ELSEVIER

Journal of Financial Economics 41 (1996) 359–399

JOURNAL OF Financial ECONOMICS

Detecting abnormal operating performance: The empirical power and specification of test statistics

Brad M. Barber*, John D. Lyon

Graduate School of Management, University of California–Davis, Davis, CA 95616, USA

(Received November 1994; final version received November 1995)

Abstract

This research evaluates methods used in event studies that employ accounting-based measures of operating performance. We examine the choice of an accounting-based performance measure, a statistical test, and a model of expected operating performance. We document the impact of these choices on the test statistics designed to detect abnormal operating performance. We find that commonly used research designs yield test statistics that are misspecified in cases where sample firms have performed either unusually well or poorly. In this sampling situation, the test statistics are only well specified when sample firms are matched to control firms of similar pre-event performance.

Key words: Operating performance; Event studies; Return on assets; Return on sales
JEL classification: G30

1. Introduction

Much recent empirical research in accounting and finance focuses on the operating performance of corporations. These studies generally assess operating performance following major corporate events or decisions, such as dividend

*Corresponding author.

We have benefited from the comments of Regina Anctil, Masako Darrough, Jonathan Karpoff, Robert Holthausen, Wayne Mikkelson (the editor), Michael Maher, Mark Nelson, Stephen Penman, Jay Ritter, Ken Shah, the two reviewers, Steve Matsunaga and Steven Kaplan, and seminar participants at UC Berkeley, UC–Davis, and Cornell. All errors are our own.

initiation, stock splits, management buyouts, or security offerings.[1] Operating performance measures are based on accounting numbers and are generally evaluated relative to an industry benchmark. There is considerable variation in the measures of performance and statistical tests that empirical researchers use to detect abnormal operating performance. In addition, little is known about the specification and power of the tests.

We evaluate three choices researchers must make in designing an event study that uses operating performance. First, they need to select a measure of operating performance. Second, they need to determine a benchmark against which to measure actual performance. We refer to this step as developing a model of expected performance. Third, they need to select an appropriate statistical test. We study these three choices by analyzing the operating performance of firms listed on the New York and American Stock Exchanges at any time from 1977 through 1992. Our methods are analogous to those employed by Brown and Warner (1985) in their research on event studies using daily stock return data.

We analyze five different measures of operating performance that researchers might consider in studies of operating performance: return on book value of assets, return on book value of assets adjusted for cash balances, return on sales, return on market value of assets, and a cash-flow-based measure of return on assets. In developing models of expected operating performance, we consider whether it is important to match sample firms to control firms on the basis of a sample firm's industry, size, or past performance. Finally, we evaluate the performance of parametric t-statistics and nonparametric Wilcoxon test statistics in tests designed to detect abnormal operating performance.

We highlight some of our main results in this introduction: In the choice of statistical test, we find that nonparametric Wilcoxon test statistics are uniformly more powerful than parametric t-statistics, regardless of the operating performance measure employed. This result is attributable to the existence of extreme observations in all of our performance measures. Concerning the choice of an expectation model, we find that test statistics using the change in a firm's operating performance relative to an appropriate benchmark consistently yield more powerful test statistics than do those based on the level of a firm's operating performance relative to the same benchmark. In random samples or samples of large firms, all expectation models based on changes in a firm's performance relative to an industry benchmark are well specified and powerful.

[1] A more extensive, though not exhaustive, list of these studies is provided in Table 1. Other recent research that considers the operating performance of firms includes Loughran and Ritter (1994) who study seasoned equity offerings, Strickland, Wiles, and Zenner (1994) who study intervention by the United Shareholders Association, Mulherin and Poulsen (1994) who study proxy contests, Jain and Kini (1994) who study initial public offerings, and Denis and Denis (1995) who study leveraged recapitalizations.

Perhaps the most important result documented here is that when sample firms experience pre-event performance that is even slightly different from control firms, commonly used methods – for example, matching sample firms to control firms on industry, or industry and size – yield test statistics that are misspecified. Test statistics are well specified only when sample firms are matched to control firms with similar pre-event performance. We attribute this misspecification to the tendency for accounting-based measures of performance to mean-revert over time. Matching sample firms to control firms on industry and performance is generally much more important than matching on industry alone, or on industry and size.

The paper is organized as follows. We discuss models of expected performance in Section 2. The data set employed is introduced in Section 3. Our statistical tests are defined in Section 4. Results are presented in Section 5. The discussion in the paper focuses on results using operating income scaled by average book value of total assets. Alternative measures of performance are discussed and analyzed in Section 6. Alternative methods of modeling expected performance are discussed in Section 7. We close the paper with specific recommendations on the choice of performance benchmark, performance measure, and statistical test.

2. Modeling expected performance

In this section, we are interested first in identifying an appropriate measure of operating performance, and second in identifying an appropriate method for detecting abnormal operating performance. In Table 1, we summarize many of the recent studies of changes in operating performance that follow major corporate events. In addition to the authors and corporate events studied, we identify the performance measure and benchmark used in each study. When explicitly defined by the authors, we provide the Compustat data items used in each of the studies.

2.1. Measuring operating performance

While early studies focused on changes in earnings per share, recent studies tend to employ operating income as a performance measure. Earnings per share (most often Compustat data item 58) represents the per-share income of a company after all expenses. It includes interest expense, special items, income taxes, and minority interest, but excludes income from discontinued operations or extraordinary items. Operating income (most often Compustat data item 13) is defined as sales less cost of goods sold, and selling, general, and administrative expenses. Thus, the major difference between the two performance measures is

Table 1
Summary of studies analyzing corporate performance following corporate events or decisions

Author(s)	Corporate event studied	Performance measure (Compustat data item)	Adjustment method
Healy and Palepu (1988)	Dividend initiation/omission	$\frac{\Delta EPS\ (A58)}{P}$	Industry
Asquith, Healy, and Palepu (1989)	Stock splits	$\frac{\Delta EPS\ (Q19)}{P}$	Industry
Kaplan (1989)	Management buyouts	$\frac{Op\ Inc\ (A13)}{Sales}$; $\frac{Op\ Inc\ (A13)}{BV\ Assets}$; $\frac{Op\ Inc\ (A13) - Cap\ Exp\ (A128)}{Sales}$; $\frac{Op\ Inc\ (A13) - Cap\ Exp\ (A128)}{BV\ Assets}$	Industry, size
Healy and Palepu (1990)	Seasoned equity offers	$\frac{\Delta EPS\ (A58\&Q19)}{P}$	Industry
Dann, Masulis, and Mayers (1991)	Repurchases	$\frac{FE^{a}}{P}$; $\frac{FE}{\sigma(FE)^{b}}$	Industry, size

Denis and Denis (1993)	Leveraged recapitalizations	$\frac{Op\ Inc\ (A13)}{BV\ Assets\ (A6)}; \frac{Free\ Cash\ Flow^{c}}{BV\ Assets\ (A6)}$	Industry, size
DeGeorge and Zeckhauser (1993)	Reverse LBO	$\frac{Op\ Inc\ (10K's)}{BV\ Assets}; \frac{Op\ Inc\ (10K's)}{BV\ Assets}$	Industry, size
Mikkelson and Partch (1994)	Dual class recapitalizations and ESOP adoption	$\frac{Op\ Inc\ (A13)}{BV\ Assets}$	Industry
Mikkelson and Shah (1994)	Initial public offerings	$\frac{Op\ Inc\ (A13)}{Assets}$	Industry, performance
Healy, Palepu, and Ruback (1994)	Takeovers	$\frac{Op\ Cash\ Flow^{d}}{MV\ Equity^{e} + BV\ Debt}$	Industry
Holthausen and Larcker (1994)	Reverse LBO	$\frac{Op\ Inc\ (A13)}{BV\ Assets\ (A6)}; \frac{Op\ Cash\ Flow^{f}}{BV\ Assets\ (A6)}$	Industry, performance

[a]Forecast errors (FE) derived from annual (EBIT) and quarterly (EPS) earnings time series models.
[b]Standard deviation of pre-event forecast errors (FE).
[c]Free cash flow = operating income (A13) − (taxes (A16) + interest (A15) + prefd div (A19) + common divs (A21)).
[d]Operating cash flow = sales − COGS − sell and admin. + depn. + goodwill amort.
[e]Market value of equity excludes change in value from five days before first offer to delisting.
[f]See Section 6 for a full description.

that operating income excludes interest expense, special items, income taxes, and minority interest.

We favor the use of operating income over earnings for two reasons. First, since operating performance can be obscured by special items, tax considerations, or the accounting for minority interests, we argue that operating income is a cleaner measure than earnings of the productivity of operating assets. Second, researchers often study corporate events that result in changes in capital structure (for example, leveraged recapitalizations). Such changes affect interest expense and, consequently, earnings net of interest expense, but leave operating income unaffected (assuming the capital structure changes did not affect the firm's operations). In addition, we prefer to use unscaled operating income, rather than an income per share measure, because corporate events that a researcher might wish to study often result in changes in the number of shares outstanding (for example, equity issuance or stock splits).

To compare performance across firms, operating income must be scaled. We are interested in measuring the productivity of operating assets in place for a group of sample firms. The guiding principle that we use in this research is that to generate a performance measure, operating income in period t should be matched with the operating assets in place in period t. Consequently, we want to scale the operating income in period t by the period t value of operating assets. Unfortunately, the current value of operating assets is not reported in financial statements. As an alternative, we use the book value of total assets (Compustat item 6) to derive our major results. We divide operating income by the average of beginning- and ending-period book value of total assets, which we call 'return on assets' (ROA). This is the measure of operating performance most commonly used by the studies summarized in Table 1. Though many of the studies use end-of-period assets, when we scale operating income by end-of-period assets, the general tenor of our conclusions is unaffected. In Section 6, we evaluate several alternative measures of operating performance that a researcher might consider.

2.2. *Expected performance*

To assess whether a firm is performing unusually well or poorly, we must specify the performance we expect in the absence of an event, thus providing a benchmark against which sample firms can be compared. Note that the pre-event characteristics of firms can lead researchers to expect that sample firms will experience above(below)-average operating performance, even before they consider the impact of the event under consideration. For example, if certain industries have experienced unusual growth in ROA during the sample period, it might be reasonable to expect the sample firms in those industries to experience a similar growth in ROA. The studies we reviewed usually employ one (or more) of four different approaches to measuring expected performance.

Generally, firms in the sample are compared to firms with the same

1. two-digit SIC code,
2. four-digit SIC code,
3. two-digit SIC code and similar size,
4. two-digit SIC code and similar pre-event performance.

We refer to these four comparison groups as two-digit matched, four-digit matched, size-matched, and performance-matched and define them in Section 3.

Industry-matching assumes that some of the cross-sectional variation in operating performance can be explained by an industry benchmark. The two methods of industry-matching most often used match sample firms to other firms with either the same two-digit or the same four-digit SIC code. Obviously, the tradeoff in choosing either two-digit or four-digit matching is that researchers must either include more firms in the control group (two-digit matching), or include fewer firms, but firms that are more closely matched on industry to sample firms (four-digit matching).

The third method of developing a control group matches sample firms to other similar-size firms with the same two-digit SIC code. This method implicitly assumes that operating performance varies by industry and firm size. Recent research by Fama and French (1995) documents that small firms, on average, have lower earnings scaled by book value of equity than do larger firms. Several recent studies of operating performance have matched sample firms to similar-size firms in the same industry (for example, Kaplan, 1989; Denis and Denis, 1993; Dann, Masulis, and Mayers, 1991; DeGeorge and Zeckhauser, 1993).

The last method of developing a control group that we consider matches sample firms to other firms with the same two-digit SIC code and similar pre-event performance. Performance matching adjusts for the mean reversion in accounting data[2] that reflects a transitory component of operating income. The transitory component can be attributed to accounting methods, such as the manipulation of accounting numbers or the one-time effects of accounting changes, as well as underlying economics forces, such as nonrecurring income or expenses, or temporary shifts in product demand.

The temporary component to operating income can confound analyses of operating performance. If there is a high level of operating income for a particular firm, there is likely a temporary component to its operating income. Over time, the return on assets reverts toward a population mean as the temporary component dissipates. In short, if a firm performs well before an event, the

[2]Both Penman (1991) and Fama and French (1995) document that return on equity measures are slowly mean-reverting.

tendency for mean reversion might lead a researcher to conclude that the firm subsequently experiences poor performance, when in fact the accounting measure of performance is merely reverting to its mean in a predictable fashion. By matching sample firms to firms with similar performance before an event, we are able to control for the mean-reversion tendency of a performance measure.

It is also possible that some firms experience high or low measures of performance because of corporate strategy, managerial ability, or the nature of investment opportunities. By matching on performance, a researcher can control for various factors, unrelated to an event, that affect the operating performance of assets.

More formally, we denote P_{it} as the performance of firm i in year t. The industry comparison group for firm i in year t is PI^j_{it}. The superscript indexes the different definitions of industry comparison group enumerated above, $j = 1, 4$. Thus, the first four models of expected performance are

$$\mathrm{E}(P_{it}) = PI^j_{it}, \qquad j = 1, 4, \tag{1}$$

where $\mathrm{E}(\cdot)$ is an expectations operator.

One drawback to using the level of an industry comparison group to measure expected performance (without any pre-event performance matching) is that it ignores the history of the firm relative to the benchmark. Consider a firm that has enjoyed an unusually high *ROA* relative to its group of comparison firms (perhaps as a result of investment in unusually profitable projects). If these projects continue to earn above-average profits after an event, this firm would appear to have operating performance that exceeds the performance expected in the absence of the event.

One means of alleviating this problem is to consider the history of a firm's performance relative to its comparison group's performance. Typically, researchers have compared each firm's performance relative to an industry benchmark pre-event $(P_{i,t-1} - PI^j_{i,t-1})$ to the same performance measure post-event $(P_{it} - PI^j_{it})$. Conclusions are then based on the changes in the sample firms' performance relative to changes in the industry benchmark, $(P_{it} - P_{i,t-1}) - (PI^j_{it} - PI^j_{i,t-1})$.

To be more explicit about the assumptions underlying these comparisons, we restate this method in terms of what it implies about a firm's expected performance. The comparison between changes in performance states that a firm's expected performance is equal to its past performance plus the change in the industry's performance:

$$\mathrm{E}(P_{it}) = P_{i,t-1} + (PI^j_{it} - PI^j_{i,t-1}) \tag{2}$$

$$= P_{i,t-1} + \Delta PI^j_{it}, \qquad j = 1, 4. \tag{3}$$

This formulation provides four additional models of expected performance.

Table 2
Models of expected operating performance

Model	Expected performance model	Description	Industry comparison group
1	PI_{it}^{1}	Level of industry performance	Two-digit SIC matched
2	PI_{it}^{2}	Level of industry performance	Four-digit SIC matched
3	PI_{it}^{3}	Level of industry performance	Two-digit SIC and size-matched
4	PI_{it}^{4}	Level of industry performance	Two-digit SIC and performance-matched
5	$P_{i,t-1} + \Delta PI_{it}^{1}$	Lagged firm performance and change in industry perf.	Two-digit SIC matched
6	$P_{i,t-1} + \Delta PI_{it}^{2}$	Lagged firm performance and change in industry perf.	Four-digit SIC matched
7	$P_{i,t-1} + \Delta PI_{it}^{3}$	Lagged firm performance and change in industry perf.	Two-digit SIC and size-matched
8	$P_{i,t-1} + \Delta PI_{it}^{4}$	Lagged firm performance and change in industry perf.	Two-digit SIC and performance-matched
9	$P_{i,t-1}$	Lagged firm performance	—

The ninth model that we consider ignores the performance of comparison firms and assumes that expected performance is simply a firm's own past performance:

$$\mathrm{E}(P_{it}) = P_{i,t-1}\,. \tag{4}$$

The nine models are summarized in Table 2. Though all models are stated in terms of the level of a particular firm's performance (P_{it}), models 5 through 9 are equivalent to an analysis of the changes in a particular firm's performance ($P_{it} - P_{i,t-1}$). We refer to these models as 'change' models, and to models 1 through 4 as 'level' models.

3. Data

3.1. Sample composition

Our analysis includes all NYSE/AMEX firms with data available on Compustat. Firms that change their fiscal year during the sample period are excluded from the analysis in the year in which the change occurs. The sample period extends from 1977 through 1992.

Table 3
Descriptive statistics on return on assets for NYSE/ASE firms: 1977–1992

Return on assets is measured as operating income (item 13) scaled by the average of beginning-of-period and end-of-period book value of assets (item 6). Descriptive statistics are based on winsorized data. All observations are winsorized at the first and 99th percentiles, based on all firm-year observations. These values are − 18.9% and 46.2%, respectively.

	Return on assets (%)					
Year	Mean	25th p'tile	Median	75th p'tile	Std. dev.	Obs.
1977	16.7	10.7	16.1	22.5	9.5	1,918
1978	17.1	11.1	16.6	22.7	9.6	2,227
1979	17.3	11.1	16.6	23.0	9.7	2,149
1980	16.5	10.6	15.7	21.9	9.9	2,083
1981	15.9	10.3	15.1	21.2	9.5	2,035
1982	13.8	8.4	13.7	19.3	9.4	1,997
1983	13.7	8.4	13.6	19.3	9.6	1,982
1984	14.9	9.7	15.0	20.2	9.8	1,950
1985	13.7	8.3	13.5	19.1	10.2	1,941
1986	13.0	7.4	12.7	18.2	10.6	1,955
1987	13.5	8.2	13.4	18.4	10.4	2,011
1988	13.3	7.9	13.3	18.8	10.6	2,045
1989	13.1	7.7	12.9	18.2	10.2	2,050
1990	12.4	7.6	12.4	17.6	10.3	2,081
1991	11.4	6.3	11.8	16.7	10.4	2,121
1992	12.0	7.1	12.0	17.1	10.4	2,135
1977–92	14.3	8.8	13.9	19.7	10.2	32,680

Descriptive statistics on the return on assets are presented in Table 3. The descriptive statistics are based on winsorized data, since a few extreme observations skew the mean and standard deviation in some sample years. Winsorizing is performed by setting the observations below the first and above the 99th percentile of the distribution to the values at the first and 99th percentiles. The winsorizing is performed at the first percentile (− 18.9%) and 99th percentiles (46.2%) of the distribution of *ROA*.

The distribution of *ROA* is reasonably symmetric: The mean and median statistics are approximately equal. Mean and median *ROA* declines during the sample period, while the *ROA* cross-sectional standard deviation increases.

Our analysis includes financial firms and utilities. We reestimate all our results after excluding financial firms (SIC codes 6000–6799) and utilities (SIC codes 4900–4999). Over our sample period, mean and median *ROA* for the remaining firms are 14.8% and 15.0%, respectively, with a cross-sectional standard deviation of 10.4%. The general tenor of our results is unaffected by the exclusion of financial firms and utilities.

3.2. Defining industry comparison groups

We define four industry comparison groups for firm i in year t, based on the expectation models developed in the prior section. In all four cases, we use the median performance of the industry comparison group as our industry performance measure, PI_{it}^{j}. When change models are employed, we use the change in the median industry performance, $PI_{it}^{j} - PI_{i,t-1}^{j}$. We also estimate all the results presented in this paper using the median change in industry performance in lieu of the change in the median industry performance. The results using this alternative specification are virtually identical to those reported.

In all of the change models and also in the performance-matched level model, we hold the industry comparison group constant over time. Since sample firms must have data available in periods t and $t-1$, holding the industry comparison group constant over time places the same data requirement on control firms.

The first comparison group, *two-digit matched*, includes all firms in the same two-digit SIC code as firm i in year t, excluding firm i. We note here that we use Compustat SIC codes throughout this analysis. This, in fact, is an important issue. For example, the agreement of SIC code classifications between Compustat and CRSP at the four-digit level in a random sample of 676 firms was only 28%. The agreement at the two-digit level is 64.1%. These issues are addressed at length by Guenther and Rosman (1994). Though we are not entirely comfortable with the use of SIC codes to define industry groups, we know of no practical alternative to their use. Kahle and Walkling (1995) analyze the differences between CRSP and Compustat SIC classifications in detail.

The second comparison group, *four-digit matched*, includes all firms in the same four-digit SIC code as firm i in year t, excluding firm i. Approximately 1.8% of all firms have no other firm in their four-digit SIC code; for these, we use an alternative rule in which we match using three-digit SIC codes, and finally two-digit SIC codes.

The third comparison group, *size-matched*, includes all firms in the same two-digit SIC code as firm i in year t and similar in size to firm i. We note that when the size-matched model in the levels is employed, the size matching is performed in period t. When we use the size-matched model in changes, the size matching is performed in period $t-1$ so that the industry comparison group can be held constant over time. Thus, this benchmark is similar to our first, except that we require the comparison firms to be similar in size to the firm in question. Size is measured as the book value of assets. Firm i is matched to other firms with the same two-digit SIC code, and with book value of total assets within 70%–130% of firm i's. When firms have no firm of similar size with the same two-digit SIC code, we use an alternative rule where we find the firm with the same two-digit SIC code and of closest size to the firm in question.

We experimented with several alternative size filters (both tighter and looser). Size matching proves to be important only when firms are drawn from the

smallest third of firm size and the top third of performance (measured by return on assets). The 70%–130% size filter was selected because it yields test statistics that are well-specified in this sampling situation.

The fourth comparison group, *performance-matched*, includes all firms in the same two-digit SIC code as firm i in year t and similar in performance to firm i in year $t-1$. Thus, this benchmark is similar to our first, except that we require the comparison firms to have similar performance to the sample firm in year $t-1$. Firm i is matched to other firms with the same two-digit SIC code, and with return on assets within 90%–110% of firm i's in year $t-1$. Again, we experimented with several alternative performance filters (both tighter and looser). Performance matching is important when sample firms have historically performed either well or poorly. The 90%–110% performance filter is selected because it yields test statistics that are well-specified in these sampling situations.

When firms have no firm of similar performance in year $t-1$ with the same two-digit SIC code, we use an alternative rule with three steps. First, we attempt to match performance within the 90%–110% filter, using all firms in the same one-digit SIC code. If we still find no performance match, then we try to match performance within the 90%–110% filter using all firms without regard to SIC code. If we still find no performance match, our third step is to use the firm with performance closest to the firm in question, without regard to SIC code. (We considered, but abandoned, several variations of this alternative rule, for example, matching the firm in question with the firm in the same two-digit SIC code and closest in performance. See Section 5.)

The matching characteristics of our sample firms are summarized in Table 4. For each firm-year observation in our sample, we determine the number of firms that form a comparison group based on the four criteria that we have developed. Note the row labeled Alt., or 'alternative rule', in this table. This row represents the number of observations for which we were forced to match using the alternative rules described above. For example, in the case of four-digit matching, we were forced to match at the three- or two-digit level for 593 firms, or 0.8% of firms. Using this method, nearly 65% of all observations have an industry comparison group that is greater than five firms. Using the size-matched method forces us to match sample firms that have no available match within the prescribed 70%–130% size filter with another firm with the same two-digit SIC code and closest in size. This affects 8.7% of all firms. Using the performance-matched method, we are unable to identify a firm with the same two-digit SIC code and within the 90%–110% performance filter for 16.5% of all firms. Ultimately, however, using the 90%–110% filter, we are unable to match the performance of only 243 firms.

While our alternative rules are empirically, rather than theoretically, based, we use them for two reasons. First, without the use of some alternative rule for matching, researchers are forced to discard any firms that have no available

Table 4
Number and percentage of firms with available matching firms(s) based on various matching criteria: 1977–1992

Four matching criteria are considered for firm i in a given calendar year. First, match with all firms in the same two-digit SIC code. Second, match with all firms in the same four-digit SIC code. Third, match with all firms in the same two-digit size code and between 70%–130% of firm i's size (measured as the book value of total assets) in year t. Fourth, match with all firms in the same two-digit SIC code and between 90%–110% of firm i's performance in year $t-1$. See Table 3 for a description of the calculation of return on assets (ROA).

	Matching criteria							
	Two-digit SIC		Four-digit SIC		Two-digit and size		Two-digit and performance	
Number of matching firms	Obser-vations	% of all obs.	Obser-vations	% of all obs.	Obser-vations	% of all obs.	Obser-vations	% of all obs.
0	26	0.1	26	0.1	26	0.1	1,541[b]	4.7
1	0	0.0	1,822	5.6	2,756	8.4	3,288	10.1
2	38	0.1	2,359	7.2	2,896	8.9	2,539	7.8
3	60	0.2	2,039	6.2	3,027	9.3	2,207	6.8
4	216	0.7	2,349	7.2	2,666	8.2	1,869	5.7
5	290	0.9	2,311	7.1	2,367	7.2	1,617	4.9
> 5	32,050	98.1	21,181	64.8	16,084	49.2	14,229	43.5
Alt.[a]	n.a.	n.a.	593	1.8	2,858	8.7	5,390[c]	16.5
All obs.	32,680	100.0	32,680	100.0	32,680	100.0	32,680	100.0

n.a. = not applicable

[a]This row represents the number of firms in which an alternative matching rule was used to find a comparison group. For example, for the four-digit matched method, 593 firms were matched at the three- or two-digit level. These alternative matching rules are described in detail in the text.

[b]The 1,541 firms without an available match using the performance-matched method result from firms that do not have return on assets reported in $t-1$.

[c]Of these 5,390 firm-year observations, 1,682 were matched with firms of similar performance in the same one-digit SIC code, 3,465 were matched with firms of similar performance without regard to SIC code, and 243 were matched to the firm closest in performance.

match. This exclusion of some sample firms can lead to biases in test statistics; discarded firms tend to be unusually small (using the size-matched method) or have historically good or poor performance (using the performance-matched method). Second, without the use of some alternative rule, we cannot compare the power of test statistics across different models of expected performance, since the populations from which sample firms are drawn are dramatically different. The alternative rules allow us to keep constant the populations from which sample firms are drawn.

3.3. The explanatory power of the expected performance models

We investigate the explanatory power of the nine models of expected performance summarized in Table 2 by estimating nine cross-sectional regressions in each year. In this study, a year is considered a calendar year. Thus, firms with a fiscal year-end in January 1990 through December 1990 are considered observations for 1990.

The dependent variable in these regressions is ROA. The independent variable is either the level of the industry benchmark (models 1 through 4) or a firm's lagged performance and the change in the industry benchmark (models 5 through 9). The regressions are estimated using winsorized data for both dependent and independent variables. Extreme observations on both dependent and independent variables lead to coefficient estimates that are extreme in some years. These extreme coefficient estimates disappear when the winsorized data are employed in the regressions. The results are similar when the population is trimmed (as opposed to winsorized) at the first and 99th percentiles. The independent variables in each regression correspond to each of the nine models presented in Table 2. To evaluate the statistical significance of each of the nine models, we calculate the mean coefficient estimates and mean adjusted R^2 across the 16 annual regressions. These results are presented in Table 5.

Three noteworthy results emerge from this table. First, all nine models yield intercepts that differ from zero, and slope coefficient estimates that are less than one. We reject the null hypothesis (that each of the mean slope coefficient estimates presented in Table 5 is equal to one) at the 1% significance level in all cases. These results indicate that the expectation models yield biased forecasts of performance. If unbiased, the intercept term would be zero and slope coefficients equal to one. Second, of the four models that employ only the contemporaneous median industry performance as an explanatory variable (models 1 through 4), the model with the most explanatory power uses performance-based matching (model 4). Third, adding the change in a firm's industry performance to its lagged performance (models 5 through 8) yields only a marginal improvement in explanatory power over a model that employs only a firm's lagged performance (model 9) or performance-based matching (model 4).

Based on these results, we cannot disqualify any of the proposed models as a candidate for detecting abnormal operating performance. All nine models have significant explanatory power. The only model that could be eliminated from contention at this stage is the four-digit industry match. Matching on four-digit SIC codes in lieu of two-digit SIC codes provides no improvement in the explanatory power of regressions. Nonetheless, for the sake of completeness, we evaluate the performance of this model in the tests that follow.

Table 5
Mean coefficient estimates and adjusted R^2 from cross-sectional regressions of return on assets on various predictors of return on assets by year: 1977–1992

For each year, nine cross-sectional regressions of return on assets on various independent variables are estimated. Four different industry performance measures are considered for the ith firm: (1) PI_{it}^1 represents the median performance of firms in the same two-digit SIC code, (2) PI_{it}^2 represents the median performance of firms in the same four-digit SIC code, (3) PI_{it}^3 represents the median performance of firms in the same two-digit SIC code and between 70%–130% of Firm i's size, and (4) PI_{it}^4 represents the median performance of firms in the same two-digit SIC code and between 90%–110% of Firm i's return on assets. The independent variables employed in each of the nine regressions differ based on the industry benchmark used and whether lags of performance are employed. The mean coefficient estimates across the 16 years is then calculated. Test statistics are based on the time-series standard deviation of the coefficient estimates.

	Mean coefficient estimates on:										
Model	Inter.	$P_{i,t-1}$	PI_{it}^1	PI_{it}^2	PI_{it}^3	PI_{it}^4	ΔPI_{it}^1	ΔPI_{it}^2	ΔPI_{it}^3	ΔPI_{it}^4	Mean adj. R^2
1	0.042**	—	0.71**	—	—	—	—	—	—	—	0.07
2	0.083**	—	—	0.41**	—	—	—	—	—	—	0.07
3	0.104**	—	—	—	0.27**	—	—	—	—	—	0.04
4	0.025**	—	—	—	—	0.81**	—	—	—	—	0.52
5	0.034**	0.74**	—	—	—	—	0.30**	—	—	—	0.59
6	0.033**	0.75**	—	—	—	—	—	0.16**	—	—	0.59
7	0.033**	0.74**	—	—	—	—	—	—	0.05*	—	0.59
8	0.022	0.82**	—	—	—	—	—	—	—	0.13**	0.64
9	0.026**	0.79**	—	—	—		—	—	—	—	0.63

Significant at the 5% (*) and 1% (**) levels, two-sided test.

4. Statistical tests for abnormal operating performance

The abnormal performance of firm i in year t, AP_{it}, is defined as realized performance, P_{it}, less expected performance, $\mathrm{E}(P_{it})$:

$$AP_{it} = P_{it} - \mathrm{E}(P_{it}), \tag{5}$$

where performance is measured using return on assets and expected performance is based on one of the nine models discussed earlier. To test the null hypothesis, in which mean abnormal performance is equal to zero for a sample of size n, we employ a parametric test statistic:

$$t = \frac{\overline{AP}}{\sigma(AP_{it})/\sqrt{n}}, \tag{6}$$

where $\overline{AP}$ is the sample average and $\sigma(AP_{it})$ is the cross-sectional sample standard deviation of abnormal performance for the sample of n firms. This test

statistic follows a student's t-distribution under the null hypothesis if the sample is drawn randomly from a normal distribution. While we can reject the hypothesis that our measures of abnormal performance follow a normal distribution,[3] it remains an empirical question whether this test statistic is well specified and/or powerful.[4]

We also consider a nonparametric Wilcoxon signed-rank test statistic, which we denote T^*. The Wilcoxon signed-rank test statistic tests the null hypothesis that the median abnormal performance is equal to zero. We use the IMSL SNRNK subroutine to compute the Wilcoxon signed-rank test statistic and the associated p-values. In this subroutine, if rankings result in a tie, the average ranking of the tied observations is used.

To test the specification of the two test statistics for each of the nine expectation models, 1,000 size n random samples are drawn without replacement.[5] For each of the 1,000 random samples, the test statistics are computed as described above and compared to the critical value of the test statistic associated with the two-tailed α significance level. If a test is well specified, $1{,}000\alpha$ tests will reject the null hypothesis of no abnormal operating performance. A test is conservative if fewer than $1{,}000\alpha$ null hypotheses are rejected; a test is anticonservative if more than $1{,}000\alpha$ null hypotheses are rejected. Based on this procedure, we test the specification of each test statistic at the 1%, 5%, and 10% theoretical level of significance.

To test the power of the statistical tests, we induce a constant level of abnormal performance into every observation in each of the 1,000 random samples. On average, a powerful test is able to detect small induced levels of abnormal operating performance. We increment the induced level of abnormal return on assets by 0.01 when estimating the empirical power functions. This is equivalent to a one-cent improvement in operating performance per dollar of assets. The empirical power function can be estimated by varying the induced level of abnormal operating performance and calculating the proportion of samples that reject the null hypothesis. We estimate the power function of each test statistic at the 5% theoretical level of significance.

---

[3]Kolmogorov D-statistics allow us to reject the null hypothesis that each performance measure within a calendar year follows a normal distribution at the 1% significance level.

[4]The *Central Limit Theorem* guarantees that if the measures of abnormal performance in the cross-section of firms are independent and identically distributed drawings from finite variance distributions, the distribution of the mean abnormal performance measure converges to normality as the number of firms in the sample increase. We suspect that the assumption of identical distributions is likely violated in the measures that we employ.

[5]All our results were also estimated using samples with replacement. The results of this analysis are virtually identical to those that we report.

5. Results

5.1. Random samples

The first set of results is based on 1,000 random samples of 50 firm-years drawn from our population of over 30,000 firm-year observations. All test statistics are based on the unwinsorized data set. The specification of the various test statistics is presented in Table 6. Two observations emerge from this analysis. First, parametric t-statistics are consistently more conservative than the Wilcoxon T^*. This conservatism is a result of extreme observations in the first and 99th percentile of the ROA distribution. When these observations are winsorized, the conservatism of the parametric test statistic disappears. Second, all the tests are well specified except for those that use only a firm's change in performance (model 9). Thus, in random samples, the selection of the test statistic and expectation model does not significantly affect the specification of the test statistic.

Table 7 presents the power of the various test statistics. In random samples, recall that only test statistics using the change in a firm's performance (model 9)

Table 6
Specification (size) in random samples; percentage of 1,000 random samples of 50 firms (1977–1992) rejecting the null of no abnormal operating performance at the 1%, 5%, and 10% theoretical significance level

The numbers presented in the body of this table represent the percentage of 1,000 random samples of 50 firms that reject the null hypothesis of no abnormal operating performance at a theoretical significance level of 1%, 5%, and 10%. The model numbers (1 through 9) correspond to those presented in Table 2.

	Parametric t-statistic			Nonparametric Wilcoxon T^*		
Theoretical significance level:	1%	5%	10%	1%	5%	10%
Model no.: Description of expected performance model						
1: Two-digit matched	0.4	3.3	8.0	0.8	4.1	9.7
2: Four-digit matched	0.5	4.0	8.3	0.6	4.6	9.6
3: Two-digit and size-matched	0.6	3.7	7.5	0.7	4.4	10.7
4: Two-digit and performance-matched	0.4	3.0	9.1	1.2	4.5	9.8
5: Lagged ROA and Δ two-digit matched	0.1	3.5	9.6	0.4	4.9	10.4
6: Lagged ROA and Δ four-digit matched	0.4	3.4	9.2	1.2	4.8	10.4
7: Lagged ROA and Δ two-digit and size-matched	0.2	3.6	7.7	0.6	4.7	9.4
8: Lagged ROA and Δ two-digit and perf.-matched	0.5	3.4	8.4	1.1	4.6	9.6
9: Lagged ROA	0.9	5.0	12.3*	1.1	6.5*	11.9*

*Significantly different from the theoretical significance level at the 5% level, one-sided binomial test statistic.

Table 7
Power in random samples; percentage of 1,000 random samples of 50 firms (1977–1992) with induced abnormal performance ranging from −0.02 to 0.02 rejecting null hypothesis of no abnormal operating performance at 5% theoretical significance level

The numbers presented in the body of this table represent the percentage of 1,000 random samples that reject the null hypothesis of no abnormal operating performance at the theoretical significance level of 5% and various levels of induced abnormal operating performance. Abnormal operating performance is induced by adding a constant to the observed performance for each of the 50 randomly selected firms in all 1,000 random samples. The model numbers (1 through 9) correspond to those presented in Table 2.

	Parametric *t*-statistic				Nonparametric Wilcoxon *T**			
Induced level of abnormal operating performance:	−0.02	−0.01	0.01	0.02	−0.02	−0.01	0.01	0.02
Model no.: Description of expected performance model								
1: Two-digit matched	26.1	9.8	8.9	26.3	38.1	14.5	15.7	42.7
2: Four-digit matched	25.1	10.5	8.1	21.6	35.3	12.9	13.4	36.8
3: Two-digit and size-matched	22.4	7.2	8.1	21.9	34.2	11.8	12.0	34.2
4: Two-digit and performance-matched	57.9	22.8	10.1	41.1	83.1	36.1	23.4	73.7
5: Lagged *ROA* and Δ two-digit matched	55.8	20.8	15.1	49.5	85.3	36.3	33.3	83.5
6: Lagged *ROA* and Δ four-digit matched	46.1	15.9	13.5	43.7	74.0	28.2	28.3	76.0
7: Lagged *ROA* and Δ two-digit/size-matched	42.7	14.6	12.6	40.6	70.4	24.8	25.3	73.7
8: Lagged *ROA* and Δ two-digit/perf.-matched	56.7	22.0	10.2	42.7	82.5	35.9	25.2	74.9
9: Lagged *ROA*	66.4	33.1	9.2	38.2†	93.6	55.6	20.1	73.5†

†Empirical tests based on this statistic are anticonservative at the 1%, 5%, and/or 10% theoretical significance level (see Table 6).

are misspecified. Three observations emerge from this analysis. First, the non-parametric tests are more powerful than the parametric tests. Second, almost all of the power functions are reasonably symmetric. Third, expectations models that incorporate a firm's past performance are more powerful than those that ignore past performance.[6]

Researchers often analyze the operating performance of firms following corporate events not only for the year following the event, but also for the second and third years. The models that are most powerful in detecting abnormal performance use a firm's lagged performance in forming a measure of expected performance. Thus, we analyze the impact of using successively longer lags on the specification and power of test statistics for these models. Expectation models 5 through 9 are altered as follows:

$$\mathrm{E}(P_{it}) = P_{i,t-k}\,, \tag{7}$$

$$\mathrm{E}(P_{it}) = P_{i,t-k} + (PI^{j}_{i,t} - PI^{j}_{i,t-k}) \quad \text{for} \quad k = 2,3\,. \tag{8}$$

Thus, expected performance is based on a firm's performance lagged by two or three years. In addition, the performance-matched method in the levels (model 4) is altered because performance is matched on the basis of performance lagged by two or three years, rather than one year. Our analysis of the specification (not reported in a table), reveals that models 4 through 8 are well specified at conventional significance levels. These results also reveal that the power of the test statistics erodes as the expectation model moves from using a firm's lagged performance at one year to using that lagged by three years.

In auxiliary analyses (not reported in a table), we analyze the specification and power of test statistics in which the industry comparison group is allowed to vary from period $t-1$ to t. In these auxiliary analyses, firms are allowed to depart or enter the industry comparison group. Allowing the industry comparison group to vary over time does not alter the specification of test statistics. Furthermore, allowing the industry comparison group to vary over time does not affect the power of test statistics when the expectation models use a firm's performance lagged by one year. However, when a firm's performance is lagged over two or three years, allowing the industry comparison group to vary over time noticeably erodes the power of the test statistics considered here. Therefore, we recommend that researchers hold the industry comparison group constant over time, for two reasons: First, as previously discussed, holding the industry comparison group constant places the same data requirements on sample and

[6]Model 4 – the performance-matched method – implicitly uses the change in a firm's performance since firm's are matched in period $t-1$ with firm's of similar performance. If this matching is done well, the results, using the levels and changes in performance, should be similar. Hereafter in this paper, we group the performance-matched level model with the change models (5 through 9).

control firms. Second, the power of test statistics based on a constant industry comparison group are more powerful at successively longer lags.

Based on these results, using random samples, we recommend the use of: (1) a nonparametric test statistic, (2) an expectation model that uses a firm's past performance (models 4 through 8), and (3) a constant industry comparison group in these models. The dominance of the Wilcoxon T^* over t-statistics and change models over level models appears in all of the analyses that follow. Therefore, for the analyses that follow, we do not report the specification or power of t-statistics or models that do not consider a firm's past performance (models 1 through 3).

5.2. *Performance-based samples*

Researchers are often interested in assessing the operating performance of a sample of firms that, as a group, historically experience especially poor or good performance. For example, firms that initiate a dividend, go public, issue equity, or repurchases shares through a tender offer often do so following a period of unusually good earnings (see Healy and Palepu, 1988; Mikkelson and Shah, 1994; Loughran and Ritter, 1995; Dann, Masulis, and Mayers, 1991, respectively). To assess the specification and power of the various test statistics when a sample consists of firms that have unusually good or poor recent performance, we use the following procedure: First, we rank all firms within a calendar year on the basis of return on assets. Second, we draw 1,000 samples of 50 firms from the lowest *ROA* decile. Third, we assess the specification of the various test statistics (as was done for random samples in Table 6). Finally, we assess the power of the various test statistics (as was done for random samples in Table 7). Separately, we also analyze 1,000 samples of 50 firms from the highest *ROA* decile. To be ranked in the bottom decile of *ROA*, a firm must have an *ROA* less than 0.2% in 1992 to 6.9% in 1980. To be ranked in the top decile of *ROA*, a firm must have an *ROA* greater than 22.5% in 1992 to 30% in 1980.

The specification of the test statistics in samples from the lowest decile of performance is presented in columns two through four of Table 8. The results clearly indicate that only the performance-matched methods (models 4 and 8) yield tests that are correctly specified. The power of the test statistics based on the performance-matched methods is presented in columns two through five of Table 9. Nonparametric Wilcoxon T^* has considerably less power than the same test statistic in random samples.

The specification results for samples drawn from the top performance decile are presented in columns five through seven of Table 8. The pattern of results is similar to that in the poor-performance samples. All test statistics are clearly misspecified except for those based on the performance-matched methods. The power of those test statistics is shown in columns six through nine of Table 9.

Table 8
Specification (size) in poor/good performance samples; percentage of 1,000 samples of 50 firms (1977–1992) from the lowest (highest) decile of performance rejecting the null hypothesis of no abnormal operating performance at the 1%, 5%, and 10% theoretical significance level

Samples were drawn randomly from the lowest (highest) decile of performance based on *ROA* in year $t-1$. Performance rankings were made within year. The numbers presented in the body of this table represent the percentage of 1,000 random samples of 50 firms that reject the null hypothesis of no abnormal operating performance at the theoretical significance level of 1%, 5%, and 10%. The model numbers (1 through 9) correspond to those presented in Table 2.

	Poor performers			Good performers		
	Wilcoxon T^*			Wilcoxon T^*		
Theoretical significance level:	1%	5%	10%	1%	5%	10%
Model no.: Description of expected performance model						
4: Two-digit and performance-matched	0.9	4.8	9.4	0.8	5.1	11.1
5: Lagged *ROA* and Δ two-digit matched	69.0*	88.8*	93.1*	58.4*	81.6*	89.5*
6: Lagged *ROA* and Δ four-digit matched	60.3*	83.2*	90.7*	39.0 *	65.5*	77.4*
7: Lagged *ROA* and Δ two-digit/size-matched	54.6*	77.9*	85.7*	43.7*	69.7*	79.8*
8: Lagged *ROA* and Δ two-digit/perf.-matched	0.9	4.6	9.9	1.0	5.4	10.2
9: Lagged *ROA*	64.7*	87.4*	92.3*	73.2*	91.7*	95.1*

*Significantly different from the theoretical significance level at the 5% level, one-sided binomial test statistic.

We again observe an erosion in the power of the test statistics relative to random samples.

We consider several variations of the performance-matched methods. All of the variations concern how to handle observations for which there is no available performance match within the 90%–110% filter, in the firm's two-digit SIC code. First, we consider matching by using the firm closest in performance, in the same two-digit SIC code. Test statistics based on this alternative rule are misspecified – that is, empirical rejection rates are higher than the theoretical significance level. Second, we consider matching by using all firm's within the 90%–110% filter, without regard to SIC code. These test statistics are well specified and have power similar to those reported above. Nonetheless, we choose to use the industry-based performance-matched method because there are good economic reasons why performance might vary by industry. Third, we assess the specification and power of test statistics based only on those samples for which there were matching firms within the 90%–110% filter and in the firm's same two-digit SIC code. These samples exclude roughly 17% of all observations because of the lack of a performance match. Nonetheless, test

Table 9
Power in poor/good performance samples; percentage of 1,000 samples of 50 firms (1977–1992) from the lowest (highest) decile of performance with induced abnormal performance ranging from −0.02 to 0.02 rejecting null hypothesis of no abnormal operating performance at 5% theoretical significance level

The numbers presented in the body of this table represent the percentage of 1,000 samples from the lowest decile of performance (ROA) that reject the null hypothesis of no abnormal operating performance at a theoretical significance level of 5% and various levels of induced abnormal operating performance. Abnormal operating performance is induced by adding a constant to the observed performance for each of the 50 randomly selected firms in all 1,000 random samples. The model numbers (1 through 9) correspond to those presented in Table 2.

	Poor performers				Good performers			
	Wilcoxon T*				Wilcoxon T*			
Induced level of abnormal operating performance:	−0.02	−0.01	0.01	0.02	−0.02	−0.01	0.01	0.02
Model no.: Description of expected performance model								
4: Two-digit and performance-matched	24.9	10.6	13.1	30.8	40.7	12.9	16.6	40.3
8: Lagged ROA and Δ two-digit/perf.-matched	23.5	10.4	13.8	33.5	54.7	22.6	10.1	29.1

statistics based on this sampling scheme are well specified and are approximately as powerful as test statistics drawn from random samples. In sum, when there are performance biases in sampling, our results indicate that it is crucial to match on performance, even if this matching results in firms that are not in the same industry as the firm in question. The most powerful tests we observe are those for which all firms had an available match within the 90%–110% performance filter, and with the same two-digit SIC code.

We also analyze the specification of test statistics in performance-based samples in performance deciles 2 through 9. In Fig. 1, panel A, we plot the specification of test statistics at the 5% theoretical significance level in samples from these deciles. This figure reveals that the misspecification of test statistics that do not performance-match is not confined to the extreme deciles of firm performance. In fact, test statistics based on change models that incorporate two-digit matching (model 5), four-digit matching (model 6), or size-matching (model 7) are well specified *only* when samples are drawn from the fifth or sixth decile of firm performance. [The two-digit matched method (model 5), which has empirical rejection rate of 6.7%, is actually outside of the established confidence intervals at the 5% level of significance in the sixth performance decile.] In contrast, the performance-matched methods (models 4 and 8) are well specified in every performance decile.

The results presented in this section indicate that it is critical to performance-match when developing test statistics to detect abnormal operating performance, not only when sample firms have extremely poor or good past performance, but even when sample firms have relatively small deviations in relation to the median performance of all firms.

5.3. *Size-based samples*

In addition to performance-based samples, we assess the specification and power of test statistics in size-based samples. Here, we are interested in determining which test statistic is most appropriate when a researcher is faced with a sample of small or large firms. We follow the same approach as that used for the performance-based samples, ranking firms on the basis of size within each calendar year, and then drawing 1,000 samples of 50 firms from the smallest firm-size decile (measured as the book value of assets). Our analysis of samples of the largest firms follows an analogous procedure.

We also ranked firms on the basis of size measured as the market value of assets. The market value of assets is calculated as the book value of total assets (6) less the book value of common equity (60) plus the market value of common equity (25×199). Results based on this measure of firm size are similar to those reported.

The specification of the test statistics for samples of small firms is presented in columns two through four of Table 10. The results indicate that only the

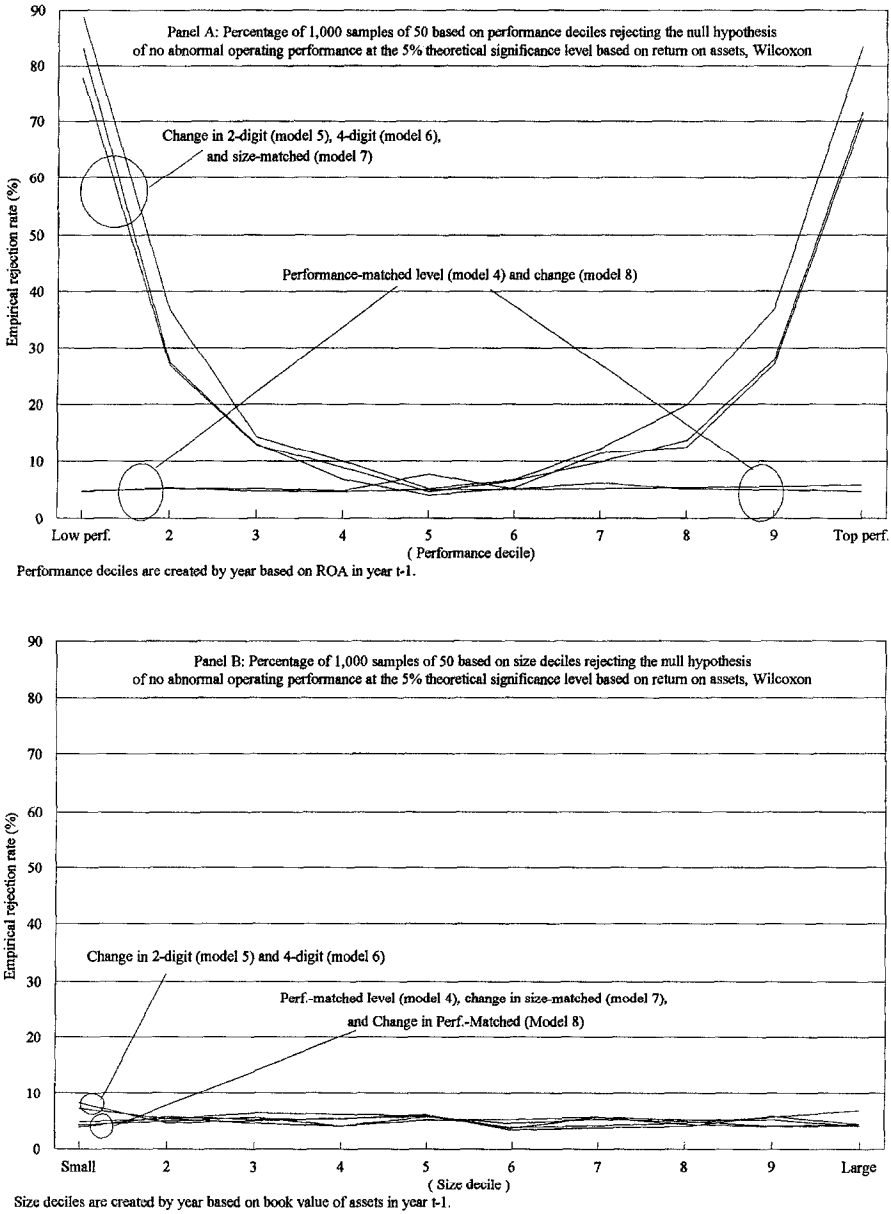


Fig. 1. The specification of Wilcoxon test statistic based on return on assets in samples partitioned into deciles on performance (panel A) and size (panel B).

Table 10
Specification (size) in small- and large-firm samples; percentage of 1,000 samples of 50 firms (1977–1992) from the smallest/largest size decile rejecting the null hypothesis of no abnormal operating performance at a 1%, 5%, and 10% theoretical significance level

Samples were drawn randomly from the smallest (largest) decile of firm size (measured as the book value of assets). Size rankings were made within year. The numbers presented in the body of this table represent the percentage of 1,000 samples of 50 firms that reject the null hypothesis of no abnormal operating performance at a theoretical significance level of 1%, 5%, and 10%. The model numbers (1 through 9) correspond to those presented in Table 2.

	Small firms			Large firms		
	Wilcoxon T*			Wilcoxon T*		
Theoretical significance level:	1%	5%	10%	1%	5%	10%
Model no.: Description of expected performance model						
4: Two-digit and performance-matched	0.7	4.8	8.2	1.5	6.9*	11.5*
5: Lagged *ROA* and Δ two-digit matched	1.7*	7.2*	13.7*	1.2	4.4	8.0
6: Lagged *ROA* and Δ four-digit matched	1.4	8.1*	13.7*	0.4	4.1	8.7
7: Lagged *ROA* and Δ two-digit/size-matched	0.4	4.3	9.6	0.4	4.2	9.7
8: Lagged *ROA* and Δ two-digit/perf.-matched	0.6	3.9	9.0	1.2	4.4	9.1
9: Lagged *ROA*	0.9	4.7	9.4	2.8	11.2*	18.4*

*Significantly different from the theoretical significance level of 5%, one-sided binomial test statistic.

methods that industry-match yield test statistics that are misspecified. Using a level model that performance-matches (model 4) or a change model with no matching (model 9), size matching (model 7), or performance matching (model 8) yields well-specified test statistics in small-firm samples. However, when compared to samples from the extreme deciles of performance, the misspecification of the industry-matched methods is minor, though statistically significant.[7]

The power of the test statistics for samples of small firms is presented in columns two through five of Table 11. All of the test statistics are noticeably less powerful in small-firm samples than in random samples. In sum, in small-firm samples, using the change in a firm's performance, performance-matching methods, or size-matching methods yields test statistics that are well specified.

The specification of the test statistics for samples of large firms is presented in columns five through seven of Table 10. These results indicate that four of the six

[7]When operating income is scaled by end-of-period assets rather than the average of beginning- and ending-period assets, both the level and change models that performance-match are slightly anticonservative, with empirical rejection rates of 1.7%, 7.3%, and 12.8% for the model 4 and 1.8%, 7.2%, and 12.9% for model 8 at the 1%, 5%, and 10% theoretical significance levels.

Table 11
Power in small- and large-firm samples; percentage of 1,000 samples of 50 firms (1977–1992) from the smallest or highest size decile with induced abnormal performance ranging from −0.02 to 0.02 rejecting null hypothesis of no abnormal operating performance at 5% theoretical significance level

The numbers presented in the body of this table represent the percentage of 1,000 samples of 50 firms from the smallest decile of firm size (measured as book value of assets) that reject the null hypothesis of no abnormal operating performance at the theoretical significance level of 5% and various levels of induced abnormal operating performance. Abnormal operating performance is induced by adding a constant to the observed performance for each of the 50 randomly selected firms in all 1,000 random samples. The model numbers (1 through 9) correspond to those presented in Table 2.

	Small firms				Large firms			
	Wilcoxon T^*				Wilcoxon T^*			
Induced level of abnormal operating performance:	−0.02	−0.01	0.01	0.02	−0.02	−0.01	0.01	0.02
Model no.: Description of expected performance model								
4: Two-digit and performance-matched	31.8	12.5	8.2	19.8	100.0	87.0	61.3	99.5†
5: Lagged *ROA* and *Δ* two-digit matched	12.6	5.5	20.9	46.1†	99.9	87.9	77.0	99.9
6: Lagged *ROA* and *Δ* four-digit matched	10.8	4.8	19.4	42.8†	99.8	82.1	71.3	99.7
7: Lagged *ROA* and *Δ* two-digit/size-matched	15.7	7.3	9.3	17.7	99.5	73.6	71.4	99.2
8: Lagged *ROA* and *Δ* two-digit/perf.-matched	32.3	13.8	8.0	18.9	100.0	82.2	70.8	99.8
9: Lagged *ROA*	19.1	6.8	14.4	37.9	100.0	97.1	59.8	98.4†

†Empirical tests based on this statistic are anticonservative at the 1%, 5%, and/or 10% theoretical significance level (see Table 16).

models considered yield tests that are well specified. As compared to small-firm samples, where changes in a firm's performance yield well-specified test statistics, this method yields misspecified test statistics in large-firm samples. The performance-matched model in the levels is also misspecified at the 5% and 10% levels. The power of the test statistics for samples of large firms is presented in columns six through nine of Table 11. The test statistics in large-firm samples are more powerful than those in random samples. Of the well-specified methods, no particular model of expected performance yields test statistics that are clearly more powerful.

We also analyze the specification of test statistics in size-based samples from size deciles 2 through 9. In Fig. 1, panel B, we plot the specification of test statistics at the 5% theoretical significance level in samples from these deciles. When contrasted with panel A of the same figure, this graph reveals that the misspecification of the industry-matched methods in small-firm samples is dwarfed by the misspecification of the methods that do not performance-match in extreme performance samples.

Since our evidence suggests that there is no relation between firm size and operating performance, we conduct auxiliary analyses to test the hypothesis that firm size and operating performance are unrelated. For each calendar year, we run a cross-sectional regression of return on assets on the log of firm size (measured as the book value of total assets). The mean coefficient estimate on firm size is positive and statistically significant across our 16 sample years. This indicates that, on average, large firms have a higher ROA than small firms, which is consistent with the findings of Fama and French (1995). However, the explanatory power of firm size is quite low; the average R^2 that we observe is 1.6%. This low explanatory power could indicate why we rarely find that size-matching is important.

In sum, in samples of unusually small or large firms, we do not find that size-matching is critical in tests designed to detect abnormal operating performance. Furthermore, the performance-matched methods analyzed are well specified and as powerful as the methods that match on firm size.

5.4. *Multiple dimensions of bias*

Our results thus far indicate that performance matching is critical in samples with pre-event performance biases. To investigate the impact of size and performance biases on the specification of test statistics, we partition our population into three performance (low, mid, and top) and three size groups (small, mid, and large). (The analyses that follow were also conducted by partitioning first on size and then on performance. The results of this alternative partitioning are analogous to the reported results.) One thousand samples are drawn from each of the nine cells that constitute the three-by-three partition on performance and size. In Fig. 2, we report the specification of models 4 through 8 at the 5% theoretical

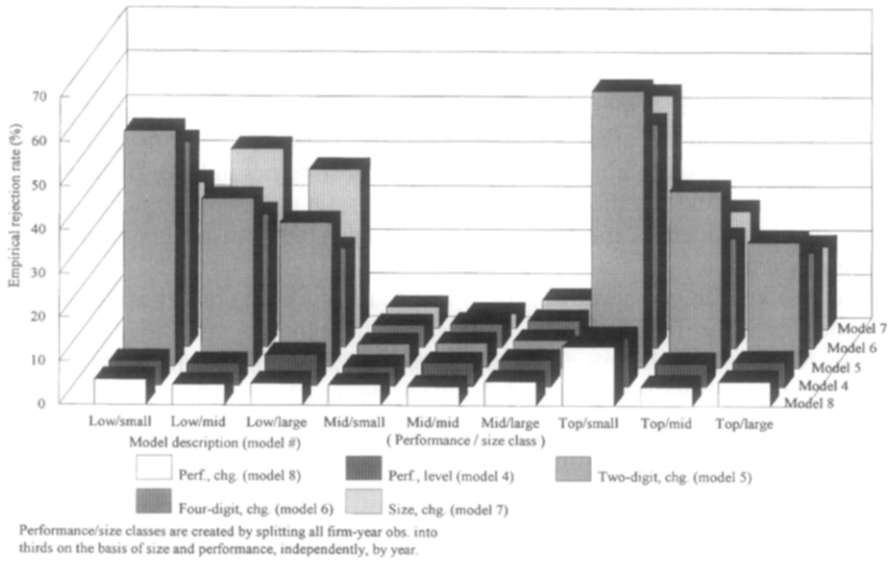


Fig. 2. Percentage of 1,000 samples of 50 based on performance/size classes rejecting the null hypothesis of no abnormal operating performance at the 5% theoretical significance level based on return on assets, Wilcoxon.

significance level. This analysis reveals that test statistics based on the performance-matched methods (models 4 and 8) are well specified in all cells except those for small firms that have had unusually good performance (top/small). The empirical rejection rate in this sampling situation, using the performance-matched level model (4), is 11.1%. Using the performance-matched change model (8), the rejection rate is 13.4%. In contrast, two-digit matching, four-digit matching, and size matching yield test statistics that are grossly misspecified, regardless of firm size, in the lowest and highest thirds of performance.

Though performance matching yields the least anticonservative test statistics, it is nonetheless disconcerting that test statistics based on the performance-matched methods are misspecified in the top performance/small firm cell. To identify a test statistic that is well specified in this sampling situation, we consider size and performance matching, without regard to industry. (We also considered size, performance, and industry matching, where industry is defined by two-digit SIC codes. However, over half of all firm-year observations could not be matched on size, performance, and industry.) First, each sample firm is matched to all firms of similar size (using the 70%–130% filter on the book value of total assets) in year $t-1$. Second, of the firms that meet the size criterion, we discard those outside of the 90%–110% filter on performance in year $t-1$. If no firms remain in the comparison group, the firm that meets the

size criterion and is closest in performance to the sample firm in year $t-1$ is used as the benchmark. We reestimate the specification of the Wilcoxon test statistic based on this matching scheme in all cells of our three-by-three partition on size and performance. This method is well specified in every cell. Furthermore, this method yields test statistics that are well specified in every decile of firm size, in every decile of performance, and in random samples.

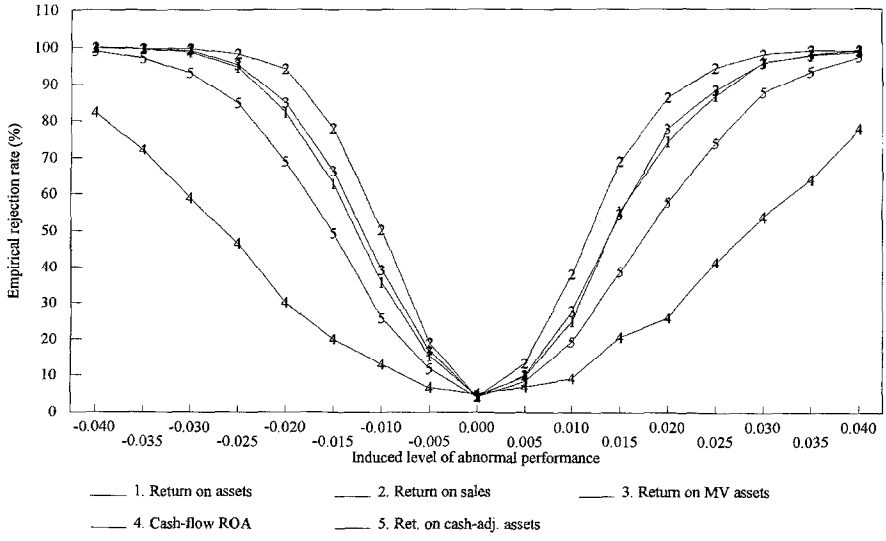
In sum, the performance-matched methods yield well-specified test statistics in all partitions of firm size and performance, with the exception of small firms with unusually good past performance. For a sample of small firms with unusually strong pre-event performance, researchers can match sample firms to other firms of similar size and performance, without regard to industry. Test statistics based on this matching scheme are well specified in all partitions on size and performance that we analyze.

5.5. *Subperiod results*

To test the robustness of the results reported in this section, we examine the empirical power and specification of test statistics based on the *ROA* across two subperiods (1977–1984 and 1985–1992) in random samples, and the extreme deciles of firm size and performance. Across both subperiods, all test statistics based on changes in performance and an industry benchmark are well specified and powerful in random samples; only the performance-matched methods are well specified in the extreme deciles of performance; and both the size-matched and performance-matched methods are well specified in the extreme deciles of firm size. In sum, the pattern of results across subperiods is similar to those reported for the full period, and does not alter our major conclusions.

6. Alternative measures of performance

In this section, we consider several alternative measures of operating performance. The *ROA* measure used up to this point has three drawbacks. First, the total assets on a firm's balance sheet are recorded at historic cost, while operating income is recorded in current dollars. The appropriate measure for the denominator would be the current or replacement cost of total assets. Second, the total-assets measure reflects all of the assets of the firm, not just operating assets. Consequently, the use of total assets could understate the true productivity of operating assets. Third, operating income is an accrual-based measure that managers could over- or understate by increasing or decreasing discretionary accruals. We label each of these drawbacks as historic cost, nonoperating assets, and earnings manipulation, respectively. Many of our alternative performance measures are designed to overcome these potential problems.



Return on assets, return on sales, and return on MV assets are operating income scaled by average book value of total assets, sales, and average market value of total assets. Cash-flow ROA is operating cash flow scaled by average book value of total assets. Ret. on cash-adj. assets is operating income scaled by average book value of total assets less cash and marketable securities.

Fig. 3. The empirical rejection rates of the Wilcoxon test statistic in random samples based on the performance-matched change model 8 using five alternative measures of operating performance.

We estimate the specification and power of test statistics based on each of the alternative performance measures in random samples, the extreme deciles of performance, and the extreme deciles of firm size. These analyses result in over 1,000 permutations of the possible combinations of performance measures, test statistics, expectation models, and sampling situations (e.g., small-firm samples or large-firm samples). Presenting all of these results is not possible. Therefore, we summarize the major findings of our analyses in this section.

In Fig. 3, we plot the empirical rejection rates in random samples, using the performance-matched change model (model 8) and the Wilcoxon signed-rank test statistic for *ROA* and the four alternative performance measures considered in this section.[8] Care should be exercised in interpreting this figure, since, for example, a one-cent improvement in operating income per dollar of *book* value of assets is not equivalent to a one-cent improvement per dollar of *market* value

[8]In an earlier version of this paper, we also considered a fifth performance measure – net income plus interest expense scaled by the book value of total assets. All of the results using this performance measure parallel those using *ROA*, but the performance measure based on net income is uniformly less powerful than that based on operating income. The mean and median *ROA* based on net income over our sample period are 7.5% and 8.3%, with a cross-sectional standard deviation of 8.4%.

of assets. Nonetheless, when we alter the incrementing to facilitate comparison across the different measures of operating performance, the general pattern observed in Fig. 3 remains.[9] In random samples, the power of our alternative performance measures is roughly equivalent, with one exception, the power of tests based on operating cash flow scaled by the book value of total assets. We discuss this point below.

In general, the results documented here are consistent with those using return on assets. Some of the results that are robust to the use of the alternative performance measures include:

1. Virtually all expectation models are well specified in random samples.
2. The Wilcoxon signed-rank test statistic is uniformly more powerful than the t-statistic.
3. Change models are uniformly more powerful than level models.
4. Performance matching is critical in samples drawn from the extreme deciles of performance.

In many situations, researchers should test the robustness of their results by using several alternative measures of performance.

6.1. *Return on cash-adjusted assets*

The return on assets measure scales operating income by the book value of total assets, which reflects all assets of the firm, both operating and nonoperating. Operating income reflects income generated by only the operating assets of the firm. To obtain a more accurate measure of the productivity of a firm's operating assets, operating income should be scaled only by the value of the operating assets.

The most important adjustment to total assets can be the deduction of cash and marketable securities from the book value of total assets. While a certain level of cash is necessary for a firm's operations, much of the time-series variation in cash balances is attributable to the financing activities of the firm. Thus, we often observe large increases in cash balances when a firm issues securities but does not immediately invest those funds. When sample firms experience a time-series variation in cash balances that is significantly different from control firms, results can be affected by deducting cash balances from total assets. This is likely to be the case for samples in which firms recently issued securities.

A separate, but related, issue is the build-up in operating assets following a securities issue. Though some firms might retain a portion of the proceeds

[9]The details of this recalibration of our incrementing scheme are available upon request.

from an issue in cash, others might invest the full amount in operating assets. Of course, this investment leads to an increase in operating assets, but in all likelihood, these assets have not been in place long enough to generate operating income. This build-up in operating assets can lead to a temporary decline in *ROA*, until the new operating assets begin to generate income. Obviously, deducting cash and marketable securities from the book value of total assets does not address this issue. However, researchers can extend their analysis to several years after an event of interest to ascertain whether an erosion in operating performance is the result of a temporary build-up in assets. Or they can use a performance measure that is unaffected by the change in the asset base (for example, return on sales).

We reestimate all of our results using an *ROA* in which assets are net of cash balances (Compustat item 1). We refer to this performance measure as return on cash-adjusted assets. (Compustat does not report cash and marketable securities for approximately 7% of the firms that comprise our population. There are 32,680 firm-year observations of return on assets, and 30,249 firm-year observations of return on cash-adjusted assets. However, virtually all firms that do not report cash and marketable securities are utilities.) Over our sample period, the mean and median returns on cash-adjusted assets are 16.3% and 15.6%, with a cross-sectional standard deviation of 13.1%.

All of the results based on cash-adjusted *ROA* are analogous to those presented earlier. (For the size-matched method, we match based on the book value of total assets less cash and marketable securities.) We also estimate the specification of test statistics when firms are drawn from the top-performance/small-firm cell of our three-by-three partition on firm performance and size (see Fig. 2). Though the empirical rejection rates of the performance-matched methods using return on cash-adjusted assets are less than those documented for *ROA*, test statistics based on the performance-matched methods and return on cash-adjusted assets are still misspecified. For example, at the 5% theoretical significance level, the empirical rejection rates for the performance-matched method based on *ROA* are 11.1% (using the performance-matched level model) and 13.4% (using the performance-matched change model). In contrast, the analogous rejection rates for return on cash-adjusted assets are 6.7% and 7.5%, respectively. Thus, though the cash adjustment improves the specification of test statistics in this sampling situation, it does not yield well-specified test statistics.

6.2. Return on sales

Scaling operating income by sales can overcome the historic cost and nonoperating assets problems associated with *ROA*. As the discussion in Section 6.1 illustrates, one problem with scaling operating income by the book value of total assets is that operating income may not be appropriately matched with the assets used to generate that income. In addition to the fact that total

assets reflect some nonoperating assets, total assets are recorded at historic cost, while operating income is reported in current dollars.

An alternative performance measure – return on sales – can be constructed by scaling operating income (Compustat item 13) by sales (item 12). The advantage of this performance measure is that both the numerator and denominator are from a firm's income statement. Consequently, they may be more appropriately matched. For example, when a securities issue results in a large increase in cash (or other asset balances), the sales reported on a firm's income statement are not affected. Similarly, sample firms can have recently acquired large amounts of operating assets and thus have higher book value of total assets than control firms *because of the recency of the acquisitions.* Consequently, the *ROA* measure for sample firms would be lower because the recent acquisitions are reflected on the balance sheet in near-current dollars. This particular problem can surface when sample firms have either issued securities to finance investment or engaged in acquisitions.

The disadvantage of using return on sales is that it does not directly measure the productivity of assets. Consider a firm that changes its operations to increase its sales (and operating income) without increasing its asset base. This firm has increased the productivity of its assets, which would be evident in a well-constructed *ROA* measure. However, this firm could have no change in return on sales, if both sales and operating income increase proportionately. Nonetheless, return on sales can detect certain types of operating performance changes – for example, reductions in selling, general, and administrative expenses, or improvements in production efficiency that reduce cost of goods sold.

We reestimate all of our results using return on sales. The mean and median returns on sales over our sample period are 16% and 12.4%, respectively, with a cross-sectional standard deviation of 18.7%. Unlike some of our other measures, data availability is not a problem for return on sales relative to return on assets. All of the results using return on sales parallel those using *ROA* with two exceptions: First, in samples drawn from the smallest decile of firm size, only the size-matched method is well specified. However, the misspecification of the industry- or performance-matched change models at the 5% theoretical significance level is not large, with empirical rejection rates ranging from 6.6% for the performance-matched change model (8) to 9.3% for the two-digit-matched change model (5).

Second, as was the case for return on cash-adjusted assets, test statistics based on return on sales are less anticonservative than those based on *ROA* (but still misspecified) in the top- performance/small-firm cell of our three-by-three partition on size and performance. Using return on sales, the empirical rejection rates are 7.3% using the performance-matched level model and 7.5% using the performance-matched change model. However, in contrast to an *ROA* in which misspecification using the performance-matched methods is observed only in this top-performance/small-firm cell of our three-by-three partition, the

performance-matched methods that use return on sales are also slightly anticonservative in the low-performance/large-firm and the mid-performance/small-firm cells. At the 5% theoretical significance level, the performance-matched level model using return on sales yields empirical rejection rates of 6.6% in the low-performance/large-firm cell and 7.6% in the mid-performance/small-firm cell, while the performance-matched change model using return on sales yields empirical rejection rates of 10.6% and 10.2%, respectively.

6.3. Return on market value of assets

Scaling operating income by the market value of assets can overcome the historic cost problem associated with return on assets. Unlike the book value of total assets, the market value of total assets can be measured at the same point in time for all firms. Thus, we alleviate the problem of sample firms acquiring assets at different times than control firms. Furthermore, the market value of assets includes off-balance-sheet and intangible assets.

The disadvantage of using the market value of assets is that it is a forward-looking measure of assets. Finance theory characterizes the market value of the firm as the present value of future cash flows. Thus, the market value of assets reflects the earnings potential of assets in place, as well as the earnings potential of all future assets that the firm is expected to acquire. Thus, firms with unusually high earnings potential and growth in earnings will have lower returns on market value of assets. In sum, it is appropriate to use the market value of assets in lieu of the book value of assets if sample firms and control firms acquired assets on their balance sheets at different points in time, but had similar prospects for earnings growth.

We measure the market value of total assets as the book value of total assets (item 6) less the book value of common equity (item 60) plus the market value of common equity (item 25 times item 199). This calculation assumes that the major difference between book and market value of total assets can be attributed to the market valuation of equity. For example, the book value of long-term debt is assumed equal to the market value of that debt. We calculated the return on market value of assets by scaling operating income (item 13) by the average of beginning- and ending-period market value of assets. Of those firms with an *ROA* measure, approximately 4% of all firm-year observations do not have the data necessary to calculate the market value of assets measure that we employ.

We reestimate all of our results using return on market value of assets. The mean and median returns on market value of assets over our sample period are 11.3% and 11.5%, respectively, while the cross-sectional standard deviation is 7.2%. The results parallel those based on *ROA*, with one exception. As with return on sales and return on cash-adjusted assets, test statistics based on return on market value of assets are less anticonservative than those based on *ROA*

(but still misspecified) in the top-performance/small-firm cell of our three-by-three partition on size and performance. The empirical rejection rates for the top-performance/small-firm cell using return on market value of assets are 5.9% using the performance-matched level model, and 8.2% using the performance-matched change model.

6.4. *Cash-flow return on assets*

Using a cash-flow-based measure of operating income can overcome the potential earnings manipulation problem associated with an accrual-based measure of operating income. If managers manipulate the recognition of revenue or expense items for personal benefit, operating income can be a biased measure of performance. For a sample of firms whose managers might have unusually strong incentives to manipulate earnings, a cash-based measure of performance could be more appropriate than the accrual-based measures. Teoh, Wong, and Rao (1994) and Teoh, Welch, and Wong (1995) present evidence indicating that prior to the issue, firms that make initial public offerings or seasoned equity offerings use accruals to overstate earnings.

We estimate operating cash flow as operating income before depreciation (item 13) plus the decrease in receivables (2), the decrease in inventory (3), the increase in accounts payable (70), the increase in other current liabilities (72), and the decrease in other current assets (68). Operating cash flow is scaled by the average of beginning- and ending-period book value of total assets (6) to yield a measure we label 'cash-flow return on assets'.

We reestimate all of our results using cash-flow *ROA*. The mean and median cash-flow *ROA*s over our sample period are 13.0% and 13.4%, respectively, while the cross-sectional standard deviation is 12.2%. The calculation of operating cash flow imposes severe data constraints, with over 17% of all firms lacking the data necessary to calculate operating cash flow. All of the results parallel those using *ROA*, with one exception. As shown in Fig. 3, test statistics based on cash-flow *ROA* are uniformly less powerful than those based on the alternative performance measures.

7. The use of percentage changes to detect abnormal operating performance

Several studies in financial economics analyze the operating performance of sample firms by comparing either the percentage change in operating income or the percentage change in *ROA* (or, in some cases return on sales) to an appropriate benchmark – usually the median percentage change in performance for other firms in the same industry (Kaplan, 1989; Lehn, Netter, and Poulsen, 1990; Denis and Denis, 1993; Jain and Kini, 1994; Denis and Denis, 1995).

There are two fundamental problems with this approach. We focus on *ROA* in the discussion that follows, but our arguments apply to any performance measure (including unscaled operating income). First, if *ROA* is negative in either year over which the percentage change is calculated, the result is nonsensical. Consequently, researchers are forced to discard firms that experience losses over the sample period under consideration. Using adjacent-year *ROA* to calculate percentage change over our sample period, we encounter negative *ROA* in at least one of the adjacent years for 2,030 firm-year observations (approximately 7% of all firm-year observations). Discarding the firms with poor performance not only diminishes the power of statistical tests, but can also lead to biases in test statistics.

Second, using the percentage change metric, changes in operating performance are implicitly assumed to be proportional to the level of pre-event *ROA*. For example, consider two firms (A and B), both of which have one million dollars in operating assets. Firm A has a pre-event operating income of $150,000 for a pre-event *ROA* of 15%. Firm B has a pre-event operating income of $40,000 for a pre-event *ROA* of 4%. In the methods analyzed in this paper, we have assumed that changes in operating performance are proportional to the assets in place. Thus, with no change in the asset base of either firm, a $0.01/$1.00 of operating assets improvement in operating performance would represent an improvement in operating income of $10,000 for both firms. In contrast, using the percentage change metric, a $10,000 increase in operating income represents a 6.7% increase in *ROA* for firm A and a 25% increase in *ROA* for firm B. We believe that it is more reasonable to assume that changes or erosions in operating performance following major corporate events are proportional to the asset base employed, rather than the level of pre-event performance.

Despite these fundamental objections, we reestimate the power and specification of the Wilcoxon ranked-sign test statistic in random samples and the extreme deciles of performance and size, using the percentage change of return on assets.[10] All of the test statistics are well specified in random samples, except for those based on two- or four-digit industry matching. However, even test statistics based on two- and four-digit matching are only slightly anticonservative, with empirical rejection rates of 11.9% and 12.2%, respectively, at the 10% theoretical significance levels. All methods are well specified at the 1% and 5% theoretical significance levels. In the extreme deciles of performance, as was the case for the change models, only the percentage change model that performance-matches yields test statistics that are well specified. All other test statistics are grossly misspecified. In the samples of small firms, only the size-matched

[10]The details of this estimation are available on request.

method yields test statistics that are well specified. However, the performance-matched method is only slightly anticonservative, with an empirical rejection rate of 8.4% at the 5% theoretical significance level. Though we object to the use of the percentage change metric for the two reasons cited, the general tenor of the results applies to the percentage change metric.[11]

8. Conclusion

We evaluate the specification and power of tests designed to detect abnormal operating performance. We consider three choices in designing a test. First, we compare five measures of operating performance: return on assets (operating income scaled by the book value of assets), return on cash-adjusted assets (operating income scaled by the book value of assets less cash and marketable securities), return on sales (operating income scaled by sales), return on market value of assets (operating income scaled by the market value of assets), and cash-flow return on assets (operating cash flow scaled by the book value of assets). Second, we compare two statistical tests: parametric t-statistic and nonparametric Wilcoxon signed-rank T^*. Third, we compare nine models of expected operating performance (see Table 8) used in recent empirical work in the academic finance and accounting literature.

Here, we provide specific recommendations that are based on two criteria. First, a test must be well specified, meaning empirical rejection rates approximate theoretical rejection rates. Second, if more than one test is well specified, we opt for the test that is the most powerful.

8.1. Parametric or nonparametric test statistic?

Perhaps the clearest result to emerge from our analysis is that nonparametric Wilcoxon signed-rank T^* test statistics are uniformly more powerful than parametric t-statistics. We attribute this result to the existence of extreme observations in the distribution of the operating performance measures analyzed. In auxiliary analyses not reported in a table, we observe that the power of Wilcoxon T^* and t-statistics are similar in random samples when the performance measures are winsorized at the first and 99th percentiles of its distribution. Because of the power advantage that the Wilcoxon T^* offers, we recommend its use in all sampling situations that we consider.

[11] We also analyze the power functions using the percentage change methods, the details of which are available on request. The percentage change models yield test statistics that are uniformly less powerful than those based on the change models.

8.2. *Which model of expected performance?*

Clearly, the most important choice is the model of expected operating performance. Without exception, the models that yield well specified, powerful test statistics incorporate a firm's past performance. Though we are confident that change models always dominate level models in detecting abnormal operating performance, it is often informative for researchers to report the levels of operating performance both before and after an event. Indeed, in virtually all of the studies of operating performance we encounter, researchers report either the mean or median level of performance for sample and industry firms over time. However, our results indicate that inferences about abnormal operating performance should not be based on the levels of performance over time, but rather on an expectation model that incorporates a firm's pre-event performance.

Our results indicate that several models work well in either random samples or samples of large firms. For example, analyzing the change in a firm's performance relative to the change in the median performance of firms in its two-digit SIC code yields test statistics that are both well specified and powerful. Other expectation models, such as industry matching based on four-digit SIC code, size matching within two-digit SIC code, or performance matching within two-digit SIC code, also yield well-specified, powerful test statistics.

Perhaps our most interesting result is that only expectation models that match sample firms to firms of similar pre-event performance are well specified in samples with performance-based biases. This result is robust to all of the performance measures that we consider. The misspecification of test statistics that do not performance match is large and occurs when sample firms have past performance that differs only slightly from the performance of population firms.

Though many recent studies use methods that match sample firms to control firms of similar size (see Table 1), we document that the performance-matched methods that we analyze perform as well as size-matched methods in samples of unusually small firms. For example, in samples of firms from the smallest decile of firm size, models based on size matching and performance matching yield well-specified test statistics.

Finally, we document that all of the expectation models that we consider are misspecified in samples of small firms with historically strong performance. This bias is much less severe, but still present, when the performance-matched models are employed. The bias is still less severe, but present, when return on sales, return on cash-adjusted assets, or return on market value of assets are used in lieu of return on assets. Nonetheless, in this sampling situation, only when we match sample firms to firms of similar size *and* pre-event performance, without regard to industry, do we obtain test statistics that are well specified.

In sum, the performance-matched methods that we analyze yield test statistics that are generally well specified and at least as powerful as the alternative models of expected performance. In general, we would recommend their use in most sampling situations. Finally, the one method that yields test statistics that are well specified in every sampling situation that we analyze is to match sample firms to control firms on size and pre-event performance, without regard to industry.

We are confident that our proposed performance-matched methods control well for the average tendency for mean reversion of accounting-based performance measures. However, there can be cross-sectional variation in the tendency for mean reversion in these performance measures. For example, Fama and French (1995) document that small firms have return on equity measures that mean-revert more quickly than similar measures for large firms. We suspect this is the reason why it is important to performance- *and* size-match in samples of small firms with historically strong performance. In short, researchers should carefully consider whether the performance measures of sample firms have more or less tendency to mean-revert than control firms.

8.3. Which performance measure?

From a statistical standpoint, the choice of performance measure is generally inconsequential, with one exception. Test statistics based on a cash-flow measure of operating income (i.e., cash-flow return on assets) are uniformly less powerful than those based on the other performance measures considered here.

However, because of the nature of a particular research question, the choice of performance measure can be critical. We conclude by presenting two examples of how a research question should affect the choice of performance measure. These examples illustrate that the results documented in this research should not be applied without careful consideration to the research question at hand.

First, consider firms that have recently issued securities. These firms can have a large increase in the book value of their assets as they invest in additional operating assets, but no commensurate increase in their operating income, since the new assets are not yet generating income. In this situation, a researcher should track the performance of sample firms for several years following the event of interest, or else use a performance measure (for example, return on sales) that is unaffected by the changes in a firm's operating assets.

Second, in certain situations (for example, firms going public), sample firms can be motivated to overstate their reported earnings. In this situation, a researcher should use a cash-based, rather than accrual-based, performance measure. An accrual-based performance measure can lead a researcher to conclude erroneously that sample firms have experienced an erosion in performance post-event, when sample firms are reporting lower income merely as

a result of their use of accruals to overstate earnings pre-event. Though we document that a cash-based performance measure is generally less powerful than an accrual-based performance measure, in the sampling situation described here, the cash-based performance measure allows the researcher to ascertain whether an erosion in performance is the result of an erosion in operating performance or the reversal of pre-event accruals.

References

Asquith, Paul, Paul Healy, and Krishna Palepu, 1989, Earnings and stock splits, The Accounting Review 64, 387–403.

Brown, Stephen J. and Jerold B. Warner, 1985, Using daily stock returns: The case of event studies, Journal of Financial Economics 14, 205–258.

Dann, Larry Y., Ron Masulis, and David Mayers, 1991, Repurchase tender offers and earnings information, Journal of Accounting and Economics 14, 217–251.

DeGeorge, Francois and Richard Zeckhauser, 1993, The reverse LBO decision and firm performance: Theory and evidence, Journal of Finance 48, 1323–1348.

Denis, David J. and Diane K. Denis, 1993, Managerial discretion, organization structure, and corporate performance: A study of leveraged recapitalizations, Journal of Accounting and Economics 16, 209–236.

Denis, David J. and Diane K. Denis, 1995, Causes of financial distress following leveraged recapitalizations, Journal of Financial Economics 37, 129–158.

Fama, Eugene F. and Kenneth French, 1995, Size and book-to-market factors in earnings and returns, Journal of Finance 50, 131–155.

Guenther, David A. and Andrew J. Rosman, 1994, Differences between Compustat and CRSP SIC codes and related effects on research, Journal of Accounting and Economics 18, 115–128.

Healy, Paul and Krishna Palepu, 1988, Earnings information conveyed by dividend initiations and omissions, Journal of Financial Economics 21, 149–176.

Healy, Paul and Krishna Palepu, 1990, Earnings and risk changes surrounding primary stock offers, Journal of Accounting Research 28, 25–48.

Healy, Paul, Krishna Palepu, and Richard Ruback, 1994, Which takeovers are profitable Strategic or financial?, Working paper (Massachusetts Institute of Technology, Cambridge, MA).

Holthausen, Robert and David Larcker, 1994, The financial performance of reverse leveraged buyouts, Working paper (The Wharton School, Philadelphia, PA).

Jain, Bharat A. and Omesh Kini, 1994, The post-issue operating performance of IPO firms, Journal of Finance 49, 1699–1726.

Kahle, Kathleen M. and Ralph A. Walkling, 1995, The impact of industry classifications on financial research, Working paper (Ohio State University, Columbus, OH).

Kaplan, Steven, 1989, The effect of management buyouts on operating performance and value, Journal of Financial Economics 24, 217–254.

Lehn, Kenneth, Jeffrey Netter, and Annette Poulsen, 1990, Consolidating corporate control: Dual-class recapitalizations versus leveraged buyouts, Journal of Financial Economics 27, 577–580.

Loughran, Tim and Jay Ritter, 1995, The operating performance of firms conducting seasoned equity offerings, Working paper (University of Illinois, Urbana, IL).

Mikkelson, Wayne and Megan Partch, 1994, Consequences of unbundling managers' voting rights and equity claims, Journal of Corporate Finance 1, 175–200.

Mikkelson, Wayne and Ken Shah, 1994, Performance of companies around initial public offerings, Working paper (University of Oregon, Eugene, OR).

Mulherin, J. Harold and Annette B. Poulsen, 1994, Proxy contests, shareholder wealth, and operating performance, Working paper (University of Georgia, Athens, GA).

Penman, Stephen, 1991, An evaluation of accounting rate of return, Journal of Accounting, Auditing, and Finance 6, 233–255.

Strickland, Deon, Kenneth W. Wiles, and Marc Zenner, 1994, A requiem for the USA: Is small shareholder monitoring effective, Working paper (University of North Carolina, Chapel Hill, NC).

Teoh, Siew Hong, T.J. Wong, and Gita R. Rao, 1994, Incentives and opportunities for earnings management in initial public offerings, Working paper (University of Michigan, Ann Arbor, MI).

Teoh, Siew Hong, Ivo Welch, and T.J. Wong, 1995, Earnings management in seasoned equity offerings, Working paper (University of Michigan, Ann Arbor, MI).