THE JOURNAL OF FINANCE • VOL. LVIII, NO. 3 • JUNE 2003

Long-run Performance after Stock Splits: 1927 to 1996

JINHO BYUN and MICHAEL S. ROZEFF*

ABSTRACT

We measure the postsplit performance of 12,747 stock splits from 1927 to 1996 using two methods to measure abnormal returns: size and book-to-market reference portfolios with bootstrapping, and calendar-time abnormal returns combined with factor models. Between 1927 and 1996, neither method applied to splits 25 percent or larger finds performance significantly different from zero. Over selected subperiods, subsamples of 2-1 splits restricted by book-to-market availability requirements display positive abnormal returns using some methods. However, these samples show small or negligible abnormal returns using the calendar-time method. Overall, the stock split evidence against market efficiency is neither pervasive nor compelling.

THE EFFICIENT-MARKET CONCEPTION of near-perfect capital markets that render only fleeting and nonsystematic gain and loss opportunities to investors has been criticized in recent years by the behavioral finance literature, which offers evidence that stock transactions are often executed (in relation to known events such as stock issues, stock splits, and repurchases) at price levels that imply predictably high or low risk-adjusted returns. If these findings are factually correct, they pose a challenge to the efficient market hypothesis, which predicts a lack of capital market profit and loss opportunities due to the abilities of investors rapidly and unbiasedly to interpret information according to correct assessments of the underlying economic processes. The behavioral literature attributes its findings to various investor biases. Supporters of efficient markets deflate the importance of the behavioral finance findings by noting that risk-adjustment methods are imperfect, by suggesting that data mining may have occurred, by noting that all the behavioral anomalies taken together suggest an unbiased market at work, and by asking for behavioral models that explain a broader range of evidence (see Fama (1998) for a cogent summary). Haugen (1999, 2002) ably responds from the behavioral camp by pointing out the superior power of capital market phenomena like momentum to predict and explain returns.

We believe that certain issues of fact are not yet settled and deserve further examination, one of these being the stock split. Since splits are widely reported and noted, a stock split anomaly would be a particularly flagrant violation of

*Byun is from the Ewha Women's University, and Rozeff is from the University at Buffalo. The helpful comments of Eugene F. Fama, Dave Ikenberry, an anonymous referee, and the editor, Rick Green, are gratefully acknowledged. All errors are our own.

market efficiency. We ask whether returns after stock splits actually do allow investors to capture abnormal returns. Our paper suggests that the stock split does not provide evidence against efficient markets when the entire record is examined.

The case of stock splits is especially interesting for several reasons. First, there is a strong contradiction between earlier and later empirical findings. Fama et al. (1969) (FFJR) find no abnormal performance subsequent to stock splits, whereas both Ikenberry, Rankine, and Stice (1996) (IRS) and Desai and Jain (1997) (DJ) report abnormal returns of seven to eight percent in the 12 months following stock splits. Second, the stock split is a relatively uncomplicated event whose informational implications probably can be gauged rather easily by traders. As such, if traders fail to trade split stocks at correct prices, then judgmental errors may be deemed more likely to prevail in other more complex informational situations such as new issues and repurchases. Third, there is ample data to examine the issue historically, since data on stock splits go back to 1926, and the total number of splits since that date runs into the thousands.

Data availability is an important consideration. Fama (1998) argues that the recent findings cannot be construed as implying market inefficiency because they are confined to the 1975 to 1991 period, whereas FFJR find no abnormal stock returns between 1927 and 1959. We address this issue by using the longest period, 1927 to 1996, available to us on the CRSP files. We reexamine the two subperiods studied earlier and we look at two new subperiods, 1960 to 1974 and 1991 to 1996, which function as holdout periods and therefore provide new evidence that none of the prior studies has developed.

Mitchell and Stafford (1998), who examine acquisitions, equity issues, and equity repurchases, conclude that methodological variations produce nontrivial differences in estimates of long-term abnormal returns. Since earlier and later stock split studies employ very different methods, we alleviate the incomparability by uniformly applying a broad set of up-to-date abnormal return and statistical testing procedures to all the subperiods. Although no completely satisfactory method of measuring long-run performance exists, Lyon, Barber, and Tsai (1999) recommend several improved approaches. We apply these and other techniques so as to obtain an intelligible picture of stock returns after splits. The methods used include matching by size only, matching by size and book-to-market ratio (each of the latter in conjunction with bootstrapping statistical tests), as well as the Fama–French (1993) three-factor and the Carhart (1997) four-factor regression models combined with the calendar time abnormal return method of Jaffe (1974) and Mandelker (1974).

Yet another difficulty in assessing long-term performance arises from sampling variation. Mitchell and Stafford (1998) find that "comparison of our estimates to those of other researchers reveals that slight modifications to either the sample or the methodology can produce dramatically different results." This study addresses the issue of sampling in several ways. We examine both two-for-one (2-1) splits (as IRS have) and all splits and stock dividends of size 25 percent or greater (as FFJR and DJ have). We use samples of firms that traded on all the major venues: New York Stock Exchange (NYSE), American Stock Exchange (AMEX),

and Nasdaq. Both IRS and DJ, who employ size and book-to-market matching, use samples that are restricted by the requirement that book values be available on COMPUSTAT. To overcome, at least partially, the resulting confinement of the empirical tests to recent years, we collect book value data for 1927–1974 and extend size and book-to-market matching to earlier periods.

We and others who use COMPUSTAT data for recent years to obtain book-to-market ratios sacrifice sample size because of limitations in COMPUSTAT's coverage. To supplement size and book-to-market matching, we employ another approach that garners a significant sample expansion, namely, size matching only. While size matching gives up controlling for return variation due to book-to-market variation, it gains power by expanding sample size. Size-matched samples typically exceed size- and book-to-market-matched samples by over 30 percent.[1]

We make no claim that earlier findings contain error. Indeed, in the period 1975 to 1990 in which IRS find long-run positive performance after splits, we similarly find that, for 2-1 splits, the postsplit value-weighted cumulative abnormal returns (CAR) (using size and book-to-market matching) is 3.06 percent with a significance level of 0.007. Such a result not only confirms IRS but also shows that their finding is robust to two changes in their methods: (1) a change in return computation [from buy-and-hold abnormal returns (BHAR) to CAR] and (2) a change in weighting (from equal to value weighting). Nevertheless, particular methodological choices do have a marked influence on outcomes. The use by IRS of 2-1 splits together with book-to-market matching gives a restricted sample of 1,802 observations. However, we find 6,918 splits of size greater than 25 percent between 1975 and 1990. We can evaluate all of these if we use size matching only, since the latter does not require that book values be available on COMPUSTAT. For the 6,918 splits, the control and split firms differ by merely 0.55 percent, an inconsequential and insignificant difference.

In the current state of the art, one can find reason to criticize virtually any method of estimating long-run performance. Barber and Lyon (1997) and Kothari and Warner (1997) point out that certain test statistics are misspecified. We use nonparametric bootstrapping methods (IRS (1996) and Lyon et al. (1999)) to overcome the latter problem. Lyon et al. recommend BHARs, while Fama (1998), who favors CARs, notes that BHARs grow with return horizon even if there is no abnormal return after the first period. In the case of stock splits, we find that no large differences occur in using CARs or BHARs. Fama and Lyon et al. recommend using the Jaffe (1974)–Mandelker (1974) calendar-time portfolio technique to overcome cross-sectional dependence and heteroskedasticity. On the other hand, Loughran and Ritter (2000) suggest that calendar-time abnormal returns (CTARs) lack power. We use both matching methods and CTARs and find that

[1] Other dimensions of sampling may cause variations in estimates of long-term abnormal performance. These include COMPUSTAT source files used, treatment of dropout firms, restriction of samples to ordinary common stocks, and restriction on allowable holdout months. Ascertaining the correct contemporaneous exchange code for a security also has an impact because of its influence on cutoffs for benchmark portfolios. See Section V for further discussion of these issues.

both give similar results. Another controversy centers on the use of equally weighted versus value-weighted portfolios, with Fama leaning toward value-weighted portfolios because they more realistically reflect investor experience, and Loughran and Ritter leaning toward equally weighted portfolios because they better reflect the experience of the average firm. We examine evidence using both weightings.

Based on all the evidence in this paper, we conclude that stock splits are *not* followed by abnormally positive (or negative) returns, and that investors have *not* systematically underreacted (or overreacted) to stock splits. Using the most comprehensive sample of 12,747 split events over the years 1927 to 1996, we find that one-year value-weighted postsplit cumulative abnormal returns differ by only 0.03 percent between split firms and control firms matched on size. Equally weighted buy-and-hold abnormal returns also exhibit insignificant stock price performance. In this instance, although the split firms' abnormal returns average 2.73 percent higher than those of size-matched control firms, the significance level is 18.2 percent. These and other results we present provide new evidence that the stock market is efficient with respect to stock splits.

The rest of the paper is organized as follows. Section I discusses the samples. Section II describes the research methods of computing abnormal returns. Results of the matching-bootstrapping and CTAR performance tests are presented in Section III. Section IV analyzes sampling-related issues, and Section V concludes the paper.

I. Samples

A. Sample Selection

The *all-split* or *size* sample consists of all stock splits and dividends (hereafter simply called splits) during the period from 1927 to 1996, as contained on the Center for Research in Security Prices (CRSP) files, that have a split factor greater than or equal to 25 percent and meet the following criteria: (1) Market capitalization data on the splitting firm is available at the end of month $t - 1$, where month t is the month the split takes effect; (2) the splitting shares are ordinary common shares (we omit ADRs, SBIs, REITs, and closed-end funds); and (3) at least two months of returns are available in the 12-month period subsequent to the split. The resulting all-split size sample consists of 12,747 splits. We also work with a *two-for-one* (2-1) subsample of 5,012 splits. By construction, the size samples can be analyzed using size-matched benchmark portfolios. Therefore these samples are called size-matched or simply size samples.

To construct subsamples that utilize size and book-to-market matching, the prior fiscal year's book value of equity (BE) must be available. Use of prior year's data avoids look-ahead bias (Banz and Breen (1986)). From 1962 onwards, we use all CRSP firms that have BE data on COMPUSTAT files of industrials and non-industrials. Specifically, we use COMPUSTAT primary/secondary/tertiary files, over-the-counter files, and research tapes for 1995, 1978, and 1969. For 1996, we use COMPUSTAT PC. Negative BE firms are excluded (see also Lyon et al. (1999)). We

adopt Fama and French's (1993) definition of book value of equity.[2] Between 1926 and 1961, when COMPUSTAT data on BE are unavailable, we collect BE from Moody's Industrial Manuals. To keep the data collection effort manageable, we do not collect data for financial firms and utilities. We call the resulting samples *size-BE/ME* samples, because matching is by size and by book equity/market equity (BE/ME).

Table I shows the distribution of splits by year and by matching variables. Splits tend to be more frequent in periods of high market returns. Also, as one would expect, splits tend to be more frequent in larger size deciles. Prior to 1959, many years possess fewer than 10 2-1 splits. BE/ME matching dramatically reduces the total size sample (splits of 25 percent or greater) from 12,747 to 8,777. The number of 2-1 split events declines from 5,012 to 3,561 with BE/ME matching. Our sample of 978 events (all splits 25 percent or greater) between 1927 and 1959 is similar to FFJR's sample size of 940, the discrepancy being due to different sampling criteria (see Sec. IV). The new periods studied in this paper, 1960 to 1974 and 1991 to 1996, provide size samples of 2,134 and 2,717 splits, respectively. Between 1975 and 1990, we examine 1,802 size-BE/ME matched 2-1 splits, compared to the 1,275 studied by IRS; we gain sample size due to our inclusion of Nasdaq splits. Our all-split size-BE/ME sample in this period contains 4,454 events, and the all-split size sample grows to 6,918 events.

We can characterize the split samples used in this study in several other ways, such as by trading venue. Between 1927 and 1974, NYSE/AMEX splits dominate both the size and size-BE/ME samples. Exchanges account for 2,932 splits and Nasdaq has but 180 events, since CRSP's coverage of Nasdaq firms begins in 1968. Size-BE/ME samples, in similar fashion, have 2,343 listed splits and 110 Nasdaq splits before 1974. After 1974, the relative amounts of Nasdaq and exchange-listed splits vary among samples. In the size samples, exchange-listed splits number 4,028 compared to 5,607 Nasdaq splits, while in the size-BE/ME samples, exchange-listed splits outnumber Nasdaq splits by 3,513 to 2,811.

Size decile and BE/ME quintile afford another means of characterizing samples. With respect to size decile, the smallest three deciles, not unexpectedly, account for less than 30 percent of the splits, typically 12 to 18 percent together. Each of the remaining seven deciles typically collects 10 to 14 percent of the splits in something of a humped pattern, the maximum percentage of splits (14 to 16 percent) usually occurring in size deciles seven or eight. With respect to book-to-market quintile, a clear pattern emerges in each subperiod. The greatest percentage of splits occurs in the lowest BE/ME (glamor) quintile, approximately 40 percent. This percentage declines regularly as BE/ME increases (roughly 25 percent, 17 percent, 11 percent, and 7 percent). Lastly, stock splits have a bit of a

[2] BE is total shareholders' equity (A216), minus preferred stock, plus deferred taxes (A35) when available, plus investment tax credit (A208) when available, plus postretirement benefit liabilities (A330) when available. Preferred stock is redemption value (A56), liquidation (A10), or carrying value (A130; in this order), depending on availability. If total shareholders' equity is missing, we substitute total assets (A6) minus total liabilities (A181).

Table I

Number of Stock Splits per Year between 1927 and 1996

Stock splits in the columns headed "Size" are identified from the CRSP files using distribution codes 5523, 5533, 5543, and 5552 and the criterion that the split factor be greater than or equal to 25 percent. Three other criteria are as follows: (1) Market capitalization data on the splitting firm is available at the end of month $t-1$, where month t is the month the split takes effect. (2) The splitting shares are ordinary common shares that have share codes 10 or 11 (we omit ADRs, SBIs, REITs, and closed-end funds). (3) At least two months returns are available in the 12-month period subsequent to the split. Stock splits in the columns headed "Size-BE/ME" are those that meet a fourth criterion, namely, availability of positive book value information either on COMPUSTAT or through data collection in Moody's Industrial Manuals between 1927 and 1961. The annual S&P 500 returns are drawn from Ibbotson and Associates.

	Size		Size-BE/ME		S&P 500		Size		Size-BE/ME		S&P 500
Year	≥25%	2-1	≥25%	2-1	Return	Year	≥25%	2-1	≥25%	2-1	Return
1927	29	2	25	2	0.375	1962	75	30	44	19	−0.087
1928	26	2	22	1	0.436	1963	76	37	48	23	0.228
1929	52	2	44	2	−0.084	1964	127	60	93	45	0.165
1930	13	3	9	2	−0.249	1965	159	87	112	59	0.125
1931	2	0	1	0	−0.433	1966	177	115	132	84	−0.101
1932	0	0	0	0	−0.082	1967	131	71	110	61	0.240
1933	1	0	1	0	0.540	1968	255	123	208	105	0.111
1934	7	0	6	0	−0.014	1969	247	138	200	116	−0.085
1935	4	1	4	1	0.477	1970	71	23	63	18	0.040
1936	10	0	7	0	0.339	1971	142	50	111	41	0.143
1937	22	4	21	4	−0.350	1972	202	71	166	60	0.190
1938	6	0	5	0	0.311	1973	229	102	160	75	−0.147
1939	4	0	3	0	−0.400	1974	106	40	89	38	−0.265
1940	4	1	2	1	−0.098	1975	158	57	123	44	0.372
1941	3	0	2	0	−0.116	1976	298	96	255	81	0.238
1942	0	0	0	0	0.203	1977	306	106	197	75	−0.072
1943	3	0	3	0	0.259	1978	429	141	269	97	0.066
1944	11	3	9	2	0.198	1979	349	114	224	77	0.184
1945	40	11	36	10	0.364	1980	530	207	339	146	0.324
1946	79	17	68	16	−0.081	1981	595	217	356	139	−0.049
1947	48	10	43	8	0.057	1982	303	92	175	58	0.214
1948	29	8	26	6	0.055	1983	850	377	506	244	0.225
1949	20	8	17	7	0.188	1984	377	148	244	100	0.063
1950	53	16	45	13	0.317	1985	549	192	329	123	0.322
1951	56	12	50	10	0.240	1986	779	340	481	205	0.185
1952	36	6	29	6	0.184	1987	601	269	390	182	0.052
1953	24	6	17	6	−0.010	1988	234	71	150	52	0.168
1954	45	10	39	8	0.526	1989	331	117	240	95	0.315
1955	90	17	74	13	0.316	1990	229	103	176	84	−0.032
1956	95	27	75	20	0.066	1991	277	103	206	78	0.305
1957	42	10	34	8	−0.108	1992	451	177	349	148	0.077
1958	17	2	12	1	0.434	1993	515	221	349	157	0.100
1959	107	28	82	19	0.120	1994	383	164	303	127	0.013
1960	70	27	54	21	0.005	1995	491	227	318	149	0.374
1961	67	18	52	13	0.269	1996	600	275	345	156	0.231

seasonal, with the highest percentages (12–13 percent) occurring in each of May and June and roughly equal percentages in the remaining months.

The calendar-time samples utilize the above samples with additional restrictions. A CTAR is an abnormal return calculated monthly for all sample firms that have effected a split within the prior 12 months. The calculations involve time-series regressions using factor models. We use a maximum of 49 months of return data centered on the month of the split. We require a minimum of 24 months of returns to estimate each individual firm model. As a second requirement, since this method estimates portfolio residual variances, we follow Lyon et al. (1999) and require a minimum of 10 firms in each monthly event portfolio. This constraint is binding only over the 1927 to 1959 period. The number of 2-1 splits is less than 10 in every year from 1927 to 1944. Thereafter, it is less than 10 in 1948, 1949, 1952, 1953, and 1958. The result is to reduce the monthly CTAR observations for this subsample from a potential maximum of 396 months to 98. Our sample of all splits of 25 percent or greater contains 247 observations.

B. Matching Samples

We construct 10 size-matched benchmark (control or reference) portfolios as follows. All the NYSE/AMEX firms on the CRSP file at time $t-1$ are ranked by market value of equity and 10 breakpoints determined according to the 10 deciles of market value. Each Nasdaq firm is assigned to the benchmark portfolio that contains its market value of equity at time $t-1$. The resulting benchmark portfolios do not contain equal numbers of firms because Nasdaq firms tend to be smaller than exchange-listed firms. Later in the paper, robustness tests are conducted in which all firms, including Nasdaq, are allocated to deciles and to 50 groups. The median numbers of firms in the size-matched benchmark portfolios range from 102 in the 1927 to 1959 subperiod to 539 in the 1991 to 1996 subperiod.

To construct size-BE/ME benchmark portfolios, each of the 10 size-matched benchmark portfolios is divided into five BE/ME quintiles. The individual firm BE/ME values are determined by dividing ME values at the end of the month prior to the test period by the prior fiscal year's BE values. To eliminate benchmark contamination bias (Loughran and Ritter (2000)), stock split firms are then eliminated from the benchmark portfolios. The result is 50 size-BE/ME benchmark portfolios.

II. Research Methods

A. BHARs, CARs, and Bootstrapping

For a 12-month period, individual firm buy-and-hold abnormal returns are calculated as

$$BHAR_i = \prod_{t=1}^{12}(1+R_{it}) - \prod_{t=1}^{12}(1+R_{bt}), \tag{1}$$

where R_{it} is the total rate of return of stock i in month t, and R_{bt} is the average of the benchmark portfolio firms' returns in month t.[3] As an averaging method for the benchmark portfolio firms, we use value weighting. Each splitting stock is matched to its appropriate benchmark portfolio using size as a criterion or size-BE/ME as criteria. The benchmark portfolios are repeatedly sampled to produce bootstrapped pseudoportfolios. When bootstrapping p values are obtained, almost identical values occur if the portfolio firms are equally weighted because sample means and bootstrapped portfolio means are calculated using the same benchmark portfolio.[4] After $BHAR_i$ is obtained for each of the N firms in a sample, we compute sample mean BHARs using either equal weights or value weights:

$$\overline{BHAR} = \sum_{i=1}^{N} w_i \cdot BHAR_i, \tag{2}$$

where $w_i = 1/N$ for the equal-weighting case and $w_i =$ market value of stock i at time $t - 1$ divided by the total market value of all stocks in the market index at time $t-1$.

The cumulative abnormal return of firm i in a 12-month period is

$$CAR_i = \sum_{i=1}^{12} (R_{it} - R_{bt}) \tag{3}$$

The benchmark portfolio return is the value-weighted average return of the size or size-BE/ME portfolio that matches the split stock. Sample average CARs are computed using both equal-weighting and value-weighting.

To assess significance levels of the BHARs and CARs, we use bootstrapping. Corresponding to each firm in a sample, a firm is randomly sampled with replacement from the firms comprising the sample firm's matching benchmark portfolio. We calculate the BHAR or the CAR of the randomly sampled firm via equations (1) and (3), respectively. This procedure yields a set of BHARs (or CARs) for one pseudoportfolio whose firms possess the same matching characteristics of size or size-BE/ME as the sample firms. Additionally, the pseudoportfolio has the same number of observations as the sample and its returns are calculated over the same calendar months. Applying (2) to the pseudoportfolio gives a single mean BHAR (or CAR). We repeat these steps 5,000 times to obtain 5,000 mean BHARs (or CARs). The p-value of the sample is calculated as the fraction of the mean BHARs (or CARs) of the pseudoportfolios that are larger in magnitude than the mean BHAR (or CAR) of the sample being tested.

[3] The pertinent 12-month period for a split begins in the month after the split is effected. This holdout period excludes any possible announcement effect, record period effect, and ex-dividend date effect. Nayar and Rozeff (2001) report an average ex-dividend date effect of 1.55 to 2.13 percent in 3,336 NYSE/AMEX splits between 1963 and 1993. They also report a record period effect of approximately negative one percent after split announcements and prior to the ex-date effect.

[4] Lyon et al. (1999, p. 175) make this point. We verify it empirically in Section IV and find negligible differences.

B. CTARs

We also test postsplit long-run performance using the technique of calendar-time abnormal returns. This method requires a time-series model of returns. We select the Fama–French (1993) three-factor model and the Carhart (1997) four-factor model. Since stocks that split experience high returns before they split, price momentum may relate to subsequent returns. This makes Carhart's model, which includes a price momentum factor, a sensible way to disentangle momentum and stock split effects.

We follow the procedure of Mitchell and Stafford (1998). At month t, $CTAR_t$ is the average abnormal return for all sample firms that have effected a split within the prior 12 months: $CTAR_t = R_{pt} - E(R_{pt})$, where R_{pt} is the monthly return on the portfolio of event firms at time t, and $E(R_{pt})$ is the expected return on the event portfolio at time t. The expected return on the event portfolio is measured by the factor model in the following way.[5] First, for each sample firm in month t, we calculate a time-series regression of the firm's excess returns on the factors:

$$R_{it} - R_{ft} = \alpha_i + \beta_i(R_{mt} - R_{ft}) + s_i SMB_t + h_i HML_t + m_i PR1YR_t + \varepsilon_{it}, \quad (4)$$

where R_{it} is firm i's monthly return including dividends in month t, R_{ft} is the one-month Treasury bill return, and R_{mt} is the return on the CRSP value-weighted portfolio of all NYSE, AMEX, and Nasdaq stocks. The size and book-to-market factors devised by Fama and French (1993), consisting of return differences of portfolios of small and large size and high and low BE/ME, are SMB_t and HML_t, respectively. The price momentum variable, $PR1YR_t$, is defined as in Carhart (1997) as an equally weighted portfolio return of stocks with highest returns less an equally weighted portfolio return of stocks with lowest returns in months $t-12$ to $t-2$. To assure continuity of these factors over the entire time period, we construct these factors ourselves following the procedures laid out by Fama and French and Carhart.

III. Postsplit Performance

A. Matching Methods—Whole Period

Table II summarizes the postsplit performance over the entire period 1927 to 1996 as measured by the two matching methods: size-BE/ME and size. When we compare the abnormal returns (and p-values) in any given line of the table, we observe only small differences between the BHAR and CAR performance measures. The largest differences are about 0.5 percent (in favor of the BHARs) using equally weighted samples, but often the significance levels are higher using CARs. To simplify matters, the following discusses only BHARs.

The most potent evidence of underreaction occurs in the equally weighted size-BE/ME sample of 2-1 splits (line 1 of Panel A) in which the BHARs average 3.74 percent with a p-value of 0.031. However, when this sample of 3,561 is enlarged to

[5] We use the Carhart model for illustration. The Fama–French model is the same except that it lacks the price momentum variable.

Table II

Performance Measures in the 12-month Period after Stock Splits over the Period 1927 to 1996

Sample size is *N*. The mean buy-and-hold abnormal return, in percent, is BHAR, and CAR is the mean cumulative abnormal return. All *p*-values are obtained using bootstrapping method with 5,000 replications. The *p*-value is the fraction of bootstrapped portfolios with values higher than the sample portfolio. Reference portfolios are 50 size and book equity/market equity (size-BE/ME) portfolios, or 10 size portfolios. Mean BHARs and CARs are found using either equal weighting (EQ) or value weighting (VW).

	N	BHAR(%)	*p*-value	CAR(%)	*p*-value
Panel A: Size-BE/ME Matching					
2-1 splits					
(1) EQ	3,561	3.74	0.031	3.20	0.025
(2) VW	3,560	1.55	0.251	1.59	0.209
All splits $\geq$ 0.25					
(3) EQ	8,777	2.61	0.085	2.01	0.064
(4) VW	8,777	1.11	0.245	0.80	0.264
Panel B: Size Matching					
2-1 splits					
(1) EQ	5,012	2.47	0.142	2.34	0.111
(2) VW	5,010	0.94	0.296	1.11	0.254
All splits $\geq$ 0.25					
(3) EQ	12,747	2.73	0.182	2.34	0.241
(4) VW	12,747	0.45	0.289	0.03	0.356

8,777 by including all splits greater than 25 percent, the BHARs average 2.61 percent with a *p*-value of 0.085 (line 3 of Panel A), not significant at the conventional five percent level. Without there being a persuasive reason to ignore the majority of splits in evaluating postsplit performance, we interpret the evidence against market efficiency as borderline even when we use the method (equally weighted BHARs) that tends to provide the strongest evidence against it. Loughran and Ritter (2000) emphasize equal weighting because it measures the abnormal returns of a typical event. However, this weighting yields only marginal whole-period evidence of market inefficiency after splits using the size-BE/ME matching.

One uses value weighting if one wishes to measure the aggregate wealth effects (or average investor experience) of the stock split event. Lines 2 and 4 of Panel A provide the results of value weighting sample portfolios and size-BE/ME matching. For all splits, BHARs average 1.11 percent (p-value = 0.245); they are slightly higher (1.55 percent) for 2-1 splits (p-value = 0.251). We cannot conclude that value-weighted abnormal returns differ from zero after the 8,777 stock splits of size greater than 25 percent or the 3,560 2-1 splits. Even if our best estimate of the subsequent returns is about one percent, this number likely falls within the range of error produced by inadequate modeling and testing of abnormal returns as well as transactions costs.

To augment the appraisal of postsplit performance, this paper reports tests using size matching only, because this method can be used to test the complete range of split observations. Although size matching gives up testing power by not controlling for normal return variation associated with the BE/ME factor, it gains power by expanding the sample size. Specifically, the number of observations of 2-1 splits rises from 3,561 to 5,012, and the number of all splits rises from 8,777 to 12,747. Size-BE/ME and size-only matching contrast in another way. The number of securities in the size-matched reference portfolios exceeds the number in the size-BE/ME portfolios, both because there are more size-BE/ME portfolios and because of relative data availability. For example, the median size-BE/ME reference portfolio contains 31 securities between 1960 and 1974 as compared with 208 in the size-matched reference portfolio, while between 1991 and 1996 the median numbers are 52 and 539, respectively. If larger sample sizes within given breakpoints of size and BE/ME produce more accurate return benchmarks within the reference portfolios, then the size-matched reference portfolios may contain less bias in measuring expected returns.

Panel B of Table II shows the test results over 1927 to 1996 using size matching. The abnormal returns of equally weighted samples cluster near 2.5 percent with p-values ranging from 0.111 to 0.241. As in the size-BE/ME case, value-weighted samples have even lower abnormal returns, from 0.03 to 1.11 percent, with the lowest p-value being 0.254. A proponent of underreaction may wish to emphasize the experience of the average firm via BHARs with equally weighted returns. However, although the BHAR is 2.73 percent in this instance for the 12,747 firms (line 3 of Panel B), the p-value is only 0.182. A market efficiency supporter who emphasizes aggregate wealth effects will point to the lower value-weighted BHAR and CAR of 0.45 percent and 0.03 percent, with p-values of 0.289 and 0.356, respectively.

B. Matching Methods—Subperiods

Table III shows postsplit returns and significance levels in four subperiods using size-BE/ME matching. Subperiod 1 is the FFJR period, 1927 to 1959. Subperiod 3, 1975 to 1990, is the IRS test period; the DJ test period is nearly the same, 1976 to 1991. Subperiod 2, covering 1960 to 1974, has not previously been examined, and subperiod 4 is the most recent six-year period, 1991 to 1996. The tests in these subperiods provide a useful perspective on the differences between the classic FFJR and later IRS and DJ findings.

The value-weighted samples with one exception, the 1975 to 1990 period and the 2-1 split sample utilized by IRS, provide no evidence against market efficiency. For 2-1 splits, subperiods 1, 2, and 4 all display abnormal returns near zero with insignificant p-values between 0.221 and 0.465. For all-split samples, which are easier to defend as they contain the total firm and investor experience, none of the subperiods, including 1975 to 1990, shows significant long-run performance. The average investor who consistently purchased stocks after they split, regardless of split size, did not experience abnormal returns in any subperiod as shown by the lack of significance among value-weighted samples. The average investor

Table III

Size-BE/ME Performance Measures in the 12-month Period after Stock Splits over Four Subperiods between 1927 and 1996

Sample size is *N*. The mean buy-and-hold abnormal return, in percent, is BHAR, and CAR is the mean cumulative abnormal return. All *p*-values are obtained using bootstrapping method with 5,000 replications. The *p*-value is the fraction of bootstrapped portfolios with values higher than the sample portfolio. Reference portfolios are 50 size and book equity/market equity (size-BE/ME) portfolios. Mean BHARs and CARs are found using either equal weighting (EQ) or value weighting (VW).

	2-1 Splits			Splits ≥ 0.25		
	N	BHAR(%) (*p*-value)	CAR(%) (*p*-value)	*N*	BHAR(%) (*p*-value)	CAR(%) (*p*-value)
1927 to 1959						
(1) EQ	166	2.97	2.51	811	1.36	1.35
		(0.063)	(0.072)		(0.065)	(0.048)
(2) VW	165	1.37	1.63	811	2.16	1.38
		(0.246)	(0.221)		(0.075)	(0.167)
1960 to 1974						
(3) EQ	778	1.87	2.22	1,642	0.00	0.00
		(0.060)	(0.024)		(0.273)	(0.204)
(4) VW	778	0.00	0.43	1,642	−0.43	−0.27
		(0.465)	(0.379)		(0.571)	(0.542)
1975 to 1990						
(5) EQ	1,802	4.32	3.67	4,454	3.03	2.29
		(0.000)	(0.000)		(0.000)	(0.000)
(6) VW	1,802	3.63	3.06	4,454	1.57	1.08
		(0.003)	(0.007)		(0.157)	(0.208)
1991 to 1996						
(7) EQ	815	4.39	3.22	1,870	3.97	2.77
		(0.002)	(0.005)		(0.001)	(0.003)
(8) VW	815	−0.17	0.35	1,870	0.42	0.75
		(0.289)	(0.227)		(0.176)	(0.139)

who purchased all the 2-1 splits between 1975 and 1990 received an abnormal return in excess of three percent (value-weighted), but this profitable investing policy yielded an average BHAR of −0.17 percent (value-weighted) in the following six years.

In the all-split samples, the experience of the typical firm, as measured by equal-weighted samples, is mixed. Some findings suggest cases in which prices do not immediately reflect the pricing implications of the split, whether because of trading frictions, asset pricing model problems, or investor biases. Specifically, in subperiods 3 and 4, equally weighted abnormal returns cluster around three percent at high levels of significance. In subperiod 1, performance is positive, but the confidence levels are five to seven percent. Furthermore, the magnitude of abnormal return for all splits is near one percent. Still considering all splits, subperiod 2 shows no evidence whatever of abnormal returns. In short, the subperiod evidence for all splits provides a mixed picture that does not suggest a

Table IV

Size-Matched Performance Measures in the 12-month Period after Stock Splits over Four Subperiods between 1927 and 1996

Sample size is *N*. The mean buy-and-hold abnormal return, in percent, is BHAR, and CAR is the mean cumulative abnormal return. All *p*-values are obtained using bootstrapping method with 5,000 replications. The *p*-value is the fraction of bootstrapped portfolios with values higher than the sample portfolio. Reference portfolios are 10 size portfolios. Mean BHARs and CARs are found using either equal weighting (EQ) or value weighting (VW).

	2-1 Splits			Splits ≥ 0.25		
	N	BHAR(%) (*p*-value)	CAR(%) (*p*-value)	*N*	BHAR(%) (*p*-value)	CAR(%) (*p*-value)
1927 to 1959						
(1) EQ	206	2.14	1.68	978	0.66	0.97
		(0.082)	(0.092)		(0.136)	(0.097)
(2) VW	205	1.85	2.16	978	1.50	0.05
		(0.253)	(0.218)		(0.110)	(0.244)
1960 to 1974						
(3) EQ	992	0.53	0.68	2,134	− 0.13	− 0.68
		(0.413)	(0.319)		(0.593)	(0.870)
(4) VW	992	− 0.96	− 0.41	2,134	− 1.63	− 1.77
		(0.671)	(0.593)		(0.820)	(0.848)
1975 to 1990						
(5) EQ	2,647	3.50	3.09	6,918	3.40	2.46
		(0.000)	(0.000)		(0.000)	(0.000)
(6) VW	2,647	2.45	2.12	6,918	0.93	0.55
		(0.012)	(0.018)		(0.108)	(0.176)
1991 to 1996						
(7) EQ	1,167	1.84	2.16	2,717	3.91	3.27
		(0.072)	(0.032)		(0.000)	(0.000)
(8) VW	1,167	0.17	0.71	2,717	0.65	0.09
		(0.249)	(0.188)		(0.116)	(0.156)

verdict of market inefficiency. As for 2-1 samples, the overall period conclusions apply reasonably well to each subperiod. Value-weighted samples show no abnormal performance except between 1975 and 1990. Equally weighted samples reject market efficiency strongly between 1975 and 1990 and 1991 to 1996. In the earlier subperiods, confidence levels tend to be six to seven percent.

Table IV provides further subperiod evidence using the larger samples possible with size matching. The size-matched tests reinforce those shown in Table III, that is, a lack of evidence against market efficiency in the value-weighted samples and positive performance in the latest two subperiods in equally weighted samples. The all-split samples show no evidence of outperforming or underperforming their size-matched benchmarks in subperiods 1 and 2, even with equal weighting. One other phenomenon is evident in comparing Tables III and IV. For 2-1 splits, the mean levels of abnormal returns are lower in size-matched performance measures. For instance, the 4.39 percent equally weighted, size-BE/ME

BHAR of 815 events in 1991 to 1996 falls to 1.84 percent for its size-matched counterpart of 1,167 events. This indicates that the 2-1 splits for which BE/ME data are unavailable actually tend to underperform their benchmarks.[6] Book-to-market sample restrictions arising because of COMPUSTAT use or other reasons seem to have a substantial impact on mean estimated abnormal returns.

C. CTAR Tests

Tables V and VI present the results of estimating abnormal returns using the Fama–French three-factor model and Carhart four-factor model, respectively. These tests involve portfolios of stocks that have experienced splits in any month in the previous 12 months. Although the estimation method uses monthly portfolio returns, the tables report returns in annualized form. The CTARs we report are time-series averages of monthly calendar time abnormal returns, while the *t*-statistics are calculated from the time series of monthly standardized CTARs. Thus it is possible for the mean CTAR to be associated with a *t*-statistic of the opposite sign.

In Table V, the main evidence against market efficiency is that, over the entire period, the equal-weighted method for either 2-1 splits or all splits produces significant abnormal returns in two-tailed tests (*t*-values of 2.06 and 1.99, respectively). This evidence lacks sting, however, because the abnormal returns are only 1.68 and 1.21 percent, respectively. As in the matching tests shown earlier, value-weighted portfolios show small and insignificant abnormal returns between 1927 and 1996. Unlike the bootstrapped reference portfolio method, subperiod evidence of postsplit abnormal returns is confined to the 1975 to 1990 subperiod and then only for the 2-1 sample. For this subperiod and sample, we observe positive abnormal returns in both split samples using either equal or value weighting. In periods outside 1975 to 1990, there is no systematic evidence of significantly positive performance measures even using equal-weighted samples. For example, between 1927 and 1959, the equal-weighted 2-1 sample exhibits an average abnormal return of -0.96 percent (t-value $= -0.64$).

The results of combining the CTAR method with the Carhart (1997) model are shown in Table VI. Equal-weighted abnormal returns average 0.60 percent between 1927 and 1996 ($t = 0.93$) in the 2-1 sample with similar results for the all-split sample. Value-weighted abnormal returns range between -0.84 percent for the 2-1 sample and 0.48 percent for the all-split sample, both being insignificant. Hence, even the slight evidence against market efficiency found using the Fama–French (1993) model disappears with the Carhart model. Note however, that the Carhart model provides positive and significant outcomes in the 1975 to 1990 period except for the equal-weighted sample of all splits. Interestingly, in each subperiod, Carhart-model average abnormal returns decline slightly, by about one percent or less, compared to Fama–French model returns. The decrease in estimated returns indicates that momentum may be a positive factor influencing

[6] Since 815 firms average 4.39 percent and 1,167 firms average 1.84 percent, the added 352 firms underperform by -4.06 percent.

Table V

Calendar-time Abnormal Returns (CTARs) Using Fama–French (1993) Three-Factor Model over the 12-month Period after Stock Splits between 1927 and 1996

Each month the sample portfolio contains firms that effected a split in the prior 12 months. A minimum of 10 events per month is required. For each firm in a monthly portfolio, the three-factor model is estimated over a 49-month period centered on that month. Individual factor loadings are averaged to obtain monthly portfolio factor loadings and monthly CTARs. Mean CTARs and standard errors are calculated from the time series of monthly CTARs. The time series of monthly standardized CTARs is used to calculate *t*-statistics.

	2-1 Splits			Splits ≥ 0.25		
	N	CTAR(%)	*t*-stat.	*N*	CTAR(%)	*t*-stat.
1927 to 1959						
(1) EQ	98	− 0.96	− 0.64	247	1.92	0.79
(2) VW	98	− 3.72	− 1.27	247	0.60	− 0.33
1960 to 1974						
(3) EQ	180	1.80	0.99	180	0.72	0.70
(4) VW	180	2.76	1.84	180	1.08	0.90
1975 to 1990						
(5) EQ	192	3.36	2.77	192	1.56	1.77
(6) VW	192	3.86	2.80	192	1.69	1.84
1991 to 1996						
(7) EQ	72	− 0.36	1.50	72	0.48	1.30
(8) VW	72	− 0.30	0.88	72	0.35	0.86
1927 to 1996						
(9) EQ	542	1.68	2.06	691	1.21	1.99
(10) VW	542	0.12	0.57	691	0.84	0.37

returns of stocks that split, and that a part of the occasional abnormal returns observed after stock splits is owed to momentum.

D. Robustness Tests

All samples in the preceding analyses used decile size breakpoints determined with NYSE/AMEX stocks. This, the conventional technique, may be called Method 1. To test the sensitivity of the findings to this benchmark sample construction technique, we replicate several tests with breakpoints chosen by ranking *all* the stocks at our disposal, that is, NYSE/AMEX and Nasdaq issues, and using all the stocks in creating breakpoints.[7] Method 2 uses the breakpoints to create deciles that contain equal numbers of securities using all the NYSE/AMEX and Nasdaq stocks. A possibly objectionable aspect of Method 2 is that the lower deciles are dominated by Nasdaq issues. Consequently, we conduct further experiments by dividing all the stocks into 50 groups. The latter is called

[7] We thank the referee for this suggestion. Since this work uses a newer version of the CRSP files, the sample sizes are somewhat higher than those used earlier in the paper.

Table VI

Calendar-time Performance Measures Using Carhart (1997) Four-Factor Model over the 12-month Period after Stock Splits between 1927 and 1996

Each month the sample portfolio contains firms that effected a split in the prior 12 months. A minimum of 10 events per month is required. For each firm in a monthly portfolio, the four-factor model is estimated over a 49-month period centered on that month. Individual factor loadings are averaged to obtain monthly portfolio factor loadings and monthly CTARs. Mean CTARs and standard errors are calculated from the time series of monthly CTARs. The time series of monthly standardized CTARs is used to calculate *t*-statistics.

	2-1 Splits			Splits ≥ 0.25		
	N	CTAR(%)	*t*-stat.	*N*	CTAR(%)	*t*-stat.
1927 to 1959						
(1) EQ	98	− 1.68	− 0.98	247	1.32	0.03
(2) VW	98	− 4.68	− 1.67	247	0.48	− 0.59
1960 to 1974						
(3) EQ	180	0.60	0.20	180	− 0.36	− 0.34
(4) VW	180	1.56	1.09	180	0.12	0.06
1975 to 1990						
(5) EQ	192	2.28	2.03	192	0.96	1.12
(6) VW	192	3.09	2.47	192	1.33	1.63
1991 to 1996						
(7) EQ	72	− 1.92	0.94	72	− 0.24	0.93
(8) VW	72	− 1.34	0.51	72	− 0.35	0.59
1927 to 1996						
(9) EQ	542	0.60	0.93	691	0.60	0.51
(10) VW	542	− 0.84	− 0.39	691	0.48	− 0.33

Method 3. Finally, to remove any possible mismatch between stocks used to create breakpoints and stocks sorted into the groups, we conduct tests solely on the NYSE/AMEX stocks. Method 4 uses only these NYSE/AMEX stocks both in forming decile breakpoints and in creating samples.

Table VII outlines the cumulative abnormal return bootstrapping results using these four methods as applied to 2-1 splits and equally weighted returns. We use 2-1 splits in keeping with previous studies in these time periods. We use equal-weighting because it provides stronger evidence against market efficiency than value-weighting, as we have seen. The time periods analyzed are between 1975 and 1996, since these are the years when Nasdaq data are available. Line 1 of Panel A and line 1 of Panel B reproduce the earlier findings, that is, those of Method 1.

In the 1975 to 1990 subperiod, shifting to Methods 2 and 3 does not create any major change from the earlier findings of Method 1. In particular, abnormal returns following splits average slightly over three percent in all three methods, and this level is highly significant. Method 4 also provides evidence that buying after splits would have been a profitable strategy. In this case, the abnormal return for listed issues rises to 3.871 percent. This occurs because the rise in the sample firms' split returns to 16.23 percent is offset by a somewhat smaller rise

TableVII

Size-adjusted Performance Measures in the 12-month Period after Stock Splits over 1975 to 1990 and 1991 to 1996 Subperiods under Four Methods

Method 1 uses all firms for sampling and benchmarks, NYSE/AMEX and NASDAQ, and uses NYSE/AMEX in creating size decile breakpoints. Method 2 uses all firms and chooses benchmarks for deciles using all firms. Method 3 is like Method 2 but creates 50 portfolios. Method 4 creates deciles and samples using only NYSE/AMEX securities. The splits analyzed are 2-1 splits. Sample size is *N*, and CAR is the mean cumulative abnormal return. All *p*-values are obtained using bootstrapping method with 5000 replications. The *p*-value is the fraction of bootstrapped portfolios with values higher than the sample portfolio. All portfolios use equal weighting of component securities.

	N	Split Return (%)	Benchmark Return (%)	Abnormal Return (%)	*p*-value
			Panel A: 1975 to 1990		
Method 1	2,772	14.40	11.32	3.08	0.0000
Method 2	2,772	14.40	11.21	3.19	0.0000
Method 3	2,772	14.40	11.21	3.19	0.0000
Method 4	1,346	16.23	12.36	3.87	0.0000
			Panel B: 1991 to 1996		
Method 1	1,175	17.28	15.65	1.63	0.0832
Method 2	1,175	17.28	15.46	1.82	0.0584
Method 3	1,175	17.28	15.33	1.95	0.0426
Method 4	474	16.78	16.13	0.65	0.3166

in the benchmark return to 12.36 percent. We can say that any of these methods confirms what earlier researchers have found over this subperiod. Since the method of using only NYSE/AMEX stocks to create breakpoints gives the same sort of abnormal return and significance level as using all stocks to create breakpoints, this method of selecting breakpoints appears robust.

Turning to the holdout period, the 1991 to 1996 interval, Method 1 indicates an abnormal return of 1.63 percent with a significance level of 8.32 percent. Methods 2 and 3 show slightly higher abnormal returns, of 1.82 percent and 1.95 percent, respectively, with significance levels of 5.84 percent and 4.26 percent. Under Method 4, the abnormal return drops to 0.65 percent, which is far from being significantly different from zero. In terms of abnormal returns, the findings under Methods 2 to 4 are similar to those using Method 1, being within one percent. In terms of levels of significance, Methods 1 to 3 cluster around the six percent level of significance. In this subperiod, Method 4 provides a different result than Methods 1 to 3. Since the breakpoint method is unlikely to be the cause, the main factor appears to be that no Nasdaq stocks are used in Method 4. This suggests that the Nasdaq stocks are the source of the positive and marginally significant abnormal returns in this subperiod.

We should also ask whether the 1991 to 1996 subperiod findings confirm those that occurred in 1975 to 1990. Methods 1 and 4 do not show significant abnormal

returns following splits. The levels of significance of Methods 2 and 3 are at the cutoff point of conventional significance. Any judgment depends a good deal on one's priors. Clearly the magnitude of abnormal returns under all the methods is smaller in 1991 to 1996 than in the 1975 to 1990 subperiod. Our own judgment is that the 1991 to 1996 subperiod does not provide convincing evidence against market inefficiency nor strong support for the 1975 to 1990 findings, in that the significant confidence levels are marginal and the size of the abnormal returns is less than two percent. If there is evidence of market inefficiency, it is located in the Nasdaq issues, not the listed issues, since the latter display no abnormal returns whatever. However, trading frictions are likely to be higher for the Nasdaq issues.

IV. Sampling and Comparison Issues

In this section, we raise certain issues of sampling, such as COMPUSTAT source files, treatment of returns of firms subsequent to the time they drop out of samples, restriction of samples to ordinary common stocks, restriction on allowable holdout months, and use of the correct exchange code for a security. Each of these sampling alternatives causes variations of unknown size in estimates of long-term abnormal performance. Since a full treatment of these issues requires a separate research paper, we address them only briefly.

Where comparisons are feasible, namely, 1927 to 1959 and 1975 to 1990, we find that our results in most important respects compare well with those reported earlier. However, certain differences crop up in the 1975 to 1990 period that we attempt to elucidate. While we do not uncover the precise sources of variation, the effort appears worthwhile in that we narrow the field down and provide clues as to directions for future research. We speculate that sampling issues regarding reference portfolios that employ book values drawn from COMPUSTAT sources can have quite important effects on the measurement of abnormal returns.

A. FFJR (1969)

The FFJR (1969) sample size of all splits is 940, whereas our all-split sample size is 978 over the same period. However, FFJR invoke a different criterion, namely, all observations must be listed for at least 12 months before the split and 12 months after the split. Later versions of the CRSP file also may differ from earlier versions.

FFJR use the market model (and log returns) in conjunction with CARs and conclude that abnormal returns after splits are randomly distributed about zero. Noting that the market model is not the same as size-matching, the closest of our methods that compares to FFJR's is the measurement of CARs using the equal-weight size-matched method. We find (Table IV) the average CAR in this case to be 0.97 percent with a p-value of 0.097. This is the *most* significant p-value, those for BHARs and value weighting being 0.110 to 0.244. These findings we construe as broadly consistent with those of FFJR, despite the clearly different methods of risk adjustment along with some differences in sampling.

If FFJR had estimated CARs using size-BE/ME matching and bootstrapping to estimate *p*-values, they presumably would have found an abnormal return of 1.35 percent with a *p*-value of 0.048 (see Table III) on a reduced sample size of 811. Loughran and Ritter (2000), employing their favored BHAR and equal weighting, would have found a BHAR of 1.36 percent and a *p*-value of 0.065. These methods produce quite similar results, although ironically each goes in a direction mildly unfriendly to the favored hypothesis of these researchers. Would such findings have altered FFJR's conclusion? Doubtful, because of the borderline degree of significance and a quite small abnormal return that cannot be regarded as being of economic significance. Furthermore, following the advice of a latter-day Fama (1998) and using value weighting, the CAR of 1.38 percent has an insignificant *p*-value of 0.167. And the CTARs of both the Fama–French (1993) and Carhart (1997) models applied to this early period produce nothing more than very small abnormal returns and *t*-values. Considering all the evidence, we believe that in the 1927 to 1959 period there is only one possible conclusion: Stock splits are bereft of significantly positive performance of any substantial magnitude.

B. IRS (1996)

We confirm IRS (1996) in finding that the 1975 to 1990 period is one in which, by some measures, the 2-1 stock split sample did not perform randomly subsequent to the split event. However, we report a 4.32 percent BHAR between 1975 and 1990 for our 2-1 sample of 1,802 NYSE/AMEX/Nasdaq splits, while IRS estimate a 7.93 percent BHAR for their 2-1 sample of 1,275 NYSE/AMEX firms. Our results differ by 3.61 percent. This raises interesting sampling and methodological issues that we explore next.

One difference in the two studies is that IRS use an equal-weighted reference portfolio, while ours is value weighted. As explained earlier, this introduces negligible variation in results. To verify this statement, we recompute reference returns using equal weighting. The equal-weighted benchmark return is 14.04 percent, nearly identical to our value-weighted benchmark return of 14.08 percent.

Another difference lies in the treatment of dropout firm returns. We fill with the market index return, while IRS fill with the size-BE/ME reference portfolio return. While this factor causes some variation, it is unlikely to cause a difference of 3.61 percent.[8] Furthermore, it is doubtful that proponents of underreaction wish to base their claim upon a mechanical matter such as how one reinvests proceeds of dropout firms.

An obvious experimental discrepancy of greater potential importance is that our sample includes Nasdaq firms, whereas the IRS sample does not. Eliminating the Nasdaq firms from our sample leaves 1,135 events, 140 shy of the IRS sample. This is brought about by differences in sampling criteria. Our splitting shares are ordinary common shares (we omit ADRs, SBIs, REITs, and closed-end funds) that

[8] If five percent of firms drop out and they earn an abnormal return of seven percent above the market portfolio, the abnormal return contribution is 35 basis points.

Table VIII

Components of Equally Weighted BHARs for 2-1 Splits between 1975 and 1990

For each period, returns of three samples are shown. Rows labeled "IRS" show results reported in Ikenberry, Rankine, and Stice (1996), which are for NYSE/AMEX 2-1 splits. Rows labeled "BR (1)" are results for the NYSE/AMEX sample of 2-1 splits of this study. Rows labeled "BR (2)" are results for the NYSE/AMEX/Nasdaq sample of 2-1 splits of this study. All returns are equally weighted buy-and-hold returns. Split returns are averages of the returns of splitting stocks in the 12 months subsequent to the split. Reference returns are estimated using the size-BE/ME method with bootstrapping. BHAR is the average annual abnormal return, computed as the difference between the split and reference returns.

Period	Study	Sample Size	Split Returns (%)	Reference Returns (%)	BHAR (%)
1975 to 1990	IRS	1,275	19.11	11.18	7.93
	BR (1)	1,135	17.59	12.28	5.30
	BR (2)	1,802	18.39	14.08	4.32
1975 to 1980	IRS	405	25.88	19.18	6.70
	BR (1)	339	25.48	21.05	4.43
	BR (2)	520	26.67	23.51	3.16
1981 to 1985	IRS	461	19.56	7.76	11.79
	BR (1)	406	16.65	10.02	6.63
	BR (2)	664	17.42	11.13	6.30
1986 to 1990	IRS	409	11.90	7.09	4.82
	BR (1)	390	11.72	8.75	2.97
	BR (2)	618	12.47	9.30	3.16

must have at least two months returns available in the 12-month period subsequent to the split. IRS do not mention similar criteria.

To gauge the importance of eliminating Nasdaq firms, we repeat our BHAR tests using NYSE/AMEX samples. This includes redoing the reference portfolios so that they include only NYSE/AMEX firms. The findings are reported in Table VIII along with results drawn from IRS (1996).

In the overall sample period, our (restricted) NYSE/AMEX 2-1 split sample averages a return of 17.59 percent compared to a reference return of 12.28 percent, for a difference of 5.31 percent. This is nearly one percent closer to the 7.93 percent reported by IRS (1996). This suggests that the IRS abnormal performance measure would have declined by about one percent if they had widened their sample to include Nasdaq firms.

The results are extremely interesting in that in two subperiods, 1975 to 1980 and 1986 to 1990, the average returns of their and our split firms are nearly identical. In the 1986 to 1990 subperiod, IRS report 11.90 percent and we find 11.72 percent for splitting firms' average returns after splits. This suggests that differences in split firm samples do not drive the discrepancies. Rather, a primary factor driving the differing results in these subperiods appears to be the difference in reference portfolio returns, their returns being lower than ours in all three subperiods. We do not know the source of these differences. We speculate that

the availability of book values on COMPUSTAT and the definition of book value adopted in a given study are factors of some importance. This is because the book values are used to determine the breakpoints of the reference portfolios.

In one subperiod, split samples do make a large difference. Between 1981 and 1985, the average returns of the splitting sample of IRS is 19.56 percent compared to our 16.65 percent. Again, we do not know the source of this difference. We believe that our findings display internal consistency. We observe that when Nasdaq firms are removed, our splitting returns decline by 1.19 percent in 1975 to 1980 and by 0.75 percent in 1986 to 1990, the two periods in which there is near agreement with IRS's splitting firm returns. Between 1981 and 1985, our splitting firm returns decline by 0.77 percent, in line with the declines in the other two time periods.

In private correspondence, both David Ikenberry and Hemang Desai inform us that their papers accumulate returns from the month of the split, whereas we accumulate returns in the month after the split becomes effective. Both suggest that this accounts for most of the experimental discrepancy between their studies and ours (over their sample periods). If this is so, then the observed long-term investor underreaction to stock splits is actually a short-term phenomenon.

We point out the experimental discrepancies, not to resolve them, which would take us far afield and could prove quite difficult without access to a sequence of different COMPUSTAT tapes, but to suggest that sampling issues, both in creating the splitting firm samples and in creating the reference portfolios, can assume importance in studies of long-term abnormal returns. Since COMPUSTAT coverage of firms alters through time, the numbers of firms with available book value data alter. This creates sampling variation in benchmark returns because the cutoff points of benchmark portfolios depend on BE/ME rankings.

C. DJ (1997)

Comparable to our all-split criterion, Desai and Jain (1997) sample splits and dividends of all sizes greater than 25 percent. DJ find that their 5,596 splits outperform their reference portfolios by 7.05 percent using equal-weighted BHARs, their time period covering 1976 to 1991. We recalculate split and reference portfolio returns over the period 1976 to 1991. Our all-split sample of 4,454 outperforms its size-BE/ME reference portfolios by 3.30 percent, also using equal-weighted BHARs. Close to one-fifth of the 3.75 percent difference between their 7.05 percent and our 3.30 percent is explained by the returns of the splitting firms. Our split sample averages 19.62 percent compared to 20.40 percent for DJ's sample. Four-fifths of the difference arises in the reference portfolio returns. Our reference firms average 16.32 percent compared to DJ's 13.35 percent.

In addition to the difference discussed above in when the abnormal return cumulation is begun, another possible factor is that the 50 DJ size-BE/ME reference portfolios (used in conjunction with the bootstrapping methodology) are additionally divided into three momentum portfolios each based on returns in the six months prior to the split. That is, DJ use a total of 150 reference portfolios. It is unlikely that momentum matching by itself is the source of a three percent

plus return discrepancy, the reason being that we find only small differences between the three-factor and four-factor (momentum) models using CTARs. It is also possible that sampling and breakpoint differences affect the findings, since our samples differ in size. In our initial *unrestricted* sample of 7,589 splits, we find 2,270 NYSE splits while DJ report 2,740 NYSE splits in their sample of 5,596 splits. Since we have been unable to resolve this discrepancy, its importance in affecting size deciles and reference portfolio breakpoints is unknown.

V. Summary and Conclusions

Previous research reaches contradictory conclusions regarding long-term market efficiency with respect to stock splits. Our study provides further evidence on this issue by examining more time periods and by applying a uniform set of up-to-date procedures. Our results have implications for other studies claiming to discover market inefficiencies in long-term abnormal returns, because we find that the appearance of significant abnormal returns is sensitive to time period, method of estimation, and sampling.

We examine 12,747 stock splits of size 25 percent or larger over the period 1927 through 1996. One can isolate specific subperiods and methods of estimation that yield significantly positive returns. Nevertheless, our overall results, based on a variety of subperiods and methodologies, indicate that buyers and sellers of splitting stocks do not, on average, earn abnormal returns that are significantly different from zero.

Comparison of the long-term abnormal returns produced by the buy-and-hold and cumulative return methods reveals that they are not very different, nor is one systematically larger than the other. Value-weighted abnormal returns show a more noticeable tendency to be somewhat less than equally weighted abnormal returns, although this does not occur in every subperiod or with every method. Although the three-factor and four-factor models estimated with calendar time returns produce similar results in each subperiod, it is notable that the small but significant overall period returns shown by the three-factor model diminish in size and significance in the four-factor model. This suggests that even the small amount of abnormal return occasionally observed for stock splits when using a three-factor model is related to momentum.

Analysis of our findings in comparison with others' suggests several issues that should be addressed in future research. These include sampling restrictions when COMPUSTAT–CRSP samples are used, book value definitions, and book value cutoffs in the creation of reference portfolios. Because of the problems inherent in using book values, we believe that a comparison of test power for size-matching only and size and book-to-market matching is a worthwhile topic for future research.

Our findings, although limited to stock splits, provide a message of caution and skepticism toward the claim that long-run abnormal performance pervades financial markets in response to publically announced events. Before we abandon market efficiency in favor of long-run market underreaction (or overreaction) to

various events, prudence demands several commonsense steps. These include evaluating evidence gathered over many, many years and kinds of markets and making efforts to evaluate the fallibilities of our methods of modeling, sampling, and measuring long-run abnormal returns. If we do conclude that long-run abnormal performance exists after a given event, prudence demands, for several reasons, that we not rashly conclude that markets are so inefficient that prices are routinely biased by large amounts, such as 50 to 100 percent. First, if investor biases are at work, they are likely eventually to be followed by losses and learning. Second, a large amount of useful economic reasoning is based on rationality. Third, it is a fact that model errors and benchmark and sampling issues affect abnormal return measures. Fourth, even if market prices reflect the imperfections of humans as information processors, market imperfections such as arbitrage costs, default risks, and short-selling costs also may rationally limit how traders respond in their attempts to profit and by their actions to affect market prices.

REFERENCES

Banz, Rolf W., and William J. Breen, 1986, Sample-dependent results using accounting and market data: Some evidence, *Journal of Finance* 41, 779–794.

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

Carhart, Mark, 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57–82.

Desai, Hemang, and Prem C. Jain, 1997, Long-run common stock returns following stock splits and reverse splits, *Journal of Business* 70, 409–433.

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

Fama, Eugene F., Lawrence Fisher, Michael C. Jensen, and Richard Roll, 1969, The adjustment of stock prices to new information, *International Economic Review* 10, 1–21.

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

Haugen, Robert A., 1999, *The New Finance* (Prentice-Hall, Upper Saddle River, NJ).

Haugen, Robert A., 2002, *The Inefficient Stock Market* (Prentice-Hall, Upper Saddle River, NJ).

Ikenberry, David L., Graeme Rankine, and Earl K. Stice, 1996, What do stock splits really signal? *Journal of Financial and Quantitative Analysis* 31, 357–375.

Jaffe, Jeffrey, 1974. Special information and insider trading, *Journal of Business* 47, 411–428.

Kothari, S.P., and Jerold B. Warner, 1997, Measuring long-horizon security performance, *Journal of Financial Economics* 43, 301–339.

Loughran, Tom, and Jay Ritter, 2000, Uniformly least powerful tests of market efficiency, *Journal of Financial Economics* 55, 361–389.

Lyon, John D., Brad Barber, and Chih-Ling Tsai, 1999, Improved methods for tests of long-run abnormal stock returns, *Journal of Finance* 54, 165–201.

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

Mitchell, Mark L., and Erik Stafford, 1998, Managerial decisions and long-term stock price performance, Unpublished working paper, University of Chicago School of Business.

Nayar, Nandkumar, and Michael S. Rozeff, 2001, Record date, when-issued, and ex-date effects in stock splits, *Journal of Financial and Quantitative Analysis* 36, 119–139.