



American Finance Association

Sample-Dependent Results Using Accounting and Market Data: Some Evidence

Author(s): Rolf W. Banz and William J. Breen

Source: *The Journal of Finance*, Vol. 41, No. 4 (Sep., 1986), pp. 779-793

Published by: Blackwell Publishing for the American Finance Association

Stable URL: <http://www.jstor.org/stable/2328228>

Accessed: 16/07/2010 12:02

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=black>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



Blackwell Publishing and American Finance Association are collaborating with JSTOR to digitize, preserve and extend access to *The Journal of Finance*.

<http://www.jstor.org>

Sample-Dependent Results Using Accounting and Market Data: Some Evidence

ROLF W. BANZ and WILLIAM J. BREEN*

ABSTRACT

Studies relating accounting and price data often use the COMPUSTAT or related PDE data base as the source for the accounting data. This practice may introduce a look-ahead bias and an ex-post-selection bias into the study. We examine this problem by comparing results from the standard COMPUSTAT data base with those from a data base which suffers from neither bias. We find that rates of return from portfolios chosen on the basis of accounting data from the two data bases differ significantly. Further, we find that these differences imply different conclusions when we test a specific hypothesis relating accounting and price data. Finally, we propose a number of remedies which may reduce the bias when the standard COMPUSTAT data base is used.

THE RELATIONS AMONG THE economic activities of the firm, the accounting measures of these activities, and the market returns on the debt and the equity of the firm are of central interest to financial economists. Recently, there has been a renewed interest in the empirical relation between market return to equity and basic characteristics of the firm, such as the size and earnings yield of the firm.¹

Studies relating accounting and price data typically derive the basic accounting measures from the COMPUSTAT data base and use CRSP data for equity returns.² It is well known that there are several potential problems with the use of the COMPUSTAT data base. We focus on two potential problems, *the ex-post-selection bias* and *the look-ahead bias*.

The ex-post-selection bias arises because the current COMPUSTAT data base contains only those companies which are currently viable entities. Thus, companies which have merged, filed for bankruptcy, or have otherwise ceased to exist are excluded from the sample.³ Further, new companies often enter the data base with a full history, which introduces data not available on the file at an earlier time.

* INSEAD, Fontainebleau and J. L. Kellogg Graduate School of Management, Northwestern University, respectively. An earlier version of this paper was presented at the meeting of the European Finance Association, September 1983. We would like to thank Larry Glosten, Robert Hodrick, Ravi Jagannathan, Peter Rossi, and Rex Siquel for helpful comments.

¹ See, e.g., Banz [1], Basu [3, 4], and Reinganum [7].

² The COMPUSTAT files and its derivatives are copyrighted and distributed by Standard & Poor's Compustat Services, Inc. The CRSP files are prepared by the Center for Research in Security Prices of the University of Chicago.

³ Some researchers use the so-called merged COMPUSTAT file for the PDE (price-dividend-earnings) file which includes all firms which were on the file at any time during the sample period. This file is obviously not subject to the survivor bias.

The *look-ahead bias* is due to a dating problem. Data reported for a particular point in time, say at the end of the year, typically are not actually available to the investor until sometime later in the next year. Computing earnings yields with year-end prices and earnings may imply the ability of the investor to forecast future reported earnings without error. For example, the annual COMPUSTAT file reports earnings of \$1.24 per share for Zenith for year end 1978. The 12-month earnings per share actually observed by the investor as of December 31, 1978 was \$0.85 per share. At a December 31, 1978 price of \$12.87, the earnings yield computed using the COMPUSTAT data file was 9.6%, whereas the earnings yield using observed data was 6.6%. As might be expected, the price of Zenith stock went from the year-end price of \$12.87 to a March ending price (when the new earnings were known to investors) of \$15.00.

Empirical researchers have long been aware of these potential problems. Until now, there has been no practical way of measuring the size of the biases introduced. Some studies have ignored the problems, others have used various measures designed to reduce the biases, while some have claimed that the biases are of a negligible magnitude.⁴

The purpose of this paper is to examine the effect of the described idiosyncracies of the COMPUSTAT data base using two empirical relations, the "P/E effect" and the "small firm effect" as examples. We show that there are significant differences in returns to portfolios formed using the COMPUSTAT data base and returns to portfolios formed using a data source which does not have the *look-ahead bias* or the *ex-post-selection bias*. We further show that the independent "P/E effect" is evident when the current COMPUSTAT data base provides the accounting data, but when accounting data are taken from the data base which is free from these biases, the independent "P/E effect" is not supported by the data.

I. Data and Methodology

For the last several years, we have been collecting certain COMPUSTAT items on a monthly basis. The data base which has resulted suffers from none of the potential problems mentioned previously. It contains data on each company which was available to the investor from COMPUSTAT at the end of each month. Thus, the data base includes all companies which subsequently went bankrupt, merged, liquidated, or otherwise disappeared from the COMPUSTAT data base. New companies enter the data base when data first became available on the COMPUSTAT file.

Earnings, sales, etc. are reported as the most recently available 12-month data. Thus, the reported accounting results reflect exactly the most recent published data available to the investor. There is no look-ahead implied by this data base.

With this data base, we examine the size and severity of the biases discussed above. We accomplish this by contrasting results using the unbiased data base ("the sequentially collected COMPUSTAT file") with results using the current COMPUSTAT file.

⁴ Some recent examples of studies addressing the P/E effect include Basu [3, 4], Reinganum [7], and Peavy and Goodman [6]. All of these studies exclude firms with negative earnings and/or require firms to have fiscal years ending in December.

All firms which appear at any time in our data base and the daily CRSP returns file are included, i.e., all NYSE and AMEX stocks. The total sample contains over 2,500 securities. (CRSP daily returns are compounded to arrive at monthly returns.) The sample period is from January, 1974 to December, 1981.

In our tests, we form portfolios based on price and accounting data and examine returns from these portfolios. Specifically, we use market size and earnings yield as criteria for portfolio formulation.

Portfolios are formed as follows: at the end of each year, all companies in the sample are ranked by market value. Size quintiles are formed from this ranking. Each size quintile is then ranked separately on the basis of earnings yield. Earnings yield quintiles within each size group are formed with those securities with positive earnings. Companies with negative earnings form a sixth portfolio within each size quintile. Monthly rates of return are computed for each of the portfolios for the next 12 months. This procedure yields thirty portfolios, each with 96 monthly returns. We perform this construction using each of the two data bases.⁵ Each contiguous set of six portfolios contains the six earnings yield categories for a given size class. Portfolio groups are ranked in descending order of size (1-6: largest quintile; . . . ; 25-30: smallest quintile). Within each size group, the first five portfolios are ranked in decreasing order of earnings yield (which is equivalent to ranking in terms of increasing price-earnings ratio). The sixth portfolio contains the stocks of firms with negative earnings. Thus, portfolio 1 contains the securities with the largest earnings yields (lowest P/E ratios) within the largest size (market value) group.

Table I shows the minimum, maximum, and mean number of securities in portfolios formed from each of the two data bases. As expected, there are more firms with negative 12-month earnings among the small firms than among the large firms.⁶ Note the much smaller sample size for the current COMPUSTAT file which excludes non-survivors. Many of those non-survivors are small companies.

II. Testing the Basic Hypothesis: Equality of Returns of Portfolios Drawn from the Two Data Bases

A. Testing for the Total Difference

We test our basic hypothesis by looking at the differences in return between the equivalent portfolios drawn from the sequentially collected and current COMPUSTAT files. Our testing procedure is based on a multivariate perspective: formally, we test whether the mean differences in return of portfolios drawn from the different data sources are equal to zero for all thirty portfolios simultaneously. Both raw and risk-adjusted differences in return are examined. The estimate of

⁵ Market values are obtained from the COMPUSTAT tape. All stocks which are in the CRSP file on the last trading date of the year are included. Firms which subsequently go bankrupt or are merged into another firm are assumed to be sold at the final price (which may be zero), and the proceeds are reinvested in the remaining securities in the portfolio.

⁶ We expect to see more earnings variability, and thus more negative earnings associated with smaller capitalization companies for several reasons. Primarily, we expect that the earnings variability, or the negative earnings, to impact the market value of the equity adversely.

Table I**Number of Securities in Equally Weighted Portfolios: Sample Period 1974–1981**

| Portfolio | Sequentially Collected COMPUSTAT | | | Current COMPUSTAT | | |
|-----------|----------------------------------|---------|------|-------------------|---------|------|
| | Maximum | Minimum | Mean | Maximum | Minimum | Mean |
| 1 | 95 | 67 | 83 | 70 | 48 | 65 |
| 2 | 95 | 69 | 82 | 70 | 47 | 65 |
| 3 | 95 | 69 | 83 | 70 | 48 | 65 |
| 4 | 95 | 69 | 83 | 70 | 48 | 65 |
| 5 | 96 | 69 | 83 | 73 | 49 | 68 |
| 6 | 5 | 1 | 3 | 5 | 1 | 2 |
| 7 | 92 | 66 | 80 | 69 | 46 | 65 |
| 8 | 92 | 67 | 80 | 69 | 47 | 64 |
| 9 | 92 | 66 | 80 | 69 | 47 | 64 |
| 10 | 92 | 63 | 79 | 69 | 47 | 65 |
| 11 | 96 | 70 | 81 | 71 | 51 | 67 |
| 12 | 25 | 1 | 10 | 12 | 1 | 6 |
| 13 | 89 | 64 | 77 | 68 | 47 | 63 |
| 14 | 89 | 66 | 78 | 68 | 47 | 63 |
| 15 | 89 | 66 | 77 | 68 | 47 | 63 |
| 16 | 89 | 66 | 76 | 68 | 47 | 63 |
| 17 | 92 | 64 | 77 | 70 | 47 | 65 |
| 18 | 50 | 8 | 22 | 28 | 7 | 14 |
| 19 | 85 | 62 | 72 | 64 | 45 | 59 |
| 20 | 85 | 60 | 72 | 64 | 45 | 59 |
| 21 | 85 | 61 | 72 | 64 | 45 | 59 |
| 22 | 85 | 60 | 71 | 64 | 45 | 59 |
| 23 | 84 | 63 | 73 | 68 | 48 | 61 |
| 24 | 88 | 26 | 47 | 59 | 14 | 34 |
| 25 | 71 | 47 | 58 | 57 | 41 | 51 |
| 26 | 71 | 48 | 58 | 58 | 41 | 51 |
| 27 | 71 | 46 | 58 | 58 | 41 | 51 |
| 28 | 71 | 45 | 57 | 58 | 41 | 51 |
| 29 | 70 | 50 | 60 | 58 | 41 | 53 |
| 30 | 189 | 76 | 113 | 119 | 37 | 75 |

Notes: Portfolios represent five size groups containing six portfolios (1–6, . . . , 25–30). The six portfolios within each size group are made up of five ranked from highest to lowest earnings yield, plus a sixth portfolio made up of all negative earnings companies within that size group. The first five portfolios in each size group contain roughly equal number of securities at portfolio formation time.

the risk-adjusted difference is the intercept estimate of the regression of the raw difference on the excess return of a market index (the difference between the index return and the risk-free rate).

All of our tests can be viewed as tests of the significance of restrictions imposed on the parameters of a series of seemingly unrelated regressions.⁷ We form thirty

⁷ See Theil [8, pp. 313–14]. Since all equations in our system of seemingly unrelated regressions share the same independent variable(s), the actual computations are greatly simplified (see Theil [8, pp. 308–9]).

portfolios based on size (market capitalization) and earnings yield from the sequentially collected COMPUSTAT and thirty portfolios from the current COMPUSTAT. We then test whether the returns over the 96 months for each portfolio are different. Let $RC_i(t)$ be the return vector to the i^{th} portfolio ($i = 1, 30$) drawn from the current COMPUSTAT and $RS_i(t)$ be the return to the corresponding (the i^{th}) portfolio drawn from the sequential COMPUSTAT file, and

$$D_i(t) = RS_i(t) - RC_i(t).$$

We then wish to test for zero means of the $D_i(t)$ series for all i . Formally, write

$$\begin{aligned} D_1(t) &= \alpha_1 + \eta_1 \\ D_2(t) &= \alpha_2 + \eta_2 \\ &\vdots \\ D_{30}(t) &= \alpha_{30} + \eta_{30} \end{aligned}$$

where each α is the mean of the difference of the two return series, and each η is a zero mean random variable. Our means test then amounts to testing whether or not all the α 's are zero. We assume that the non-contemporaneous covariances are zero. If we could also assume that $E\eta_i\eta_j = 0$, we could test $\alpha_i = 0$ in each equation separately. However, there is no *a priori* reason for such an assumption. This is especially true for the means test, since all of the portfolios are probably correlated with the market portfolio. The possibility that $E\eta_i\eta_j \neq 0$ suggests that a multivariate approach is appropriate. Formally, we test the significance of the restriction that $\alpha_i = 0$ for all i simultaneously. In the means case we are discussing, this is equivalent to the Hotelling T^2 test.

In the risk-adjusted test, we can write $RC_i = \alpha c_i + \beta c_i Rm$, and $RS_i = \alpha s_i + \beta s_i Rm$, where Rm is the risk premium on the CRSP index. Let $\Delta_i = RC_i - RS_i$. The system of equations of interest is then:

$$\begin{aligned} \Delta_1 &= (\alpha c_1 - \alpha s_1) + (\beta c_1 - \beta s_1)Rm + \eta_1 \\ \Delta_2 &= (\alpha c_2 - \alpha s_2) + (\beta c_2 - \beta s_2)Rm + \eta_2 \\ &\vdots \\ \Delta_{30} &= (\alpha c_{30} - \alpha s_{30}) + (\beta c_{30} - \beta s_{30})Rm + \eta_{30} \end{aligned}$$

with η_i again being a zero mean random variable ($i = 1, 30$). In this case, we test for $(\alpha c_i - \alpha s_i) = 0$, for $i = 1, 30$ as the basic equivalent of the means test. We also test for $(\beta c_i - \beta s_i) = 0$ for $i = 1, 30$, and for zero differences for the α 's and β 's simultaneously. Again, the tests amount to testing for prior restrictions on a set of simultaneous equation estimates. The first step computes the regression of the return differences on a constant (for the raw return differences) and on the excess return on the market index (for the risk-adjusted differences) for each of the thirty portfolios. We use the residuals from these OLS regressions to estimate the contemporaneous covariance matrix of the thirty portfolios. Then, we impose the necessary linear constraints to test whether all thirty differences or selected subsets thereof are zero. The appropriate test statistic has an F -

Table II

Difference in Portfolio Returns/Sequentially Collected COMPUSTAT Minus
Current COMPUSTAT: Sample Period 1974–1981—Groups of Portfolios

| I. Raw Returns: Tests of Zero Mean Differences | | |
|---|---------------------|--|
| Portfolio | <i>F</i> -Statistic | Degrees of Freedom |
| All | 5.3194* | 30,2850 |
| 1–6 | 3.7653* | 6,2850 |
| 7–12 | 6.3701* | 6,2850 |
| 13–18 | 14.2885* | 6,2850 |
| 19–24 | 11.3287* | 6,2850 |
| 25–30 | 3.2145* | 6,2850 |
| II. Risk-adjusted Returns: Tests of Zero Differences, Both α 's and β 's | | |
| Portfolio | <i>F</i> -Statistic | Degrees of Freedom |
| All | 5.3576* | 60,2820 (Test α Only = 4.8595*) |
| 1–6 | 5.3000* | 12,2820 |
| 7–12 | 3.4300* | 12,2820 |
| 13–18 | 6.0300* | 12,2820 |
| 19–24 | 4.2200* | 12,2820 |
| 25–30 | 2.7300* | 12,2820 |

* Denotes significance at the 1% level.

distribution with q and $Ln - Lk$ degrees of freedom, where q is the number of constraints, L is the number of portfolios, n is the number of observations per portfolio, and k is the number of independent variables in the regression (one for raw differences, two for risk-adjusted differences).

We present results for equally weighted portfolios only.⁸ We use the CRSP equally weighted index as a proxy for the market portfolio and the Treasury bill rate from Ibbotson and Sinquefeld [5] as the risk-free rate.

Table II shows the results. We report the overall test of zero difference in return of all thirty portfolios simultaneously, and we also present results for the five size groups in the form of the appropriate F -statistics for the differences in mean portfolio returns. For our risk-corrected results, we report the F -statistic for differences in both α 's and β 's for the whole sample, and each of the five size groups. We also show the F -statistic for the α only for the whole thirty portfolios. (This statistic tests the restrictions that all thirty α differences based on returns from the two samples are zero). In Table III, we report t -statistics for the individual portfolio means, and for individual α 's and β 's. However, the reader should bear in mind that the returns on which these t -statistics are based are, in general, not independent across portfolios.⁹ For our sample, the results of the t -test and the F -test (which is the appropriate test given our multivariate approach) are very similar.

Table II shows that we can reject the hypothesis of no difference in return between the portfolios of the two data bases for both raw and risk-adjusted

⁸ We have replicated all of the experiments in this study using value-weighted portfolios. The results using value-weighted portfolios are virtually identical to those using equal-weighted portfolios, and thus are not presented separately.

⁹ The importance of the multivariate view was pointed out to us by an anonymous referee.

returns.¹⁰ From Table III, looking at the rows labeled "TOTAL DIFFERENCE," we also note a clear pattern in the signs of the point estimates: the difference for the first portfolio in each size group (highest positive E/P ratio or lowest positive P/E ratio) is positive, while that of the fifth portfolio in each size group (highest positive P/E ratio) is negative. We will return to this observation later.

These differences represent the compound effect of the look-ahead bias and the *ex-post-selection bias*. In order to show the effects of these biases individually, additional tests are required.

B. Testing for the Effect of the Ex-post-selection Bias

We test for the effect of the *ex-post-selection bias* in the following way. The sequentially collected COMPUSTAT file is used to form two sets of portfolios. The first is the original set of portfolios, while the second contains only those securities which are also in the current COMPUSTAT file. Thus, firms in this second set of portfolios have not been merged, liquidated, or otherwise deleted from the COMPUSTAT file. We use the same definitions of quintile limits for size and earnings yields for this second set of portfolios as were used in the first set. Thus, any difference in portfolio return is unambiguously due to the *ex-post-selection bias*, since both sets of portfolios use the same price and earnings data.

Table IV shows the mean differences in raw and risk-adjusted returns from the two portfolios in the same format as Table II. These data show that the *ex-post-selection bias* does contribute significantly to the total difference in return given in Table II. The *F*-statistic is significant at the one percent level. The "SELECTION BIAS" rows in Table III show a preponderance of negative signs in the individual portfolio differences. Thus, the sample containing all companies shows higher returns than the sample containing survivor-only companies, on average. Some portfolio pairs show substantial differences in risk, as many differences in portfolio beta are significant. There is no *a priori* reason for this particular pattern of differences. It is possible that, in other periods or for portfolios formed on the basis of other accounting information, an altered pattern of differences might be found. What is important is the potential for the survival bias to appear and to affect the outcome of statistical tests.

C. Testing the Effect of the Look-ahead Bias

In order to examine the effects of the *look-ahead bias* separately, we compare the returns of portfolios drawn from the current COMPUSTAT file with returns of portfolios drawn from the survivor file described previously. Since both files contain the same securities and differ only in the timing of the earnings numbers, any difference in return can be ascribed unambiguously to the *look-ahead bias*.

The results are given in Table V, which is in the same format as Tables II and IV. We find that the hypothesis of zero difference in returns can be rejected at the one percent level. From Table III, the same pattern of positive differences

¹⁰ Given the large sample size, the one percent significance level is appropriate for the *F*-statistics. We do, however, mark levels of the *F* and *t* statistics which are significant at the five and one percent levels in our tables.

Table III
Individual Portfolios: t -Statistics for Individual Portfolio Differences—Raw Returns

| | Earnings Yield | | | | | |
|----------|--------------------|----------|-----------|-----------|-----------|----------|
| | Highest | | Lowest | | Negative | |
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Largest | 1 TOTAL DIFFERENCE | -0.2271 | -2.0239* | -1.6560 | -2.7777** | 0.3880 |
| | SELECTION BIAS | -0.5691 | -1.9636* | -1.2246 | -1.9382 | -0.3602 |
| | LOOK-AHEAD BIAS | 1.2692 | 1.0386 | -1.1742 | 0.1187 | 0.7864 |
| Size | 2 TOTAL DIFFERENCE | 1.1778 | -1.3333 | -1.0187 | -2.9647** | -0.0272 |
| | SELECTION BIAS | -0.2615 | -3.3981** | -1.8057 | -1.7939 | 0.3447 |
| | LOOK-AHEAD BIAS | 2.3040* | 1.4258 | 0.4647 | -1.7366 | -0.2814 |
| Medium | 3 TOTAL DIFFERENCE | 0.3257 | -1.8511 | -3.6961** | -0.4028 | 0.4421 |
| | SELECTION BIAS | -2.4944 | -2.3865* | -1.7972 | -0.8845 | 0.4912 |
| | LOOK-AHEAD BIAS | 3.3231** | 0.5039 | -3.0706** | 0.7335 | 0.1361 |
| Smallest | 4 TOTAL DIFFERENCE | 2.5168* | -1.0905 | -2.0887 | -3.0710** | -1.0214 |
| | SELECTION BIAS | -1.1422 | -1.1491 | -3.3849** | -1.4564 | 0.7423 |
| | LOOK-AHEAD BIAS | 2.6840** | -0.0864 | 0.7481 | -2.2027* | -1.4628 |
| Largest | 5 TOTAL DIFFERENCE | 2.4960* | 1.0593 | 0.0729 | -2.1361* | 0.0787 |
| | SELECTION BIAS | -0.1032 | -1.8882 | -2.3535* | 0.7554 | 3.9040** |
| | LOOK-AHEAD BIAS | 1.8299 | 2.3313* | 1.8978 | -3.0984 | -1.5433 |

Individual Portfolios: t -Statistics for Differences in α 's and β 's—Risk Adjusted

| | Earnings Yield | | | | | | |
|----------|--------------------|---------|----------|---------|----------|---------|---------|
| | Highest | | Lowest | | Negative | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | |
| Largest | α | β | α | β | α | β | |
| | 1 TOTAL DIFFERENCE | 0.50 | 0.57 | 0.01 | 0.57 | -1.37 | -2.80** |
| | SELECTION BIAS | -0.11 | -1.91 | -1.58 | -1.58 | -1.13 | -0.21 |
| Medium | α | β | α | β | α | β | |
| | 2 TOTAL DIFFERENCE | 0.70 | 2.42* | 1.03 | -0.12 | -1.13 | -2.40* |
| | SELECTION BIAS | 0.70 | 2.42* | 1.03 | -0.12 | -1.13 | -2.40* |
| Smallest | α | β | α | β | α | β | |
| | 3 TOTAL DIFFERENCE | 0.50 | 0.57 | 0.01 | 0.57 | -1.37 | -2.80** |
| | SELECTION BIAS | -0.11 | -1.91 | -1.58 | -1.58 | -1.13 | -0.21 |
| Largest | α | β | α | β | α | β | |
| | 4 TOTAL DIFFERENCE | 0.50 | 0.57 | 0.01 | 0.57 | -1.37 | -2.80** |
| | SELECTION BIAS | -0.11 | -1.91 | -1.58 | -1.58 | -1.13 | -0.21 |
| Medium | α | β | α | β | α | β | |
| | 5 TOTAL DIFFERENCE | 0.50 | 0.57 | 0.01 | 0.57 | -1.37 | -2.80** |
| | SELECTION BIAS | -0.11 | -1.91 | -1.58 | -1.58 | -1.13 | -0.21 |
| Smallest | α | β | α | β | α | β | |
| | 6 TOTAL DIFFERENCE | 0.50 | 0.57 | 0.01 | 0.57 | -1.37 | -2.80** |
| | SELECTION BIAS | -0.11 | -1.91 | -1.58 | -1.58 | -1.13 | -0.21 |

| | | | | | | | | | | | | | |
|----------|------------------|--------|---------|---------|---------|---------|---------|---------|---------|---------|-------|--------|---------|
| 2 | TOTAL DIFFERENCE | 1.67 | -2.17* | -1.21 | -0.30 | -0.54 | -1.97 | -2.21* | -3.53** | -2.21* | -0.77 | 0.02 | -0.21 |
| | SELECTION BIAS | 0.54 | -3.45** | -2.79** | -2.55** | -1.82 | 0.35 | -0.95 | -3.93** | -1.32 | 0.19 | 0.16 | 0.74 |
| | LOOK-AHEAD BIAS | 1.91 | 1.43 | 0.97 | 1.81 | 1.11 | -2.76** | -1.60 | -0.29 | -0.78 | -1.08 | -0.09 | -0.76 |
| 3 | TOTAL DIFFERENCE | 1.49 | -4.87** | -1.52 | -1.19 | -3.31** | -1.15 | 0.06 | -1.94 | -3.80** | -1.55 | 0.40 | 0.11 |
| | SELECTION BIAS | -1.72 | -3.59** | -1.70 | -3.03** | -1.06 | -3.29** | 0.60 | -1.12 | -1.87 | -0.80 | 0.54 | -0.27 |
| | LOOK-AHEAD BIAS | 3.82** | -2.30* | 0.02 | 2.04* | -3.50** | 2.06* | 0.95 | -1.04 | -2.99** | -1.19 | 0.05 | 0.33 |
| 4 | TOTAL DIFFERENCE | 2.49* | -0.29 | -0.77 | -1.22 | -2.15* | 0.59 | -2.66** | -1.47 | -3.39** | 0.44 | -0.78 | -0.87 |
| | SELECTION BIAS | 1.23 | 0.55 | -1.31 | 0.98 | -3.20** | -0.23 | -1.47 | 0.28 | -1.53 | -0.85 | 0.77 | -0.25 |
| | LOOK-AHEAD BIAS | 2.76** | -0.55 | 0.33 | -1.80 | 0.49 | 1.00 | -1.72 | -1.89 | -2.39* | 1.17 | -1.28 | -0.56 |
| Smallest | TOTAL DIFFERENCE | 2.29* | 0.48 | 0.54 | 2.19* | -0.01 | 0.73 | -1.90 | -0.70 | -0.84 | 1.92 | 0.90 | -3.50** |
| | SELECTION BIAS | 0.58 | -2.92** | -1.17 | -3.19** | -1.91 | -1.68 | 1.13 | -1.70 | 1.41 | 1.34 | 3.62** | 0.62 |
| | LOOK-AHEAD BIAS | 1.18 | 2.81** | 1.41 | 4.62** | 1.38 | 2.12* | -3.10** | 0.48 | -1.77 | 1.16 | -2.02* | -5.09** |

* Denotes significance at the 5% level.

** Denotes significance at the 1% level.

Table IV

Difference in Portfolio Returns/Sequentially Collected COMPUSTAT/12 Fiscal Year Minus All Companies: Sample Period 1974–1981—Groups of Portfolios

| I. Raw Returns: Tests of Zero Mean Differences | | |
|---|-------------|--|
| Portfolio | F-Statistic | Degrees of Freedom |
| All | 4.2226* | 30,2850 |
| 1–6 | 3.6416* | 6,2850 |
| 7–12 | 4.7726* | 6,2850 |
| 13–18 | 9.2333* | 6,2850 |
| 19–24 | 4.8641* | 6,2850 |
| 25–30 | 8.2517* | 6,2850 |
| II. Risk-adjusted Returns: Tests of Zero Means, Both α 's and β 's | | |
| Portfolio | F-Statistic | Degrees of Freedom |
| All | 5.5494* | 60,2820 (Test α Only = 3.7843)* |
| 1–6 | 3.6900* | 12,2820 |
| 7–12 | 4.4200* | 12,2820 |
| 13–18 | 6.6900* | 12,2820 |
| 19–24 | 1.7900** | 12,2820 |
| 25–30 | 5.0100* | 12,2820 |

* Denotes significance at the 1% level.

** Denotes significance at the 5% level.

Table V

Difference in Portfolio Returns/Current COMPUSTAT Minus Sequentially Collected COMPUSTAT/12 Fiscal Year: Sample Period 1974–1981—Groups of Portfolios

| I. Raw Returns: Tests of Zero Mean Differences | | |
|--|-------------|--|
| Portfolio | F-Statistic | Degrees of Freedom |
| All | 4.4404* | 30,2850 |
| 1–6 | 2.5590** | 6,2850 |
| 7–12 | 5.3882* | 6,2850 |
| 13–18 | 8.5253* | 6,2850 |
| 19–24 | 7.0265* | 6,2850 |
| 25–30 | 9.4781* | 6,2850 |
| II. Risk-adjusted Returns: Tests of Zero Differences in α 's and β 's | | |
| Portfolio | F-Statistic | Degrees of Freedom |
| All | 5.2225* | 60,2820 (Test α Only = 3.9947)* |
| 1–6 | 2.0100** | 12,2820 |
| 7–12 | 2.6900* | 12,2820 |
| 13–18 | 4.2500* | 12,2820 |
| 19–24 | 3.2400* | 12,2820 |
| 25–30 | 7.8799* | 12,2820 |

* Denotes significance at the 1% level.

** Denotes significance at the 5% level.

for low P/E stocks and negative ones for high P/E stocks are again evident in the "LOOK-AHEAD BIAS" rows.

This pattern of the signs of the point estimates of the differences for the low and high P/E portfolios is consistent with the existence of a *look-ahead bias*. Low

P/E portfolios have higher returns in the current COMPUSTAT data base which is subject to the look-ahead bias, while high P/E portfolios have lower returns in that data base. Consider a firm which has a certain P/E ratio based on the then available earnings information. If this firm has unexpectedly high earnings (announced at some time in the new year), its share price is likely to increase even though its P/E ratio may remain unchanged. But if we use the actual (but unavailable) earnings figure to calculate the P/E ratio as of the previous year end, it will suddenly appear to be a low P/E firm. The example of Zenith given previously illustrates this possibility.

The opposite mechanism works for firms with unexpectedly low earnings; they appear to be high P/E firms using the new (but yet unavailable) earnings data.

We have shown that there are significant differences in portfolio returns when the same grouping rules are used with two different data sources. Both the *look-ahead bias* and the *ex-post-selection bias* contribute to the observed differences. In the next section, we show that the observed differences change the outcome of a specific test relating earnings yields and portfolio size.

III. Testing the Impact on the P/E and Size Effects

Some researchers have shown that there is a positive relation between earnings yield and return, independent of the relationship between size and return.¹¹ They typically use a sample similar to the current COMPUSTAT file. Their tests are often equivalent to examining the difference in return between portfolios at the extremes of the earnings yield and/or size range. That is, they test whether “high” earnings-price (low P/E) portfolios outperform “low” earnings-yield (high P/E) portfolios within a given size class. Thus, one test involves buying the high earnings-yield portfolio, selling short the low earnings-yield portfolio and recording the total gain.

Given the findings of such studies, we expect to find a significant gain from buying the highest earnings-yield portfolios within each size group while selling the lowest (positive) earnings-yield portfolio within each size group. The question we examine is whether this result holds for the current COMPUSTAT and the sequentially collected COMPUSTAT file, or only for the current COMPUSTAT file.

Table VI presents evidence on this question. In Table VI, we show both raw and risk-adjusted returns to the strategy of buying the highest earnings-yield decile and selling the lowest earnings-yield decile within each size group. We present returns for this strategy for data drawn from each of the two data bases.

For returns of portfolios constructed using the current COMPUSTAT file, we can reject the null hypothesis of zero mean difference vector between portfolios. Thus, if we use the current COMPUSTAT data file, we conclude that there is a significant relation between earnings-yield and return within size groups.

On the basis of the sequentially collected COMPUSTAT file, we come to a different conclusion. There are no significant differences for either raw or risk-adjusted returns. Thus, we conclude that no relation exists between earnings yield and return, when the experiment controls for size effects. A similar conclu-

¹¹ See, e.g., Basu [3, 4].

Table VI

**Test of Strategy: Buy Highest E/P Portfolio, Sell Lowest E/P Portfolio—
Current COMPUSTAT Data Base**

I. Raw Returns

| Size Group | Portfolio | Return (High E/P Portfolio—Low E/P Portfolio) | F-Statistic | t-Statistic |
|------------|-----------|---|-------------|-------------|
| All | | $F = 3.61^*$ (5,2850 DF) | | |
| 1 | 1-5 | 0.0105 | 6.54** | 2.54** |
| 2 | 7-11 | 0.0066 | 2.99 | 1.72 |
| 3 | 13-17 | 0.0088 | 7.65* | 2.75* |
| 4 | 19-23 | 0.0114 | 6.90* | 2.98* |
| 5 | 25-30 | 0.0082 | 1.70 | 2.69* |

II. Risk-adjusted Returns

| Size Group | Portfolio | α (t-Statistic) | β (t-Statistic) | F-Statistic (Total) |
|------------|--------------|------------------------|-----------------------|---------------------|
| All | $F = 2.34^*$ | $F = 4.5209^*$ | $F = 2.7226$ | |
| 1 | 1-5 | 0.0090 (2.16)** | 0.0882 (1.52) | 4.46** |
| 2 | 7-11 | 0.0069 (1.76) | -0.0183 (-0.34) | 1.54 |
| 3 | 13-17 | 0.0098 (3.00)* | -0.0565 (-1.25) | 4.62* |
| 4 | 29-23 | 0.0128 (3.33)* | -0.0860 (-1.61) | 5.85* |
| 5 | 24-30 | 0.0077 (2.49)** | 0.0278 (0.65) | 3.85** |

**Test of Strategy: Buy Highest E/P Portfolio, Sell Lowest E/P Portfolio—
Sequentially Collected COMPUSTAT Data Base: Sample Period 1974-1981**

I. Raw Returns

| Size Group | Portfolio | Return (High E/P Portfolio—Low E/P Portfolio) | F-Statistic | t-Statistic |
|------------|-----------|---|-------------|-------------|
| All | | $F = 0.94$ (5,2850 DF) | | |
| 1 | 1-5 | 0.0071 | 2.63 | 1.61 |
| 2 | 7-11 | 0.0024 | 0.41 | 0.64 |
| 3 | 13-17 | 0.0022 | 0.51 | 0.71 |
| 4 | 19-23 | 0.0019 | 1.23 | 0.51 |
| 5 | 25-30 | 0.0008 | 0.14 | 0.25 |

II. Risk-adjusted Returns: Test of Individual Size Groups

| Size Group | Portfolio | α (t-Statistic) | β (t-Statistic) | F-Statistic (Total) |
|------------|--------------|------------------------|-----------------------|---------------------|
| All | $F = 1.2353$ | $F = 0.6927$ | $F = 1.6255$ | |
| 1 | 1-5 | 0.0066 (1.48) | 0.0249 (0.40) | 1.38 |
| 2 | 7-11 | 0.0024 (0.61) | 0.0037 (0.07) | 0.21 |
| 3 | 13-17 | 0.0022 (0.70) | -0.0016 (-0.04) | 0.25 |
| 4 | 29-23 | 0.0030 (0.81) | -0.0690 (-1.34) | 1.01 |
| 5 | 24-30 | -0.0004 (-0.13) | 0.0757 (1.62) | 1.34 |

Table VI—Continued

Test of Strategy: Buy Highest E/P Portfolio, Sell Lowest E/P Portfolio—
Sequentially Collected COMPUSTAT/12 Fiscal Year Data Base: Sample
Period 1974–1981

I. Raw Returns

| Size Group | Portfolio | Return (High E/P Portfolio—Low E/P Portfolio) | F-Statistic | t-Statistic |
|------------|-----------|---|-------------|-------------|
| All | | $F = 1.48$ (5,2850 DF) | | |
| 1 | 1–5 | 0.0083 | 3.43 | 1.84 |
| 2 | 7–11 | 0.0039 | 0.84 | 0.91 |
| 3 | 13–17 | 0.0016 | 0.29 | 0.53 |
| 4 | 19–23 | 0.0013 | 1.26 | 0.43 |
| 5 | 25–30 | -0.0021 | 1.69 | -0.51 |

II. Risk-adjusted Returns

| Size Group | Portfolio | α (t-Statistic) | β (t-Statistic) | F-Statistic (Total) |
|------------|--------------|---------------------------|--------------------------|------------------------|
| All | $F = 1.6655$ | $F = 0.6037$ | $F = 2.2373$ | |
| 1 | 1–5 | 0.0074 (1.60) | 0.0562 (0.88) | 2.09 |
| 2 | 7–11 | 0.0047 (1.09) | -0.0485 (-0.81) | 0.75 |
| 3 | 13–17 | 0.0024 (0.76) | -0.0444 (-1.03) | 0.67 |
| 4 | 29–23 | 0.0018 (0.62) | -0.0346 (-0.84) | 0.44 |
| 5 | 24–30 | -0.0008 (-0.19) | -0.0771 (-1.35) | 1.03 |

* Denotes significance at the 1% level.

** Denotes significance at the 5% level.

sion is reached for the survivors data base. Thus, when we remove the *look-ahead bias* from the current COMPUSTAT file, there is no relation between earnings yield and return for any size group. Thus, differences in our findings from those findings of other researchers clearly are attributable to the biases in the data source(s) used by these other researchers.

Given the striking difference in these results, it is clear that the choice of the sample is an important consideration for empirical studies involving accounting data. The next section discusses a number of measures to avoid the biases introduced by the use of the current COMPUSTAT file.

IV. Reducing the Biases in the COMPUSTAT File

Most researchers do not have access to the sequentially collected COMPUSTAT file. We, therefore, examine a number of corrective measures which can be applied to the current COMPUSTAT file to reduce the magnitude of the biases.

The use of the “merged” or “research” version of the current COMPUSTAT file reduces the *ex-post-selection bias*, but may not eliminate it completely. It

includes the non-survivors, but contains data for time periods before a firm was first included on the file. For example, among all firms which begin public trading in a year, only the successful ones will be added to the current COMPUSTAT at some time in the future. If the accounting information of those firms for time periods prior to their inclusion is used in a study, a bias is introduced. This bias is a part of the *ex-post-selection bias* discussed previously. Little can be done to correct for this problem.

A number of researchers have attempted to eliminate the *look-ahead bias* by including in their sample only data of firms with fiscal years ending in December. Then the investment period is chosen from March to December. This procedure assumes that the information in the file, while not available at the end of December, will be known 3 months later and will be fully reflected in the price of the security.

This procedure does hold promise, but there are some problems. First, a portfolio of December fiscal year companies produces significantly different returns (at the one percent level) than does the whole sequentially collected COMPUSTAT file. There are also large differences in the portfolio betas between these two samples. There is less of a difference for the current COMPUSTAT file. The F -statistics for the mean raw differences and the α 's of the entire sample are not significantly different at the one percent level, but the F -statistics for the equality of α 's and β 's are significantly different.

These differences seem to be associated with an industry bias in the December fiscal year companies. Certain industry sectors, such as retailers, are seldom included in this sample.

When we replicate the P/E, size effects experiment using only December fiscal year companies and the current COMPUSTAT file, we cannot reject the hypothesis of no difference between the returns of low and high P/E portfolios with or without risk adjustment. Therefore, using only December fiscal year companies, and forming portfolios based on year-end earnings and March-ending prices produces similar conclusions for the current COMPUSTAT file and the sequentially collected COMPUSTAT file.¹²

This result suggests that studies, the primary focus of which is the determination of the risks and returns of a feasible investment strategy, should limit the firm sample to December fiscal year companies and measure prices as of the end of March. By implication, this method could be extended to using prices dated 3 months after the firm's fiscal year end. This sample construction should reduce or eliminate the *look-ahead bias*.

In studies where the arbitrary exclusion of non-December fiscal year companies cannot be justified, an appeal to the sequentially collected data file seems to be in order.

Finally, studies which test an asset pricing model using accounting data cannot use these suggested remedies, due to the arbitrary exclusion of part of the sample or the arbitrary timing of information assumed by the remedy or both.

¹² These tests cover the period from April 1974 to December 1981 for a sample size of 93. The degrees of freedom of the appropriate F -statistic are reduced accordingly. The supporting data for these tests are available from the authors.

V. Summary and Conclusions

We have examined the relation between the choice of data source for accounting data and the returns to portfolios formed on the basis of this data source. Specifically, we compared portfolios formed on the basis of the current version of the COMPUSTAT files to those formed on the basis of a series of historical COMPUSTAT files. We find significant differences in returns between these two data files when portfolios are formed on the basis of size and earnings yield.

We find that the effect of low price-earnings multiples on return, after adjusting for a size effect, is quite evident when the current COMPUSTAT file is the basic data source. However, an independent "low P/E" effect is not evident in the sequentially collected COMPUSTAT file. Thus, the *ex-post-selection bias* and the *look-ahead bias* appear to create the "low P/E" effect. Using earnings data which are not yet known to investors to form portfolios tends to put high return (positive earnings surprise) companies in low P/E portfolios and low return (negative earnings surprise) companies in high P/E portfolios.

The general implication of this research is that attention must be given to matching the date of assumed equity or debt prices to the date of availability of accounting information which is hypothesized to influence the equity or debt price.

We have shown that, at least for earnings data, using December fiscal year companies along with March prices and March portfolio formulation appears to be an effective method of eliminating the *look-ahead bias*. However, this remedy will not work for studies which require a complete sample of companies or a random sample of companies. This is because it appears that December fiscal year companies are not a random sample of all available firms.

REFERENCES

1. Ray Ball and Phillip Brown. "An Empirical Evaluation of Accounting Income Numbers." *Journal of Accounting Research* 6 (Autumn 1968), 157-78.
2. Rolf W. Banz. "The Relation between Return and Market Value of Common Stocks." *Journal of Financial Economics* 9 (March 1981), 3-18.
3. Sanjoy Basu. "Investment Performance of Common Stocks in Relation to their Price-earnings Ratios: A Test of the Efficient Market Hypothesis." *Journal of Finance* 32 (September 1977), 663-82.
4. ———. "The Relationship between Earnings Yield, Market Value and the Return for NYSE Stocks: Further Evidence." *Journal of Financial Economics* 12 (June 1983), 129-56.
5. Roger G. Ibbotson and Rex A. Sinquefeld. *Stocks, Bonds, Bills and Inflation: The Past and the Future*, Charlottesville, VA.: Financial Analysts Research Foundation, 1982.
6. John W. Peavy, III and David A. Goodman. "The Significance of P/E's for Portfolio Returns." *Journal of Portfolio Management* 10 (Spring 1983), 43-47.
7. Marc R. Reinganum. "Misspecification of Capital Asset Pricing: Anomalies Based on Earnings Yields and Market Values." *Journal of Financial Economics* 9 (March 1981), 19-46.
8. Henri Theil. *Principles of Econometrics*, New York: John Wiley & Sons, Inc., 1971.