Econometrics of Event Studies

S.P. Kothari
Sloan School of Management, MIT

Jerold B. Warner
William E. Simon Graduate School of Business Administration
University of Rochester

October 20, 2004

This article will appear in the Handbook of Corporate Finance: Empirical Corporate Finance (Elsevier/North-Holland), which is edited by B. Espen Eckbo.

We thank Jon Lewellen and Adam Kolasinski for insightful comments, and Irfan Safdar and Alan Wancier for research assistance.

ABSTRACT

The number of published event studies exceeds 500, and the literature continues to grow. We provide an overview of event study methods. Short-horizon methods are quite reliable. While long-horizon methods have improved, serious limitations remain. A challenge is to continue to refine long-horizon methods. We present new evidence that properties of event study methods can vary by calendar time period and can depend on event sample firm characteristics such as volatility. This reinforces the importance of examining event study statistical properties for non-randomly selected samples.

Table of Contents

1. Introduction and Background

This chapter focuses on the design and statistical properties of event study methods. Event studies examine the behavior of firms' stock prices around corporate events.[1] A vast event vast literature written over the past several decades has become an important part of financial economics. Prior to that time, "there was little evidence on the central issues of corporate finance. Now we are overwhelmed with results, mostly from event studies" (Fama, 1991, p. 1600). In a corporate context, the usefulness of event studies arises from the fact that the magnitude of abnormal performance at the time of an event provides a measure of the (unanticipated) impact of this type of event on the wealth of the firms' claimholders. Thus, event studies focusing on announcement effects for a short-horizon around an event provide evidence relevant for understanding corporate policy decisions.

Event studies also serve an important purpose in capital market research as the principle means of testing market efficiency. Systematically nonzero abnormal security returns that persist after a particular type of corporate event are inconsistent with market efficiency. Accordingly, event studies focusing on long-horizons following an event can provide key evidence on market efficiency (Brown and Warner, 1980, and Fama, 1991).

Beyond financial economics, event studies are useful in related areas. For example, in the accounting literature, the effect of earnings announcements on stock prices has received much attention. In the field of law and economics, event studies are

[1] We discuss event studies that focus only on the mean stock price effects. Many other types of event studies also appear in the literature, including event studies that examine return variances (e.g., Beaver, 1968, and Patell, 1976), trading volume (e.g., Beaver, 1968, and Campbell and Wasley, 1996), operating (accounting) performance (e.g., Barber and Lyon, 1996), and earnings management via discretionary accruals (e.g., Dechow, Sloan, and Sweeney, 1995, and Kothari, Leone, and Wasley, 2005).

used to examine the effect of regulation, as well as to assess damages in legal liability cases.

The number of published event studies easily exceeds 500 (see section 2), and continues to grow. A second and parallel literature, which concentrates on the methodology of event studies, began in the 1980's. Dozens of papers have now explicitly studied statistical properties of event study methods. Both literatures are mature.

From the methodology papers, much is known about how to do – and how not to do – an event study. While the profession's thinking about event study methods has evolved over time, there seems to be relatively little controversy about statistical properties of event study methods. The conditions under which event studies provide information and permit reliable inferences are well-understood.

This chapter highlights key econometric issues in event study methods, and summarizes what we know about the statistical design and the interpretation of event study experiments. Based on the theoretical and empirical findings of the methodology literature, we provide clear guidelines both for producers and consumers of event studies. Rather than provide a comprehensive survey of event study methods, we seek to sift through and synthesize existing work on the subject. We provide many references and borrow heavily from the contributions of published papers. Two early papers that cover a wide range of issues are by Brown and Warner (1980, 1985). More recently, an excellent chapter in the textbook of Campbell, Lo, and MacKinlay (1997) is a careful and broad outline of key research design issues. These standard references are recommended reading, but predate important advances in our understanding of event study methods, in

particular on long horizon methods. We provide an updated and much needed overview, and include a bit of new evidence as well.

Although much emphasis will be on the statistical issues, we do not view our mission as narrowly technical. As financial economists, our ultimate interest is in how to best specify and test interesting economic hypotheses using event studies. Thus, the econometric and economic issues are interrelated, and we will try to keep sight of the interrelation.

In section 2, we briefly review the event study literature and describe the changes in event study methodology over time. Section 3 we discuss how to use events studies to test economic hypotheses. We also touch upon the properties of the event study tests and examine the determinants of the properties as a function of firm characteristics, sample size, and event clustering, etc. Section 4 is devoted to issues most likely encountered when conducting long-horizon event studies. The main issues are risk adjustment, cross-correlation in returns, and changes in volatility during the event period. A brief summary of the chapter appears in section 5.

2. The Event Study Literature: Basic Facts

2.1 The stock and flow of event studies

To quantify the enormity of the event study literature, we conducted a census of event studies published in 5 leading journals: the Journal of Business (JB), Journal of Finance (JF), Journal of Financial Economics (JFE), Journal of Financial and Quantitative Analysis (JFQA), and the Review of Financial Studies (RFS). We began in 1974, the first year the JFE was published.

Table 1 reports the results for the years 1974 through 2000. The total number of papers reporting event study results is 565. Since many academic and practitioner-oriented journals are excluded, these figures provide a lower bound on the size of the literature. The number of papers published per year increased in the 1980's, and the flow of papers has since been stable. The peak years are 1983 (38 papers), 1990 (37 papers), and 2000 (37 papers). All five journals have significant representation. The JFE and JF lead, with over 200 papers each.

Table 1 makes no distinction between long horizon and short horizon studies. While the exact definition of "long horizon" is arbitrary, it generally applies to event windows of 1 year or more. Approximately 200 of the 565 event studies listed in Table 1 use a maximum window length of 12 months or more, with no obvious time trend in the year by year proportion of studies reporting a long-horizon result.

No survey of these 565 event study papers is attempted here. For the interested reader, the following are some examples of event study surveys. MacKinlay (1997) and Campbell, Lo, and MacKinlay (1997) document the origins and breadth of event studies. The relation of event studies to tests of market efficiency receives considerable attention in Fama (1991), and in recent summaries of long-horizon tests in Kothari and Warner (1997) and Fama (1998). Smith (1986) presents reviews of event studies of financing decisions. Jensen and Ruback (1983), Jensen and Warner (1988) and Jarrell, Brickley and Netter (1988) survey corporate control events. Recently, Kothari (2001) reviews event studies in the accounting literature.

2.2 Changes in event study methods: the big picture

Even the most cursory perusal of event studies done over the past 30 years reveals a striking fact: the basic statistical format of event studies has not changed over time. It is still based on the table layout in the classic stock split event study of Fama, Fisher, Jensen, and Roll (1969). The key focus is still on measuring the sample securities' mean and cumulative mean abnormal return around the time of an event.

Two main changes in methodology have taken place, however. First, the use of daily (and sometimes intraday) rather than monthly security return data has become prevalent, which permits more precise measurement of abnormal returns and more informative studies of announcement effects. Second, the methods used to estimate abnormal returns and calibrate their statistical significance have become more sophisticated. This second change is of particular importance for long-horizon event studies. The changes in long-horizon event study methods reflect new findings in the late 1990s on the statistical properties of long-horizon security returns. The change also parallels developments in the asset pricing literature, particularly the Fama-French 3-factor model.

While long-horizon methods have improved, serious limitations of long-horizon methods have been brought to light and still remain. We now know that inferences from long-horizon tests "require extreme caution" (Kothari and Warner, 1997, p. 301) and even using the best methods "the analysis of long-run abnormal returns is treacherous" (Lyon, Barber, and Tsai, 1999, p. 165). These developments underscore and dramatically strengthen earlier warnings (e.g., Brown and Warner, 1980, p. 225) about the reliability – or lack of reliability - of long-horizon methods. This contrasts with short-horizon

methods, which are relatively straightforward and trouble-free. As a result, we can have more confidence and put more weight on the results of short-horizon tests than long-horizon tests. Short-horizon tests represent the "cleanest evidence we have on efficiency" (Fama, 1991, p.1602), but the interpretation of long-horizon results is problematic. As discussed later, long-horizon tests are highly susceptible to the joint-test problem (see sections 3.4 and 3.5), and have lower power.

Of course these statements about properties of event study tests are very general. To provide a meaningful basis for assessing the usefulness of event studies - both short- and long-horizon – it is necessary to have a framework that specifies: i) the economic and statistical hypotheses in an event study, and ii) an objective basis for measuring and comparing the performance of event study methods. Section 3 lays out this framework, and summarizes general conclusions from the methodology literature. In the remainder of the chapter, additional issues and problems are considered with more specificity.

3. Characterizing Event Study Methods

3.1 An event study: the model

An event study typically tries to examine return behavior for a sample of firms experiencing a common type of event (e.g., a stock split). The event might take place at different points in calendar time or it might be clustered at a particular date (e.g., a regulatory event affecting an industry or a subset of the population of firms). Let t = 0 represent the time of the event. For each sample security i, the return on the security for time period t relative to the event, $R_{it,}$ is:

$$R_{it} = K_{it} + e_{it} \qquad (1)$$

where K_{it} is the "normal" (i.e., expected or predicted return given a particular model of expected returns), and e_{it} is the component of returns which is abnormal or unexpected.[2] Given this return decomposition, the abnormal return, e_{it}, is the difference between the observed return and the predicted return:

$$e_{it} = R_{it} - K_{it} \quad (2)$$

Equivalently, e_{it} is the difference between the return conditional on the event and the expected return unconditional on the event. Thus, the abnormal return is a direct measure of the (unexpected) change in securityholder wealth associated with the event. The security is typically a common stock, although some event studies look at wealth changes for firms' preferred or debt claims.

A model of normal returns (i.e., expected returns unconditional on the event but conditional on other information) must be specified before an abnormal return can be defined. A variety of expected return models (e.g., market model, constant expected returns model, capital asset pricing model) have been used in event studies.[3] Across alternative methods, both the bias and precision of the expected return measure can differ, affecting the properties of the abnormal return measures. Properties of different methods have been studied extensively, and are discussed later.

3.2 Statistical and economic hypotheses

Cross-sectional aggregation. An event study seeks to establish whether the cross-sectional distribution of returns at the time of an event is abnormal (i.e., systematically different from predicted). Such an exercise can be conducted in many

[2] This framework is from Brown and Warner (1980) and Campbell, Lo, and MacKinlay (1997).

[3] For descriptions of each of these models, see Brown and Warner (1985) or Campbell, Lo, and MacKinlay (1997).

ways. One could, for example, examine the entire distribution of abnormal returns. This is equivalent comparing the distributions of actual with the distribution of predicted returns and asking whether the distributions are the same. In the event study literature, the focus almost always is on the mean of the distribution of abnormal returns. Typically, the specific null hypothesis to be tested is whether the mean abnormal return (sometimes referred to as the average residual, AR) at time t is equal to zero. Other parameters of the cross-sectional distribution (e.g., median, variance) and determinants of the cross-sectional variation in abnormal returns are sometimes studied as well. The focus on mean effects, i.e., the first moment of the return distribution, makes sense if one wants to understand whether the event is, on average, associated with a change in security holder wealth, and if one is testing economic models and alternative hypotheses that predict the sign of the average effect. For a sample of *N* securities, the cross-sectional mean abnormal return for any event month t is:

$$AR_t = \tfrac{1}{N} \sum_{i=1}^{N} e_{it} \quad . \tag{3}$$

Time-series aggregation. It is also of interest to examine whether mean abnormal returns for periods around the event are equal to zero. First, if the event is partially anticipated, some of the abnormal return behavior related to the event should show up in the pre-event period. Second, in testing market efficiency, the speed of adjustment to the information revealed at the time of the event is an empirical question. Thus, examination of post-event returns provides information on market efficiency.

In estimating the performance measure over any multi-period interval (e.g., time 0 through +6), there are a number of methods for time-series aggregation over the period of

interest. The cumulative average residual method (CAR) uses as the abnormal performance measure the sum of each month's average abnormal performance. Later, we also consider the buy-and-hold method, which first compounds each security's abnormal returns and then uses the mean compounded abnormal return as the performance measure. The CAR starting at time t_1 through time t_2 (i.e., horizon length $L = t_2 - t_1 + 1$) is defined as:

$$CAR(t_1, t_2) = \sum_{t=t_1}^{t_2} AR_t \ . \qquad (4)$$

Both CAR and buy-and-hold methods test the null hypothesis that abnormal performance is equal to zero. Under each method, the abnormal return measured is the same as the returns to a trading rule which buys sample securities at the beginning of the first period, and holds through the end of the last period. CARs and buy-and-hold abnormal returns correspond to security holder wealth changes around an event. Further, when applied to post-event periods, tests using these measures provide information about market efficiency, since systematically nonzero abnormal returns following an event are inconsistent with efficiency and imply a profitable trading rule (ignoring trading costs).

3.3 Sampling distributions of test statistics

For a given performance measure, such as the CAR, a test statistic is typically computed and compared to its assumed distribution under the null hypothesis that mean abnormal performance equals zero.[4] The null hypothesis is rejected if the test statistic exceeds a critical value, typically corresponding to the 5% or 1% tail region (i.e., the test level or size of the test is 0.05 or 0.01). The test statistic is a random variable because

[4] Standard tests are "classical" rather than "Bayesian." A Bayesian treatment of event studies is beyond the scope of this chapter.

abnormal returns are measured with error. Two factors contribute to this error. First, predictions about securities' unconditional expected returns are imprecise. Second, individual firms' realized returns at the time of an event are affected for reasons unrelated to the event, and this component of the abnormal return does not average to literally zero in the cross-section.

For the CAR shown in equation (4), a standard test statistic is the CAR divided by an estimate of its standard deviation.[5] Many alternative ways to estimate this standard deviation have been examined in the literature (see, for example, Campbell, Lo, and MacKinlay, 1997). The test statistic is given by:

$$\frac{CAR(t_1, t_2)}{[s^2(t_1, t_2)]^{1/2}}, \tag{5}$$

where

$$s^2(t_1, t_2) = L\, s^2(AR_t) \tag{6}$$

and $s^2(AR_t)$ is the variance of the one-period mean abnormal return.. Equation (6) simply says that the CAR has a higher variance the longer is *L*, and assumes time-series independence of the one-period mean abnormal return. The test statistic is typically assumed unit normal in the absence of abnormal performance. This is only an approximation, however, since estimates of the standard deviation are used.

The test statistic in eq. (5) is well-specified provided the variance of one-period mean abnormal return is estimated correctly. Event-time clustering renders the

[5] An alternative would be a test statistic that aggregates standardized abnormal returns, which means each observation is weighted in inverse proportion of the standard deviation of the estimated abnormal return. The standard deviation of abnormal returns is estimated using time-series return data on each firm. While a test using standardized abnormal returns is in principle superior under certain conditions, empirically in short-horizon event studies it typically makes little difference (see Brown and Warner, 1980, and 1985).

independence assumption for the abnormal returns in the cross-section incorrect (see Collins and Dent, 1984, and Bernard, 1987, and more detailed discussion in section 4 below). This would bias the estimated standard deviation estimate downward and the test statistic given in eq. (5) upward. To address the bias, the significance of the event-period average abnormal return can be and often is gauged using the variability of the time series of event portfolio returns in the period preceding or after the event date. That is, the researcher constructs a portfolio of event firms and obtains a time series of abnormal returns on the portfolio for a number of days (e.g., 180 days) around the event date. The standard deviation of the portfolio returns can be used to assess the significance of the event-window average abnormal return. The cross-sectional dependence is accounted for because the variability of the portfolio returns through time incorporates whatever cross-dependence that exists among the returns on individual event securities.

The portfolio return approach, however, has a drawback. To the extent the event period is associated with increased uncertainty, i.e., greater return variability, the use of historical or post-event time-series variability might understate the true variability of the event-period abnormal performance. An increase in event-period return variability is economically intuitive. The event might have been triggered by uncertainty-increasing factors and/or the event itself causes uncertainty in the economic environment for the firm. In either case, the event-period return variability is likely to exceed that during other time periods for the event firms. Therefore, the statistical significance of the event-window abnormal performance would be overstated if it is evaluated on the basis of historical variability of the event-firm portfolio returns (see Brown and Warner, 1980, and 1985, and Collins and Dent, 1984). One means of estimating the likely increase in

the variability of event-period returns is to estimate the cross-sectional variability of returns during the event and non-event periods. The ratio of the variances during the event period and non-event periods might serve as an estimate of the degree of increase in the variability of returns during the event period, which can be used to adjust for the bias in the test statistic calculated ignoring the increased event-period uncertainty.[6]

3.4 Criteria for "reliable" event study tests

Using the test statistics, errors of inference are of two types. A Type I error occurs when the null hypothesis is falsely rejected. A Type II error occurs when the null is falsely accepted. Accordingly, two key properties of event study tests have been investigated. The first is whether the test statistic is correctly specified. A correctly-specified test statistic yields a Type I error probability equal to the assumed size of the test. The second concern is power, i.e., a test's ability to detect abnormal performance when it is present. Power can be measured as one minus the probability of a Type II error. Alternatively, it can be measured as the probability that the null hypothesis will be rejected given a level of Type I error and level of abnormal performance. When comparing tests that are well-specified, those with higher power are preferred.

3.5 Determining specification and power

The joint-test problem. While the specification and power of a test can be statistically determined, economic interpretation is not straightforward because all tests are joint tests. That is, event study tests are well-specified only to the extent that the assumptions underlying their estimation are correct. This poses a significant challenge because event study tests are joint tests of whether abnormal returns are zero and of

[6] Use of non-parametric tests of significance, as suggested in Corrado (1989), might also be effective in performing well-specified tests in the presence of increased event-period uncertainty.

whether the assumed model of expected returns (i.e. the CAPM, market model, etc.) is correct. Moreover, an additional set of assumptions concerning the statistical properties of the abnormal return measures must also be correct. For example, a standard t-test for mean abnormal performance assumes, among other things, that the mean abnormal performance for the cross-section of securities is normally distributed. Depending on the specific t-test, there may be additional assumptions that the abnormal return data are independent in time-series or cross-section. The validity of these assumptions is often an empirical question. This is particularly true for small samples, where one cannot rely on asymptotic results or the central limit theorem.

Brown-Warner simulation. To directly address the issue of event study properties, the standard tool in event study methodology research is to employ simulation procedures that use actual security return data. The motivation and specific research design is initially laid out in Brown and Warner (1980, 1985), and has been followed in almost all subsequent methodology research.

Much of what is known about general properties of event study tests comes from such large-scale simulations. The basic idea behind the event study simulations is simple and intuitive.[7] Different event study methods are simulated by repeated application of each method to samples that have been constructed through a random selection of securities and random selection of an event date to each. If performance is measured correctly, these samples should show no abnormal performance, on average. This makes it possible to study test statistic specification, that is, the probability of rejecting the null hypothesis when it is known to be true. Further, various levels of abnormal performance

[7] This characterization of simulation is from Brown and Warner (1985, p. 4).

can be artificially introduced into the samples. This permits direct study of the power of event study tests, that is, the ability to detect a given level of abnormal performance.

Analytical methods. Simulation methods seem both natural and necessary to determine whether event study test statistics are well-specified. Once it has been established using simulation methods that a particular test statistic is well-specified, analytical procedures have also been used to complement simulation procedures. Although deriving a power function analytically for different levels of abnormal performance requires additional distributional assumptions, the evidence in Brown and Warner (1985, p. 13) is that analytical and simulation methods yield similar power functions for a well-specified test statistic. As illustrated below, these analytical procedures provide a quick and simple way to study power.

3.6 A quick summary of our knowledge

Qualitative properties. Table 2 highlights, in qualitative terms, what is known about the properties of event study tests. The table shows the characteristics of event study methods along three dimensions: specification, power against specific types of alternative hypotheses, and the sensitivity of specification to assumptions about the return generating process. The table also shows how these properties can differ sharply for short and long horizon studies. Much of the remainder of the chapter deals with the full details of this table.

From Table 2, horizon length has a big impact on event study test properties. First, short-horizon event study methods are generally well-specified, but long-horizon methods are sometimes very poorly specified. While much is understood about how to reduce misspecification in long horizon studies (see section 4), no procedure in whose

specification researchers can have complete confidence has yet been developed. Second, short-horizon methods are quite powerful if (but only if) the abnormal performance is concentrated in the event window. For example, a precise event date is known for earnings announcements, but insider trading events might be known to have occurred only sometime during a one-month window. In contrast to the short-horizon tests, long-horizon event studies (even when they are well-specified) generally have low power to detect abnormal performance, both when it is concentrated in the event window and when it is not. That power to detect a given level of abnormal performance is decreasing in horizon length is not surprising, but the empirical magnitudes are dramatic (see below). Third, with short-horizon methods the test statistic specification is not highly sensitive to the benchmark model of normal returns or assumptions about the cross-sectional or time-series dependence of abnormal returns. This contrasts with long-horizon methods, where specification is quite sensitive to assumptions about the return generating process.

Along several lines, however, short- and long-horizon tests show similarities, and these results are easy to show using either simulation or analytical procedures. First, a common problem shared by both short- and long-horizon studies is that when the variance of a security's abnormal returns conditional on the event increases, test statistics can easily be misspecified, and reject the null hypothesis too often. This problem was first brought to light and has been studied mainly in the context of short-horizon studies (Brown and Warner, 1985, and Corrado, 1989). A variance increase is indistinguishable from abnormal returns differing across sample securities at the time of an event, and would be expected for an event. Thus, this issue is likely to be empirically relevant both in a short- and long-horizon context as well. Second, power is higher with increasing

sample size, regardless of horizon length. Third, power depends on the characteristics of firms in the event study sample. In particular, firms experiencing a particular event can have nonrandom size and industry characteristics. This is relevant because individual security variances (and abnormal return variances) exhibit an inverse relation to firm size and can vary systematically by industry. Power is inversely related to sample security variance: the noisier the returns, the harder to extract a given signal. As shown below, differences in power by sample type can be dramatic.

Quantitative results. To provide additional texture on Table 2, below we show specific quantitative estimates of power. We do so using the test statistic shown previously in equations (5) and (6), using two-tailed tests at the 0.05 significance level.[8] Since this test statistic is well-specified, at least at short horizons, the power functions are generated using analytic (rather than simulation) procedures. The estimates are for illustrative purposes only, however, and only represent "back of the envelope" estimates. The figures and the test statistic on which they are based assume independence of the returns (both through time and in the cross-section), and that all securities within a sample have the same standard deviation. The power functions also assume that return and abnormal return variances are the same (i.e., the model of abnormal returns is the "mean-adjusted returns" model of Brown and Warner, 1980).

Volatility. In calculating the test statistic in an event study, a key input required here is the individual security return (or abnormal return) variance (or standard deviation). To determine a reasonable range of standard deviations, we estimate daily

[8] This format for displaying power functions is similar to Campbell, Lo, and MacKinlay (1997, pp. 168-172). Our test statistic and procedures are the same as for their test statistic J1, but as discussed below we use updated variance inputs.

standard deviations for all CRSP listed firms from 1990 to 2002. Specifically, for each year, we: i) calculate each stock's standard deviation, and ii) assign firms to deciles ranked by standard deviation. From each decile, the averages of each year's mean and median values are reported in Table 3. The mean daily standard deviation for all firms is 0.053. This is somewhat higher than the value of 0.026 reported by Brown and Warner (1985, p. 9) for NYSE/AMEX firms and the value of 0.035 reported by Campbell and Wasley (1993, p. 79) for NASDAQ firms. The differences reflect that individual stocks have become more volatile over time (Campbell, Lettau, Malkiel, Xu, 2001)). This is highly relevant because it suggests that the power to detect abnormal performance for events over 1990-2002 is lower than for earlier periods. From Table 3, there is wide variation across the deciles. Firms in decile 1 have a mean daily standard deviation of 0.014, compared to 0.118 for decile 10. The figure of 0.118 for decile 10 seems very high, although this is likely to represent both very small firms and those with low stock prices. Further, there is a strong negative empirical relation between volatility and size. Our qualitative results apply if ranking is by firm size, so table 3 is not simply picking up measurement error in volatility.

Results. Figure 1 shows how, for a sample comprised of securities of average risk and 10% abnormal performance, the power to detect abnormal performance falls with horizon length. This level of abnormal performance seems economically highly significant. If the abnormal performance is concentrated entirely in one day (and the day in known with certainty), a sample of only six stocks detects this level of abnormal performance 100% of the time. In contrast, if the same abnormal performance occurs over six months, a sample size of 200 is required to detect the abnormal performance

even 65% of the time. These various rejection frequencies are lower than those using pre-1990 volatilities (not reported), although this is not surprising.

Figure 2a through 2c show related results using a one-day horizon for samples whose individual security standard deviations correspond to the average standard deviation for: the lowest decile (Figure 2a); all firms (Figure 2b); and the highest decile (Figure 2c). For decile 1 firms, with 1% abnormal performance a 90% rejection rate requires only 21 stocks. For firms in decile 10, even with 5% abnormal performance a 90% rejection rate requires 60 stocks. These comparisons may distort the differences in actual power if high variance firms are less closely followed and events are bigger surprises. When the effect of events differs cross-sectionally, analysis of test properties (i.e., power and specification) is more complicated.

Collectively, our results illustrate that power against alternative hypotheses can be sensitive to calendar time period and sample firm characteristics, and highlight the importance already recognized in the profession of studying test statistic properties for nonrandom samples. A complete analysis of these issues would focus on abnormal return (rather than return) volatility, and study how specification (and abnormal return distributional properties such as skewness) varies across time and firm characteristics.

4. Long-Horizon Event Studies

All event studies, regardless of horizon length, must deal with several basic issues. These include risk adjustment and expected/abnormal return modeling (Section 4.2), the aggregation of security-specific abnormal returns (Section 4.3), and the calibration of the statistical significance of abnormal returns (Section 4.4). These issues

become critically important with long horizons. The remainder of this chapter focuses on efforts in the long-horizon literature to deal with the issues

4.1 Background

Long-horizon event studies have a long history, including the original stock split event study by Fama, Fisher, Jensen, and Roll (1969). As evidence inconsistent with the efficient markets hypothesis started to accumulate in the late seventies and early eighties, interest in long-horizon studies continued. Evidence on the post-earnings announcement effect (see Ball and Brown, 1968, and Jones and Litzenberger, 1970), size effect (Banz, 1981), and earnings yield effect (Basu, 1977 and 1983) contributed to skepticism about the CAPM as well as market efficiency. This evidence prompted researchers to develop hypotheses about market inefficiency stemming from investors' information processing biases (see DeBondt and Thaler, 1985 and 1987) and limits to arbitrage (see DeLong et al., 1990a and 1990b, and Shleifer and Vishny, 1997).

The "anomalies" literature and the attempts to model the anomalies as market inefficiencies has led to a burgeoning field known as behavioral finance. Research in this field formalizes (and tests) the security pricing implications of investors' information processing biases.[9] Because the behavioral biases might be persistent and arbitrage forces might take a long time to correct the mispricing, a vast body of literature hypothesizes and studies abnormal performance over long horizons of one-to-five years following a wide range of corporate events. The events might be one-time (unpredictable) phenomena like an initial public offering or a seasoned equity offering, or they may be recurring events such as earnings announcements.

[9] See Shleifer (2000), Barberis, Shleifer, and Vishny (1998), Daniel, Hirshleifer, and Subramanyam (1998), Daniel, Hirshleifer, and Teoh (2002), Hirshleifer (2001), and Hong and Stein (1999).

Many long-horizon studies document apparent abnormal returns spread over long horizons. The literature on long-horizon security price performance following corporate events is summarized extensively in many studies, including Fama (1998), Kothari and Warner (1997), Schwert (2001), and Kothari (2001). Whether the apparent abnormal returns are due to mispricing, or simply the result of measurement problems, is a contentious and unresolved issue among financial economists. The methodological research in the area is important because it demonstrates how easy it is to conclude there is abnormal performance when none exists. Before questions on mispricing can be answered, better methods than currently exist are required.

We summarize some of the salient difficulties and the state-of-the-art event study methods for estimating long-horizon security price performance. More detailed discussions appear in Barber and Lyon (1997), Kothari and Warner (1997), Fama (1998), Brav (2000), Lyon, Barber, and Tsai (1999), Mitchell and Stafford (2000), Jegadeesh and Karceski (2004), and Viswanathan and Wei (2004).

4.2 Risk adjustment and expected returns

In long-horizon tests, appropriate adjustment for risk is critical in calculating abnormal price performance. This is in sharp contrast to short-horizon tests in which risk adjustment is straightforward and typically unimportant. The error in calculating abnormal performance due to errors in adjusting for risk a short-horizon test is likely to be small. Daily expected returns are about 0.05% (i.e., annualized about 12-13%). Therefore, even if the event firm portfolio's beta risk is misestimated by as much as 0.4 (e.g., estimated beta risk of 1.0 when true beta risk is 1.4), the abnormal return would be misestimated only by 0.02% per day. If the event-window is 3 days, then the event

portfolio's abnormal return would be misestimated by about 0.06%, which is economically small, especially compared to the abnormal return of 1% or more that is typically documented in short-window event studies. Not surprisingly, Brown and Warner (1985) conclude that simple risk-adjustment approaches to conducting short-window event studies are quite effective in detecting abnormal performance.

In multi-year long-horizon tests, risk-adjusted return measurement is the Achilles heel for at least two reasons. First, even a small error in risk adjustment can make an economically large difference when calculating abnormal returns over horizons of one year or longer, whereas such errors make little difference for short horizons. Thus the precision of the risk adjustment becomes far more important in long-horizon event studies. Second, it is unclear which expected return model is correct, and estimates of abnormal returns over long horizons are highly sensitive to model choice. We now discuss each of these problems in turn.

Errors in risk adjustment. Such errors can make an economically non-trivial difference in measured abnormal performance over one-year or longer periods. The problem of risk adjustment error is exacerbated in long-horizon event studies because the potential for such error is greater for longer horizons. In many event studies, (i) the event follows unusual prior performance (e.g., stock splits follow good performance), or (ii) the event sample consists of firms with extreme (economic) characteristics (e.g., low market capitalization syocks, low-priced stocks, or extreme book-to-market stocks), or (iii) the event is defined on the basis of unusual prior performance (e.g., contrarian investment strategies in DeBondt and Thaler, 1985, and Lakonishok, Shleifer, and Vishny, 1994). Under these circumstances, accurate risk estimation is difficult, with historical estimates

being notoriously biased because economic performance negatively impacts the risk of a security. Therefore, in long-horizon event studies, it is crucial that abnormal-performance measurement be on the basis of post-event, not historical risk estimates (see Ball and Kothari, 1989, Chan, 1988, and Ball, Kothari, and Shanken, 1995, and Chopra, Lakonishok, and Ritter, 1992). However, how the post-event risk should be estimated is itself a subject of considerable debate, which we summarize below in an attempt to offer guidance to researchers.

Model for expected returns. The question of which model of expected returns is appropriate remains an unresolved, contentious issue. As noted earlier, event studies are joint tests of market efficiency and a model of expected returns (e.g., Fama, 1970). On a somewhat depressing note, Fama (1998, p. 291) concludes that "all models for expected returns are incomplete descriptions of the systematic patterns in average returns," which can lead to spurious indications of abnormal performance in an event study. With the CAPM as a model of expected returns being thoroughly discredited as a result of the voluminous anomalies evidence, a quest for a better-and-improved model began. The search culminated in the Fama and French (1993) three-factor model, further modified by Carhart (1997) to incorporate the momentum factor. However, absent a sound economic rationale motivating the inclusion of the size, book-to-market, and momentum factors, whether these factors represent equilibrium compensation for risk or they are an indication of market inefficiency has not been satisfactorily resolved in the literature (see, e.g., Brav and Gompers, 1997). Fortunately, from the standpoint of event study analysis, this flaw is not fatal. Regardless of whether the size, book-to-market, and momentum factors proxy for risk or indicate inefficiency, it is essential to use them when measuring

abnormal performance. The purpose of an event study is to isolate the incremental impact of an event on security price performance. Since the price performance associated with the size, book-to-market, and momentum characteristics is applicable to all stocks sharing those characteristics, not just the sample of firms experiencing the event (e.g., a stock split), the performance associated with the event itself must be distinguished from that associated with other known determinants of performance, such as the aforementioned four factors.[10]

4.3 Approaches to abnormal performance measurement

While post-event risk-adjusted performance measurement is crucial in long-horizon tests, actual measurement is not straightforward. Two main methods for assessing and calibrating post-event risk-adjusted performance are used: characteristic-based matching approach and the Jensen's alpha approach, which is also known as the calendar-time portfolio approach (see Fama, 1998 or Mitchell and Stafford, 2000). Analysis and comparison of the methods is detailed below. Despite an extensive literature, there is still no clear winner in a horse race. Both have low power against economically interesting null hypotheses, and neither is immune to misspecification.

4.3.1 BHAR approach

In recent years, following the works of Ikenberry, Lakonishok, and Vermaelen (1995), Barber and Lyon (1997), Lyon et al. (1999), the characteristic-based matching approach (or also known as the buy-and-hold abnormal returns, BHAR) has been widely used. Mitchell and Stafford (2000, p. 296) describe BHAR returns as "the average multiyear return from a strategy of investing in all firms that complete an event and

[10] See Kothari, Leone, and Wasley (2005) for an extended discussion.

selling at the end of a prespecified holding period versus a comparable strategy using otherwise similar nonevent firms." An appealing feature of using BHAR is that buy-and-hold returns better resemble investors' actual investment experience than periodic (monthly) rebalancing entailed in other approaches to measuring risk-adjusted performance.[11] The joint-test problem remains in that any inference on the basis of BHAR hinges on the validity of the assumption that event firms differ from the "otherwise similar nonevent firms" only in that they experience the event. The researcher implicitly assumes an expected return model in which the matched characteristics (e.g., size and book-to-market) perfectly proxy for the expected return on a security. Since corporate events themselves are unlikely to be random occurrences, i.e., they are unlikely to be exogenous with respect to past performance and expected returns, there is a danger that the event and nonevent samples differ systematically in their expected returns notwithstanding the matching on certain firm characteristics. This makes matching on (unobservable) expected returns more difficult, especially in the case of event firms experiencing extreme prior performance.

Once a matching firm or portfolio is identified, BHAR calculation is straightforward. A T-month BHAR for event firm i is defined as:

$$BHAR_i(t, T) = \prod_{t = 1 \text{ to } T} (1 + R_{i,t}) - \prod_{t = 1 \text{ to } T} (1 + R_{B,t}) \qquad (7)$$

where R_B is the return on either a non-event firm that is matched to the event firm i, or it is the return on a matched (benchmark) portfolio.[12] If the researcher believes that the

[11] Apart from similarity with the actual investment experience, the BHAR approach also avoids biases arising from security microstructure issues when portfolio performance is measured with frequent rebalancing (see Blume and Stambaugh, 1983, Roll, 1983, and Ball, Kothari, and Shanken, 1995). The latter biases are also reduced if value-weight portfolio performance is examined.

[12] See Mitchell and Stafford (2000) for details.

Carhart (1997) four-factor model is an adequate description of expected returns, then firm-specific matching might entail identifying a non-event firm that is closest to an event firm on the basis of firm size (i.e., market capitalization of equity), book-to-market ratio, and past one-year return. Alternatively, characteristic portfolio matching would identify the portfolio of all non-event stocks that share the same quintile ranking on size, book-to-market, and momentum as the event firm (see Daniel, Grinblatt, Titman, and Wermers, 1997, or Lyon, Barber, and Tsai, 1997, for details of benchmark portfolio construction). The return on the matched portfolio is the benchmark portfolio return, R_B. For the sample of event firms, the mean BHAR is calculated as the (equal- or value-weighted) average of the individual firm BHARs.

4.3.2 Jensen-alpha approach

The Jensen-alpha approach (or the calendar-time portfolio approach) to estimating risk-adjusted abnormal performance is an alternative to the BHAR calculation using a matched-firm approach to risk adjustment. Jaffe (1974) and Mandelker (1974) introduced a calendar time methodology to the financial-economics literature, and it has since been advocated by many, including Fama (1998) and Mitchell and Stafford (2000).[13] The distinguishing feature of the most recent variants of the approach is to calculate calendar-time portfolio returns for firms experiencing an event, and calibrate whether they are abnormal in a multifactor (e.g., CAPM or Faama-French three factor) regression. The estimated intercept from the regression of portfolio returns against factor returns is the post-event abnormal performance of the sample of event firms.

[13] For a variation of the Jensen-alpha approach, see Ibbotson (1975) *returns across time and securities* (RATS) methodology, which is used in Ball and Kothari (1989) and others.

To implement the Jensen-alpha approach, assume a sample of firms experiences a corporate event (e.g., an IPO or an SEO).[14] The event might be spread over several years or even many decades (the sample period). Also assume that the researcher seeks to estimate price performance over two years (T = 24 months) following the event for each sample firm. In each calendar month over the entire sample period, a portfolio is constructed comprising all firms experiencing the event within the previous T months. Because the number of event firms is not uniformly distributed over the sample period, the number of firms included in a portfolio is not constant through time. As a result, some new firms are added each month and some firms exit each month. Accordingly, the portfolios are rebalanced each month and an equal or value-weighted portfolio excess return is calculated. The resulting time series of monthly excess returns is regressed on the CAPM market factor, or the three Fama-French (1993) factors, or the four Carhart (1997) factors as follows:

$$R_{pt} - R_{ft} = a_p + b_p (R_{mt} - R_{ft}) + s_p SMB_t + h_p HML_t + m_p UMD_t + e_{pt} \qquad (8)$$

where

R_{pt} is the equal or value-weighted return for calendar month t for the portfolio of event firms that experienced the event within previous T years,

R_{ft} is the risk-free rate,

R_{mt} is the return on the CRSP value-weight market portfolio,

SMB_{pt} is the difference between the return on the portfolio of "small" stocks and "big" stocks;

HML_{pt} is the difference between the return on the portfolio of "high" and "low" book-to-market stocks;

[14] The description here is based on Mitchell and Stafford (2000).

UMD_{pt} is the difference between the return on the portfolio of past one-year "winners" and "losers,"

a_p is the average monthly abnormal return (Jensen alpha) on the portfolio of event firms over the T-month post-event period,

b_p, s_p, h_p, and m_p are sensitivities (betas) of the event portfolio to the four factors.

Inferences about the abnormal performance are on the basis of the estimated a_p and its statistical significance. Since a_p is the average monthly abnormal performance over the T-month post-event period, it can be used to calculate annualized post-event abnormal performance.

Recent work on the implications of using the Jensen-alpha approach is mixed. For example, Mitchell and Stafford (2000) and Brav and Gompers (1997) favor the Jensen-alpha approach. However, Loughran and Ritter (2000) argue against using the Jensen-alpha approach because it might be biased toward finding results consistent with market efficiency. Their rationale is that corporate executives time the events to exploit mispricing, but the Jensen-alpha approach, by forming calendar-time portfolios, under-weights managers' timing decisions and over-weights other observations. In the words of Loughran and Ritter (2000, p. 362): "If there are time-varying misvaluations that firms capitalize on by taking some action (a supply response), there will be more events involving larger misvaluations in some periods than in others……In general, tests that weight firms equally should have more power than tests that weight each time period equally." Since the Jensen-alpha (i.e., calendar-time) approach weights each period equally, it has lower power to detect abnormal performance if managers time corporate events to coincide with misvaluations. As a means of addressing the problem, Fama

(1998) recommends weighting calendar months by the number of event observations in the month, or some other suitable approach to weighting monthly observations.

4.4 Significance tests for BHAR and Jensen-alpha measures

The choice between the matched-firm BHAR approach to abnormal return measurement and the calendar time Jensen-alpha approach (also known as the calendar-time portfolio approach) hinges on the researcher's ability to accurately gauge the statistical significance of the estimated abnormal performance using the two approaches. That is, unbiased standard errors for the distribution of the event-portfolio abnormal returns are not easy to calculate, which leads to test misspecification. Assessing the statistical significance of the event portfolio's BHAR has been particularly difficult because (i) long-horizon returns depart from the normality assumption that underlies many statistical tests; (ii) long-horizon returns exhibit considerable cross-correlation because the return horizons of many event firms overlap and also because many event firms are drawn from a few industries; and (iii) volatility of the event firm returns exceeds that of matched firms because of event-induced volatility. We summarize below the econometric inferential issues encountered in performing long-horizon tests and some of the remedies put forward in recent studies.

4.4.1 Skewness

Long-horizon buy-and-hold returns, even after adjusting for the performance of a matched firm (or portfolio), tend to be right skewed. The right skewness of buy-and-hold returns is not surprising because the lower bound is -100% and returns are unbounded on the upside. Skewness in abnormal returns imparts a skewness bias to long-horizon abnormal performance test statistics (see Barber and Lyon, 1997). Brav (2000, p. 1981)

concludes that "with a skewed-right distribution of abnormal returns, the Student t-distribution is asymmetric with a mean smaller than the zero null." While the right-skewness of individual firms' long-horizon returns is undoubtedly true, the extent of skewness bias in the test statistic for the hypothesis that mean abnormal performance for the portfolio of event firms is zero is expected to decline with sample size.[15] Fortunately, the sample size in long-horizon event studies is often several hundred observations (e.g., Teoh, Welch, and Wong, 1998, and Byun and Rozeff, 2003). Therefore, if the BHAR observations for the sample firms are truly independent, as assumed in using a t-test, the Central Limit Theorem's implication that "the sum of a large number of independent random variables has a distribution that is approximately normal" should apply (Ross, 1976, p. 252). The right-skewness of the distribution of long-horizon abnormal returns on event *portfolios*, as documented in, for example, Brav (2000) and Mitchell and Stafford (2000), appears to be due largely to the lack of independence arising from overlapping long-horizon return observations in event portfolios. That is, skewness in portfolio returns is in part a by-product of cross-correlated data rather than a direct consequence of skewed firm-level buy-and-hold abnormal (or raw) returns.

4.4.2 Cross-correlation

The issue. Specification bias arising due to cross-correlation in returns is a serious problem in long-horizon tests of price performance. Brav (2000, p. 1979) attributes the misspecification to the fact that researchers conducting long-horizon tests typically "maintain the standard assumptions that abnormal returns are independent and

[15] Simulation evidence in Barber and Lyon (1997) on skewness bias is based on samples consisting of 50 firms and early concern over skewness bias as examined in Neyman and Pearson (1928) and Pearson (1929a and 1929b) also refers to skewness bias in small samples.

normally distributed although these assumptions fail to hold even approximately at long horizons."[16] The notion that that economy-wide and industry-specific factors would generate contemporaneous co-movements in security returns is the cornerstone of portfolio theory and is economically intuitive and empirically compelling. Interestingly, the cross-dependence, although muted, is also observed in risk-adjusted returns.[17] The degree of cross-dependence decreases in the effectiveness of the risk-adjustment approach and increases in the homogeneity of the sample firms examined (e.g., sample firms clustered in one industry). Cross-correlation in abnormal returns is largely irrelevant in short-window event studies when the event is not clustered in calendar time. However, in long-horizon event studies, even if the event is not clustered in calendar time, cross-correlation in abnormal returns cannot be ignored (see Brav, 2000, Mitchell and Stafford, 2000, and Jegadeesh and Karceski, 2004). Long-horizon abnormal returns tend to be cross-correlated because: (i) abnormal returns for subsets of the sample firms are likely to share a common calendar period due to the long measurement period; (ii) corporate events like mergers and share repurchases exhibit waves (for rational economic reasons as well as opportunistic actions on the part of the shareholders and/or management); and (iii) some industries might be over-represented in the event sample (e.g., merger activity among technology stocks).

If the test statistic in an event study is calculated ignoring cross-dependence in data, even a fairly small amount of cross-correlation in data will lead to serious

[16] Also see Barber and Lyon (1997), Kothari and Warner (1997), Fama (1998), Lyon, Barber, and Tsai (1999), Mitchell and Stafford (2000), and Jegadeesh and Karaceski (2004).

[17] See Schipper and Thompson (1983), Collins and Dent (1984), Sefcik and Thompson (1986), Bernard (1987), Mitchell and Stafford (2000), Brav (2000), and Jegadeesh and Karceski (2004).

misspecification of the test. In particular, the test will reject the null of no effect far more often than the size of the test (see Collins and Dent, 1984, Bernard, 1987, and Mitchell and Stafford, 2000). The overrejection is caused by the downward biased estimate of the standard deviation of the cross-sectional distribution of buy-and-hold abnormal returns for the event sample of firms.

Magnitude of bias. To get an idea of approximate magnitude of the bias, we begin with the cross-sectional standard deviation of the event firms' abnormal returns, AR, assuming equal variances and pairwise covariances across all sample firms' abnormal returns:

$$\sigma_{AR} = [(1/N)\ \sigma^2 + ((N-1)/N)\ (\rho_{i,j}\ \sigma^2\ )]^{1/2} \qquad (9)$$

where N is the number of sample firms, σ^2 is the variance of abnormal returns, which is assumed to be the same for all firms; and $\rho_{i,j}$ is the correlation between firm i and j's abnormal returns, which is also assumed to be the same across all firms. The second term in the square brackets in eq. (9) is due to the cross-dependence in the data, and it would be absent if the standard deviation is calculated assuming independence in the data. The bias in the standard deviation assuming independence is given by the ratio of the "true" standard deviation allowing for dependence to the standard deviation assuming independence:

$$\sigma_{AR}\ (\text{Dependence}) / \sigma_{AR}\ (\text{Independence}) = [1 + (N-1)\ \rho_{i,j}\ ]^{1/2}$$

(10)

The ratio in eq. (10) is the factor by which the standard error in a test for the significance of abnormal performance is understated and therefore the factor by which the test statistic (e.g., t-statistic) itself is overstated. The ratio is increasing in the

pairwise cross-correlation, $\rho_{i,j}$. Empirical estimates of the average pairwise correlation between annual BHARs of event firms are about 0.02 to 0.03 (see Mitchell and Stafford, 2000). The average pairwise correlation in multi-year BHARs is likely to be greater than that for annual returns because Bernard (1987, table 1) reports that the average cross-correlations increase with return horizon. Assuming the average pairwise cross-sectional correlation to be only 0.02, for a sample of 100, the ratio is eq. (4) is 1.73, and it increases with both sample size and the degree of cross-correlation. Since the sample size in many long-horizon event studies is a few hundred securities, and the BHAR horizon is three-to-five years, even a modest degree of average cross-correlation in the data can inflate the test statistics by a factor of two or more. Therefore, accounting for cross-correlation in abnormal returns is crucial to drawing accurate statistical inferences in long-horizon event studies. Naturally, this has been a subject of intense interest among researchers.

Potential solutions. One simple solution to the potential bias due to cross-correlation is to use the Jensen-alpha approach. It is immune to the bias arising from cross-correlated (abnormal) returns because of the use of calendar-time portfolios. Whatever the correlation among security returns, the event portfolio's time series of returns in calendar time accounts for that correlation. That is, the variability of portfolio returns is influenced by the cross-correlation in the data. The statistical significance of the Jensen alpha is based on the time-series variability of the portfolio return residuals. Since returns in an efficient market are serially (almost) uncorrelated, on this basis the independence assumption in calculating the standard error and the t-statistic for the regression intercept (i.e., the Jensen alpha) seems quite appropriate. However, the

evidence is that this method is misspecified in nonrandom samples (Lyon et al., 1999, Table 10). This is unfortunate, given that the method seems simple and direct. The reasons for the misspecification are unclear (see Lyon et al.). Appropriate calibration under calendar time methods probably warrants further investigation.

In the BHAR approach, estimating standard errors that account for the cross-correlation in long-horizon abnormal returns is not straightforward. As detailed below, there has been much discussion, and some interesting progress. Statistically precise estimates of pairwise cross-correlations are difficult to come by for the lack of availability of many time-series observations of long-horizon returns to accurately estimate the correlations (see Bernard, 1987). The difficulty is exacerbated by the fact that the only a portion of the post-event-period might overlap with other firms. Researchers have developed bootstrap and pseudoportfolio-based statistical tests that might account for the cross-correlations and lead to accurate inferences.

Cross-correlation and skewness. Lyon et al. (1999) develop a bootstrapped skewness-adjusted t-statistic to address the cross-correlation and skewness biases. The first step in the calculation is the skewness-adjusted t-statistic (see Johnson, 1978). This statistic adjusts the usual t-statistic by two terms that are a function of the skewness of the distribution of abnormal returns (see eq. 5 in Lyon et al., 1999, p. 174). Notwithstanding the skewness adjustment, the adjusted t-statistic indicates overrejection of the null and thus warrants a further refinement. The second step, therefore, is to construct a bootstrapped distribution of the skewness-adjusted t-statistic (see Sutton, 1993, and Lyon et al., 1999). To bootstrap the distribution, a researcher must draw a large number (e.g., 1,000) of resamples from the original sample of abnormal returns and calculate the

skewness-adjsuted t-statistic using each resample. The resulting empirical distribution of the test statistics is used to ascertain whether the skewness-adjusted t-statistic for the original event sample falls in the α% tails of the distribution to reject the null hypothesis of zero abnormal performance.

The pseudoportfolio-based statistical tests infer statistical significance of the event sample's abnormal performance by calibrating against an empirical distribution of abnormal performance constructed using repeatedly-sampled pseudoportfolios.[18] The empirical distribution of average abnormal returns on the pseudoportfolios is under the null hypothesis of zero abnormal performance. The empirical distribution is generated by repeatedly constructing matched firm samples with replacement. The matching is on the basis of characteristics thought to be correlated with the expected rate of return. Following the Fama and French (1993) three-factor model, matching on size and book-to-market as expected return determinants is quite common (e.g., Lyon et al., 1999, Byun and Rozeff, 2003, and Gompers and Lerner, 2003). For each matched-sample portfolio, an average buy-and-hold abnormal performance is calculated as the raw return minus the benchmark portfolio return. It's quite common to use 1,000 to 5,000 resampled portfolios to construct the empirical distribution of the average abnormal returns on the matched-firm samples. This distribution yields empirical 5 and 95% cut-off probabilities against which the event-firm sample's performance is calibrated to infer whether or not the event-firm portfolio buy-and-hold abnormal return is statistically significant.

[18] See, for example, Brock, Lakonishok, and LeBaron (1992), Ikenberry, Lakonishok, and Vermaelen (1995), Ikenberry, Rankine, and Stice (1996), Lee (1997), Lyon, Barber, and Tsai (1999), Mitchell and Stafford (2000), and Byun and Rozeff (2003).

Unfortunately, the two approaches described above, which are aimed at correcting the bias in standard errors due to cross-correlated data, are not quite successful in their intended objective. Lyon et al. find pervasive test misspecification in non-random samples. Because the sample of firms experiencing a corporate event is not selected randomly by the researcher, correcting for the bias in the standard errors stemming from the non-randomness of the event sample selection is not easy. In a strident criticism of the use of bootstrap- and pseudoportfolio-based tests, Mitchell and Stafford (2000, p. 307) conclude that long-term event studies often incorrectly "claim that bootstrapping solves all dependence problems. However, that claim is not valid. Event samples are clearly different from random samples. Event firms have chosen to participate in a major corporate action, while nonevent firms have chosen to abstain from the action. An empirical distribution created by randomly selecting firms with similar size-BE/ME characteristics does not replicate the covariance structure underlying the original event sample. In fact, the typical bootstrapping approach does not even capture the cross-sectional correlation structure related to industry effects…." Jegadeesh and Karceski (2004, pp. 1-2) also note that the Lyon et al. (1999) approach is misspecified because it "assumes that the observations are cross-sectionally uncorrelated. This assumption holds in random samples of event firms, but is violated in nonrandom samples. In nonrandom samples where the returns for event firms are positively correlated, the variability of the test statistics is larger than in a random sample. Therefore, if the empiricist calibrates the distribution of the test statistics in random samples and uses the empirical cutoff points for nonrandom samples, the tests reject the null hypothesis of no abnormal performance too often."

Autocorrelation. To overcome the weaknesses in prior tests, Jegadeesh and Karaceski (2004) propose a correlation and heteroskedasticity-consistent test. The key innovation in their approach is to estimate the cross-correlations using a monthly time-series of portfolio long-horizon returns (see Jegadeesh and Karceski, 2004, section II.A for details). Because the series is monthly, but the monthly observations contain long-horizon returns, the time-series exhibits autocorrelation that is due to overlapping return data. The autocorrelation is, of course, due to cross-correlation in return data. The autocorrelation is expected to be positive for H-1 lags, where H is the number of months in the long horizon. The length of the time-series of monthly observations depends on the sample period during which corporate events being examined take place. Because of autocorrelation in the time series of monthly observations, the usual t-statistic that is a ratio of the average abnormal return to the standard deviation of the time series of the monthly observations would be understated. To obtain an unbiased t-statistic, the covariances (i.e., the variance-covariance matrix) should be taken into account. Jegadeesh and Karceski (2004) use the Hansen and Hodrick (1980) estimator of the variance-covariance matrix assuming homoskedasticity. They also use a heteroskedasticity-consistent estimator that "generalizes White's heteroskedasticity-consistent estimator and allows for serial covariances to be non-zero" (p. 8). In both random and non-random (industry) samples the Jegadeesh and Karceski (2004) tests perform quite well, and we believe these might be the most appropriate to reduce misspecification in tests of long-horizon event studies.

4.4.3 The bottom line

Despite positive developments in BHAR calibration methods, two general long-horizon problems remain. The first concerns power. Jegadeesh and Karceski report that their tests show no increase in power relative to that of the test employed in previous research, which already had low power. For example, even with seemingly huge cumulative abnormal performance (25% over 5 years) in a sample of 200 firms, the rejection rate of the null is typically under 50% (see their Table 6).

Second, as discussed earlier (Section 3.6), events are generally likely to be associated with variance increases, which are equivalent to abnormal returns varying across sample securities. Previous literature shows that variance increases induce misspecification, and can cause the null hypothesis to be rejected far too often. Thus, whether a high level of measured abnormal performance is due to chance or mispricing (or a bad model) is still difficult to empirically determine, unless the test statistic is adjusted downward to reflect the variance shift. Solutions to the variance shift issue include such intuitive procedures as forming subsamples with common characteristics related to the level of abnormal performance (e.g., earnings increase vs. decrease subsamples). With smaller subsamples, however, specification issues unrelated to variance shifts become more relevant.

Given the various power and specification issues, a challenge that remains for the profession is to continue to refine long-horizon methods. Whether calendar time, BHAR methods or some combination can best address long-horizon issues remains an open question.

References

Ball, R., and P. Brown (1968), An empirical evaluation of accounting income numbers, Journal of Accounting Research 6: 159-177.

Ball, R., and S. Kothari (1989), Nonstationary expected returns: Implications for tests of market efficiency and serial correlation in returns, Journal of Financial Economics 25: 51-74.

Ball, R., S. Kothari and J. Shanken (1995), Problems in measuring portfolio performance: An application to contrarian investment strategies, Journal of Financial Economics 38: 79-107.

Banz, R., (1981), The relationship between return and market value of common stocks, Journal of Financial Economics 9: 3-18.

Barber, B., and J. Lyon (1996), Detecting abnormal operating performance: The empirical power and specification of test statistics, Journal of Financial Economics 41: 359-399.

Barber, B., and J. Lyon (1997), Detecting long-run abnormal stock returns: The empirical power and specification of test statistics, Journal of Financial Economics 43: 341-372.

Barberis, N., A. Shleifer and R. Vishny (1998), A model of investor sentiment, Journal of Financial Economics 49: 307-343.

Basu, S., (1977), The investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient markets hypothesis, Journal of Finance 32: 663-682.

Basu, S., (1983), The relationship between earnings yield, market value, and returns for NYSE common stocks: Further evidence, Journal of Financial Economics 12: 129-156.

Beaver, W., (1968), The information content of annual earnings announcements, Journal of Accounting Research Supplement 6: 67-92.

Bernard, V., (1987), Cross-sectional dependence and problems in inference in market-based accounting research, Journal of Accounting Research 25: 1-48.

Blume, M., and R. Stambaugh (1983), Biases in computed returns: an application to the size effect, Journal of Financial Economics 12: 387-404.

Brav, A., (2000), Inference in long-horizon event studies: A Bayesian approach with application to initial public offerings, Journal of Finance 55: 1979-2016.

Brav, A., C. Geczy and P. Gompers (2000), Is the abnormal return following equity issuances anomalous? Journal of Financial Economics 56: 209-249.

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

Brock, W., J. Lakonishok and B. LeBaron (1992), Simple trading rules and the stochastic properties of stock returns, Journal of Finance 47: 1731-1764.

Brown, S., and J. Warner (1980), Measuring security price performance, Journal of Financial Economics 8: 205-258.

Brown, S., and J. Warner (1985), Using daily stock returns: The case of event studies, Journal of Financial Economics 14: 3-31.

Byun, J., and M. Rozeff (2003), Long-run performance after stock splits: 1927 to 1996, Journal of Finance 58: 1063-1085.

Campbell, C. and C. Wasley (1996), Measuring abnormal trading volume for samples of NYSE/ASE and NASDAQ securities using parametric and nonparametric test statistics, Review of Quantitative Finance and Accounting

Campbell, C. and C. Wasley (1993), Measuring security price performance using daily NASDAQ returns, Journal of Financial Economics 33: 73-92.

Campbell, J., M. Lettau, B. Malkiel and Y. Xu (2001), Have individual stocks become more volatile? An empirical exploration of idiosyncratic risk, Journal of Finance 56: 1-43.

Campbell, J., A. Lo and A. C. MacKinlay (1997), The Econometrics of Financial Markets (Princeton University Press).

Carhart, M., (1997), On persistence in mutual fund performance, Journal of Finance 52: 57-82.

Chan, K., (1988), On the contrarian investment strategy, Journal of Business 61: 147-163.

Chan, K. and J. Lakonishok (1993), Are the reports of beta's death premature? Journal of Portfolio Management 19: 51-62.

Chopra, N., J. Lakonishok and J. Ritter (1992), Measuring abnormal performance: Does the market overreact, Journal of Financial Economics 32: 235-268.

Collins, D. and W. Dent (1984), A comparison of alternative testing methodologies used in capital market research, Journal of Accounting Research 22: 48-84.

Corrado, C., (1989), A nonparametric test for abnormal security-price performance in event studies, Journal of Financial Economics 23: 385-395.

Daniel, K., M. Grinblatt, S. Titman and R. Wermers (1997), Measuring mutual fund performance with characteristic-based benchmarks, Journal of Finance 52: 1035-1058.

Daniel, K., D. Hirshleifer and A. Subrahmanyam (1998), Investor psychology and security market under- and overreactions, Journal of Finance 53: 1839-1885.

De Long, J., A. Shleifer, R. Vishny and R. Waldman (1990a), Noise trader risk in financial markets, Journal of Political Economy 98: 703-738.

De Long, J., A. Shleifer , R. Vishny and R. Waldman (1990b), Positive feedback investment strategies and destabilizing rational speculation, Journal of Finance 45: 375-395.

DeBondt, W., Thaler, R., (1985), Does the stock market overreact? Journal of Finance 40: 793-805.

DeBondt, W., and R. Thaler (1987), Further evidence of investor overreaction and stock market seasonality, Journal of Finance 42: 557-581.

Dechow, P., R. Sloan and A. Sweeney (1995), Detecting earnings management, The Accounting Review 70: 3-42.

Fama, E., (1970), Efficient capital markets: A review of theory and empirical work, Journal of Finance 25: 383-417.

Fama, E., (1991), Efficient capital markets: II, Journal of Finance 46: 1575-1617.

Fama, E., (1998), Market efficiency, long-term returns, and behavioral finance, Journal of Financial Economics 49: 283-306.

Fama, E., L. Fisher, M. Jensen and R. Roll (1969), The adjustment of stock prices to new information, International Economic Review 10: 1-21.

Fama, E., and K. French (1993), Common risk factors in the returns on stocks and bonds, Journal of Financial Economics 33: 3-56.

Gompers, P., and J. Lerner (2003), The really long-run performance of initial public offerings: The pre-Nasdaq evidence, Journal of Fiannce 58: 1355-1392.

Hirshleifer, D., (2001) Investor psychology and asset pricing, Journal of Finance 56: 1533-1598.

Hong, H., J. Stein (1999), A unified theory of underrreaction, momentum trading, and overreaction in asset markets, Journal of Finance 54: 2143-2184.

Ibbotson, R., (1975), Price performance of common stock new issues, Journal of Financial Economics 2: 235-272.

Ikenberry, D., J. Lakonishok and T. Vermaelen (1995), The underreaction to open market share repurchases, Journal of Financial Economics 39: 181-208.

Ikenberry, D., G. Rankine, and E. Stice (1996), What do stock splits really signal? Journal of Financial and Quantitative Analysis 31: 357-376.

Jaffe, J., (1974), Special information and insider trading, Journal of Business 47: 411-428.

Jarrell, G., J. Brickley and J. Netter (1988), The market for corporate control – The empirical evidence since 1980, Journal of Economic Perspectives 2: 49-68.

Jegadeesh, N., and J. Karceski (2004), Long-run performance evaluation: Correlation and heteroskedasticity-consistent tests, working paper, Emory University.

Jensen, M., and R. Ruback, (1983), The market for corporate control – The scientific evidence, Journal of Financial Economics 11: 5-50.

Jensen, M., and J. Warner (1988), The distribution of power among corporate managers, shareholders, and directors, Journal of Financial Economics 20: 3-24.

Johnson, N., (1978), Modified t tests and confidence intervals for symmetrical populations, Journal of American Statistical Association 73: 536-544.

Jones, C., and R. Litzenberger (1970), Quarterly earnings reports and intermediate stock price trends, Journal of Finance 25: 143-148.

Kothari, S., (2001), Capital markets research in accounting, Journal of Accounting & Economics 31: 105-231.

Kothari, S., A. Leone and C. Wasley (2005), Performance-matched discretionary accruals, Journal of Accounting & Economics, forthcoming.

Kothari, S., J. Shanken, and R. Sloan (1995), Another look at the cross-section of expected returns, Journal of Finance 50: 185-224.

Kothari, S., and J. Warner (1997), Measuring long-horizon security price performance, Journal of Financial Economics 43: 301-339.

Lakonishok, J., A. Shleifer and R. Vishny (1994), Contrarian investment, extrapolation, and risk, Journal of Finance 49: 1541-1578.

Lee, I., (1997), Do firms knowingly sell overvalued equity? Journal of Finance 52: 1439-1466.

Loughran, T., and J. Ritter (2000), Uniformly least powerful tests of market efficiency, Journal of Financial Economics 55: 361-389.

Lyon, J., B. Barber and C. Tsai (1999), Improved methods of tests of long-horizon abnormal stock returns, Journal of Finance 54: 165-201.

MacKinlay, A. C., (1997), Event studies in economics and finance, Journal of Economic Literature 35: 13-39.

Mandelker, G., (1974), Risk and return: The case of merging firms, Journal of Financial Economics 1: 303-335.

Mitchell, M., and E. Stafford (2000), Managerial decisions and long-term stock price performance, Journal of Business 73: 287-329.

Neyman, J., and E. Pearson (1928), On the use and interpretation of certain test criteria for purposes of statistical inference, part I, Biometrika 20A: 175-240.

Patell, J., (1976), Corporate forecasts of earnings per share and stock price behavior: Empirical tests, Journal of Accounting Research 14: 246-276.

Pearson, E., (1929a), The distribution of frequency constants in small samples from symmetrical distributions, Biometrika 21: 356-360.

Pearson, E., (1929b), The distribution of frequency constants in small samples from non-normal symmetrical and skew populations, Biometrika 21: 259-286.

Roll, R., (1983), On computing mean returns and the small firm premium, Journal of Financial Economics 12: 371-386.

Ross, S., (1976), A First Course in Probability (Macmillan, New York).

Schipper, K., and R. Thompson (1983), The impact of merger-related relationships on the shareholders of acquiring firms, Journal of Accounting Research 21: 184-221.

Schwert, G. W., (2001), Anomalies and Market Efficiency, in: G. Constantinides, M. Harris, and R. Stulz, eds, Handbook of the Economics of Finance (North-Holland, Amsterdam) 939-974.

Sefcik, S., and R. Thompson (1986), An approach to statistical inference in cross-sectional models with security abnormal returns as dependent variables, Journal of Accounting Research 24: 316-334.

Shleifer, A., (2000), Inefficient markets: An introduction to behavioral finance (Oxford University Press).

Shliefer, A., and R. Vishny (1997), The limits of arbitrage, Journal of Finance 52: 35-55.

Smith, C., (1986), Investment banking and the capital acquisition process, Journal of Financial Economics 15: 3-29.

Sutton, C., (1999), Computer-intensive methods for tests about the mean of an asymmetrical distribution, Journal of American Statistical Association 88: 802-808.

Teoh, S., I. Welch and T. Wong (1998), Earnings management and the long-run performance of initial public offerings, Journal of Finance 53: 1935-1974.

Viswanathan, S., and B. Wei (2004), Endogenous event and long run returns, working paper, Duke University.

Table 1

Event studies, by year and journal. For each journal, all papers that contain an event study are included. Survey and methodological papers are excluded.

Year	Journal of Business	Journal of Finance	Journal of Financial Economics	Journal of Financial and Quant. Analysis	Review of Financial Studies	Grand Total
1974	2		2	1		5
1975		2	2	1		5
1976		5	1	1		7
1977		5	5	1		11
1978	1	5	4	1		11
1979		7		2		9
1980	3	4	2	2		11
1981	1	3	4	2		10
1982	1	6	2	1		10
1983	2	14	18	4		38
1984		5	5	1		11
1985	2	4	7	2		15
1986	2	7	14	4		27
1987		7	18	1		26
1988	1	4	7	5	1	18
1989		11	11	1	1	24
1990	5	17	7	6	2	37
1991	5	17	2	4	1	29
1992	4	13	9	4	1	31
1993	5	7	5	5	3	25
1994	1	10	10	5		26
1995	1	8	14	11	2	36
1996	1	7	10	5	3	26
1997	3	8	12	3		26
1998	1	14	11	3		29
1999	1	7	12	1	4	25
2000	2	15	13	5	2	37
Totals	**44**	**212**	**207**	**82**	**20**	**565**

Table 2

General characterization of properties of event study test methods.

Criterion	Length of Event Window	
	Short (< 12 months)	**Long (12 months or more)**
Specification	Good	Poor/Moderate
Power when abnormal performance is:		
Concentrated in event window	High	Low
Not concentrated in event window	Low	Low
Sensitivity of test statistic specification to assumptions about the return generating process:		
Expected returns, unconditional on event	Low	High
Cross-sectional and time-series dependence of sample abnormal returns	Low/Moderate	Moderate/High
Variance of abnormal returns, conditional on event	High	High
Sensitivity of power to:		
Sample size	High	High
Firm characteristics (e.g., size, industry)	High	High

Table 3

Standard deviation of daily returns on individual securities using all CRSP common-stock securities from 1990-2002. For each year, firms are ranked by their estimated daily standard deviation. Firms with missing observations are excluded. The numbers under mean and median columns represent the average of the annual mean and median values for the firms in each decile and for all firms. The number of firms in each decile ranges from 504 in 2002 to 673 in 1997.

Decile	Standard Deviation	
	Mean	**Median**
1	0.014	0.014
2	0.019	0.019
3	0.023	0.023
4	0.028	0.028
5	0.033	0.033
6	0.039	0.039
7	0.046	0.046
8	0.055	0.055
9	0.069	0.068
10	0.118	0.098
All Firms	**0.053**	**0.053**

Figure 1: Power of event study test statistic when abnormal return is 10%

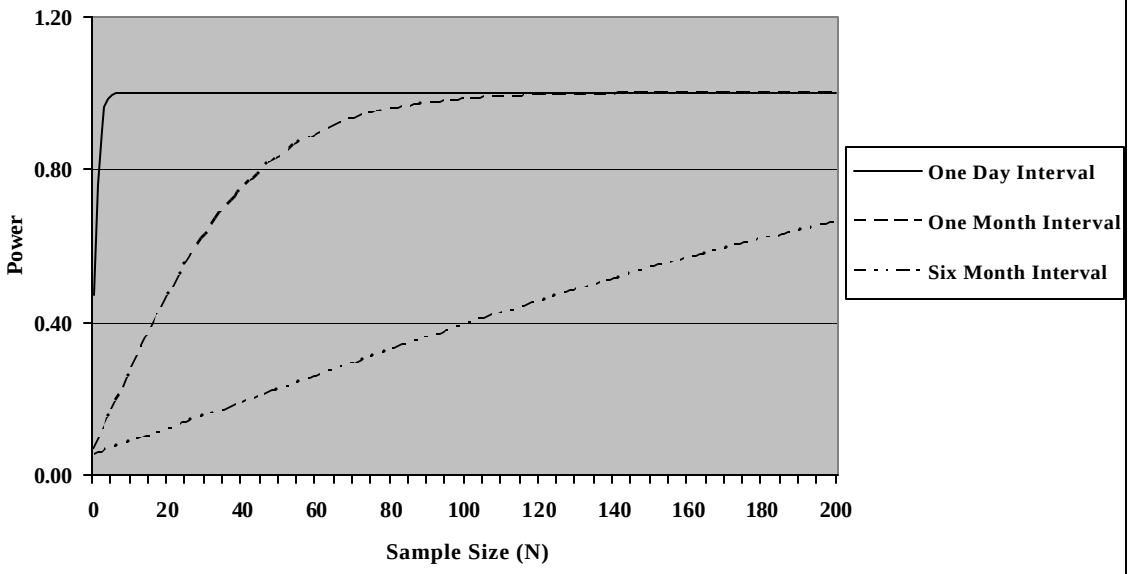

Figure 2a: Power of event study for firms in the lowest volatility decile.

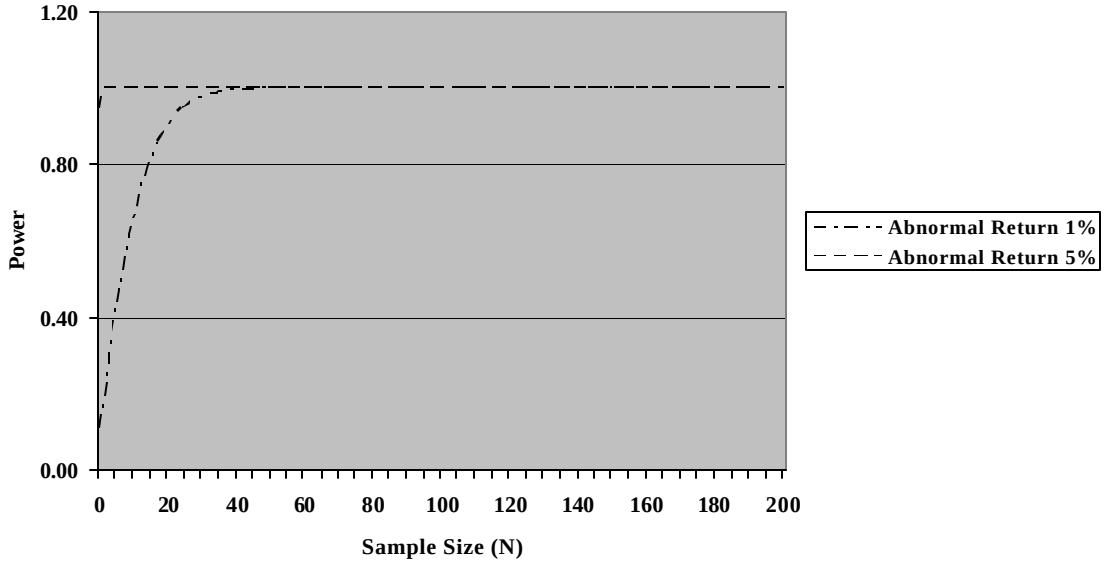

Figure 2b: Power of event study for firms with average volatility

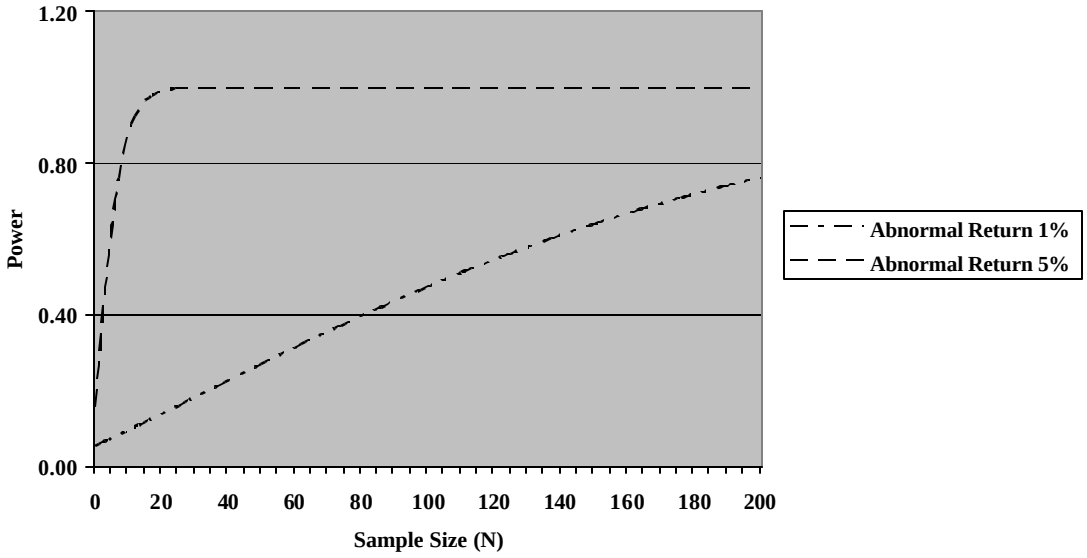

Figure 2c: Power of event study for firms in the highest volatility decile.