{\mathrm{reb}}$), (ii) the buy-and-hold return on a size/book-to-market reference portfolio ($R_{ps\tau}^{\mathrm{bh}}$), or (iii) the return on a size and book-to-market matched control firm as a proxy for the expected return for each security.[7]

We use seventy size/book-to-market reference portfolios. These portfolios are formed as follows. Fourteen size reference portfolios are created as described in Section I. Each size portfolio is further partitioned into five book-to-market quintiles (without regard to exchange) in June of year t. We calculate the τ month rebalanced and buy-and-hold return on each of these seventy portfolios as described in Section I. Note that the population mean long-horizon abnormal return calculated using the buy-and-hold reference portfolios is guaranteed to be zero by construction of the abnormal return measure, regardless of the horizon of analysis. Thus, both the new listing and rebalancing biases, which plague tests of long-run abnormal stock returns, are eliminated.

To identify a size and book-to-market matched control firm, we first identify all firms with market value of equity between 70 percent and 130 percent of the market value of equity of the sample firm; from this set of firms we choose the firm with the book-to-market ratio closest to that of the sample firm.

It is well-documented that the common stocks of small firms and firms with high book-to-market ratios earn high rates of return (Fama and French (1992), Chan, Jegadeesh, and Lakonishok (1995), Davis (1994), Barber and Lyon (1997b), Fama and French (1997)). Consequently, we consider portfolios or control firms selected on the basis of firm size and book-to-market ratio. However, we later document that the use of these two characteristics alone can yield misspecified test statistics in certain sampling situations. Nonetheless, the methods that we analyze can be applied in straightforward ways to portfolios formed on alternative characteristics (for example, pre-event return performance, sales growth, industry, or earnings yields). In fact, one of our central messages is that controlling for firm size and book-to-market ratio alone does not guarantee well-specified test statistics.

---

[7] We only report results based on abnormal returns calculated in this manner. Research in financial economics often employs cumulative abnormal returns (summed monthly abnormal returns). However, Barber and Lyon (1997a) argue that cumulative abnormal returns can lead to incorrect inference regarding long-horizon return performance. Nonetheless, the alternative methods that we analyze in this research also work well for the analysis of cumulative abnormal returns when reference portfolios are constructed for each sample observation by averaging only the monthly returns of firms listed in the initial event month (period s). We discuss this issue in detail in Section V.

III. Statistical Tests and Simulation Method

In this section, we describe the statistical tests that we analyze in tests of long-run abnormal returns. We close the section with a description of the simulation method used to evaluate the empirical specification of these tests.

A. Conventional t-statistic

To test the null hypothesis that the mean buy-and-hold abnormal return is equal to zero for a sample of n firms, we first employ a conventional t-statistic:

$$t = \frac{\overline{AR}_\tau}{\sigma(AR_\tau)/\sqrt{n}}, \tag{4}$$

where $\overline{AR}_\tau$ is the sample mean and $\sigma(AR_\tau)$ is the cross-sectional sample standard deviation of abnormal returns for the sample of n firms. We apply this conventional t-statistic to abnormal returns calculated using (i) 70 size/book-to-market rebalanced portfolios, (ii) 70 size/book-to-market buy-and-hold portfolios, and (iii) size/book-to-market matched control firms.

B. Bootstrapped Skewness-Adjusted t-statistic

Barber and Lyon (1997a) document that long-horizon buy-and-hold abnormal returns are positively skewed and that this positive skewness leads to negatively biased t-statistics. Their results are consistent with early investigations by Neyman and Pearson (1929a), and Pearson (1929a, 1929b), which indicate that skewness has a greater effect on the distribution of the t-statistic than does kurtosis and that positive skewness in the distribution from which observations arise results in the sampling distribution of t being negatively skewed. This leads to an inflated significance level for lower-tailed tests (i.e., reported p values will be smaller than they should be) and a loss of power for upper-tailed tests (i.e., reported p values will be too large).

Abnormal returns calculated using the control firm approach or buy-and-hold reference portfolios eliminate the new listing and rebalancing biases. Barber and Lyon (1997a) also document that the control firm approach eliminates the skewness bias. However, to eliminate the skewness bias when long-run abnormal returns are calculated using our buy-and-hold reference portfolios, we advocate the use of a bootstrapped skewness-adjusted t-statistic:

$$t_{sa} = \sqrt{n}\left(S + \frac{1}{3}\hat{\gamma}S^2 + \frac{1}{6n}\hat{\gamma}\right), \tag{5}$$

where

$$S = \frac{\overline{AR}_\tau}{\sigma(AR_\tau)}, \quad \text{and} \quad \hat{\gamma} = \frac{\sum_{i=1}^{n}(AR_{i\tau} - \overline{AR}_\tau)^3}{n\sigma(AR_\tau)^3}.$$

Note that $\hat{\gamma}$ is an estimate of the coefficient of skewness and $\sqrt{n}S$ is the conventional t-statistic of equation (4). This transformed test statistic, originally developed by Johnson (1978), is based on an Edgeworth Expansion and has been studied more recently by Hall (1992) and Sutton (1993). Sutton concludes that a bootstrapped application of Johnson's statistic "should be preferred to the t test when the parent distribution is asymmetrical, because it reduces the probability of type I error in cases where the t test has an inflated type I error rate and it is more powerful in other situations."

Though we evaluate the specification of the skewness-adjusted t-statistic, our results are consistent with Sutton's (1993) recommendation: Only the bootstrapped application of this skewness-adjusted test statistic yields well-specified test statistics. Bootstrapping the test statistic involves drawing b resamples of size n_b from the original sample. In general, the skewness-adjusted test statistic is calculated in each of these b bootstrapped resamples and the critical values for the transformed test statistic are calculated from the b values of the transformed statistic.

Specifically, the bootstrapping that we employ proceeds as follows: Draw 1,000 bootstrapped resamples from the original sample of size $n_b = n/4$.[8] In each resample, calculate the statistic:

$$t_{sa}{}^{b} = \sqrt{n_b}\left(S^b + \frac{1}{3}\hat{\gamma}^b S^{b2} + \frac{1}{6n_b}\hat{\gamma}^b\right), \tag{6}$$

---

[8] Our choice of $n_b = n/4$ is based on empirical analysis. The skewness adjustment results in more conservative test statistics as the size of the bootstrap resample decreases. Bootstrap resample sizes of $n/2$ also yield well-specified inferences, while bootstrap resample sizes of n do not. An analysis of resampling fewer than n observations can be found in Bickel, Gotze, and van Zwet (1997) and Shao (1996).

where

$$S^b = \frac{\overline{AR}^b_\tau - \overline{AR}_\tau}{\sigma^b(AR_\tau)}, \quad \text{and} \quad \hat{\gamma}^b = \frac{\sum\limits_{i=1}^{n_b} (AR_{i\tau}{}^b - \overline{AR}^b_\tau)^3}{n_b \sigma^b (AR_\tau)^3}.$$

Thus, ${t_{sa}}^b$, S^b, and $\hat{\gamma}^b$ are the bootstrapped resample analogues of t_{sa}, S, and $\hat{\gamma}$ from the original sample for the $b = 1,\ldots,1{,}000$ resamples. We reject the null hypothesis that the mean long-run abnormal return is zero if: $t_{sa} < x_l^*$ or $t_{sa} > x_u^*$. From the 1,000 resamples, we calculate the two critical values (x^*s) for the transformed test statistic (t_{sa}) required to reject the null hypothesis that the mean long-run abnormal return is zero at the α significance level by solving:

$$\Pr[{t_{sa}}^b \le x_l^*] = \Pr[{t_{sa}}^b \ge x_u^*] = \frac{\alpha}{2}.$$

C. Pseudoportfolios to Compute Empirical p value

The final method that we use to evaluate the statistical significance of long-run abnormal stock returns is a method employed by Brock et al. (1992), Ikenberry et al. (1995), Ikenberry, Rankine, and Stice (1996), Lee (1997), and Rau and Vermaelen (1996). In this approach, we generate the empirical distribution of long-run abnormal stock returns under the null hypothesis. Specifically, for each sample firm with an event month *t*, we randomly select with replacement a firm that is in the same size/book-to-market portfolio in event month *t*. This process continues until each firm in our original sample is represented by a control firm in this pseudoportfolio. This portfolio contains one randomly drawn firm for each sample firm, matched in time with similar size and book-to-market characteristics. After forming a single pseudoportfolio, we estimate long-run performance using the buy-and-hold size/book-to-market reference portfolios as was done for the original sample. This yields one observation of the abnormal performance obtained from randomly forming a portfolio with the same size and book-to-market characteristics as our original sample. This entire process is repeated until we have 1,000 pseudoportfolios, and thus 1,000 mean abnormal returns observations. These 1,000 mean abnormal return observations are used to approximate the empirical distribution of mean long-run abnormal returns.

Our results based on empirical *p* values are robust to abnormal returns calculated using the rebalanced reference portfolios. This is predictable because both the sample mean and pseudoportfolio sample means are calculated using the same reference portfolio. Nonetheless, we favor the use of the buy-and-hold reference portfolios, because the abnormal return measure more accurately reflects the excess return earned from investing in sample firms relative to an alternative executable trading strategy. Moreover, the population mean abnormal return is guaranteed to be zero when our buy-and-hold reference portfolios are employed, but the use of rebalanced reference portfolios does not guarantee a population mean abnormal return that is zero.

Unlike the conventional t-statistic or bootstrapped skewness-adjusted t-statistic, in which the null hypothesis is that the mean long-run abnormal return is zero, the null hypothesis tested by approximating the empirical distribution of mean long-run abnormal returns is that the mean long-run abnormal return equals the mean long-run abnormal return for the 1,000 pseudoportfolios. This hypothesis is rejected at the α significance level if: $\overline{AR}_\tau < y_l^*$ or $\overline{AR}_\tau > y_u^*$. The two y^* values are determined by solving

$$\Pr[\overline{AR}_\tau^p \leq y_l^*] = \Pr[\overline{AR}_\tau^p \geq y_u^*] = \frac{\alpha}{2},$$

where $\overline{AR}_\tau^p$ are the $p = 1,\ldots,1{,}000$ mean long-run abnormal returns generated from the pseudoportfolios.

In Table II we summarize the six statistical methods that we evaluate. The first two methods are the use of a conventional t-statistic when long-run abnormal returns are calculated using rebalanced size/book-to-market portfolios and a size/book-to-market matched control firm. Barber and Lyon (1997a) document that the former statistic is negatively biased in tests of long-run abnormal return, and the latter is well-specified. We report results for these two methods for purposes of comparison. The remaining four methods include: a conventional t-statistic, a skewness-adjusted t-statistic, a bootstrapped skewness-adjusted t-statistic, and empirical p values. All four methods rely on long-run abnormal returns calculated using our buy-and-hold reference portfolios.

D. Simulation Method

To test the specification of the test statistics based on our six statistical methods, we draw 1,000 random samples of n event months without replacement. For each of the 1,000 random samples, the test statistics are computed as described in Section III. If a test is well-specified, 1,000α tests reject the null hypothesis. A test is conservative if fewer than 1,000α null hypotheses are rejected and is anticonservative if more than 1,000α null hypotheses are rejected. Based on this procedure we test the specification of each test statistic at the 1 percent, 5 percent, and 10 percent theoretical levels of significance. A well-specified null hypothesis rejects the null at the theoretical rejection level in favor of the alternative hypothesis of negative (positive) abnormal returns in 1,000$\alpha/2$ samples. Thus, we separately document rejections of the null hypothesis in favor of the alternative hypothesis that long-run abnormal returns are positive or negative. For example, at the 1 percent theoretical significance level, we document the percentage of cal-

Table II
Summary of Statistical Methods

Method Description	Critical Values Based on	Statistic	Benchmark
Conventional t-statistic	Tabulated distribution of t-statistic	$t = \dfrac{\overline{AR_\tau}}{\sigma(AR_\tau)/\sqrt{n}}$	Rebalanced size/book-to-market portfolio
Conventional t-statistic	Tabulated distribution of t-statistic	$t = \dfrac{\overline{AR_\tau}}{\sigma(AR_\tau)/\sqrt{n}}$	Buy-and-hold size/book-to-market control firm
Conventional t-statistic	Tabulated distribution of t-statistic	$t = \dfrac{\overline{AR_\tau}}{\sigma(AR_\tau)/\sqrt{n}}$	Buy-and-hold size/book-to-market portfolios
Skewness-adjusted t-statistic	Tabulated distribution of t-statistic	$t_{sa} = \sqrt{n\left(S+\dfrac{1}{3}\,\hat{\gamma}S^2+\dfrac{1}{6n}\,\hat{\gamma}\right)}$	Buy-and-hold size/book-to-market portfolios
Bootstrapped skewness-adjusted t-statistic	Empirical distribution of t-statistic from bootstrapped resamples	$t^b_{sa} = \sqrt{n}_b\left(S^b + \dfrac{1}{3}\hat{\gamma}^b S^{b^2} + \dfrac{1}{6n_b}\,\hat{\gamma}^b\right)$	Buy-and-hold size/book-to-market portfolios
Empirical p value	Empirical distribution of sample means from pseudoportfolios	Not applicable	Buy-and-hold size/book-to-market portfolios

culated t-statistics that are less than the theoretical cumulative density function of the t-statistic at 0.5 percent and greater than the theoretical cumulative density function at 99.5 percent.

IV. Results

We begin with a discussion of the results in random samples, followed by a discussion of results in nonrandom samples. We consider nonrandom samples based on firm size, book-to-market ratio, pre-event return performance, calendar clustering of event dates, and industry clustering. We also discuss the impact of cross-sectional dependence on test statistics.

A. Random Samples

A.1. Specification

The first set of results is based on 1,000 random samples of 200 event months. The specification of the six statistical methods at one-, three-, and five-year horizons is presented in Table III. Two of our results are consistent with those reported in Barber and Lyon (1997a). First, t-statistics based on rebalanced size/book-to-market portfolios have a severe negative bias. Second, the t-statistics based on the buy-and-hold size/book-to-market reference portfolio are also negatively biased, though the magnitude of the bias is much less severe because the use of the buy-and-hold reference portfolios alleviates the new listing and rebalancing biases. The remaining negative bias can be attributed to the severe positive skewness of long-run abnormal returns. These two results indicate that the careful construction of our reference portfolios reduces much of the misspecification in test statistics. Left unresolved, however, is a means of controlling the skewness bias.

Though the skewness-adjusted t-statistic improves the specification of the test statistic, it too is negatively biased. However, when the critical values for rejection of the null hypothesis are determined using the bootstrapped procedure described in Section III, the misspecification is markedly reduced. Additionally, test statistics based on empirical p values derived from the distribution of mean long-run abnormal stock returns in pseudoportfolios also yield tests that are correctly specified in random samples. In sum, three of our six methods yield tests that are well-specified in random samples: a conventional t-statistic using size/book-to-market matched control firms, a bootstrapped skewness-adjusted test statistic using buy-and-hold size/book-to-market reference portfolios, and empirical p values derived from the distribution of mean long-run abnormal stock returns in pseudoportfolios.

The Central Limit Theorem guarantees that if the measures of abnormal returns in the cross section of firms are independent and identically distributed drawings from finite variance distributions, the distribution of the mean abnormal return measure converges to normality as the number of firms in the sample increases. Thus, we expect that the conventional t-statistic will

Table III

Specification (Size) of Alternative Test Statistics Using Buy-and-Hold Abnormal Returns (AR) in Random Samples

The numbers presented in this table represent the percentage of 1,000 random samples of 200 firms that reject the null hypothesis of no annual (Panel A), three-year (Panel B), and five-year (Panel C) buy-and-hold abnormal return (AR) at the theoretical significance levels of 1 percent, 5 percent, or 10 percent in favor of the alternative hypothesis of a significantly negative AR (i.e., calculated p value is less than 0.5 percent at the 1 percent significance level) or a significantly positive AR (calculated p value is greater than 99.5 percent at the 1 percent significance level). The alternative statistics and benchmarks are described in detail in the main text.

		Two-Tailed Theoretical Significance Level					
		1%		5%		10%	
		Theoretical Cumulative Density Function (%)					
Statistic	Benchmark	0.5	99.5	2.5	97.5	5.0	95.0
	Panel A: Annual ARs						
t-statistic	Rebalanced size/book-to-market portfolio	3.6*	0.0	10.1*	0.1	16.3*	0.7
t-statistic	Buy-and-hold size/book-to-market portfolio	1.2*	0.0	4.9*	0.9	8.2*	2.4
Skewness-adjusted t-statistic	Buy-and-hold size/book-to-market portfolio	1.0	0.3	3.7*	2.2	7.1*	4.7
t-statistic	Size/book-to-market control firm	0.3	0.0	2.3	2.0	5.3	4.1
Bootstrapped skewness-adjusted t-statistic	Buy-and-hold size/book-to-market portfolio	0.2	0.6	2.1	2.7	5.3	4.6
Empirical p value	Buy-and-hold size/book-to-market portfolio	0.4	0.9	2.6	2.5	5.0	4.8
	Panel B: Three-Year ARs						
t-statistic	Rebalanced size/book-to-market portfolio	10.6*	0.0	21.6*	0.0	30.5*	0.1
t-statistic	Buy-and-hold size/book-to-market portfolio	2.7*	0.0	6.8*	0.6	9.8*	2.6
Skewness-adjusted t-statistic	Buy-and-hold size/book-to-market portfolio	2.1*	0.3	5.3*	2.2	8.7*	5.2
t-statistic	Size/book-to-market control firm	0.3	0.2	3.2	1.2	5.5	3.0
Bootstrapped skewness-adjusted t-statistic	Buy-and-hold size/book-to-market portfolio	0.8	0.8	3.6	3.1	5.9	5.5
Empirical p value	Buy-and-hold size/book-to-market portfolio	0.8	0.9	3.4	3.1	6.8*	5.9
	Panel C: Five-Year ARs						
t-statistic	Rebalanced size/book-to-market portfolio	11.7*	0.0	23.7*	0.0	33.2*	0.2
t-statistic	Buy-and-hold size/book-to-market portfolio	2.4*	0.0	6.1*	0.5	10.5*	1.6
Skewness-adjusted t-statistic	Buy-and-hold size/book-to-market portfolio	1.4*	0.4	4.4*	1.7	8.2*	4.8
t-statistic	Size/book-to-market control firm	0.1	0.1	3.0	1.9	5.4	3.9
Bootstrapped skewness-adjusted t-statistic	Buy-and-hold size/book-to-market portfolio	0.6	1.2*	2.2	3.1	5.0	5.7
Empirical p value	Buy-and-hold size/book-to-market portfolio	0.2	1.5*	2.7	3.7*	4.9	6.3

*Significantly different from the theoretical significance level at the 1 percent level, one-sided binomial test statistic.

be well specified if the sample under consideration is sufficiently large. We consider this issue by calculating the empirical rejection rates in 1,000 simulations of sample sizes ranging from 200 to 4,000. As predicted for a positively skewed distribution, the negative bias in the conventional t-statistic declines with sample size and is well specified in samples of size 4,000. In contrast the three alternative methods are well specified in all of the sample sizes considered.

A.2. Power

To evaluate the power of the three methods that yield reasonably well-specified test statistics in random samples, we conduct the following experiment. For each sampled firm in our 1,000 simulations, we add a constant level of abnormal return to the calculated abnormal return. We document the empirical rejection rates at the 5 percent theoretical significance level of the null hypothesis that the mean sample long-run abnormal return is zero across 1,000 simulations at induced levels of abnormal returns ranging from -20 percent to $+20$ percent in increments of 5 percent. The results of this experiment are depicted graphically in Figure 1. As anticipated, the bootstrapped skewness-adjusted t-statistic and empirical p values both yield improved power in random samples relative to the control firm approach

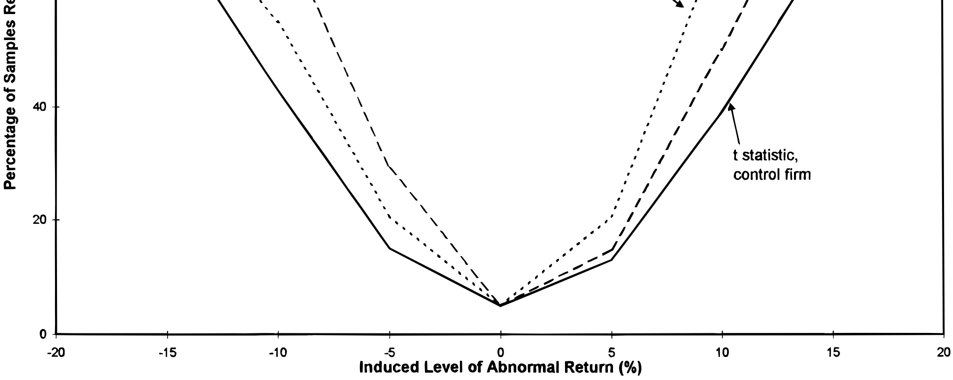

Figure 1. The power of test statistics in random samples. The percentage of 1,000 random samples of 200 firms rejecting the null hypothesis of no annual buy-and-hold abnormal return at various induced levels of abnormal return (horizontal axis) based on control firm method, bootstrapped skewness-adjusted t-statistic, and empirical p values.

advocated by Barber and Lyon (1997a). For example, in our 1,000 random samples of 200 firms, a +10 percent (-10 percent) abnormal return added to each of our sampled firms enables us to reject the null hypothesis in 43 percent (39 percent) when conventional *t*-statistics and the control firm method are used, 55 percent (70 percent) when the bootstrapped skewness-adjusted *t*-statistics and buy-and-hold reference portfolios are used, and 73 percent (50 percent) when empirical *p* values and buy-and-hold reference portfolios are used. The power advantage of the bootstrapped skewness-adjusted *t*-statistic and empirical *p* values is evident in all sampling situations that we analyze (differing sample size, horizons, and nonrandom samples).

B. Nonrandom Samples

Though our alternative methods are generally well specified and improve power for tests of long-run abnormal returns in random samples, we are also interested in how our proposed statistical methods work in nonrandom samples. Though we do not anticipate that our methods will work well in all sampling situations, a thorough understanding of the origin and magnitude of misspecification when sampling biases are present is important in developing a well-reasoned test statistic for a particular sample situation.

In general, to evaluate the impact of sampling biases we randomly draw 1,000 samples of 200 firms from a subset of our population. We then evaluate the empirical specification of four test statistics in these biased samples: (1) conventional *t*-statistic based on buy-and-hold reference portfolios, (2) conventional *t*-statistic based on size/book-to-market matched control firms, (3) bootstrapped skewness-adjusted *t*-statistic based on buy-and-hold reference portfolios, and (4) empirical *p* values based on buy-and-hold reference portfolios. We present results based on a conventional *t*-statistic in tables for purposes of illustration, though we know that it will be negatively biased in most sampling situations because of the severe positive skewness in the underlying distribution. In contrast, methods 2 through 4 are generally well-specified in random samples.

B.1. Firm Size

To assess the impact of size-based sampling biases on our statistical methods, we randomly draw 1,000 samples separately from the largest size decile (decile 10, Table I, Panel A) and smallest size decile (portfolios 1A to 1E in Table I, Panel A). The specifications of the four statistical methods in these samples are presented in Table IV. These results indicate that our three alternative methods are better specified than the conventional *t*-statistic based on buy-and-hold size/book-to-market reference portfolios.

B.2. Book-to-Market Ratio

To assess the impact of book-to-market based sampling biases on our statistical methods, we randomly draw 1,000 samples separately from the highest book-to-market decile (decile 10, Table I, Panel B) and smallest book-to-

Table IV

Specification (Size) of Alternative Test Statistics in Size-Based Samples

The numbers presented in this table represent the percentage of 1,000 random samples of 200 large (Panel A) or small (Panel B) firms that reject the null hypothesis of no one-, three-, and five-year buy-and-hold abnormal return (AR) at the theoretical significance level of 5 percent in favor of the alternative hypothesis of a significantly negative AR (i.e., calculated *p* value is less than 2.5 percent at the 5 percent significance level) or a significantly positive AR (calculated *p* value is greater than 97.5 percent at the 5 percent significance level). The statistics and benchmarks are described in detail in the main text.

		Horizon					
		1 Year		3 Years		5 Years	
		Theoretical Cumulative Density Function (%)					
Statistic	Benchmark	2.5	97.5	2.5	97.5	2.5	97.5
Panel A: Samples of Large Firms							
t-statistic	Buy-and-hold size/book-to-market portfolio	4.2*	1.2	3.8*	2.0	4.5*	1.6
t-statistic	Size/book-to-market control firm	3.9*	1.2	2.8	2.6	2.2	3.9*
Bootstrapped skewness-adjusted *t*-statistic	Buy-and-hold size/book-to-market portfolio	2.8	2.0	2.2	2.8	2.3	2.6
Empirical *p* value	Buy-and-hold size/book-to-market portfolio	3.5	3.6	2.5	3.3	3.6	2.1
Panel B: Samples of Small Firms							
t-statistic	Buy-and-hold size/book-to-market portfolio	5.7*	0.6	6.3*	0.3	8.0*	1.3
t-statistic	Size/book-to-market control firm	2.8	2.7	2.4	1.8	3.0	2.0
Bootstrapped skewness-adjusted *t*-statistic	Buy-and-hold size/book-to-market portfolio	2.5	2.8	2.4	2.8	2.8	3.3
Empirical *p* value	Buy-and-hold size/book-to-market portfolio	2.7	2.6	2.2	2.6	3.4	3.5

*Significantly different from the theoretical significance level at the 1 percent level, one-sided binomial test statistic.

market decile (decile 1, Table I, Panel B). The specifications of the four statistical methods in these samples are presented in Table V. Among firms with low book-to-market ratios, only the size/book-to-market matched control firm approach yields test statistics that are reasonably well specified.[9]

These results yield our first admonition regarding the use of reference portfolios to calculate long-run abnormal stock returns. Our methods assume that all firms constituting a particular portfolio have the same expected return. Recall that our reference portfolios are formed first by sorting firms into deciles on the basis of firm size and then by sorting these deciles into quintiles on the basis of book-to-market ratios. However, inspection of Table I, Panel B, reveals that there is a large difference between the returns on the bottom two book-to-market portfolios. Thus, when we sample from the lowest book-to-market *decile*, but (roughly) match these firms to the average return of firms from the lowest book-to-market *quintile*, the result is a negatively biased measure of mean abnormal returns and negatively biased test statistics.

This problem creates the negative bias in the methods that rely on the buy-and-hold reference portfolios (the bootstrapped skewness-adjusted t-statistic and empirical p values). However, the control firm approach alleviates this problem by matching sample firms reasonably well on book-to-market ratio. Reference portfolios formed based on a finer partition of book-to-market ratio could also alleviate the negative bias.

In contrast to the results in samples of firms with low book-to-market ratios, our three alternative methods yield reasonably well-specified test statistics among samples of firms with high book-to-market ratios. Unlike the bottom two deciles formed on the basis of book-to-market ratios, the top two deciles have similar rates of return.

B.3. Pre-Event Return Performance

A common characteristic of firms included in an event study is a period of unusually high or low stock returns preceding the event of interest. For example, equity issuance is generally preceded by a period of high stock returns, and share repurchases are generally preceded by a period of low stock returns. Moreover, Jegadeesh and Titman (1993) document persistence in stock returns, which Fama and French (1996) are unable to explain well using factors related to firm size and book-to-market ratio. Consequently, we anticipate an important factor to consider in long-run event studies to be the prior return performance of sample firms.

We calculate the preceding six-month buy-and-hold return on all firms in each month from July 1973 through December 1994. We then rank all firms on the basis of this six-month return and form deciles on the basis of the

---

[9] The negative bias at the five-year horizon is likely a result of random sampling variation. We do not observe this result at five-year horizons and the 1 percent and 10 percent theoretical significance levels.

Table V

Specification (Size) of Alternative Test Statistics in Book-to-Market Based Samples

The numbers presented in this table represent the percentage of 1,000 random samples of 200 firms with low (Panel A) or high (Panel B) book-to-market ratios that reject the null hypothesis of no one-, three-, and five-year buy-and-hold abnormal return (AR) at the theoretical significance level of 5 percent in favor of the alternative hypothesis of a significantly negative AR (i.e., calculated p value is less than 2.5 percent at the 5 percent significance level) or a significantly positive AR (calculated p value is greater than 97.5 percent at the 5 percent significance level). The statistics and benchmarks are described in detail in the main text.

		Horizon					
		1 Year		3 Years		5 Years	
		Theoretical Cumulative Density Function (%)					
Statistic	Benchmark	2.5	97.5	2.5	97.5	2.5	97.5
Panel A: Samples of Firms with Low Book-to-Market Ratios							
t-statistic	Buy-and-hold size/book-to-market portfolio	7.7*	0.4	9.1*	0.2	11.2*	0.1
t-statistic	Size/book-to-market control firm	2.8	1.5	3.1	1.6	3.8*	2.3
Bootstrapped skewness-adjusted t-statistic	Buy-and-hold size/book-to-market portfolio	3.9*	2.7	3.8*	1.8	4.0*	1.2
Empirical p value	Buy-and-hold size/book-to-market portfolio	4.8*	2.1	5.9*	2.6	6.7*	1.3
Panel B: Samples of Firms with High Book-to-Market Ratios							
t-statistic	Buy-and-hold size/book-to-market portfolio	5.8*	0.3	5.0*	0.4	5.7*	0.9
t-statistic	Size/book-to-market control firm	2.5	3.2	1.5	4.1*	1.7	2.1
Bootstrapped skewness-adjusted t-statistic	Buy-and-hold size/book-to-market portfolio	3.0	2.1	3.2	2.1	2.5	2.7
Empirical p value	Buy-and-hold size/book-to-market portfolio	3.6	1.9	2.8	2.0	2.9	2.6

*Significantly different from the theoretical significance level at the 1 percent level, one-sided binomial test statistic.

rankings in each month. Finally, we separately draw 1,000 samples of 200 firms from the high-return decile and the low-return decile.

The results of this analysis are presented in Table VI. For firms with high six-month pre-event returns, all test statistics are positively biased at an annual horizon, but negatively biased at a three- or five-year horizon. The annual results are consistent with those of Jegadeesh and Titman (1993), who document return persistence for up to one year followed by significant reversals. These reversals are sufficiently large to render test statistics at three- and five-year horizons negatively biased. For firms with low six-month pre-event returns, all test statistics are negatively biased at an annual horizon. The magnitude of the negative bias declines at a three-year horizon and is virtually nonexistent at a five-year horizon.

These results indicate that an important dimension to consider, one that is commonly overlooked in tests of long-run abnormal returns, is the pre-event return performance. Though we do not analyze reference portfolios formed on the basis of prior return performance, the methods used to construct our size/book-to-market portfolios could be applied to create portfolios formed on the basis of recent return performance. Similarly, we anticipate that matching sample firms to firms of similar pre-event return performance would also control well for the misspecification documented here. Lee (1997) and Ikenberry et al. (1996) employ sensible variations of the empirical p-value approach which control for pre-event returns.

B.4. Industry Clustering

Assume that expected returns vary by industry and we have an imprecise asset pricing model. We would reject the null hypothesis of zero long-run abnormal returns in favor of both positive and negative long-run abnormal returns too often. We assess the impact of industry clustering of sample firms by drawing 1,000 samples of 200 firms such that each of the 200 firms in a particular sample has the same two-digit SIC code. This sampling procedure involves two steps. First, we randomly select a two-digit SIC code; second, we draw a sample of 200 firms from this two-digit SIC code. If the two-digit SIC code contains fewer than 200 firms over our period of analysis, we complete the sample with firms from a second, randomly selected, two-digit SIC code.[10] The results of this analysis are presented in Table VII.

---

[10] The same firm is allowed to appear more than once within a particular sample of 200 firms. However, multiple occurrences of the same firm are not allowed to have overlapping periods for the calculation of returns. For example, if an event date of January 1981 were selected for IBM in our analysis of annual abnormal returns, subsequent randomly selected observations would preclude the sampling of IBM from February 1980 through November 1982. Thus, our analysis isolates the impact of sampling from one industry rather than the impact of overlapping returns, which we discuss in the next section.

Table VI

Specification (Size) of Alternative Test Statistics in Samples Based on Pre-Event Return Performance

The numbers presented in this table represent the percentage of 1,000 random samples of 200 firms with high six-month pre-event returns (Panel A) or low six-month pre-event returns (Panel B) that reject the null hypothesis of no one-, three-, and five-year buy-and-hold abnormal return (AR) at the theoretical significance level of 5 percent in favor of the alternative hypothesis of a significantly negative AR (i.e., calculated p value is less than 2.5 percent at the 5 percent significance level) or a significantly positive AR (calculated p value is greater than 97.5 percent at the 5 percent significance level). The statistics and benchmarks are described in detail in the main text.

		Horizon					
		1 Year		3 Years		5 Years	
		Theoretical Cumulative Density Function (%)					
Statistic	Benchmark	2.5	97.5	2.5	97.5	2.5	97.5
Panel A: Samples of Firms with High Six-Month Pre-Event Returns							
t-statistic	Buy-and-hold size/book-to-market portfolio	1.1	3.8*	14.4*	0.1	20.8*	0.1
t-statistic	Size/book-to-market control firm	0.4	6.3*	4.5*	1.3	7.3*	0.6
Bootstrapped skewness-adjusted t-statistic	Buy-and-hold size/book-to-market port.	0.4	6.8*	7.4*	0.4	9.1*	0.4
Empirical p value	Buy-and-hold size/book-to-market portfolio	0.4	5.2*	10.4*	0.6	14.8*	0.3
Panel B: Samples of Firms with Low Six-Month Pre-Event Returns							
t-statistic	Buy-and-hold size/book-to-market portfolio	21.0*	0.1	8.5*	0.3	5.2*	0.4
t-statistic	Size/book-to-market control firm	11.0*	0.3	5.0*	0.8	3.1	1.2
Bootstrapped skewness-adjusted t-statistic	Buy-and-hold size/book-to-market portfolio	10.3*	0.9	3.0	2.6	1.6	3.0
Empirical p value	Buy-and-hold size/book-to-market portfolio	22.1*	2.3	6.2*	4.3*	3.4	3.8*

*Significantly different from the theoretical significance level at the 1 percent level, one-sided binomial test statistic.

Table VII

Specification (Size) of Alternative Test Statistics in Samples with Industry Clustering

The numbers presented in this table represent the percentage of 1,000 random samples of 200 firms with a common two-digit SIC code that reject the null hypothesis of no one-, three-, and five-year buy-and-hold abnormal return (AR) at the theoretical significance level of 5 percent in favor of the alternative hypothesis of a significantly negative AR (i.e., calculated p value is less than 2.5 percent at the 5 percent significance level) or a significantly positive AR (calculated p value is greater than 97.5 percent at the 5 percent significance level). If 200 firms cannot be drawn from the same two-digit SIC code, a second two-digit SIC code is used to fill the sample of 200 firms. The statistics and benchmarks are described in detail in the main text.

		Horizon					
		1 Year		3 Years		5 Years	
		Theoretical Cumulative Density Function (%)					
Statistic	Benchmark	2.5	97.5	2.5	97.5	2.5	97.5
t-statistic	Buy-and-hold size/book-to-market portfolio	9.5*	3.3	13.2*	8.2*	20.4*	11.0*
t-statistic	Size/book-to-market control firm	2.8	4.7*	6.1*	7.6*	7.8*	10.6*
Bootstrapped skewness-adjusted t-statistic	Buy-and-hold size/book-to-market portfolio	6.4*	6.0*	7.7*	10.7*	10.5*	15.9*
Empirical p value	Buy-and-hold size/book-to-market portfolio	6.9*	5.1*	9.4*	9.4*	14.2*	12.9*

*Significantly different from the theoretical significance level at the 1 percent level, one-sided binomial test statistic.

Our analysis reveals that controlling for size and book-to-market alone is not sufficient when samples are drawn from a single two-digit SIC code. All of the empirical methods analyzed yield empirical rejection levels that exceed theoretical rejection levels. Auxiliary analyses (not reported in a table) reveal that this misspecification disappears when samples are evenly distributed among four or more two-digit SIC codes. Thus, only extreme industry clustering leads to the misspecification that we document.

C. Cross-Sectional Dependence of Sample Observations

Brav (1997) argues that cross-sectional dependence in sample observations can lead to misspecified test statistics. Cross-sectional dependence inflates test statistics because the number of sample firms overstates the number of independent observations. To assess the magnitude of the problem of cross-sectional dependence we consider two extreme sample situations: (1) calendar clustering and (2) overlapping return calculations.

C.1. Calendar Clustering

We make the reasonable assumption that the contemporaneous returns of firms are more likely to be cross-sectionally related than returns from different periods. If true, the problem of cross-sectional dependence will be most severe when all sample firms share the same event date. We assess the impact of calendar clustering of event dates by drawing 1,000 samples of 200 firms such that all of the 200 firms in a particular sample have the same event date. The results of this analysis, presented in Table VIII, indicate that the three alternative methods we analyze control well for calendar clustering of event dates, though there is some evidence of a positive bias when the size/book-to-market matched control firm method is used at an annual horizon.

Auxiliary analyses (not reported in a table) indicate that when samples are composed of small firms, large firms, low book-to-market firms, or high book-to-market firms the three alternative methods yield test statistics with rejection levels similar to those reported in Table IV and Table V, even when all sample firms share a common event date. When samples are composed of firms from a single industry or of firms with unusually high or low pre-event returns, the empirical rejection rates of the three alternative methods are approximately 50 percent greater than those reported in Table VI and Table VII. In sum, the three alternative methods that we analyze control reasonably well for cross-sectional dependence that arises from the relation between firm size, book-to-market ratios, and returns. However, when samples are calendar clustered and exhibit unusual pre-event return performance or industry clustering, these methods yield misspecified test statistics. This misspecification can be attributed to cross-sectional dependence and/or the use of a poor asset pricing model—an issue that we discuss in more detail in the conclusion.

Table VIII

Specification (Size) of Alternative Test Statistics in Samples with a Common Event Month (Calendar Clustering)

The numbers presented in this table represent the percentage of 1,000 random samples of 200 firms with a common event month that reject the null hypothesis of no one-, three-, and five-year buy-and-hold abnormal return (AR) at the theoretical significance level of 5 percent in favor of the alternative hypothesis of a significantly negative AR (i.e., calculated p value is less than 2.5 percent at the 5 percent significance level) or a significantly positive AR (calculated p value is greater than 97.5 percent at the 5 percent significance level). The statistics and benchmarks are described in detail in the main text.

		Horizon					
		1 Year		3 Years		5 Years	
		Theoretical Cumulative Density Function (%)					
Statistic	Benchmark	2.5	97.5	2.5	97.5	2.5	97.5
t-statistic	Buy-and-hold size/book-to-market portfolio	4.5*	1.3	3.8*	0.9	7.2*	0.6
t-statistic	Size/book-to-market control firm	2.8	5.9*	2.5	1.9	2.0	1.4
Bootstrapped skewness-adjusted *t*-statistic	Buy-and-hold size/book-to-market portfolio	3.0	3.4	1.8	2.7	3.1	2.5
Empirical *p* value	Buy-and-hold size/book-to-market portfolio	3.1	3.3	1.9	2.2	2.8	2.8

*Significantly different from the theoretical significance level at the 1 percent level, one-sided binomial test statistic.

C.2. Overlapping Return Calculations

A common problem in event studies that analyze long-run abnormal returns is overlapping periods of return calculation for the same firm—for example, Microsoft's common stock split in April 1990, June 1991, and June 1992. Clearly, the three- or five-year returns calculated relative to each of these event months are not independent because these returns share several months of overlapping returns. This is the most severe form of cross-sectional dependence that a researcher could encounter in an event study of long-run abnormal stock returns.

To assess the impact of overlapping return calculations on tests of long-run abnormal stock returns, we proceed in two steps. First, we randomly sample 100 firms from our population. Second, for these *same* 100 firms, we randomly select a second event month that is within $\tau - 1$ periods of the original event month (either before or after), where τ is the period over which returns are compounded (12, 36, or 60 months). Combining these two sets of 100 observations yields one sample of 200 event months for the 100 firms. We repeat this procedure 1,000 times. The results of this analysis are reported in Table IX.

As anticipated, the lack of independence generated by overlapping returns yields misspecified test statistics. Note that these samples are random with respect to firm size, book-to-market ratio, pre-event return performance, and industry. Thus, this experiment best demonstrates the problem of cross-sectional dependence discussed in Brav (1997), since it is unlikely that a poor model of asset pricing is leading to overrejection in these samples. Cowan and Sergeant (1996) also document that overlapping return calculations yield misspecified test statistics in random samples when the horizon of analysis is long (three or five years) and sample sizes are large, since the probability of samples containing overlapping return calculations in these sampling situations is large. The only ready solution to this source of bias in event studies of long-run buy-and-hold abnormal stock returns is to purge the sample of observations of overlapping returns. For example, in their analysis of seasoned equity offerings, Loughran and Ritter (1995) and Speiss and Affleck-Graves (1995) require sample firms to have a five-year pre-event period with no equity issuance. Speiss and Affleck-Graves (1996) impose an analogous requirement in their analysis of debt issuance. This pre-event screen would solve the overlapping return problem that we document here.

C.3. Adjustments to the Variance-Covariance Matrix

Conceptually, a researcher could adjust test statistics for cross-sectional dependence if for a sample of n firms an appropriate $\boldsymbol{n} \times \boldsymbol{n}$ variance-covariance matrix could be estimated. We investigated several possible methods for estimating the variance-covariance matrix of event-time abnormal returns, which we discuss in the Appendix. In short, though these methods reduce the misspecification in samples with overlapping return calculations, they do not eliminate the problem of cross-sectional dependence.

Table IX

Specification (Size) of Alternative Test Statistics in Samples with Overlapping Returns

The analysis is this table is based on 1,000 random samples of 200 event month for 100 firms. The sampling is conducted in two steps. First, 100 event months are randomly selected. For each of these 100 event months, a second event month is randomly selected within τ periods of the original event month (either before or after) to guarantee the presence of overlapping returns. The numbers presented in this table represent the percentage of the 1,000 random samples that reject the null hypothesis of no one-, three-, and five-year buy-and-hold abnormal return (AR) at the theoretical significance level of 5 percent in favor of the alternative hypothesis of a significantly negative AR (i.e., calculated p value is less than 2.5 percent at the 5 percent significance level) or a significantly positive AR (calculated p value is greater than 97.5 percent at the 5 percent significance level). The statistics and benchmarks are described in detail in the main text.

		Horizon					
		1 Year		3 Years		5 Years	
		Theoretical Cumulative Density Function (%)					
Statistic	Benchmark	2.5	97.5	2.5	97.5	2.5	97.5
t-statistic	Buy-and-hold size/book-to-market portfolio	8.8*	3.4	9.5*	2.4	10.6*	1.9
t-statistic	Size/book-to-market control firm	4.5*	5.2*	4.5*	3.4	4.3*	4.1*
Bootstrapped skewness-adjusted t-statistic	Buy-and-hold size/book-to-market portfolio	5.7*	6.3*	5.5*	5.2*	6.1*	6.7*
Empirical p value	Buy-and-hold size/book-to-market portfolio	6.5*	5.1*	6.1*	5.3*	5.1*	4.8*

*Significantly different from the theoretical significance level at the 1 percent level, one-sided binomial test statistic.

V. The Use of Cumulative Abnormal Returns and Calendar-Time Portfolio Methods

The analysis to this point has focused on buy-and-hold abnormal returns. The analysis of buy-and-hold abnormal returns is warranted if a researcher is interested in answering the question of whether sample firms earned abnormal stock returns over a particular horizon of analysis. A related, but slightly different question, can be answered by analyzing cumulative abnormal returns or mean monthly abnormal returns over a long-horizon (or, equivalently, mean monthly returns): Do sample firms persistently earn abnormal monthly returns? Though correlated, cumulative abnormal returns are a biased predictor of buy-and-hold abnormal returns (Barber and Lyon (1997a)). Nonetheless, cumulative or mean monthly abnormal returns might be employed because they are less skewed and therefore less problematic statistically.

A. Cumulative Abnormal Returns

We reestimate many of our results using cumulative abnormal returns, but in the interest of parsimony summarize the major results of the analysis here. First, in random samples all of the methods that yield well-specified test statistics for buy-and-hold abnormal returns also yield well-specified test statistics for cumulative abnormal returns: the control firm approach, the bootstrapped skewness-adjusted t-statistic, and the pseudoportfolio approach.

Second, since cumulative abnormal returns are less skewed than buy-and-hold abnormal returns, conventional t-statistics also yield well-specified test statistics. It is important to note that conventional t-statistics are only well specified when reference portfolios are purged of the new listing or survivor bias. Thus, the τ-period cumulative abnormal return ($CAR_{i\tau}$) for firm i beginning in period s is calculated as:

$$CAR_{i\tau} = \sum_{t=s}^{s+\tau}\left[R_{it} - \frac{1}{n_t^s}\sum_{j=1}^{n_t^s} R_{jt}\right], \tag{7}$$

where R_{it} is the simple monthly return for sample firm i, R_{jt} is the simple monthly return for the $j = 1,\ldots,n_t^s$ firms that are in the same size/book-to-market reference portfolio as firm i, which are also publicly traded in *both* period s and t. Cumulative abnormal returns based on conventional rebalanced size/book-to-market portfolios yield positively biased test statistics (Kothari and Warner (1997), Barber and Lyon (1997a)).

Third, cumulative abnormal returns are affected by sampling biases (size, book-to-market, pre-event returns, calendar clustering, industry clustering, and overlapping returns) in an analogous fashion to buy-and-hold abnormal returns.

B. Calendar-Time Portfolio Methods

Loughran and Ritter (1995), Brav and Gompers (1997), and Brav et al. (1995) employ the Fama–French three-factor model to analyze returns on calendar-time portfolios of firms that issue equity. Jaffe (1974) and Mandelker (1974) use variations of this calendar-time portfolio method. We consider two variations of calendar-time portfolio methods: one based on the use of the three-factor model developed by Fama and French (1993) and one based on the use of mean monthly calendar-time abnormal returns.

The calendar-time portfolio methods offer some advantages over tests that employ either cumulative or buy-and-hold abnormal returns. First, this approach eliminates the problem of cross-sectional dependence among sample firms because the returns on sample firms are aggregated into a single portfolio. Second, the calendar-time portfolio methods yield more robust test statistics in nonrandom samples. Nonetheless, in nonrandom samples, the calendar-time portfolio methods often yield misspecified test statistics.

B.1. The Fama–French Three-Factor Model and Calendar-Time Portfolios

Assume the event period of interest is five years. For each calendar month, calculate the return on a portfolio composed of firms that had an event (e.g., issued equity) within the last five years of the calendar month. The calendar-time return on this portfolio is used to estimate the following regression:

$$R_{pt} - R_{ft} = \alpha_i + \beta_i (R_{mt} - R_{ft}) + s_i SMB_t + h_i HML_t + \epsilon_{it}, \tag{8}$$

where R_{pt} is the simple monthly return on the calendar-time portfolio (either equally weighted or value-weighted), R_{ft} is the monthly return on three-month Treasury bills, R_{mt} is the return on a value-weighted market index, SMB_t is the difference in the returns of value-weighted portfolios of small stocks and big stocks, HML_t is the difference in the returns of value-weighted portfolios of high book-to-market stocks and low book-to-market stocks.[11] The regression yields parameter estimates of α_i, β_i, s_i, and h_i. The error term in the regression is denoted by ϵ_{it}. The estimate of the intercept term (α_i) provides a test of the null hypothesis that the mean monthly excess return on the calendar-time portfolio is zero.[12]

---

[11] The construction of these factors is discussed in detail in Fama and French (1993). We thank Kenneth French for providing us with these data.

[12] The error term in this regression may be heteroskedastic, since the number of securities in the calendar-time portfolio varies from one month to the next. We find that this heteroskedasticity does not significantly affect the specification of the intercept test in random samples. However, a correction for heteroskedasticity can be performed using weighted least squares estimation, where the weighting factor is based on the number of securities in the portfolio in each calendar month.

We evaluate the empirical specification of the calendar-time portfolio approach by randomly drawing 1,000 samples of 200 event months. For each sample, we estimate an equally weighted and value-weighted (by market capitalization) calendar-time portfolio return, which assumes sample firms are held in the portfolio for either 12, 36, or 60 months after the randomly selected event month. The number of firms in the calendar-time portfolio varies from month to month. If in a particular calendar month there are no firms in the portfolio, that month is dropped when estimating equation (8). We also analyze the specification of the calendar-time portfolio approach in nonrandom samples. The results of these analyses are presented in Table X.

The calendar-time portfolio methods are well specified in random samples. However, the calendar-time portfolio methods are generally misspecified in nonrandom samples. For example, the calendar-time portfolio method does not perform as well as test statistics based on reference portfolios in samples with size-based or book-to-market-based biases. On the other hand, the calendar-time portfolio method performs well when cross-sectional dependence is severe (e.g., when return calculations are overlapping). Note also that the calendar-time portfolio approach reduces, but does not eliminate, misspecification when samples are drawn from a single industry. This result indicates that the misspecification that results from industry clustering is at least partially attributable to the bad model problem. In sum, the calendar-time portfolio method controls well for the problem of cross-sectional dependence, but remains sensitive to the bad model problem.

B.2. Mean Monthly Calendar-Time Abnormal Returns

Assume the event period of interest is five years. For each calendar month, calculate the abnormal return (AR_{it}) for each security using the returns on the seventy size/book-to-market reference portfolios (R_{pt}):

$$AR_{it} = R_{it} - R_{pt}.$$

In each calendar month t, calculate a mean abnormal return (MAR_t) across firms in the portfolio:

$$MAR_t = \sum_{i=1}^{n_t} x_{it} AR_{it},$$

where n_t is the number of firms in the portfolio in month t. The weight, x_{it}, is $1/n_t$ when abnormal returns are equally weighted and $MV_{it}/\Sigma MV_{it}$ when abnormal returns are value-weighted. A grand mean monthly abnormal returns ($MMAR$) is calculated:

$$MMAR = \frac{1}{T} \sum_{t=1}^{T} MAR_t,$$

Table X

Specification (Size) of Intercept Tests from Regressions of Monthly Calendar-Time Portfolio Returns on Market, Size, and Book-to-Market Factors

The analysis is this table is based on 1,000 random samples of 200 event months. Each of the 200 securities is included in the calendar-time portfolio for 12, 36, or 60 months following the randomly selected event month. The following regression is estimated:

$$R_{pt} - R_{ft} = \alpha_i + \beta_i(R_{mt} - R_{ft}) + s_i SMB_t + h_i HML_t + \epsilon_{it},$$

where R_{pt} is the simple return on the calendar-time portfolio (either equally weighted or value-weighted, R_{ft} is the return on three-month Treasury bills, R_{mt} is the return on a value-weighted market index, SMB_t is the difference in the returns of a value-weighted portfolio of small stocks and big stocks, HML_t is the difference in the returns of a value-weighted portfolio of high book-to-market stocks and low book-to-market stocks. The estimate of the intercept term (α_i) provides a test of the null hypothesis that the mean monthly excess return on the calendar-time portfolio is zero. The numbers presented in the first row of each panel represent the percentage of the 1,000 random samples that reject the null hypothesis of no mean monthly excess return when the holding period is one, three, or five years at the theoretical significance level of 5 percent in favor of the alternative hypothesis of a significantly negative intercept (i.e., calculated p value is less than 2.5 percent at the 5 percent significance level) or a significantly positive intercept (calculated p value is greater than 97.5 percent at the 5 percent significance level). The remaining rows of each panel represent the rejection levels in nonrandom samples, which are described in detail in Section IV.B and IV.C and Tables IV through Table IX.

			Holding Period			
	12 Months		36 Months		60 Months	
			Theoretical Cumulative Density Function (%)			
Sample Characteristics	2.5	97.5	2.5	97.5	2.5	97.5
Panel A: Equally-Weighted Calendar-Time Portfolios						
Random samples	2.1	1.7	1.5	2.1	0.9	1.8
Small firms	1.4	2.4	0.5	2.1	0.2	3.1
Large firms	1.9	2.0	1.3	3.2	0.7	2.9
Low book-to-market ratio	22.8*	0.0	22.5*	0.0	16.8*	0.0
High book-to-market ratio	0.0	7.2*	0.0	17.1*	0.0	15.0*
Poor pre-event returns	5.2*	0.0	1.2	0.2	1.2	0.2
Good pre-event returns	0.5	5.1*	2.2	1.0	1.0	1.0
Industry clustering	3.3	3.6	3.5	6.0*	3.9*	7.5*
Overlapping returns	2.3	1.4	1.3	1.0	0.2	2.2
Calendar clustering	5.6*	3.4	3.7*	6.7*	3.1	7.8*
Panel B: Value-Weighted Calendar-Time Portfolios						
Random samples	2.7	1.3	2.6	1.1	3.3	1.2
Small firms	5.3*	0.9	3.8*	0.7	2.9	1.0
Large firms	2.1	1.9	2.1	3.3	1.7	2.7
Low book-to-market ratio	10.8*	0.1	9.6*	0.2	9.9*	0.4
High book-to-market ratio	1.9	1.6	0.9	4.6*	1.6	4.1*
Poor pre-event returns	24.9*	0.0	12.0*	0.1	10.1*	0.2
Good pre-event returns	1.0	4.3*	2.9	1.7	4.0*	1.3
Industry clustering	5.0*	2.2	4.4*	2.7	4.9*	3.9*
Overlapping returns	3.7*	1.2	2.8	2.3	3.6	1.4
Calendar clustering	2.8	3.2	3.2	3.4	2.0	3.2

*Significantly different from the theoretical significance level at the 1 percent level, one-sided binomial test statistic.

Table XI

Specification (Size) of Monthly Calendar-Time Portfolio Abnormal Returns

The analysis is this table is based on 1,000 random samples of 200 event months. Each of the 200 securities is included in the calendar-time portfolio for 12, 36, or 60 months following the randomly selected event month. Monthly abnormal returns (AR_{it}) for each security are calculated using the returns on the seventy size/book-to-market reference portfolios (R_{pt}): $AR_{it} = R_{it} - R_{pt}$. In each calendar month, a mean abnormal return of firms in the portfolio is calculated as $MAR_t = \sum_{i=1}^{n_t} x_{it} AR_{it}$. For the analysis of equally weighted abnormal returns, $x_{it} = 1/n_t$; for the value-weighted analysis, $x_{it} = MV_{it}/\sum_{i=1}^{n_t} MV_{it}$. A grand mean monthly abnormal return is calculated as MMAR $= (1/T)\sum_{t=1}^{T} MAR_t$. To test the null hypothesis of zero mean monthly abnormal returns, a t-statistic is calculated using the time-series standard deviation of the mean monthly abnormal returns: $t(MMAR) = MMAR/[\sigma(MAR_t)/\sqrt{T}]$. The numbers presented in the first row of each panel represent the percentage of the 1,000 random samples that reject the null hypothesis of no mean monthly excess return when the holding period is one, three, or five years at the theoretical significance level of 5 percent in favor of the alternative hypothesis of a significantly negative intercept (i.e., calculated p value is less than 2.5 percent at the 5 percent significance level) or a significantly positive intercept (calculated p value is greater than 97.5 percent at the 5 percent significance level). The remaining rows of each panel represent the rejection levels in nonrandom samples, which are described in detail in Section IV.B and IV.C and Table IV through Table IX.

	Holding Period					
	12 Months		36 Months		60 Months	
	Theoretical Cumulative Density Function (%)					
Sample Characteristics	2.5	97.5	2.5	97.5	2.5	97.5
Panel A: Equally-Weighted Calendar-Time Portfolios						
Random samples	2.2	1.8	0.9	1.6	1.3	2.3
Small firms	1.9	1.8	1.1	1.3	1.1	1.9
Large firms	1.6	1.4	0.7	2.8	0.0	2.1
Low book-to-market ratio	4.0*	1.1	2.4	0.6	1.0	1.2
High book-to-market ratio	1.5	1.2	1.1	1.3	1.9	0.4
Poor pre-event returns	2.7	0.7	0.3	1.2	0.5	2.2
Good pre-event returns	0.6	2.8	2.3	1.0	1.6	0.9
Industry clustering	2.6	3.9*	3.0	5.2*	2.8	6.9*
Overlapping returns	1.9	1.3	1.8	0.9	1.1	1.4
Calendar clustering	2.5	2.2	0.9	2.4	1.1	3.5
Panel B: Value-Weighted Calendar-Time Portfolios						
Random samples	2.4	1.2	3.3	1.5	3.4	1.0
Small firms	3.4	0.9	4.5*	0.7	2.2	1.4
Large firms	2.4	1.8	2.1	1.1	1.1	0.8
Low book-to-market ratio	4.9*	0.8	4.4*	1.0	4.0*	1.8
High book-to-market ratio	3.4	0.7	2.6	1.2	2.8	0.7
Poor pre-event returns	10.3*	0.2	5.3*	0.3	4.5*	0.7
Good pre-event returns	1.0	4.0*	3.9*	1.1	4.7*	0.9
Industry clustering	3.3	1.7	2.6	1.9	1.9	2.1
Overlapping returns	2.5	0.9	2.9	1.0	3.0	1.2
Calendar clustering	3.7*	2.7	2.3	1.6	2.7	2.0

*Significantly different from the theoretical significance level at the 1 percent level, one-sided binomial test statistic.

where T is the total number of calendar months. To test the null hypothesis of zero mean monthly abnormal returns, a t-statistic is calculated using the time-series standard deviation of the mean monthly abnormal returns:

$$t(MMAR) = \frac{MMAR}{\sigma(MAR_t)/\sqrt{T}}.$$

Our analysis of the empirical specification of the calendar-time portfolio methods based on reference portfolios is reported in Table XI. When contrasted with the results based on the Fama–French three-factor model (Table X), the empirical rejection levels for test statistics based on calendar-time abnormal returns calculated using reference portfolios are generally more conservative. For example, in samples of firms with extreme book-to-market or pre-event return characteristics, the empirical rejection levels reported in Table XI (calendar-time abnormal returns based on reference portfolios) are uniformly lower than those based on the Fama–French three-factor model. Though the issues involved are complex, we suspect that the calendar-time portfolio methods based on reference-portfolio abnormal returns generally dominate those based on the Fama–French three-factor model for two reasons. First, the latter implicitly assumes linearity in the constructed market, size, and book-to-market factors. Inspection of Table I reveals that this assumption is unlikely to be the case for the size and book-to-market factors, at least during our period of analysis. Second, the Fama–French three-factor model assumes there is no interaction between the three factors. During our period of analysis, this assumption is also likely violated because the relation between book-to-market ratio and returns is most pronounced for small firms (Loughran (1997)).[13]

VI. Conclusion

In this research, we analyze various methods to test for long-run abnormal stock returns. Generally, misspecification of test statistics can be traced to (1) the new listing bias, (2) the rebalancing bias, (3) the skewness bias, (4) cross-sectional dependence, and/or (5) a bad model of asset pricing. How and whether these factors affect the misspecification of test statistics depend on the methods used to calculate abnormal returns.

To recommend a particular approach to test for long-run abnormal returns, we consider it a necessary condition that the method yield well-specified test statistics in random samples. Ultimately, we identify two general approaches that satisfy this condition. Though both offer advantages and disadvantages, a pragmatic solution for a researcher who is analyzing long-run abnormal returns would be to use both.

---

[13] There is a tendency for negative rejection when the abnormal returns are value-weighted (Panel B, Table XI). We suspect that this results from the fact that the benchmark portfolios are equally weighted, giving more weight to the relatively high returns of small firms. Thus, large firms, which receive more weight in the calculation of a value-weighted mean, likely have negative abnormal returns.

The first approach relies on a traditional event study framework and the calculation of buy-and-hold abnormal returns using a carefully constructed reference portfolio, such that the population mean abnormal return is guaranteed to be zero. Inference is based on either a bootstrapped skewness-adjusted *t*-statistic or the empirically generated distribution of mean long-run abnormal stock returns from pseudoportfolios. The advantage of this approach is that it yields an abnormal return measure that accurately represents investor experience. The disadvantage of this approach is that it is more sensitive to the problem of cross-sectional dependence among sample firms and a poorly specified asset pricing model.

The second method relies on the calculation of calendar-time portfolio abnormal returns (either equally weighted or value-weighted). The advantage of this approach is that it controls well for cross-sectional dependence among sample firms and is generally less sensitive to a poorly specified asset pricing model. The disadvantage of this approach is that it yields an abnormal return measure that does not precisely measure investor experience.

Even these two methods, which yield well-specified test statistics in random samples, often yield misspecified test statistics in nonrandom samples (e.g., in samples with unusual pre-event returns or samples concentrated in one industry). Though we are able to control for many sources of misspecification, ultimately, the misspecification that remains can be attributed to the inability of firm size and book-to-market ratio to capture all of the misspecifications of the Capital Asset Pricing Model. Though firm size and book-to-market ratio have received considerable attention in the recent research in financial economics, some would argue that other variables explain the cross section of stock returns (e.g., recent return performance, recent quarterly earnings surprises, price-to-earnings ratios). To address this issue, we recommend that researchers compare sample firms to the general population on the basis of these (and perhaps other) characteristics. A thoughtful descriptive analysis should provide insights regarding the important dimensions on which researchers should develop a performance benchmark.

Our central message is that the analysis of long-run abnormal returns is treacherous. As such, we recommend that the study of long-run abnormal returns be subjected to stringent out-of-sample testing, for example in different time periods or across many financial markets. Furthermore, we recommend that such studies be strongly rooted in theory, which might, for example, emanate from traditional models of asset pricing or from the systematic cognitive biases of market participants.

Appendix

This appendix documents methods that we considered for estimating the variance-covariance matrix of event-time abnormal returns. The results presented in Table III through Table IX assume cross-sectional independence of sample observations. Thus, it is assumed that the variance-covariance matrix of event-time abnormal returns (Σ) is diagonal, where the cross-sectional standard deviation of sample observations is the estimate of the n diagonal elements of this matrix. In matrix form, the conventional t-statistic is estimated as $\overline{AR}_\tau[\mathbf{1}'\hat{\Sigma}\mathbf{1}/n^2]^{-1/2}$, where $\mathbf{1}$ is a $n \times 1$ vector of ones and $\overline{AR}_\tau$ is the mean abnormal return for the sample.

A.I. Buy-and-Hold Abnormal Returns

When abnormal returns are calculated using either a single control firm or the bootstrapped skewness-adjusted test statistic, in principle a researcher could calculate test statistics assuming a variance-covariance matrix that is not diagonal (i.e., sample observations are not assumed independent). We estimate the variance-covariance matrix (Σ) as follows. Consider two firms: firm i has an abnormal return calculated from period s to $s + \tau$, firm j has an abnormal return calculated from period $s + a$ to $s + \tau + a$. If $a \geq \tau$, we assume the covariance between the abnormal returns is zero. If $a < \tau$, we estimate the i,jth element of the variance-covariance matrix (σ_{ij}) as:

$$\frac{1}{\tau - a - 1}\sum_{t=s+a}^{s+\tau}(AR_{it} - \overline{AR}_i)(AR_{jt} - \overline{AR}_j),$$

where AR_{it} is the monthly abnormal return for firm i. The mean abnormal returns used in the estimate of the covariance are the means for the $\tau - a$ overlap periods.

We test the specification of test statistics in samples with overlapping returns calculations (see Table IX) based on (1) a conventional t-statistic and the use of a single control firm and (2) a bootstrapped skewness-adjusted t-statistic using the buy-and-hold reference portfolios. We focus on the samples with overlapping returns because that situation is where cross-sectional dependence is most severe. The details of the bootstrap procedure are available on request. The empirical p value approach cannot be adjusted using the estimated variance-covariance matrix. The method described here reduces, but does not eliminate, the misspecification in samples with overlapping return calculations.

A.II. Cumulative Abnormal Returns

We also apply the method described in the preceding section to the calculation of test statistics based on cumulative abnormal returns calculated using conventional t-statistics and abnormal returns calculated using (1) a single control firm and (2) reference portfolios free of the new listing bias (see Section VI.A). Thus, the diagonal elements of the variance-covariance matrix are estimated using the cross-sectional standard deviation of sample observations and the off-diagonal covariance elements are estimated as described before. This adjustment for cross-sectional dependence again reduces but does not eliminate the bias documented in Table IX.

Finally, we estimate the n diagonal elements of the variance-covariance matrix separately using the time-series standard deviation of monthly abnormal returns for each sample firm during the event period. The covariance elements are estimated as before. This adjustment again reduces, but does not eliminate, the bias that results from cross-sectional dependence.

In summary, severe cross-sectional dependence can lead to overrejection of the null hypothesis. When researchers are faced with a sample where cross-sectional dependence is likely to be a problem (e.g., when return calculations involve overlapping periods or there is severe industry clustering), the calendar-time portfolio methods described in Section VI.B would be preferred.

REFERENCES

Ball, Ray, S. P. Kothari, and Charles E. Wasley, 1995, Can we implement research on stock trading rules?, *Journal of Portfolio Management* 21, 54–63.

Barber, Brad M., and John D. Lyon, 1997a, Detecting long-run abnormal stock returns: The empirical power and specification of test statistics, *Journal of Financial Economics* 43, 341–372.

Barber, Brad M., and John D. Lyon, 1997b, Firm size, book-to-market ratio, and security returns: A holdout sample of financial firms, *Journal of Finance* 52, 875–884.

Bickel, Peter J., Friedrich Götze, and Willem R. van Zwet, 1997, Resampling fewer than n observations: Gains, losses, and remedies for losses, *Statistica Sinica* 7, 1–31.

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

Brav, Alon, 1997, Inference in long-horizon event studies: A Bayesian approach with application to initial public offerings, Working paper, University of Chicago.

Brav, Alon, Chris Géczy, and Paul A. Gompers, 1995, Underperformance of seasoned equity offerings revisited, Working paper, Harvard University.

Brav, Alon, and Paul A. Gompers, 1997, Myth or reality? The long-run underperformance of initial public offerings: Evidence from venture and nonventure capital-backed companies, *Journal of Finance* 52, 1791–1821.

Brock, William, Josef Lakonishok, and Blake LeBaron, 1992, Simple technical trading rules and the stochastic properties of stock returns, *Journal of Finance* 47, 1731–1764.

Canina, Linda, Roni Michaely, Richard Thaler, and Kent Womack, 1998, Caveat compounder: A warning about using the daily CRSP equally-weighted index to compute long-run excess returns, *Journal of Finance* 53, 403–416.

Chan, Louis K. C., Narasimhin Jegadeesh, and Josef Lakonishok, 1995, Evaluating the performance of value versus glamour stocks: The impact of selection bias, *Journal of Financial Economics* 38, 269–296.

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

Cowan, Arnold R., and Anne M. A. Sergeant, 1996, Interacting biases, non-normal return distributions and the performance of parametric and bootstrap tests of long-horizon event studies, Working paper, Iowa State University.

Davis, James L., 1994, The cross-section of realized stock returns: The pre-Compustat evidence, *Journal of Finance* 49, 1579–1593.

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

Fama, Eugene F., and Kenneth R. French, 1992, The cross-section of expected stock returns, *Journal of Finance* 47, 427–465.

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

Fama, Eugene F., and Kenneth R. French, 1996, Multifactor explanations of asset pricing anomalies, *Journal of Finance* 51, 55–84.

Fama, Eugene F., and Kenneth R. French, 1998, Value versus growth: The international evidence, *Journal of Finance* 53, 1975–1999.

Hall, Peter, 1992, On the removal of skewness by transformation, *Journal of the Royal Statistical Society*, Series B 54, 221–228.

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

Ikenberry, David, Graeme Rankine, and Earl Stice, 1996, What do stock splits really signal?, *Journal of Financial and Quantitative Analysis* 31, 357–376.

Jaffe, Jeffrey F., 1974, Special information and insider trading, *Journal of Business* 47, 410–428.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65–91.

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

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

Lee, Inmoo, 1997, Do firms knowingly sell overvalued equity?, *Journal of Finance* 52, 1439-1466.

Loughran, Tim, 1997, Book-to-market across firm size, exchange, and seasonality: Is there an effect?, *Journal of Financial and Quantitative Analysis* 32, 249–268.

Loughran, Tim, and Jay Ritter, 1995, The new issues puzzle, *Journal of Finance* 50, 23–52.

Mandelker, Gershon, 1974, Risk and return: the case of merging firms, *Journal of Financial Economics* 1, 303–336.

Neyman, Jerzy, and Egon S. Pearson, 1928, On the use and interpretation of certain test criteria for purposes of statistical inference, part I, *Biometrika* 20A, 175–240.

Pearson, Egon S., 1929a, The distribution of frequency constants in small samples from symmetrical distributions, *Biometrika* 21, 356–360.

Pearson, Egon S., 1929b, The distribution of frequency constants in small samples from nonnormal symmetrical and skew populations, *Biometrika* 21, 259–286.

Rau, P. Rahavendra, and Theo Vermaelen, 1996, Glamour, value and the post-acquisition performance of acquiring firms, Working paper, INSEAD.

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

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

Shao, Jun, 1996, Bootstrap model selection, *Journal of the American Statistical Association* 91, 655–665.

Speiss, D. Katherine, and John Affleck-Graves, 1995, Underperformance in long-run stock returns following seasoned equity offerings, *Journal of Financial Economics* 28, 243–268.

Speiss, D. Katherine, and John Affleck-Graves, 1996, The long-run performance of stock returns following debt offers, Working paper, University of Notre Dame.

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