Journal of Financial Economics 8 (1980) 205–258. © North-Holland Publishing Company

# MEASURING SECURITY PRICE PERFORMANCE*

Stephen J. BROWN

*Bell Laboratories, Murray Hill, NJ 07974, USA*

Jerold B. WARNER

*University of Rochester, Rochester, NY 14627, USA*

Received January 1980, revised version received April 1980

Event studies focus on the impact of particular types of firm-specific events on the prices of the affected firms' securities. In this paper, observed stock return data are employed to examine various methodologies which are used in event studies to measure security price performance. Abnormal performance is introduced into this data. We find that a simple methodology based on the market model performs well under a wide variety of conditions. In some situations, even simpler methods which do not explicitly adjust for marketwide factors or for risk perform no worse than the market model. We also show how misuse of any of the methodologies can result in false inferences about the presence of abnormal performance.

## 1. Introduction and summary

The impact of particular types of firm-specific events (e.g., stock splits, earnings reports) on the prices of the affected firms' securities has been the subject of a number of studies. A major concern in those 'event' studies has been to assess the extent to which security price performance around the time of the event has been abnormal — that is, the extent to which security returns were different from those which would have been appropriate, given the model determining equilibrium expected returns.

Event studies provide a direct test of market efficiency. Systematically nonzero abnormal security returns which persist after a particular type of event are inconsistent with the hypothesis that security prices adjust quickly to fully reflect new information. In addition, to the extent that the event is unanticipated, the magnitude of abnormal performance at the time the event actually occurs is a measure of the impact of that type of event on the wealth of the firms' claimholders. Any such abnormal performance is consistent with market efficiency, however, since the abnormal returns would only have been

*Financial support for this research was provided to J.B. Warner by the Managerial Economics Research Center, University of Rochester, and the Institute for Quantitative Research in Finance, Columbia University. We are indebted to numerous colleagues for their help on this paper. We are especially grateful to Michael Gibbons and Michael Jensen for their assistance.

attainable by an investor if the occurrence of the event could have been predicted with certainty.

In this paper, observed stock return data are employed to examine various methodologies which are used in event studies to measure security price performance. Abnormal performance is introduced into this data. We assess the likelihood that various methodologies will lead to Type I errors — rejecting the null hypothesis of no abnormal performance when it is true, and Type II errors — failing to reject the null hypothesis of no abnormal performance when it is false. Our concern is with the power of the various methodologies. Power is the probability, for a given level of Type I error and a given level of abnormal performance, that the hypothesis of no abnormal performance will be rejected. Since a test's power indicates its ability to discern the presence of abnormal performance, then all other things equal, a more powerful test is preferred to a less powerful test.

The use of various methodologies is simulated by repeated application of each methodology to samples which have been constructed by random selection of securities and random assignment of an 'event-date' to each. Randomly selected securities should not, on average, exhibit any abnormal performance. Thus, for a large number of applications of a given methodology, we examine the frequency of Type I errors. Abnormal performance is then artificially introduced by transforming each sample security's return around the time it experiences a hypothetical event. Each methodology is then applied to a number of samples where the return data have thus been transformed. For each methodology, and for various levels of abnormal performance, this technique provides direct measures of the frequency of Type II errors. Since, for any given level of abnormal performance, the power of a test is equal to one minus the probability of a Type II error, this technique thus allows us to examine the power of the methodologies and the ability to detect abnormal performance when it is present.

*Overview of the paper*

General considerations which are relevant to measuring abnormal security price performance are discussed in section 2. Performance measures in event studies are classified into several categories: Mean Adjusted Returns, Market Adjusted Returns, and Market and Risk Adjusted Returns. In section 3, we specify methodologies which are based on each of these performance measures and which are representative of current practice. We then devise a simulation procedure for studying and comparing these methods, and their numerous variations.

Initial results are presented in section 4. For each methodology, the probability of Type I and Type II errors is assessed for both parametric and non-parametric significance tests. In addition, the distributional properties of the test statistics generated by each methodology are examined. We also



focus on different ways in which actual event studies take into account the systematic risk of the sample securities. The risk adjustment methods we compare are based on market model residuals, Fama–MacBeth residuals, and what we call Control Portfolios.

In section 5, we discuss the effect of imprecise prior information about the timing of the event on the power of the tests. The use of the Cumulative Average Residual procedure suggested by Fama, Fisher, Jensen and Roll (1969) is also investigated.

In section 6, two forms of sample security 'clustering' are examined. We first look at the calendar time clustering of events, and examine the characteristics of the tests when all sample securities experience an event during the same calendar time period. We then examine how the tests are affected when all sample securities have higher than average (or lower than average) systematic risk.

Section 7 examines the effect of the choice of market index on the various tests. Section 8 reports additional simulation results. The sensitivity of earlier simulation results to the number of sample securities is investigated. Evidence is also presented on the likelihood that the various test methods will, for a given sample, lead to the same inference.

Our conclusions, along with a summary of the paper's major results, are presented in section 9; an appendix contains a more detailed discussion of the specific performance assessment methods used in the study.

## 2. Measuring abnormal performance: General considerations

### *2.1. Defining abnormal performance for a security*

A security's price performance can only be considered 'abnormal' relative to a particular benchmark. Thus, it is necessary to specify a model generating 'normal' returns before abnormal returns can be measured. In this paper, we will concentrate on three general models of the process generating *ex ante* expected returns. These models are general representations of the models which have been assumed in event studies. For each model, the abnormal return for a given security in any time period $t$ is defined as the difference between its actual *ex post* return and that which is predicted under the assumed return-generating process, The three models are as follows.

(1) *Mean Adjusted Returns*

The Mean Adjusted Returns model assumes that the *ex ante* expected return for a given security $i$ is equal to a constant $K_i$ which can differ across securities: $E(\tilde{R}_i)=K_i$. The predicted *ex post* return on security $i$ in time

period $t$ is equal to $K_t$. The abnormal return $\varepsilon_{it}$ is equal to the difference between the observed return, $R_{it}$, and the predicted return $K_t$: $\varepsilon_{it}=R_{it}-K_t$.

The Mean Adjusted Returns model is consistent with the Capital Asset Pricing Model; under the assumption that a security has constant systematic risk and that the efficient frontier is stationary, the Asset Pricing Model also predicts that a security's expected return is constant.

(2) *Market Adjusted Returns*

This model assumes that *ex ante* expected returns are equal across securities, but not necessarily constant for a given security. Since the market portfolio of risky assets $M$ is a linear combination of all securities, it follows that $E(\tilde{R}_{it})=E(\tilde{R}_{mt})=K_t$ for any security $i$. The *ex post* abnormal return on any security $i$ is given by the difference between its return and that on the market portfolio: $\varepsilon_{it}=R_{it}-R_{mt}$. The Market Adjusted Returns model is also consistent with the Asset Pricing model if all securities have systematic risk of unity.

(3) *Market and Risk Adjusted Returns*

This model presumes that some version of the Capital Asset Pricing Model generates expected returns. For example, in the Black (1972) two-parameter Asset Pricing Model, $E(\tilde{R}_{it})=E(\tilde{R}_{zt})+\beta_i[E(\tilde{R}_{mt})-E(\tilde{R}_{zt})]=K_{it}$ for any security $i$, where $R_{zt}$ is the return on a minimum variance portfolio of risky assets which is uncorrelated with the market portfolio. In the Black model, the abnormal return $\varepsilon_{it}$ is equal to $R_{it}-[R_{zt}(1-\beta_i)+\beta_i R_{mt}]$.

For each of these three models, the return which will be realized on security $i$ in period $t$, $\tilde{R}_{it}$, is given by

$$\tilde{R}_{it}=K_{it}+\tilde{\varepsilon}_{it},$$

where $K_{it}$ is the expected return given by the particular model, and $\tilde{\varepsilon}_{it}$, which is unknown at the beginning of period $t$, is the component which is abnormal or unexpected.

### 2.2. Evaluating alternative performance measures

Under each model of the return generating process, there will be times when the realized return on a given security is different from that which was predicted. However, returns in an efficient market cannot systematically differ from those which are predicted. That is, the expected value of the unexpected component, $\tilde{\varepsilon}_{it}$, of a security's return cannot systematically differ from zero.

Let $I$ be an integer which is equal to 0 when no 'event' takes place, and



equal to 1 when a particular event does take place. In an efficient market, the abnormal return measure $\varepsilon_{it}$, if correctly specified, must be such that

$$E(\tilde{\varepsilon}_{it})=[E(\tilde{\varepsilon}_{it}\,|\,I=0)]p(I=0)+[E(\tilde{\varepsilon}_{it}\,|\,I=1)]p(I=1)=0;$$

abnormal returns conditional on the event can systematically be non-zero, as can abnormal returns conditional on no event. The only restriction is that a security's abnormal return, weighted by its magnitude and probability of occurrence, have an expected value of zero. Under each model just discussed, the abnormal performance measure for every security has an unconditional mean of 0 if the model is correct. In that sense, the abnormal performance measures are unbiased for each model.

Of course, another major purpose of specifying the 'correct' model for expected returns is to reduce the variance of the abnormal return component $\varepsilon_{it}$. For example, in the Market and Risk Adjusted Returns model, a contemporaneous relationship between realized security returns and realized market returns is predicted by the *ex ante* model. In an event study, where the market return which was observed at the time of each firm's event is known, the variance of the abnormal component of returns will be lower if a model takes into account the *ex post* relationship between a security's return and that of the market. When the *ex post* return generating process is correctly specified, abnormal performance, which is just the difference between returns conditional on the event and returns unconditional on the event, should be easier to detect.[1] Thus, if the Capital Asset Pricing model is correct, then the Market and Risk Adjusted Returns method, by bringing to bear additional information about the determinants of realized returns, such as the security's systematic risk and the market's return, could increase the power of the tests over the Mean Adjusted Returns method.[2]

[1]Our definition of abnormal performance as the difference between conditional (expected) and unconditional (expected) returns is consistent with the abnormal performance metric used in studies where the event is associated with either good news ($I=1$) or bad news ($I=0$) [e.g., Ball and Brown (1968)]. In such studies, abnormal performance is often measured as the average of the deviation from unconditional returns when there is good news and the deviation from unconditional returns when there is bad news, where the deviation from unconditional returns when there is bad news is first multiplied by $-1$. It can be shown that this abnormal performance measure is equal to our definition of abnormal performance conditional on good news, multiplied by twice the probability of good news. If good news has probability 0.5, the two abnormal performance measures will be identical; in general, the two measures differ only by a factor of proportionality. See Patell (1979, p. 536) for a related discussion.

[2]This line of argument has been pushed still further. The Asset Pricing model allows a security's return to be contemporaneously related to additional 'factors' as well. For example, a security's realized return could be related to the return on the securities of a particular industry. Even though there is no 'industry factor' in the *ex ante* Asset Pricing model, under certain conditions taking into account such an *ex post* relationship leads to more powerful tests. For a further discussion, see Warner (1977, p. 259) and Langetieg (1978). Fama and MacBeth (1973, pp. 634–635) and Roll and Ross (1979) also discuss related issues in the context of multiple factor models.

### 2.3. On the role of simulation

Unfortunately, specifying a more precise model of the processs generating realized returns is not sufficient for that model to generate a more powerful test for abnormal performance. Even if the Capital Asset Pricing model is the correct specification of the return generating process, it does not follow that a performance measure based upon that model will dominate performance measures based on the Mean Adjusted Returns method.

First, there is measurement error in each of the variables upon which abnormal returns depend in the Asset Pricing model. Not only is a security's risk measured with error, but, as Roll (1977) has argued, the market portfolio cannot be observed directly. Such measurement error need not introduce any systematic bias in event studies.[3] However, with small samples, the measurement error in these variables may be so large that it renders inconsequential any potential efficiency gains from more precise specification of the return-generating process.[4]

Second, the efficiency of using a particular model of the return-generating process will depend critically on the appropriateness of the additional peripheral assumptions about the $\tilde{\varepsilon}_{it}$ which must be made in order to test the hypothesis of 'no abnormal performance' conditional on a particular event. For example, with each method, a test statistic such as a $t$-statistic must be computed and compared to the distribution of test statistics which is assumed to obtain under the null hypothesis. To the extent that the assumed sampling distribution under the null hypothesis differs from the true distribution, false inferences can result. If the assumed properties of the test statistic under the Mean Adjusted Returns Method are more appropriate than those under the Market and Risk Adjusted Returns Method, the Mean Adjusted Returns Method can be preferred even if the second method is 'correct'.

Finally, there are a variety of ways of measuring abnormal returns under different variants of the Asset Pricing model. These include market model residuals, Fama–MacBeth residuals, and control portfolios. The differences in the predictive ability of such alternative methods could be substantial; the usefulness of the Asset Pricing model is not independent of the specific method of implementing the Market and Risk Adjusted Returns model.

Even if it were possible to analytically derive and compare the properties of alternative methods for measuring abnormal performance in event studies, conclusions from the comparison would not necessarily be valid if the actual data used in event studies were generated by a process which differed from that which the comparison assumed. For this reason, the performance of the

[3]See Mayers and Rice (1979) for a more detailed discussion of how the unobservability of the true market portfolio affects the measures of abnormal performance. Bawa, Brown, and Klein (1979) present an extensive discussion of how measurement error can be taken into account by using the predictive distribution of returns [see also Patell (1976, p. 256)].

[4]Brenner (1979) makes a similar point.



alternative methods is an empirical question. To address the question, we will employ simulation techniques which use actual security return data (presumably generated by the 'true' process) to examine the characteristics of various methodologies for measuring abnormal performance.

## 3. The experimental design

### 3.1. Sample construction

Our study concentrates on abnormal performance measurement using monthly data.[5] To simulate methodologies based on the three general models just discussed, we first construct 250 samples, each containing 50 securities. The securities are selected at random and with replacement from a population consisting of all securities for which monthly return data are available on the files of the Center for Research in Security Prices (CRSP) at the University of Chicago.[6] For each security, we generate a hypothetical 'event' month. Events are assumed to occur with equal probability in each month from June 1944 through February 1971.[7] Events can occur in different

[5]For our simulation study, monthly data offers several advantages over daily data. The use of monthly data enables us to consider those studies which have employed Fama–MacBeth (1973) residuals; the data necessary for readily computing daily Fama–MacBeth residuals are not to our knowledge available, and such daily residuals have not been used in any event study.

Furthermore, the use of daily data involves complications whose treatment is largely beyond the scope of this paper. Daily stock returns depart more from normality than do monthly returns [Fama (1976, ch. 1)]. In addition, the estimation of parameters (such as systematic risk) from daily data is a non-trivial matter due to the non-synchronous trading problem [see Scholes and Williams (1977)]. Any conclusions from simulations using daily data could be sensitive to specific procedures we employed to handle the complications associated with non-normality and non-synchronous trading.

We have no strong reason to believe that our conclusions about the relative performance of various methods for measuring abnormal performance would be altered by the use of daily data. However, in the absence of problems such as non-normality and non-synchronous trading, all of the methods for measuring abnormal performance are potentially more powerful with daily data. *First, daily returns have smaller standard deviations than do monthly returns. The mean* standard deviation of monthly returns for randomly selected securities is about 7.8% [Fama (1976, p. 123)], whereas the corresponding mean standard deviation of daily returns will be approximately 1.8% if daily returns are serially independent. In addition, as we later indicate, the power of all the methodologies increases with knowledge about precisely when an event occurs; use of daily data is potentially useful in that it permits the researcher to take advantage of prior information about the specific day of the month on which an event took place. Performance measurement with daily data is the subject of a separate study we are currently undertaking.

[6]We used a combination congruential and Tausworthe (shift register) algorithm to generate uniformly distributed random numbers on the [0, 1) interval. See Marsaglia, Ananthanarayanan and Paul (1973) for a description of the algorithm.

[7]Given the other data requirements we will discuss, including the requirement that Fama–MacBeth residuals be computed, these are the calendar months whose selection maximizes the length of the calendar time period over which our simulations can be performed. With the exception of mutual funds, all CRSP listed securities are eligible for selection. Each CRSP security initially has the same probability of being selected, subject to data availability. A security can be selected more than once for inclusion in a given sample or in a different sample. In both cases, whenever a security already selected is again selected, it is treated as a 'different' security in the sense that a new event-date is generated.

calendar months for different securities. This set of sample securities and hypothetical event dates will be used in most of the present study.

Define month '0' as the month in which the firm has been assigned an event. For a given sample, we use 100 return observations on each security for the period around the time of the event. We use 100 months of data, from month −89 through month +10.[8]

*Introducing abnormal performance*

Return data for the 250 samples which have been chosen is based on randomly selected securities and event dates, and, as indicated in section 2, should not systematically exhibit any abnormal performance. However, an important question we want to investigate is how different methodologies perform when some abnormal performance is present. It is thus necessary to specify a procedure for introducing a known level of abnormal performance into the sample securities.

A particular level of abnormal performance is artificially introduced into a given sample by transforming its actual return data. To introduce, say, 5% abnormal performance for each security of a sample, 0.05 is added to the actual return on each sample security in the particular calendar month in which its event is assumed to occur. Abnormal performance is thus introduced by adding a constant to a security's observed return.[9]

[8]If a security does not have this 100 months of return data surrounding its event-date, it is not included in the sample. To handle such cases, we continue to select securities and event-dates until, for a given sample, we have found 50 securities with a sufficient amount of data. With this selection procedure, the probability of being included in our sample will depend upon the amount of data which is available for a security. For example, a security with continuous return data from 1935 through 1971 will be included with a higher frequency than one with a smaller amount of available data. Thus, our data requirements introduce a bias towards including only surviving companies; none of our simulation results suggest that the bias is of importance.

[9]Three points about the procedure for introducing abnormal performance are worth mentioning. First, note that the level of abnormal performance associated with an actual event could itself be stochastic; an event could thus affect not only the conditional mean of a security's return, but higher-order moments as well. Introducing a constant represents a simple case which enables us to focus on the detection of mean shifts when an event takes place, holding constant the conditional variance. The detection of mean shifts is the relevant phenomenon to study when investigating how well different methodologies pick up the impact of an event on the value of the firm.

Second, although it is not critical for our purposes, it should also be noted that if for a given security there is positive abnormal performance conditional on an event, there should also be negative abnormal performance conditional on no event. Otherwise, the unconditional expected return on the security will be abnormal, which is inconsistent with an efficient market. However, for simulations introducing positive abnormal performance in month '0', the appropriate downward adjustment to security returns in those months when the event does not occur is not obvious. The adjustment which leaves expected returns unaltered will depend upon the *ex ante* probability of the event, which in an actual event study is unobservable.

For all results reported in this paper, in order to leave mean returns unaltered across all levels of abnormal performance, for each sample security the observed return for each month in the (−89, +10) period is reduced by the level of abnormal performance divided by 100. Roughly speaking, this transformation presumes that for each sample security the *ex ante* probability of the event in any one month is 0.01. Simulations have also been carried out with no such adjustment, and the results do not appear to be sensitive to whether or not such an adjustment procedure is used.

Finally, it should be noted that our simulations are directly applicable to the case where there is 'good news' or 'bad news'. We are implicitly examining abnormal performance for those securities which had good news; if month '0' had unconditional abnormal performance equal to zero, then there need be no adjustment to returns in the (−89, +10) period. Furthermore, we have also simulated a situation where, for a given sample security, good news (positive abnormal performance) or bad news (negative abnormal performance) occur with equal probability at month '0', and where the abnormal performance measure conditional on a bad news realization is multiplied by −1 before the null hypothesis of no abnormal sample security returns is tested. The results from such alternative simulations are quite similar to those reported in the paper, although there is a slight reduction in the degree of misspecification in the non-parametric tests.



*3.2. Abnormal performance measures for a given sample*

For every sample, we have a set of security returns which is transformed to reflect various levels of abnormal performance. For each sample, we calculate performance measures based on the three models of the return-generating process discussed in section 2. The performance measures are briefly summarized here; further details are contained in the appendix.

(a) *Mean Adjusted Returns* — To implement this model, we focus on the returns to each sample security around the time of its event. We examine whether or not the returns on the sample securities in month '0' are statistically significantly different from the returns on the securities in the time period surrounding the event. As discussed below, several different significance tests are used. The Mean Adjusted Returns method is used by Masulis (1978).

(b) *Market Adjusted Returns* — Unlike the Mean Adjusted Returns methodology, this method takes into account marketwide movements which occurred at the same time that the sample firms experienced events. The variable of interest is the *difference* between the return on a sample security and the corresponding return on the market index. We initially use the Fisher Equally Weighted Index to represent the market portfolio, and we will later examine the results when the CRSP Value Weighted Index is employed. The performance measures are the differences between the sample security returns and the market index in month '0'. Again, the statistical significance of the measures is assessed in several different ways. The Market Adjusted Returns method is used by Cowles (1933) and Latane and Jones (1979).

(c) *Market and Risk Adjusted Returns* — This method takes into account both market-wide factors and the systematic risk of each sample security. Although we will examine a number of different variations of this

method, we initially use the 'market model'.[10] For each sample security, we use ordinary least squares to regress its return over the period around the event against the returns on the Equally Weighted Index for the corresponding calendar months. The 'market model' regression which is performed yields a residual in each event related month for each sample security. The significance of the month '0' market model residuals is then examined.

*Detecting Type I and Type II errors for a given sample*

For a given sample, when no abnormal performance has been introduced we test whether or not, under each performance measure, the hypothesis of no abnormal performance is rejected. This null hypothesis should indeed be true if randomly selected securities do not, on average, exhibit any abnormal performance given a particular benchmark. We classify rejection of the null hypothesis here as a Type I error — rejecting it when it is true.

We then investigate how the methodologies perform when the null hypothesis is not true for the sample, that is, when the returns of the sample securities have been transformed to reflect abnormal performance. For a given level of abnormal performance introduced into every sample security, each methodology is applied and the hypothesis of no abnormal performance then tested. If the null hypothesis fails to be rejected, this is classified as a Type II error — failure to reject the null hypothesis of no abnormal performance when it is false.

## 4. Simulating the methodologies across samples: Procedure and initial results

Whether a particular performance measure happens to result in a Type I or Type II error for a given sample and a given level of abnormal performance yields little insight into the likelihood that a particular type of error will *systematically* be made with a given methodology. To get direct measures of the *ex ante* probability of Type I and Type II errors, the procedure of introducing abnormal performance and then testing for it must be applied to each of the 250 samples. For a specific level of abnormal performance introduced into each security of every sample, we examine the overall performance of a methodology when it is applied to each sample — that is, when the methodology is replicated 250 times. We concentrate on the frequency of Type I and Type II errors in these 250 trials.

For each methodology, table 1 shows the frequency with which the hypothesis of no abnormal performance in month '0' is rejected using several different significance tests. The results are reported for 0, 1, 5, 15 and 50% levels of abnormal performance introduced into each security of every sample

[10]See Fama (1976, chs. 3 and 4) for a discussion of the market model.



in month '0'.[11] The frequency of rejections is reported when the null hypothesis is tested at both the 0.05 and 0.01 significance levels using a one-tailed test.[12]

### 4.1. Rejection frequencies using t-tests

One set of significance tests for which results are reported in table 1 are *t*-tests.[13] When there is no abnormal performance, for all of the performance measurement methods the *t*-tests reject the null hypothesis at approximately the significance level of the test. For example, for the tests at the 0.05 level, the rejection rates range from 3.2% for the Market Adjusted Returns method

Table 1

A comparison of alternative performance measures. Percentage of 250 replications where the null hypothesis is rejected. One-tailed test. $H_0$: mean abnormal performance in month '0' = 0.0. Sample size = 50 securities.

| | Test level: $\alpha = 0.05$ | | | Test level: $\alpha = 0.01$ | | |
|---|---|---|---|---|---|---|
| | Actual level of abnormal performance in month '0' | | | | | |
| Method | 0% | 1% | 5% | 0% | 1% | 5% |
| *Mean Adjusted Returns* | | | | | | |
| *t*-test | 4.0 | 26.0 | 100.0 | 1.6 | 8.8 | 99.2 |
| Sign test | 0.8 | 6.4 | 96.0 | 0.0 | 1.6 | 90.8 |
| Wilcoxon signed rank test | 1.6 | 12.8 | 99.6 | 0.4 | 4.4 | 97.6 |
| *Market[a] Adjusted Returns* | | | | | | |
| *t*-test | 3.2 | 19.6 | 100.0 | 1.6 | 5.2 | 96.4 |
| Sign test | 0.0 | 9.2 | 99.2 | 0.0 | 2.0 | 97.6 |
| Wilcoxon signed rank test | 1.6 | 17.2 | 99.6 | 0.4 | 4.4 | 98.8 |
| *Market[a] and Risk Adjusted Returns* | | | | | | |
| *t*-test | 4.4 | 22.8 | 100.0 | 1.2 | 6.8 | 98.4 |
| Sign test | 0.4 | 7.2 | 99.6 | 0.0 | 2.4 | 98.0 |
| Wilcoxon signed rank test | 2.8 | 16.4 | 100.0 | 0.0 | 4.4 | 99.2 |

[a]Fisher Equally Weighted Index. Note that for 15 and 50% levels of abnormal performance, the percentage of rejections is 100% for all methods.

[11]The range of these levels of abnormal performance corresponds roughly to the range of estimated abnormal performance reported in Fama, Fisher, Jensen and Roll (1969, table 2). For example, for their sample of stock splits followed by divided increases, estimated abnormal performance ranged from about 1% in month '0' to 38% when the performance measure is cumulated over a 30-month period before and including month '0'.

[12]Throughout most of the paper, results will be reported for one-tailed tests. In a one-tailed test at any significance level $\alpha$, the critical value of the test statistic at or above which the null hypothesis is rejected is given by the $(1-\alpha)$ fractile of the frequency distribution of the test statistic which is assumed to obtain under the null. In section 5, results for two-tailed tests will be discussed.

[13]The assumptions underlying the *t*-tests are discussed in the appendix. For an example of *t*-tests in event studies, see Jaffe (1974). Although different variations of the *t*-tests are examined in section 6, results for the initial simulations are not sensitive to the specific variation employed.

to 4.4% for the Market and Risk Adjusted Returns method; for tests at the 0.01 level of significance, the rejection rates for the three methods range from 1.2 to 1.6%.[14]

With 1% abnormal performance, using $t$-tests the Mean Adjusted Returns method rejects the null hypothesis in 26.0% of the 250 replications when testing at the 0.05 level of significance. This compares to a 22.8% rejection rate with the Market and Risk Adjusted Returns method, and a 19.6% rejection rate with the Market Adjusted Returns method. This result is striking: it suggests that the simplest method, the Mean Adjusted Returns method, is no less likely than either of the other two to detect abnormal performance when it is present.[15]

Furthermore, the results which obtain with 1% abnormal performance are robust with respect to seemingly minor variations in the simulation procedure. For example, the relative rankings of the tests do not seem to be very sensitive to the significance level at which the null hypothesis is tested: at the 0.01 level of significance, the Mean Adjusted Returns method rejects 8.8% of the time, compared to a rejection rate of 6.8% for the Market and Risk Adjusted Returns method and a 5.2% rate for the Market Adjusted Returns method. It should also be emphasized that our conclusions about the relative

[14]Even if the empirical sampling distribution of a particular test statistic corresponds exactly to the assumed theoretical distribution, the proportion of rejections when the null hypothesis is true will not be exactly equal to the test level: The proportion of rejections is itself a random variable with a sampling distribution. Suppose that, under the null hypothesis, the outcomes of the hypothesis tests for each of the 250 replications are independent. Then at the 0.05 test level, the proportion of rejections for such a Bernoulli process has a mean of 0.05 and a standard deviation of 0.014. If the proportion of rejections is normally distributed, then the percentage of rejections reported in table 1 for 0% abnormal performance should, if the test statistics are properly specified, be between 2 and 8% approximately 95% of the time when testing at the 0.05 level. At the 0.01 level, the proportion of rejections should be between 0 and 2.2% approximately 95% of the time.

In calculating the proportion of rejections to be observed under the null hypothesis, it should be kept in mind that our 250 samples or 'trials' cannot be regarded as literally independent. A given security can be included in more than 1 of the 250 replications. To investigate the degree of dependence, for each of the 250 samples we computed an equally weighted average return for the sample securities for event months −89 through +10. We then computed the 31125 pairwise correlation coefficients for the 250 samples. The correlation coefficient between sample 1 and sample 2, for example, is computed from the 100 equally weighted returns on each sample in event time.

The largest of the 31125 pairwise correlation coefficients is 0.42, and the smallest is −0.34. Using a two-tailed test, only 485, or about 1.5% of the correlation coefficients are significant at the 0.01 level, compared to an expected proportion of 1%. While the hypothesis that the samples are pairwise independent is rejected, the degree of linear dependence appears to be small.

[15]*In comparing rejection frequencies across methodologies, it is necessary to gauge the magnitude of the differences in rejection proportions, either pairwise or jointly. If, for each replication, the results for two different test methods are independent of each other, then the difference in the proportion of rejections in 250 replications could be as large as about 4% merely due to chance; hence the difference between the 26.0% rejection rate for Mean Adjusted Returns need not be regarded as significantly different from the 22.8% rejection rate for the Market and Risk Adjusted Returns method.*



performance of the Market Adjusted Returns and Market and Risk Adjusted Returns methods have not been induced by the use of the Equally Weighted Index. For example, with 1% abnormal performance, the rejection rate we obtain for the Market and Risk Adjusted Returns with the Equally Weighted Index is 22.8%; with the Value-Weighted Index, the rejection rate is even lower, 15.2%. Differences between the use of the Equally Weighted and Value Weighted Indices are examined in detail in section 7.[16]

When the level of abnormal performance is increased from 1 to 5% in each sample security, all three methods detect the abnormal performance almost all of the time: at the 0.05 significance level, all three methods reject the null hypothesis 100% of the time, and at the 0.01 level, the minimum rejection rate is 96.4% for the Market Adjusted Returns method. Similarly, when the level of abnormal performance is again increased first to 15% and *then to 50%, all three methods reject virtually 100% of the time.* While this high frequency of rejections suggests that the tests for abnormal performance are quite powerful when there is 5% or more abnormal performance, it should be kept in mind, as we will later discuss, that these results are critically dependent on the assumption that the precise time at which the abnormal performance occurs is known with certainty. Furthermore, as we will also discuss, the relatively favorable performance of the Mean Adjusted Returns method will not obtain under all experimental conditions.

### 4.2. *Parametric vs. non-parametric significance tests*

Implicit in the $t$-tests which are used to assess abnormal performance are a number of strong assumptions: for example, in order for the test statistics to be distributed Student-$t$ in the Mean Adjusted Returns method, security returns must be normally distributed. If such an assumption is not met, then the sampling distribution of test statistics assumed for the hypothesis tests could differ from the actual distribution, and false inferences could result. If the distribution of the test statistic is misspecified, then the null hypothesis, when true, could be rejected with some frequency other than that given by the significance level of the test.

To examine the usefulness of significance tests which make less restrictive assumptions than the $t$-tests, we also employ two non-parametric tests of the

[16]In an Asset Pricing model context, there is no clear *a priori* justification for use of an equally weighted index. However, even if their use is viewed as an *ad hoc* procedure, the fact that such indices are employed in actual event studies [e.g., Fama, Fisher, Jensen and Roll (1969), Watts (1978)] suggests that the consequences of their use are of interest. In addition, there are strong reasons for reporting initial simulation results with the Equally Weighted Index. As we later discuss, some of the performance measures under study can actually be biased when used with the Value-Weighted Index. If we reported our initial simulation results using the Value-Weighted Index, biases associated with the use of that index would make it difficult to standardize the level of Type I errors across test methods; valid comparisons of the power of different methodologies would thus not be possible with our simulation procedure.

performance measures which have been used in actual event studies: (1) a sign test, and (2) a Wilcoxon signed rank test.[17] In the sign test for a given sample, the null hypothesis is that the proportion of sample securities baving positive measures of abnormal performance (e.g., positive residuals) is equal to 0.5; the alternative hypothesis (for any particular level of abnormal performance) is that the proportion of sample securities having positive performance measures is greater than 0.5. In the Wilcoxon test, both the sign *and* the magnitude of the abnormal performance are taken into account in computing the test statistic.[18]

Table 1 indicates the frequency with which the two non-parametric tests reject the hypothesis of no abnormal performance in month '0' for each methodology and for 0, 1, 5, 15 and 50% abnormal performance. From the results with 0% abnormal performance, it appears that there is a serious problem with the use of these non-parametric tests: under the null bypothesis, the tests do not reject at the 'correct' level. For example, for tests at the 0.05 level, the rejection rates range from a low of 0% for the sign test in the Market Adjusted Returns method to a high of 2.8% for the Wilcoxon test used in conjunction with the Market and Risk Adjusted Returns method. For tests at the 0.01 level, four of the six rejection rates are equal to 0%, and the other two are equal to 0.4%. Compared to the significance level of the test, the sign and Wilcoxon tests do not appear to reject the null hypothesis often enough. Although they are used to avoid the problem of possible misspecification of the $t$-tests, it appears that the non-parametric tests themselves suffer from such a problem of misspecification.

### *Distributional properties of the test statistics*

To further examine the properties of the $t$, sign, and Wilcoxon tests, in table 2 we report summary measures for the actual frequency distribution of each test statistic, based on the 250 replications. Even when there is no abnormal performance, in many cases there appear to be significant differences between the empirical sampling distribution of the test statistic and the distribution which is assumed for the hypothesis tests. That such differences are substantial implies that tests for abnormal security price performance can be misleading and must be interpreted with great caution.

For the $t$-tests, the differences between the actual and assumed distribution of the test statistics seem small. For example, when there is no abnormal performance, the average $t$-statistics are approximately 0, ranging from a low of $-0.13$ in the Mean Adjusted Returns method to a high of $-0.04$ in the Market Adjusted Returns method. There is also evidence that the $t$-statistics

[17]See, for example, Kaplan and Roll (1972), Ball, Brown and Finn (1977), and Collins and Dent (1979).

[18]Details of the calculation of these test statistics are contained in the appendix.



are leptokurtic and slightly skewed to the right. However, at the 0.05 significance level it is only for the Mean Adjusted Returns method that the Kolmogorov–Smirnov test rejects the hypothesis that the distribution of $t$-values is indeed $t$. For both the Mean Adjusted Returns and Market and Risk Adjusted Returns methods, it also appears that the $t$-tests result in slightly 'too many' extreme negative $t$-values.[19]

For the sign and Wilcoxon tests, in large samples the test statistics should be distributed unit normal. However, table 2 indicates that the mean of the test statistics is generally significantly less than 0 under the null hypothesis; the mean test statistics ranges from a low of $-0.52$ for tbe sign test in the Market and Risk Adjusted Returns method to a high of $-0.42$ for the Wilcoxon test in Market and Risk Adjusted Returns method; the $\chi^2$ and Kolmogorov–Smirnov tests reject the hypothesis of normality with mean 0 for all the tests.

Our finding for the non-parametric tests that the average test statistic is significantly negative is not difficult to explain: the sign and Wilcoxon tests assume that the distribution of a security specific performance measure (such as a market model residual) is symmetric, with half of the observations above the mean and half below the mean. However, there is evidence of right skewness in security specific performance measures such as market model residuals [Fama, Fisher, Jensen and Roll (1969, p. 6)]. With fewer positive than negative performance measures, the median performance measure will

[19]There are two related points about the frequency distributions of the $t$-statistics, summarized in table 2, which should be mentioned. First, note that the $t$-statistics in the Mean Adjusted Returns method have an estimated variance of 1.32, higher than the variance of the $t$-statistics of either of the other methods. The higher variance is indicative of the troubling behavior of the Mean Adjusted Returns $t$-tests in the left-hand tail region. There, 21 of the 250 $t$-statistics fall in the 5% left-hand tail of a $t$ distribution, compared to an expected number of 12.5, and 41 of the test statistics fall in the 10% lower tail of a $t$ distribution, compared to an expected number of 25. This large fraction of test statistics in the lower tail region implies that a test of the hypothesis that there is no abnormal performance (compared to an alternative hypothesis that abnormal performance is negative) will result in rejection of that hypothesis at a rate almost twice that of the significance level of the test when the null hypothesis is true. Similar left-hand tail behavior is obtained in later simulations where the Value-Weighted Index is employed for computing market model residuals. Use of the Jaffe–Mandelker dependence adjustment procedure, which we will later discuss, also yields such left-hand tail behavior in the $t$-statistics even when market model residuals are computed from the Equally Weighted Index.

Second, note that when there is 1% abnormal performance, the distributions of the $t$-statistics for all methods are quite different from the $t$ distribution, which is the distribution which should obtain under the null hypothesis; that there are such differences merely indicates that the $t$-tests do in fact pick up abnormal performance when it is present. When there is abnormal performance, one could also compare the distributions of test statistics to the non-central $t$, which is the distribution which would be expected under the alternative hypothesis if the test statistics were correctly specified. However, since even the null distributions are at least slightly misspecified, it also seems reasonable to anticipate some misspecification in the distribution which should obtain under the alternative hypothesis. Given such misspecification, analytically deriving power functions under the assumptions of the various tests is not a reliable way of understanding the actual power functions for the tests. A simulation technique such as ours is necessary.

Table 2

Summary measures for the actual frequency distribution of each test statistic, based on the 250 replications. Upper and lower lines indicate 0 and 1% abnormal performance, respectively.

| Method | Mean | Variance | $t$-statistic for mean | $\beta_1 = \mu_3^2/\mu_2^3$ | Kurtosis | Pearson skewness | $\chi^2$ statistic (20 equally spaced intervals) | $\chi^2$ statistic (9 tail region intervals)[a] | Kolmogorov–Smirnov $D$-statistic |
|---|---|---|---|---|---|---|---|---|---|
| *Mean Adjusted Returns* | | | | | | | | | |
| $t$-test values 0% abnormal performance | −0.13 | 1.32 | −1.80 | 0.02 | 3.20 | 0.06 | 29.0 | 24.1 | 0.09 |
| 1% abnormal performance | 0.92 | 1.30 | 12.8 | 0.03 | 3.18 | 0.07 | 284.0 | 280.0 | 0.33 |
| Sign test values | −0.51 | 0.83 | −8.89 | 0.00 | 3.01 | 0.03 | 325.0 | 90.2 | 0.29 |
| | 0.25 | 0.79 | 4.45 | 0.00 | 3.01 | 0.02 | 282.0 | 13.0 | 0.20 |
| Wilcoxon test values | −0.42 | 0.96 | −6.87 | 0.01 | 2.78 | 0.05 | 88.2 | 64.5 | 0.18 |
| | 0.66 | 0.96 | 9.14 | 0.00 | 2.75 | 0.01 | 107.0 | 60.7 | 0.25 |
| *Market Adjusted Returns* | | | | | | | | | |
| $t$-values | −0.04 | 1.04 | −0.59 | 0.01 | 4.18 | 0.04 | 13.6 | 5.65 | 0.06 |
| | 0.92 | 1.06 | 14.1 | 0.04 | 4.17 | 0.06 | 246.8 | 212.2 | 0.35 |
| Sign test values | −0.52 | 0.87 | −8.83 | 0.11 | 2.71 | 0.24 | 304.5 | 86.6 | 0.26 |
| | 0.39 | 0.79 | 6.84 | 0.05 | 2.98 | 0.12 | 283.1 | 27.5 | 0.24 |
| Wilcoxon test values | −0.43 | 1.01 | −6.66 | 0.02 | 2.79 | 0.07 | 75.1 | 81.5 | 0.17 |
| | 0.69 | 1.01 | 1.08 | 0.05 | 2.93 | 0.12 | 149.1 | 103.9 | 0.30 |
| *Market and Risk Adjusted Returns* | | | | | | | | | |
| $t$-values | −0.05 | 1.01 | −0.77 | 0.07 | 3.72 | 0.09 | 20.1 | 9.06 | 0.06 |
| | 0.91 | 1.03 | 14.1 | 0.08 | 3.61 | 0.10 | 259.4 | 227.0 | 0.34 |
| Sign test values | −0.48 | 0.82 | −8.51 | 0.10 | 2.93 | 0.18 | 347.7 | 84.3 | 0.29 |
| | 0.48 | 0.73 | 8.91 | 0.00 | 3.01 | 0.03 | 287.1 | 18.8 | 0.30 |
| Wilcoxon test valuex | −0.42 | 1.02 | −6.66 | 0.00 | 2.77 | 0.00 | 79.9 | 61.5 | 0.18 |
| | 0.72 | 0.97 | 11.6 | 0.02 | 2.78 | 0.07 | 166.6 | 115.1 | 0.30 |

[a]For tests concentrating on the tail regions, the 9 intervals are: 0–0.01, 0.01–0.02, 0.02–0.05, 0.05–0.1, 0.1–0.9, 0.9–0.95, 0.95–0.98, 0.98–0.99, 0.99–1.0.

Upper percentage points

| | 0.95 | 0.99 |
|---|---|---|
| $\chi^2(8)$ | 15.5 | 20.1 |
| $\chi^2(19)$ | 30.1 | 36.2 |
| $D$ ($N = 250$) | 0.086 | 0.103 |
| $\beta_1$ (assuming normality, $N = 250$) | 0.063 | 0.129 |
| Kurtosis (normality, $N = 250$) | 3.52 | 3.87 |

be negative even when the average performance measure is equal to 0. The non-parametric tests will tend to reject the null 'too often' (compared to the significance level of the test) when testing for negative abnormal performance and 'not often enough' when testing for positive abnormal performance.[20]

The non-parametric tests could, in principle, take asymmetry in the distribution of the performance measure into account and test the null hypothesis that the proportion of positive performance measures is equal to some number other than 0.5. However, such a test would first require a procedure for determining the proportion of positive security-specific performance measures which obtains in the absence of abnormal performance. We know of no event study which has employed such a test.[21]

### 4.3. Different risk adjustment methods

In the initial simulations reported in table 1, we concluded that tests which used risk-adjusted returns were no more powerful than tests which used returns which had not been adjusted for systematic risk. However, that conclusion was predicated on the assumption that the 'market model residual' method we chose represented the appropriate method of risk adjustment. To investigate the robustness of those earlier results, it is useful to simulate other risk adjustment methods which have also been used in actual event studies. We will examine two alternative methods; specific details of each method are discussed in the appendix.

*Fama–MacBeth Residuals* — Instead of computing a market model residual for each sample security, a 'Fama–MacBeth' (1973) residual is computed instead. Average residuals are then computed and abnormal performance is assessed in the same way as with the Market Model Residual method.[22]

Market model residuals are an appropriate performance measure if security returns are multivariate normal. For Fama–MacBeth residuals to be an appropriate performance measure, it is also necessary for equilibrium expected returns to be generated according to the Black (1972) version of the Asset Pricing model. A comparison of the performance of the market model and Fama–MacBeth residual techniques will indicate the benefits, if any, which are associated with the restrictive assumptions (and additional data requirements) implicit in using the Fama–MacBeth residuals.

*Control Portfolios* — This method forms the sample securities into a portfolio with an estimated $\beta$ of 1. Regardless of the risk level of each sample security, the portfolio thus formed should have the same risk as the market portfolio. Those securities comprising the market portfolio become a 'control portfolio' in the sense that the market portfolio has the same risk level as the sample securities, but is not experiencing the 'event' under study. The performance measure for month '0' is the difference between the return on a portfolio of sample securities (formed so that $\beta = 1$) and the average return on the market portfolio in the calendar months in which the sample securities experience events.

Variations of the Control Portfolio technique have been used by, for example, Black and Scholes (1973), Gonedes, Dopuch and Penman (1976), Warner (1977) and Watts (1978).[23] By concentrating on the difference in mean returns, this method makes no particular assumption about which version of the Asset Pricing model is correct.

### Simulation results for alternative risk adjustment methods

To compare the different methods for risk adjustment, table 3 indicates the simulation results for 250 replications of each risk adjustment method with 0, 1 and 5% levels of abnormal performance. Two important results emerge from the simulation.

Compared to using Market Model residuals, the use of Fama–MacBeth residuals does not increase the power of the tests. Earlier, for example, using the Equally Weighted Index and with 1% abnormal performance, the Market Model Residual method rejected 22.8% of the time; the rejection rate using Fama–MacBeth residuals is 21.6%. Even if the Black model is corrent, there appears to be sufficient measurement error in the parameter estimates on which Fama–MacBeth residuals are based so that the tests based on those residuals are no more useful than those based on the multivariate normality assumption of the Market Model. Furthermore, use of the Control Portfolio method also results in no increase in the proportion of rejections which take place under the alternative hypotheses: With 1% abnormal performance, the Control Portfolio method rejects the null hy-

[20]Even a small degree of asymmetry will lead to such a result. For example, if a sample of 50 securities has 27 negative and 23 positive market model residuals in month '0', the test statistic in the sign test will be approximately $-0.5$. This is about equal to the average value of $-0.48$ reported in table 2. Note that the use of continuously compounded (rather than arithmetic) returns is likely to reduce the extent of the asymmetry in market model residuals.

[21]Residual based techniques focusing on median (rather than mean) residuals could presumably use estimation procedures other than ordinary least squares to perform the market model regressions [see Bassett and Koenker (1978), and Cornell and Dietrich (1978)]. However, even if the non-parametric tests were properly calibrated by focusing on differences from medians, it is not obvious that the tests would be more powerful (against specific alternatives) than the $t$-tests, particularly since the $t$-tests, with their additional restrictions, seem reasonably well specified. But it should be kept in mind that there do exist distributions of the security-specific performance measures for which tests such as the sign test will be more efficient, particularly distributions with sufficiently heavy tails. See Lehman (1975, pp. 171–175) for a further discussion of the power of the $t$, sign, and Wilcoxon tests.

[22]Fama–MacBeth residuals have been used by, for example, Jaffe (1974) and Mandelker (1974).

[23]The Control Portfolio technique has also been used to control for factors other than systematic risk. See Gonedes, Dopuch and Penman (1976, p. 113) for a discussion.

pothesis in 18.0% of the 250 replications. These results for alternative risk adjustment procedures are consistent with our earlier conclusion that the Mean Adjusted Returns method performs no worse than those methods which explicitly adjust for systematic risk.[24]

Table 3

Different methods for risk adjustment. Percentage of 250 replications where the null hypothesis is rejected ($\alpha = 0.05$). One-tailed $t$-test results. $H_0$: mean abnormal performance in month '0' = 0.0. Sample size = 50 securities.

| Method | Actual level of abnormal performance in month '0' | | | Mean $t$-statistic with 1% abnormal performance |
|---|---|---|---|---|
| | 0% | 1% | 5% | |
| *Methods making no explicit risk adjustment* | | | | |
| Mean Adjusted Returns | 4.0 | 26.0 | 100.0 | 0.92 |
| Market Adjusted Returns | 3.2 | 19.6 | 100.0 | 0.92 |
| *Methods with market and risk-adjusted returns* | | | | |
| Market Model Residuals | 4.4 | 22.8 | 100.0 | 0.91 |
| Fama–MacBeth Residuals | 4.0 | 21.6 | 100.0 | 0.89 |
| Control Portfolio | 4.4 | 18.0 | 100.0 | 0.86 |

## 5. The use of prior information

The simulations which have been performed thus far make the strong assumption that the time at which abnormal security price performance occurs is known with complete certainty. However, if it is only known when, for example, the *Wall Street Journal* announced that the 'event' had taken place, then the calendar date of the event cannot be pinpointed exactly and the date itself becomes a random variable; in that case, abnormal returns for a number of periods before the 'announcement date' will typically be scrutinized for evidence of 'abnormal' performance. Similarly, even when it can be established with certainty when the event occurred, one is often concerned with whether or not there exists a profitable trading rule which could be implemented conditional on an event. In such a situation, it is necessary to study abnormal price performance for the period following time '0'.

[24]We have also examined the properties of the test statistics generated with Fama–MacBeth residuals and the Control Portfolio method. For both methods, the distribution of $t$-statistics is reasonably close to Student-$t$, and the properties of the test statistics are very similar to those reported in table 2 for the market model residual methodology.



### 5.1. Assessing abnormal performance when its precise date is unknown

We now examine how uncertainty about the precise date of the abnormal performance affects the power of the tests. For every security in each of the 250 samples, abnormal performance is generated in one specific month in the interval from month −10 through +10. The event month of abnormal performance can differ across securities; for a given security, the event month of abnormal performance is a drawing from a uniform distribution.[25] In this experiment, 0, 1, 5, 15% and 50% abnormal performance is introduced for each security for one month in the (−10, +10) interval. This experimental situation corresponds to one where abnormal performance occurs (a) at some time in the 21-month interval up to and including month '0', or (b) at some time in the 21-month interval including and following the event. The null hypothesis to be tested is that the mean level of abnormal performance over the entire 21-month interval is equal to 0.

Table 4 shows the frequency with which each test method results in a rejection of the null hypothesis of no abnormal performance. The results are dramatic: even at high levels of abnormal performance, the hypothesis of no abnormal performance often fails to be rejected. For example, with 5% abnormal performance, the rejection rates range from 16.0% with the Control Portfolio method to a high of 28.4% with the Mean Adjusted Returns method. With 15% abnormal performance, the rejection rates increase and are on the order of 70 to 80% for the various test methods; however, these rejection rates are still much lower than those obtained in the earlier simulations, where the precise date of abnormal performance was known with certainty. There, using $t$-tests even 5% abnormal performance was detected 100% of the time by all of the test methods.[26]

To further illustrate how prior information can be used to increase the power of the tests, in table 4 we also show the results of a simulation where all abnormal performance occurs in the (−5, +5) interval and is uniformly distributed. When prior information can be used to narrow the time interval in which the abnormal performance could have occurred, in this case from (−10, +10) to (−5, +5), the rejection rates increase substantially in the presence of a given level of abnormal performance. With 5% abnormal performance, the rejection rates increase from 28.4 to 35.2% for the Mean Adjusted Returns method, and from 24.4 to 39.6% using Market Model residuals.

[25]When other distributions (e.g., normal, exponential) were used, the qualitative conclusions of this section remained unchanged.

[26]Furthermore, the rejection rates in table 4 cannot be markedly increased if the researcher is merely willing to tolerate a slightly higher probability of Type I error — that is, if one is willing to conduct the hypothesis test at a higher significance level. For example, in the Mean Adjusted Returns method, the rejection rate with 5% abnormal performance is 28.4%. To obtain a rejection rate of 50%, the significance level would have to be increased to about 0.20; to obtain a 75% rejection rate, the significance level would have to be increased to about 0.35.

Table 4

Alternative performance measures when the precise date of the abnormal performance is unknown.[a] Percentage of 250 replications where the null hypothesis is rejected ($\alpha = 0.05$). One-tailed $t$-test results using Equally Weighted Index. $H_0$: mean abnormal performance in the interval $(-10, +10) = 0.0$

| | Actual level of abnormal performance in interval $(-10, +10)$ | | | | |
|---|---|---|---|---|---|
| Method | 0% | 1% | 5% | 15% | 50% |
| Mean Adjusted Returns | 7.6 | 9.2 | 28.4 | 82.0 | 100.0 |
| | (9.2) | (13.6) | (35.2) | (94.4) | (100.0) |
| Market Adjusted Returns | 3.6 | 5.6 | 18.4 | 73.2 | 100.0 |
| | (5.2) | (6.8) | (35.2) | (96.4) | (100.0) |
| Market Model Residuals | 7.2 | 10.8 | 24.4 | 86.4 | 100.0 |
| | (7.6) | (10.4) | (39.6) | (96.8) | (100.0) |
| Fama–MacBeth Residuals | 3.6 | 5.2 | 16.4 | 74.8 | 100.0 |
| | (8.8) | (15.6) | (45.6) | (97.6) | (100.0) |
| Control Portfolio | 4.8 | 6.4 | 16.0 | 70.0 | 100.0 |
| | (5.6) | (6.8) | (32.0) | (91.2) | (100.0) |

[a]For each security, abnormal performance is introduced for one month in the interval $(-10, +10)$ with each month having an equal probability of being selected. The rejection rates shown in brackets are for the case where (1) for each security, abnormal performance is introduced for one month in the $(-5, +5)$ interval, with each month having an equal probability of being selected, and (2) the null hypothesis is that the mean abnormal performance in the $(-5, +5)$ interval is equal to 0.

Table 5

The behavior of two-tailed tests. Percentage of replications, for the 0.025 significance level, where a one-tailed $t$-test rejects the null hypothesis of no abnormal performance. This rejection rate is identical to the percentage of replications where a two-tailed test at the 0.05 level rejects the null and detects positive abnormal performance. Rejection rates from table 4, for a one-tailed test with $\alpha = 0.05$, are shown in brackets. $H_0$: mean abnormal performance in the interval $(-10, +10) = 0.0$.[a]

| | Actual level of abnormal performance in interval $(-10, +10)$ | | | |
|---|---|---|---|---|
| Method | 0% | 1% | 5% | 15% |
| Mean Adjusted Returns | 4.8 | 5.6 | 17.6 | 74.4 |
| | (7.6) | (9.2) | (28.4) | (82.0) |
| Market Adjusted Returns | 1.0 | 2.4 | 9.2 | 58.0 |
| | (3.6) | (5.6) | (18.4) | (73.2) |
| Market Model Residuals | 3.2 | 5.2 | 18.4 | 78.8 |
| | (7.2) | (10.8) | (24.4) | (86.4) |
| Fama-MacBeth Residuals | 2.0 | 2.0 | 9.2 | 60.4 |
| | (3.6) | (5.2) | (16.4) | (74.8) |
| Control Portfolio | 2.0 | 2.4 | 8.0 | 48.4 |
| | (4.8) | (6.4) | (16.0) | (70.0) |

[a]For each security, abnormal performance is introduced for one month in the $(-10, +10)$ interval, with each month having an equal probability of being selected.



*Rejection rates for two-tailed significance tests*

There is yet another assumption about prior information which all of our simulations make and whose consequences can also be studied: the hypothesis tests we perform throughout this paper are one-tailed tests. An implicit assumption in such tests is that the sign of the abnormal performance is also known. However, if one cannot use prior information to impose this restriction, the appropriate test is two-tailed. For a given significance level, the power of the tests is thus reduced.[27]

In table 5, we report rejection rates for one-tailed tests conducted at both the 0.05 and 0.025 significance levels. The rejection rate for a one-tailed test at the 0.025 level also represents the percentage of replications in which a two-tailed test at the 0.05 level will pick up positive abnormal performance. Thus, comparing the rejection rates for one-tailed tests at the 0.05 and 0.025 levels is equivalent to comparing the frequency with which one-tailed and two-tailed tests, each conducted at the 0.05 level, will lead the researcher to conclude that positive abnormal performance is present.

When the sign of the abnormal performance is not known *a priori*, the ability to discern abnormal performance is reduced markedly. For example, with 5% abnormal performance in the $(-10, +10)$ interval, a two-tailed test at the 0.05 significance level picks up positive abnormal performance 17.6% of the time for the Mean Adjusted Returns method, compared to a rate of 28.4% for the corresponding one-tailed test. With 15% abnormal performance in the $(-10, +10)$ interval, a two-tailed test with that method detects positive abnormal performance in 74.4% of the replications, compared to a rate of 82.0% for a one-tailed test. While such results are hardly surprising, they serve to underscore the importance of using all available prior information in testing for abnormal performance.

### 5.2. Using cumulative average residuals

One method frequently used to investigate abnormal performance when there is incomplete prior information about when it occurs is the 'cumulative average residual' (CAR) technique employed by Fama, Fisher, Jensen and Roll (1969).[28] The technique focuses on the average market model residuals of the sample securities for a number of periods around the event. The

[27]Similarly, to obtain a particular rejection frequency when there is abnormal performance of a given sign and magnitude, the level of Type I error must be increased in moving from a one-tailed to a two-tailed test. For example, two-tailed tests would have to be conducted at the 0.1 level to pick up positive abnormal performance with the same frequency as that which has been reported throughout this paper for one-tailed testes at the 0.05 significance level.

[28]A similar technique involves construction of an Abnormal Performance Index [e.g., Ball and Brown (1968)]. In simulations not reported here, the abnormal performance measures of Ball–Brown, Pettit, and Beaver–Dukes [see Ohlson (1978, p. 184) for a description of these measures] were also examined. The properties of the confidence bands traced out by such alternative metrics were similar to those discussed for the CARs.

cumulative average residual for a given event-related month $t$ is defined as the value of the cumulative average residual in the previous event-month plus the current value of the average residual, $AR_t$,

$$CAR_t = CAR_{t-1} + AR_t. \quad (1)$$

Examining the CAR of a set of sample securities as of any given event-related month $t$ is a way of looking at whether or not the values of the average residuals, starting from the month of cumulation and up to that point, are systematically different from 0.[29]

To simulate the CAR technique for various levels of abnormal performance, we use the values of the average market model residuals which were obtained for the simulations reported in table 4, where abnormal performance is uniformly distributed in the $(-10, +10)$ interval. For a given sample and a given level of abnormal performance, we take the average market model residuals and begin cumulating them in month $-10$; cumulation then continues for every month through month $+10$. For each sample, the procedure yields a set of 21 cumulative average residuals, one for each event-related month from $-10$ through $+10$. For a given event-related month, repeated application of the procedure to each of the 250 samples yields 250 cumulative average residuals.

*Cumulative average residuals when there is no abnormal performance*

To understand the properties of CARs under the null hypothesis, in fig. 1 we trace selected fractiles of the 250 CARs in each event-related month for the case where no abnormal performance is introduced. As the figure indicates, the 0.05 and 0.95 fractiles of the 250 CARs depart more and more from 0 as the cumulation process continues. By the end of month $+10$, the 0.95 fractile takes on a value of over 9%, and the 0.05 fractile takes on a value of about $-9\%$. This suggests that the CAR for a given sample could appear to wander a great deal from 0, even in the absence of abnormal performance.[30]

The behavior of the CAR is consistent with a simple explanation. As eq. (1) indicates, the CAR for a given sample is by construction a random walk.[31] Like any process which follows a random walk, the CAR can easily

[29]Examining the CAR as of any event month is equivalent to examining the significance of the mean average residual over the cumulation period. However, looking at the entire set of event-time CARs on a month by month basis is not very meaningful unless the significance test explicitly takes into account the fact that CARs are, by construction, highly serially dependent.

[30]In table 2, we presented evidence that average residuals were skewed to the right. The slight apparent downward drift in the 0.5 fractile of the CAR would thus be expected.

[31]CARs will be a random walk if the average residuals in event time are independent and identically distributed. A confidence band such as that traced out by the 0.05 and 0.95 fractiles in fig. 1 should increase with the square root of the number of months over which cumulation takes place.



give the *appearance* of 'significant' positive or negative drift, when none is present. However, even if no abnormal performance were present, neither the seemingly significant upward drift indicated by the 0.95 fractile or the downward drift of the 0.05 fractile could be considered outside of the realm of chance. Indeed, in 5% of the 250 samples, the value of the CAR exceeds the values taken on by the 0.95 fractile reported in fig. 1; in another 5% of the samples, the value of the CAR is less than that taken on by the 0.05 fractile. The pattern of CAR fractiles in fig. 1 serves to underscore the necessity for statistical tests on the performance measures, since merely looking at a picture of CARs can easily result in Type I errors.

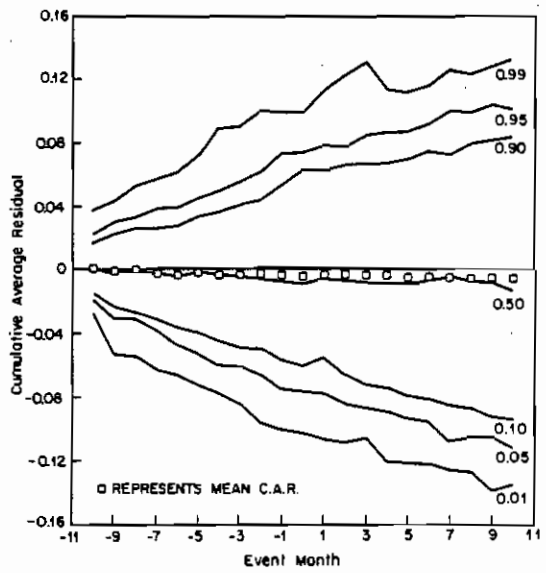


Fig. 1. Fractiles of cumulative average residual under null hypothesis of no abnormal performance.

*Cumulative average residuals when abnormal performance is present*

In fig. 2, we show selected fractiles of the CARs for the case where 5% abnormal performance occurs for each sample security, and the month of abnormal performance is uniformly distributed in the $(-10, +10)$ interval.

The value of each fractile as of month +10 is higher by approximately 0.05 than the corresponding value in fig. 1, when no abnormal performance was present; however, the 0.5 fractile, that is, the median CAR as of month +10 still falls well within the bounds which were shown in fig. 1, and which obtain under the null hypothesis. Moreover, since for a given sample the

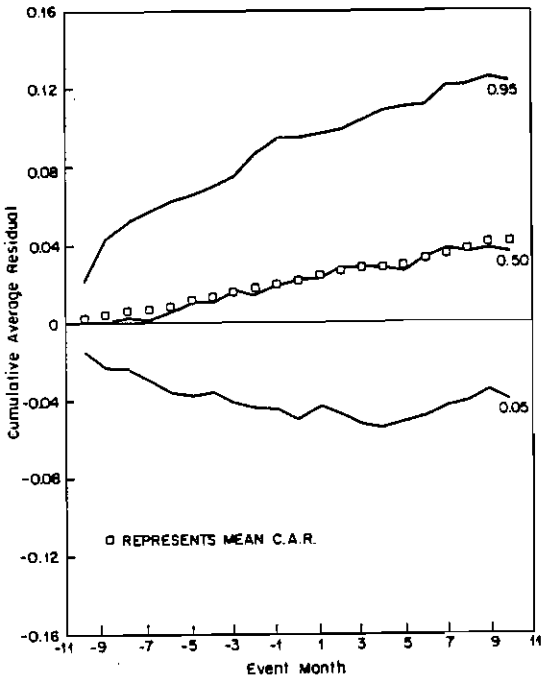


Fig. 2. Fractiles of cumulative average residual with 5% excess return distributed uniformly on months −10 to +10.

month of abnormal performance is uniformly distributed across securities and not on average large in any one month, CAR plots for the individual samples would tend to show a pattern not strikingly different from what would be expected under the null hypothesis. In such a case, there is little information which the CAR for the sample provides in helping to decide whether abnormal performance is present.



However, in fig. 3 we show fractiles of CARs when 5% abnormal performance occurs in month '0' for all sample securities. Although the fractiles at the end of the cumulation period take on values similar to those shown in fig. 2, there is an apparent 'spike' at month '0'; such a spike shows up not only in the selected fractiles, but in the CAR plot for any given

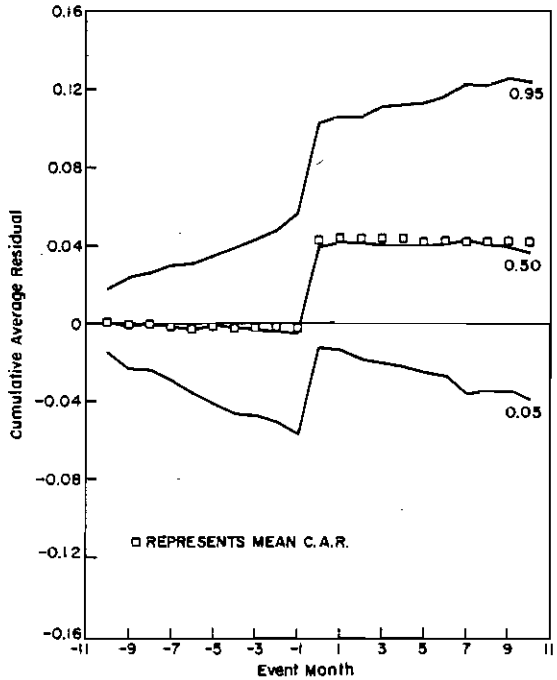


Fig. 3. Fractiles of cumulative average residual with 5% excess return in month zero.

sample. Given some prior information that month '0' is of particular interest to focus on, the existence of such a spike in the CAR pattern could reasonably suggest to the researcher that a hypothesis test for abnormal performance in month '0' rather than the entire (−10, +10) period would be appropriate. The test on month '0' picks up abnormal performance 100% of the time, as indicated in table 1, whereas the test in the (−10, +10) interval

only yields the rejection rate of 24.4% shown in table 4. Thus, when the timing of the abnormal performance is not uniform over the period under study, the precise pattern of estimated abnormal returns is conveniently summarized by a CAR plot; the pattern can provide useful information beyond that given by the value of the CAR at the end of an arbitrary 'cumulation period'.[32]

## 6. The effect of clustering

### 6.1. Event month clustering

The securities of a sample will frequently each experience an event during the same calendar time period. For example, as Schwert (1978) and Foster (1980) discuss, government regulation or mandated accounting procedures will often have a simultaneous impact on a number of different securities whose price performance around the time of an event is being examined. We refer to the close or simultaneous spacing of events as event month clustering.

Clustering has implications for the characteristics of the test methods being examined in this paper. The general impact of clustering is to lower the number of securities whose month '0' behavior is independent. The month '0' dependence is important for two reasons. First, if performance measures such as the deviation from historical mean returns or market model residuals are positively correlated across securities in calendar time, then such clustering will increase the variance of the performance measures (e.g., the average residual) and hence lower the power of the tests. Secondly, the month '0' dependence in security-specific performance measures must explicitly be taken into account in testing the null hypothesis of no abnormal performance. Otherwise, even in the absence of abnormal performance, the null

[32]Our discussion of CAR plots touches on a deeper set of issues which arises in connection with all tests for abnormal performance: *a priori*, the researcher often does not know when the abnormal performance would have occurred, nor perhaps even the frequency distribution of the time of the abnormal performance. Lacking that information, the choice of time period over which to conduct the hypothesis test is arbitrary, and one's inferences can be sensitive to the choice.

For example, if for all securities the abnormal performance occurs in month '0' or a few months surrounding it, a hypothesis test conducted over the entire $(-10, +10)$ period is less likely to pick up abnormal performance than one concentrating on month '0'. Conversely, if the event month of abnormal performance is uniformly distributed over the $(-10, +10)$ period, but the hypothesis test is performed for month '0', the abnormal performance is also less likely to be detected. In general, the hypothesis test should give weights to event months which are the same as those implicit in the actual frequency distribution of the time of abnormal performance. However, in the absence of a plausible *a priori* reason for doing so, it is dangerous to infer the frequency distribution of the time of abnormal performance by merely looking at CAR plots and the estimated level of abnormal performance in each event-related month: if one puts enough weight on 'outliers', the null can always be rejected even when it is true.



hypothesis will be rejected too frequently if security specific performance measures are positively correlated.[33]

#### Inducing clustering in event-dates of the sample securities

To examine the effect of clustering, we must specify a new procedure for generating event-dates. For each of the securities in a given sample, month '0' is restricted to fall in a particular calendar time month which is common to all securities in the sample. The month is randomly selected. For each sample, a new calendar month is selected.[34] The effect of clustering is simulated with the same levels of abnormal performance which were used previously: 0, 1, 5 and 15%. All abnormal performance is generated in month '0'.

#### Testing for abnormal performance when there is clustering

For each performance measurement method, $t$-tests for month '0' abnormal performance are conducted using three different methods, each of which is discussed in the appendix. The $t$-tests are first conducted under the assumption that the performance measures (e.g., residuals) are independent across securities. The $t$-tests are also performed using two different methods for taking into account cross-sectional dependence.

One procedure, which we call 'Crude Dependence Adjustment', focuses on the series of event-time average performance measures (e.g., average residuals). A second, more complicated method we simulate has been employed by Jaffe (1974) and Mandelker (1974); that method forms various portfolios of sample securities in calendar time; as discussed in the appendix, the portfolios are formed so as to make their performance measures independent and homoscedastic, and the hypothesis test is actually performed on the independent performance measures of the portfolios. In table 6, we present simulation results for the different tests. From the numbers presented in the table, several results are apparent.

First, event-month clustering has a substantial impact on rejection frequencies for the Mean Adjusted Returns method. For example, when the $t$-tests ignore cross-sectional dependence in security specific performance measures,

[33]This consequence of the independence assumption has recently been discussed by, for example, Beaver (1979) and Collins and Dent (1978). Note that for the simulations presented thus far, cross-sectional dependence is unimportant because the degree of clustering is small: Since event dates are independent draws from a uniform distribution comprised of over 300 different calendar months, the likelihood that many securities in a given sample will have a common event date is small. Previous simulations have been performed both with and without dependence adjustment of the types to be discussed in this section; the differences in Type I error frequencies under the null hypothesis are generally small.

[34]The month is between June 1944 and March 1968. For a given sample, any month in that period has an equal probability of being selected; for the purpose of maintaining sample to sample independence, the selection of calendar months is carried out without replacement.

the Mean Adjusted Returns method rejects the null hypothesis 32.4% of the time when there is no abnormal performance;[35] this compares to rejection rates of about 3% using either Crude Dependence Adjustment or the Jaffe–Mandelker procedure. Clustering also reduces the power of the Mean Adjusted Returns method against specific alternative hypotheses: with highly correlated security specific performance measures, the variance of the mean

Table 6

The effect of event-month clustering.[a] Percentage of 250 replications where the null hypothesis is rejected ($\alpha = 0.05$). One-tailed $t$-test results. $H_0$: mean abnormal performance in month '0' = 0.0.

| Method | Actual level of abnormal performance in month '0' | | |
|---|---|---|---|
| | 0% | 1% | 5% |
| *Mean Adjusted Returns* | | | |
| No Dependence Adjustment | 32.4 | 44.4 | 74.0 |
| Crude Adjustment | 3.2 | 4.8 | 31.6 |
| Jaffe–Mandelker | 3.6 | 5.6 | 34.4 |
| *Market Adjusted Returns* | | | |
| No Dependence Adjustment | 3.6 | 22.8 | 99.6 |
| Crude Adjustment | 4.0 | 23.6 | 99.6 |
| Jaffe–Mandelker | 5.2 | 24.4 | 99.6 |
| *Market Model Residuals* | | | |
| No Dependence Adjustment | 4.0 | 23.2 | 99.2 |
| Crude Adjustment | 5.6 | 24.8 | 99.6 |
| Jaffe–Mandelker | 6.0 | 26.8 | 99.6 |
| *Fama–MacBeth Residuals* | | | |
| No Dependence Adjustment | 4.0 | 25.2 | 100.0 |
| Crude Adjustment | 4.8 | 24.4 | 98.8 |
| Jaffe–Mandelker | 4.8 | 26.4 | 99.2 |
| *Control Portfolio* | | | |
| Crude Adjustment | 4.4 | 23.2 | 99.2 |

[a]For a given replication, month '0' falls on the same calendar date for each security. The calendar date differs from replication to replication. Equally Weighted Index.
[b]Methodology not readily adapted to other dependence adjustment procedures.

[35]Event month clustering is not the only stochastic process generating events which would lead to 'too many' rejections for the Mean Adjusted Returns method. When there is no clustering, but the event tends to occur only in those months when the market return is abnormally high, then rejection rates using Mean Adjusted Returns will also be too high with a one-tailed test for positive abnormal performance. Furthermore, in an actual event study involving only one security, an analogous situation arises. For that case, the Mean Adjusted Returns method will not yield the probability of Type I error assumed in the hypothesis tests if the market return at the time of the event happened to have been abnormally high (or abnormally low).



performance measure, computed across sample securities, is higher, and the power of the tests is expected to be lower. With 5% abnormal performance, the Mean Adjusted Returns method rejects the null hypothesis 31.6% of the time using Crude Dependence Adjustment, and 34.4% of the time using the Jaffe–Mandelker procedure. These numbers are much lower than the rejection rate of 100% we obtained in the analogous earlier simulation without clustering.

Secondly, in marked contrast to the results for the Mean Adjusted Returns method, clustering appears to have little impact on rejection frequencies for any of the other performance measurement methods, and thus our earlier conclusions about the relatively favorable performance of Mean Adjusted Returns do not apply if there is clustering. When the null hypothesis is true, for our simulations it makes little difference whether or not cross-sectional dependence is taken into account. For example, in the Market Adjusted Returns method, the rejections rate with no dependence adjustment is 3.6%, compared to rejection rates of 4.0% using Crude Dependence Adjustment and 5.2% with the Jaffe–Mandelker procedure.

Furthermore, when abnormal performance is present, the rejection rates when there is clustering are not markedly different from those when there is no clustering: With 1% abnormal performance, the rejection rates with clustering are on the order of 20 to 25%, slightly higher than was the case in earlier simulations without clustering. It thus appears that for all methods taking into account market-wide factors, with the Equally Weighted Index the degree of cross-sectional dependence in the performance measures is negligible for randomly selected securities.[36] However, in an actual event study, a sample of securities whose events are clustered in calendar time may be nonrandom; the sample securities might be drawn from a common industry group having positively correlated performance measures. In such a case, the power of the tests is reduced even if a particular methodology abstracts from the market, and taking into account cross-sectional dependence in order to assure the 'correct' proportion of rejections is appropriate in such a case.

Third, it appears that the differences in simulation results between the Crude Adjustment procedure and the Jaffe–Mandelker procedure are small. In the presence of abnormal performance, there is a slight increase in rejection frequencies with the Jaffe–Mandelker procedure, and the increase

[36]Note that our finding of negligible cross-sectional dependence is specific to the Equally Weighted Index. With that index, randomly selected securities would not be expected to exhibit systematic positive cross-sectional dependence in performance measures. For example, if all securities in the market had positively correlated market model residuals, then the market model has really not abstracted from marketwide influences. With the Value Weighted Index, the unweighted average of pairwise covariances between residuals can be positive; in fact, simulations of clustering with the Value Weighted Index result in rejection rates under the null of about 15% (for $\alpha = 0.05$) when cross-sectional dependence is ignored.

takes place for every test method. The increase is consistent with our discussion in the appendix, where we suggest that the Jaffe–Mandelker procedure will be more precise than Crude Dependence Adjustment.

### 6.2. Security risk clustering

Another form of clustering which is pertinent to our study is clustering by systematic risk: a particular sample may consist of securities which tend to have higher than average (or lower than average) systematic risk. Since for individual securities there is a positive empirical relationship between the variance of returns and systematic risk (as well as between market model residual variance and systematic risk),[37] it seems reasonable to expect that tests for abnormal performance will be more powerful for low risk securities than for high risk securities; the intuition is simply that a given level of abnormal performance should be easier to detect when 'normal' fluctuations in sample security returns (and the standard errors of parameter estimates such as $\hat{\beta}$) are small rather than large.

#### *Security risk clustering: Sample selection procedure*

To see the effect of security risk clustering, we construct two sets of 250 samples, where all 500 samples have 50 securities each. We call the first 250 samples low-risk samples, and the second 250 samples high-risk samples.

The samples are constructed as follows. Securities are picked and event dates generated as discussed in section 3. In addition to data availability requirements imposed earlier, it is also required that a security have data from event-months −149 through −90. Based on an estimate of $\beta$ in that 60-month period, a security is assigned to a sample in either the high-risk or low-risk set, depending on whether its estimated $\beta$ is greater than 1 or less than 1.[38] The procedure of picking securities and event dates, and then of assigning securities to samples based on $\hat{\beta}$, is continued until both the low-risk set of samples and the high-risk set of samples each has 250 samples of 50 securities.

#### *Simulation results for risk-clustered samples*

For each set of 250 samples, various methodologies are simulated as in previous experiments not involving either security-risk or event-month clus-

[37] See Fama (1976, pp. 121–124). Note that empirically there is also a negative relationship between $\hat{\beta}$ (and hence variance of returns) and firm size. We have not examined the separate, independent effect of firm size on the power of the tests. However, to the extent that $\beta$ merely proxies for firm size, tests for abnormal performance would be expected to be more powerful for large firms than for small firms.

[38] By selecting on the basis of previous $\hat{\beta}$, the expected value of the measurement error in $\hat{\beta}$ over the (−89, −11) period should be zero for both the high-$\hat{\beta}$ and low-$\hat{\beta}$ samples. See Black, Jensen and Scholes (1972) for a discussion.



Table 7

The effect of clustering by risk.[a] Percentage of 250 replications where the null hypothesis is rejected ($\alpha = 0.05$). One-tailed $t$-test results. $H_0$: mean abnormal performance in month '0' = 0.0.

| | Rejection rates | | | | Mean $t$-statistics | | | |
|---|---|---|---|---|---|---|---|---|
| | Abnormal performance | | | | Abnormal performance | | | |
| | 0% | | 1% | | 0% | | 1% | |
| Method | $\beta<1$ | $\beta>1$ | $\beta<1$ | $\beta>1$ | $\beta<1$ | $\beta>1$ | $\beta<1$ | $\beta>1$ |
| Mean Adjusted Returns | 6.8 | 6.0 | 26.8 | 24.4 | −0.03 | −0.02 | 1.18 | 0.89 |
| Market Adjusted Returns | 7.6 | 7.2 | 31.6 | 24.8 | −0.02 | 0.07 | 1.14 | 0.99 |
| Market Model Residuals | 8.0 | 5.6 | 29.6 | 21.6 | −0.02 | 0.01 | 1.18 | 0.91 |
| Fama–MacBeth Residuals | 8.4 | 5.6 | 30.0 | 20.0 | 0.02 | −0.02 | 1.19 | 0.88 |
| Control Portfolio | 8.4 | 5.2 | 17.6 | 14.0 | 0.11 | 0.03 | 0.74 | 0.54 |

[a] CRSP Equally Weighted Index. There are 500 samples, each having 50 securities. $\beta<1$ refers to the 250 samples formed from securities with estimated $\beta$s less than 1. $\beta>1$ refers to the 250 samples formed from securities with estimated $\beta$s greater than 1.

| | Mean $\beta$ ($N = 12500$) |
|---|---|
| 250 samples with $\beta<1$ | 0.81 |
| 250 samples with $\beta>1$ | 1.24 |

tering. In table 7, for both the low-$\hat{\beta}$ set of samples and the high-$\hat{\beta}$ set of samples, we show rejection rates and mean $t$-statistics when various methodologies are applied to each set of 250 samples, and when all abnormal performance is introduced in month '0'.

When there is no abnormal performance, neither the rejection rates nor the mean $t$-statistics seem particularly out of line for any of the test methods. It is a bit surprising that the Market Adjusted Returns method does not reject 'too often' for the $\hat{\beta}>1$ samples and 'not often enough' when $\hat{\beta}<1$; however, it should be kept in mind that the rejection proportions shown in the table are merely estimates of the true proportions; in addition, there is some overlap in the sets of samples in the sense that individual securities in the high-$\hat{\beta}$ set can have true $\beta$s less than 1, and securities in the low-$\hat{\beta}$ set can have $\beta$s greater than 1.

With 1% abnormal performance, for all test methods both the rejection rates and the mean $t$-statistics are higher for the $\hat{\beta}<1$ set of samples than for the $\hat{\beta}>1$ set of samples. But the rejection rates for the high-$\hat{\beta}$ and low-$\hat{\beta}$ samples are generally not markedly different from each other, averaging 27.1% (across test methods) for the $\hat{\beta}<1$ samples and 21.0% for the $\hat{\beta}>1$ samples. Nor are these rejection rates much different from the 21.6% average rejection rate (across test methods) we obtained earlier for samples where the average $\hat{\beta}$ was approximately 1. Furthermore, the Mean Adjusted Returns method continues to perform well, rejecting the null 26.8% of the time for the $\hat{\beta}<1$ samples, and 24.4% of the time for the $\hat{\beta}>1$ samples.

Our risk clustering simulations also provide some new insight into the relative efficacy of various test methods. A careful look at table 7 reveals that while the rejection rates and mean $t$-statistics for the Control Portfolio method are indeed higher for $\hat{\beta}<1$ sample than for $\hat{\beta}>1$ samples when there is abnormal performance, both sets of numbers are lower than for previous samples with $\beta$s averaging 1. For example, using the Control Portfolio method, the mean $t$-statistic in our earlier results was 0.86 with 1% abnormal performance; the figures are 0.74 and 0.54, respectively, for the $\hat{\beta}<1$ and $\hat{\beta}>1$ sets of samples. While the relative rankings of most of the test methods are not very sensitive to sample security $\beta$, the version of the Control Portfolio method we have simulated performs noticeably worse, relative to both itself and to the other methods, when $\beta$ departs from 1.

The unfavorable performance of the Control Portfolio method when average $\hat{\beta}$ is different from 1 is related to the manner in which the methodology forms portfolios.[39] The Control Portfolio method we simulate is likely to involve short selling of some sample securities when average $\hat{\beta}$ is much different from 1. With short selling, the weights applied to the two subportfolios of sample securities will be vastly different from each other, and the variance of returns on the portfolio thus formed will be quite large;

[39]Gonedes, Dopuch and Penman (1976) and Gonedes (1978) discuss a related issue.



compared to a situation where each subportfolio has the same positive weight, the performance measure, which is the difference between portfolio returns and market returns, will also have a higher variance.

In addition, portfolio residual variance tends to be lowest when $\beta$ is 1 [Black, Jensen and Scholes (1972, table 2)]. Portfolios of sample securities which have $\beta$s much different from 1 will have greater estimation error in subportfolio $\beta$ and hence greater estimation error in calculating appropriate weights for subportfolios. This will also increase the variance of the performance measure and lower the power of the tests.

## 7. The choice of market index

Simulation results reported thus far have been based on use of the Equally Weighted Index. However, while the Equally Weighted Index is often employed in actual event studies, the Asset Pricing Model provides no justification for its use: the Asset Pricing model specifies an *ex ante* relationship between security expected returns and systematic risk measured with respect to the Value-Weighted Index. To examine the sensitivity of our earlier results to the choice of market index, we replicate the experiment reported in table 4 using the Value-Weighted (rather than the Equally Weighted) Index. As in the table 4 simulations, the event-month of abnormal performance is a drawing from a uniform distribution in the interval from month $-10$ through $+10$. For each methodology, results using the Value Weighted Index, along with the corresponding earlier results for the Equal Weighted Index, are reported in table 8.

### 7.1. Estimates of systematic risk using different market indices

One way in which the simulation results using the Value-Weighted Index differ from those using the Equally Weighted Index is related to the estimates of sample security systematic risk. To focus on the differences, which will play an important role in our discussion, for each of the 250 replications the average and standard deviation of the market model $\hat{\beta}$s using each index are computed; summary statistics are shown at the bottom of table 8.

For the Equally Weighted Index, the mean of the 250 average $\hat{\beta}$s is equal to 0.993, which is insignificantly different from 1. That the average $\hat{\beta}$ is approximately equal to 1 is hardly surprising, since our simulation procedure involves random selection of securities.

However, with the Value-Weighted Index, estimates of sample security $\beta$s are systematically different from 1. With that Index, the mean of the 250 average $\hat{\beta}$s is 1.13, with a standard deviation of 0.031. At the 0.01 level of significance, the hypothesis that the mean average $\beta$ is equal to 1 must be rejected.

Table 8

The effect of the choice of market index. Percentage of 250 replications where the null hypothesis is rejected ($\alpha = 0.05$). One-tailed t-test results. $H_0$: mean abnormal performance in interval $(-10, +10) = 0.0$. $VW$ = CRSP Value Weighted Index. $EW$ = Equally Weighted Index.

| | Actual level of abnormal performance in the interval $(-10, +10)$[a] | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0% | | 1% | | 5% | | 15% | | | |
| Method | $VW$ | $EW$ | $VW$ | $EW$ | $VW$ | $EW$ | $VW$ | $EW$ | | |
| Mean Adjusted Returns | 7.6 | 7.6 | 9.2 | 9.2 | 28.4 | 28.4 | 82.0 | 82.0 | | |
| Market Adjusted Returns | 20.4 | 3.6 | 24.0 | 5.6 | 44.8 | 18.4 | 94.8 | 73.2 | | |
| Market Model Residuals | 4.0 | 7.2 | 6.0 | 10.8 | 18.8 | 24.4 | 76.4 | 86.4 | | |
| Fama–MacBeth Residuals | 1.2 | 3.6 | 2.0 | 5.2 | 7.2 | 16.4 | 54.8 | 74.8 | | |
| Control Portfolio | 14.0 | 4.8 | 15.6 | 6.4 | 29.6 | 16.0 | 78.0 | 70.0 | | |

[a]For each security, abnormal performance is introduced for one month in the interval $(-10, +10)$, with each month having an equal probability of being selected.

Summary Statistics, Systematic Risk Estimates

| | $VW$ | $EW$ |
|---|---|---|
| Mean estimate of $\beta$ | 1.13 | 0.993 |
| Average cross-sectional standard deviation of $\beta$ | 0.49 | 0.43 |
| Standard deviation of average $\beta$s for the 250 samples | 0.031 | 0.027 |



There is no necessity for $\hat{\beta}$s computed from a value-weighted (rather than an equally weighted) index to have an unweighted average of 1; the only requirement is that the value-weighted average of all security $\hat{\beta}$s (computed with respect to the value-weighted index) be equal to 1. For randomly selected securities, an unweighted average $\hat{\beta}$ greater than 1 would be expected if securities with low value weights have relatively high $\hat{\beta}$s, and vice versa. Hence an average $\hat{\beta}$ of 1.13 results from a particular cross-sectional distribution of $\beta$, and does not imply that our selection procedure is somehow biased toward including high risk securities.

### 7.2. Type I errors with the value-weighted index

Our finding that the unweighted average of $\hat{\beta}$s computed from the Value-Weighted Index is not equal to 1, along with the fact that not all securities have the same value weights in the market portfolio, turns out to have significant implications for the behavior of one performance measurement method under study: the Market Adjusted Returns method implicitly assumes that security $\beta$s average to one, and looks at the average differences between security returns and those on the Value-Weighted Market Index. However, since average $\beta$ is greater than 1, an equally weighted portfolio of randomly selected stocks is expected to have returns which are greater than those of the Value-Weighted Index. The unweighted average difference between security returns and returns on the Value-Weighted Market Index will tend to be positive for any sample of randomly selected securities. Using the Value-Weighted Index, the Market Adjusted Returns method will reject the null hypothesis 'too often'.

This potential problem with the Market Adjusted Returns method does not result from the use of the Value-Weighted Index itself. Rather, a potential bias is induced by the failure to appropriately value weight security returns and security specific performance measures.[40] To our knowledge no

[40]For the Market Model Residuals method, under the null hypothesis there is no bias inherent in not value-weighting the residuals: Since for every security the expected value of the residuals is zero, the average residual is expected to be zero for any set of weights applied to the individual residuals. On the other hand, for the Market Adjusted Returns method, some securities have performance measures which on average will be positive, (e.g., $\beta > 1$) and others have performance measures which will be negative (e.g., $\beta < 1$). Equal weighting of the security specific performance measures will not guarantee an average performance measure of zero because the number of securities with positive performance measures is greater than the number with negative performance measures.

For Fama–MacBeth residuals, there could well be biases in our simulations with the Value-Weighted Index. Note that Fama–MacBeth residuals are computed on the basis of estimates of $\gamma_0$ and $\gamma_1$ derived from the Equally-Weighted Index. If a security's $\beta$ is computed from the Value-Weighted Index, and Fama–MacBeth residuals then calculated from $\hat{\gamma}_0$ and $\hat{\gamma}_1$ based on the Equally Weighted Index, there is no reason for a security's Fama–MacBeth residual to have an expected value of 0 under the null hypothesis of no abnormal performance. Furthermore, deriving estimates of $\gamma_0$ and $\gamma_1$ for the Value-Weighted Index would require more than a mere replication of the Fama–MacBeth procedure; the use of value-weighting procedures (e.g., for portfolio formation) would also be indicated.

event study employing the Market Adjusted Returns method has used such value-weighting. Furthermore, as a practical matter the bias can be substantial. As table 8 indicates, when there is no abnormal performance, the Market Adjusted Returns method rejects the null hypothesis a whopping 20.4% of the time using the Value-Weighted Index, and the mean $t$-statistic for 250 replications is 0.77; if the hypothesis test is performed at the 0.1 rather than the 0.05 level, the rejection rate increases to 27.2%.[41]

From table 8, it also appears that the Control Portfolio method exhibits a positive bias in the performance measure. For that methodology, the rejection rate under the null hypothesis is 14.0% (for $\alpha = 0.05$), and the mean $t$-statistic is 0.36. However, the problem of 'too many' rejections when using the Value-Weighted Index cannot be attributed to the failure to value-weight the security-specific performance measures; this is because the Control Portfolio method, unlike the Market Adjusted Returns method, applies weights to securities such that the portfolio thus formed has a $\beta$ of 1 and an expected return equal to that of the Value-Weighted Market Index.[42]

### 7.3. *Type II errors with the value-weighted index*

Because some of the performance measurement methods do not, under the null hypothesis, reject at the significance level of the test when used with the Value-Weighted Index, legitimate comparisons of the power of the tests for the Equally Weighted versus Value-Weighted Index are not possible for those methods, since the probability of Type I errors differs according to the index being used. However, the one method which does not suffer from 'too high' a frequency of Type I errors using the Value-Weighted Index is the Market Model Residuals method. Some clue as to the relationship between

[41]The rejection frequencies reported in table 8 for the Market Adjusted Returns method will not always be applicable because the magnitude of the bias in the method is critically dependent on several parameters of the experimental situation. For example, for randomly selected securities, the propensity for rejecting 'too often' under the null will be positively related to sample size and the length of time over which abnormal performance is being measured; if a given performance measure, averaged over sample securities, is positive, whether one can reject the null hypothesis that the average performance measure is zero depends on the number of independent observations over which the average is computed. For example, when the null hypothesis being tested is that month '0' (rather than months $-10, +10$) abnormal performance is 0, the rejection rates using the Value-Weighted Index are not much different from those which obtained in simulations using the Equally Weighted Index.

Note also that the bias in the Market Adjusted Returns method is not always positive. The sign of the bias is related to the systematic risk of the sample securities. If the value-weighted average $\beta$ is greater than 1, the bias will be positive; if it is less than 1, it will be negative.

[42]The computation of the test statistic in the Control Portfolio method is discussed in the appendix. The distributional properties of the test statistic in repeated sampling will depend upon the sampling properties of the weights which are estimated and applied to the returns on individual securities; the properties could differ according to the index employed, particularly since the degree of measurement error in $\beta$ (and hence in the weights) can be a function of the index. Conditions on the weights sufficient for the test statistic to be distributed Student-$t$ have not to our knowledge been discussed in the event study literature.



the power of the tests and the specific market index is provided by the results for that method.

As shown in table 8, for each of several different levels of abnormal performance, the rejection rates using Market Model Residuals are higher with the Equally Weighted Index than with the Value-Weighted Index. With 1% abnormal performance, the rejection rates are 10.8 and 6.0%, respectively; with 5% abnormal performance, the rejection rates are 24.4 and 18.8%, respectively.

It thus appears that use of the Equally Weighted Index is no less likely, and in fact slightly more likely, to pick up abnormal performance than use of the Value-Weighted Index. Such a finding is consistent with the argument that the returns on randomly selected securities are on average more highly correlated with the Equally Weighted Index than the Value-Weighted Index. If for a majority of sample securities the precision with which $\beta$ and hence residuals are measured is higher with the Equally Weighted Index, abnormal performance would be easier to detect using that benchmark.

## 8. Additional simulation results

### 8.1. *Simulation results for different sample sizes*

All results reported thus far in this paper are for sample sizes of 50 securities. However, it is of interest to examine the sensitivity of our results to sample size. For the case where all sample securities experience abnormal performance in month '0', table 9 reports simulation results for sample sizes of 12, 25, and again for 50 securities.[43]

As would be expected, the power of the tests increases with sample size. However, the rejection frequencies are not especially sensitive to the number of sample securities. For example, with the Mean Adjusted Returns method, doubling the sample size from 12 to 25 securities increases the rejection frequency with 1% abnormal performance from 14.0 to 15.2%. Doubling sample size again to 50 securities increases the rejection frequency to 26.0%. Furthermore, the relatively favorable performance of the Mean Adjusted Returns method seems to be independent of sample size, and the rejection frequencies still do not appear to be dramatically different than those for methods which adjust returns for market performance and risk.

### 8.2. *The relationship among the tests*

In many of the simulations we have performed, the rejection frequencies are not dramatically different for different methodologies. However, even if

[43]When we attempt to examine larger samples as well, the computing costs were found to be prohibitively high. For sample sizes of 100, the cost of performing just a third of the 250 replications was in excess of $1,000.

two methods have the same rejection frequency for any level of abnormal performance, this does not imply that the two methods will always lead a researcher to the same conclusion. For example, if each of two methods rejects the null hypothesis in 50 of the 250 samples, the samples on which the first method rejects the null hypothesis need not be the same as the samples on which the second method rejects the null hypothesis.[44] To assess the likelihood that the various methods will lead to results which are consistent

Table 9

The effect of sample size on rejection frequencies. Percentage of 250 replications where the null hypothesis is rejected ($\alpha = 0.05$). One-tailed t-test results. $H_0$: mean abnormal performance in month '0' = 0.0. Equally Weighted Index.

| | Actual level of abnormal performance in month '0' | | |
|---|---|---|---|
| Method | 0% | 1% | 5% |
| *Mean Adjusted Returns* | | | |
| $N = 12$ | 5.2 | 14.0 | 79.2 |
| $N = 25$ | 6.0 | 15.2 | 94.8 |
| $N = 50$ | 4.0 | 26.0 | 100.0 |
| *Market Adjusted Returns* | | | |
| $N = 12$ | 2.8 | 8.4 | 72.4 |
| $N = 25$ | 3.6 | 13.2 | 91.6 |
| $N = 50$ | 3.2 | 19.6 | 100.0 |
| *Market Model Residuals* | | | |
| $N = 12$ | 3.2 | 9.6 | 72.4 |
| $N = 25$ | 4.4 | 13.6 | 91.6 |
| $N = 50$ | 4.4 | 22.8 | 100.0 |
| *Fama–MacBeth Residuals* | | | |
| $N = 12$ | 2.8 | 8.4 | 56.8 |
| $N = 25$ | 4.8 | 15.2 | 93.2 |
| $N = 50$ | 4.0 | 21.6 | 100.0 |
| *Control Portfolio* | | | |
| $N = 12$ | 2.8 | 8.4 | 56.8 |
| $N = 25$ | 2.4 | 12.8 | 84.8 |
| $N = 50$ | 4.4 | 18.0 | 100.0 |

for a given sample, it is necessary to examine the results of our earlier hypothesis tests in more detail. In table 10, for the case where all abnormal performance occurs in month '0', we indicate the frequency with which the results of the hypothesis tests for a given sample are the same for different methodologies.

[44]Charest (1978), Langetieg (1978), and Brenner (1979) have all conducted event studies where the results of the hypothesis tests appear to be somewhat sensitive to the particular test method which is used.



When the null hypothesis is true, it appears that the test methods typically lead to results which are somewhat, but not perfectly consistent. For example, in the simulations we presented in table 1, under the null hypothesis the Mean Adjusted Returns method rejected in 4.0% of the samples, and the Market Adjusted Returns method rejected the null hypothesis in 3.2% of the 250 samples. However, as indicated in table 10, the frequency with which *both* methods reject the null hypothesis when it is true

Table 10

The relationship among the tests. For 1% abnormal performance in month '0', the table shows the frequency (in 250 replications) with which a given combination of methods resulted in (a) at least one rejection ($R \geq 1$), and (b) an inconsistency (one rejection, one failure to reject; $R = 1$). The third entry is the frequency with which both methods reject the null hypothesis when it is true ($R = 0$). The number in parentheses is the product of the individual rejection frequencies which obtained for each method under the null hypothesis. $H_0$: mean abnormal performance in month '0' = 0.0 ($\alpha = 0.05$). One-tailed t-test results. Equally Weighted Index.

| | *Market Adjusted Returns* | *Market Model Residuals* | *Fama–MacBeth Residuals* | *Control Portfolios* |
|---|---|---|---|---|
| *Mean Adjusted Returns* | | | | |
| $R \geq 1$ | 45.6% | 48.8 | 47.6 | 44.0 |
| $R = 1$ | 33.2 | 33.2 | 33.6 | 33.2 |
| $R = 0$ | 1.6(0.13) | 2.0(0.18) | 1.6(0.16) | 1.2(0.18) |
| *Market Adjusted Returns* | | | | |
| $R \geq 1$ | | 42.4 | 41.2 | 37.6 |
| $R = 1$ | | 25.2 | 25.6 | 22.4 |
| $R = 0$ | | 2.4(0.15) | 2.8(0.13) | 2.8(0.14) |
| *Market Model Residuals* | | | | |
| $R \geq 1$ | | | 44.4 | 40.8 |
| $R = 1$ | | | 25.2 | 26.4 |
| $R = 0$ | | | 3.2(0.18) | 2.4(0.19) |
| *Fama–MacBeth Residuals* | | | | |
| $R \geq 1$ | | | | 39.6 |
| $R = 1$ | | | | 25.6 |
| $R = 0$ | | | | 2.8(0.18) |

is 1.6%, which is approximately 10 times the frequency which would be expected if the two methods were independent. Furthermore, for all of the pairwise combinations of methods shown in table 10, it appears that the results of the hypothesis tests are also highly correlated; the frequency with which two given methods reject the null hypothesis when it is true ranges from 1.2 to 3.2%. This high correlation suggests that rejecting the null hypothesis using two different methods is much more likely than would be expected by assuming independence of the test methods.

When the null hypothesis is not true, the test methods are still not perfectly consistent. With 1% abnormal performance, the likelihood that one

method will reject the null hypothesis but the other will fail to reject ranges from 22.4 to 33.6%; inconsistencies between two methods seem least likely for the combination of Control Portfolios and Market Adjusted Returns.

Our finding that the test methods are not always consistent when there is abnormal performance opens up the possibility that there are sets of methodologies which, when used jointly, are more likely to detect abnormal performance than any one method alone. For example, as table 10 indicates, with 1% abnormal performance, the frequency with which *at least* one of two methods rejects the null hypothesis of no abnormal performance ranges from 37.6 to 48.8%, which is higher than the rejection rates which typically obtain for the individual tests. However, we hasten to add that the higher rejection rates are not themselves evidence of a more powerful test. It should be kept in mind that the significance level of the test (that is, the probability of *falsely* rejecting at least one of two null hypotheses) also increases when methodologies are used in combination with each other. The probability of at least one Type I error increases with the number of tests, and cannot be assessed unless the dependence of the tests is taken into account.

## 9. Summary and conclusions

In this paper, observed monthly stock return data were employed to examine various methodologies with which event studies measure security price performance. Abnormal performance was artificially introduced into this data. Our conclusions about the performance of the different methodologies can be summarized as follows.

### 9.1. *Simulation results for the 'no clustering' case*

Initially, we simulated a situation where securities and event dates were randomly selected, and event dates for different securities were not clustered in calendar time. When abnormal performance was present, the differences between methodologies based on Mean Adjusted Returns, Market Adjusted Returns, and Market and Risk Adjusted Returns were quite small; the simplest methodology, Mean Adjusted Returns, picked up abnormal performance no less frequently than did the other methodologies, and the power of the tests did not appear to be enhanced using risk adjustment procedures suggested by the Asset Pricing model. For example, when 1% abnormal performance was introduced in month '0' for every security in a sample of 50, each of the methodologies rejected the null hypothesis of no abnormal month '0' performance about 20% of the time when performing a one-tailed $t$-test at the 0.05 level of significance. Such a result also indicates that if the researcher is working with a sample size of 50 and the event under study is not expected *a priori* to have changed the value of the affected securities by 1% or more, the use of monthly data is unlikely to detect the event's impact.



*The use of prior information*

With 5% or more abnormal performance in month '0', rejection rates for a sample size of 50 were 100% for all of the methodologies. However, that simulation result does not imply that an event study using monthly data will always pick up 5% or more abnormal performance using a sample size of 50: if the researcher is unable to identify the specific time at which the abnormal performance would have occurred, the power of the tests for abnormal performance falls off dramatically. For example, we simulated a situation where each of 50 sample securities had 5% abnormal performance in a particular month surrounding month '0', but the precise month was uncertain and different across securities. When the time of the abnormal performance could only be narrowed to an 11 month 'window', the null hypothesis of no abnormal performance over the window was rejected only 30 to 40% of the time with the different methodologies. Thus, unless the time of the abnormal performance can be narrowed using prior information, the null hypothesis often fails to be rejected even when the sample securities experience high levels of abnormal performance.

### 9.2. *Performance measurement when event dates or systematic risk estimates are clustered*

*Calendar time clustering of events*

Our conclusions about the relatively favorable performance of the Mean Adjusted Returns method were found to be highly sensitive to the specification of the stochastic process generating events. For example, when we simulated a situation in which event dates were randomly selected, but clustered in calendar time, the Mean Adjusted Returns method performed very poorly compared to those methods which explicitly adjusted for market performance and for systematic risk. In the extreme example of clustering we examined, all securities of a given sample were assigned a common event date, and the $t$-tests were adjusted to take into account cross-sectional dependence in the security-specific performance measures. The Mean Adjusted Returns method detected 5% abnormal performance in month '0' only about 35% of the time, compared to rejection rates of 98.8 to 100.0% for all the other test methods. On the basis of such results, it is difficult to argue that the use of the Mean Adjusted Returns method will always be appropriate. When there is event month clustering, methodologies which incorporate information about the market's realized return perform substantially better than Mean Adjusted Returns.

*Sample security risk clustering*

Within the class of methodologies which adjust for marketwide factors, we examined several alternatives. These included a one-factor market model, a

two-factor model utilizing Fama–MacBeth residuals, and a Control Portfolio technique in which the return on a portfolio of sample securities was compared to that of another portfolio with the same estimated systematic risk. For randomly selected securities, which were of 'average' risk, the differences between these methodologies were small, regardless of whether or not there was calendar time clustering of events. However, when securities were not randomly selected, and sample security systematic risk estimates were systematically 'clustered' and different from 1, an important difference between the methodologies emerged: with systematic risk clustering, the Control Portfolio method was much less likely to pick up a given level of abnormal performance than either a one-factor or a two-factor model. In fact, when there was risk clustering but not event month clustering, even the simple Mean Adjusted Returns method outperformed the seemingly complicated Control Portfolio method. Thus, under plausible conditions the researcher can actually be made worse off using explicit risk adjustment procedures.

### 9.3. *Additional simulation results*

#### *The choice of market index*

Although use of the Equally Weighted Index is an *ad hoc* procedure, that index led to no notable difficulties in our simulations; however, improper use of the Value-Weighted Index was shown to cause considerable problems which have not been recognized in extant event studies. For example, when some methodologies (including the widely used 'Control Portfolio' methodology) were used with the Value-Weighted Index, the null hypothesis was rejected too often, in some cases over 20% of the time (when testing at the 0.05 significance level) even when there was no abnormal performance. Furthermore, we find no evidence that the use of the Value-Weighted Index increases the power of the tests.

#### *The appropriate statistical test*

For methodologies using the Equally Weighted Index, and for many of those using the Value-Weighted Index, we found that $t$-tests focusing on the average month '0' performance measure (e.g., the average residual) are reasonably well-specified. Although stated significance levels should not be taken literally, when the null hypothesis is true the $t$-tests typically reject at approximately the significance level of the test; the differences between the empirical frequency distribution of the test statistics and the $t$-distribution are generally not large.



On the other hand, certain non-parametric tests used in event studies are not correctly specified. We indicated how the sign and Wilcoxon tests will not give the 'correct' number of rejections unless asymmetry in the distribution of security specific performance measures is taken into account; as far as we can determine, no event study using non-parametric tests has recognized how sensitive the tests can be to departures from the symmetry assumption.

### 9.4. *The bottom line: What's the best methodology?*

Our goal in this paper has not been to formulate the 'best' event study methodology, but rather, to compare different methodologies which have actually been used in event studies and which constitute current practice. Even among the methods we have studied, it is difficult to simulate every conceivable variation of each methodology, and every plausible experimental situation; while we cannot, therefore, indicate the 'best' methodology (given some set of criteria), our simulations do provide a useful basis for discriminating between alternative procedures.

A 'bottom line' that emerges from our study is this: beyond a simple, one-factor market model, there is no evidence that more complicated methodologies convey any benefit. In fact, we have presented evidence that more complicated methodologies can actually make the researcher worse off, both compared to the market model and to even simpler methods, like Mean Adjusted Returns, which make no explicit risk adjustment. This is not to say that existing techniques cannot be improved; indeed, our results have led us to suggest a number of ways in which such improvements can be made. But even if the researcher doing an event study has a strong comparative advantage at improving existing methods, a good use of his time is still in reading old issues of the *Wall Street Journal* to more accurately determine event dates.

## Appendix: Methodologies for measuring security price performance

In this appendix, we discuss in more detail the different methods for measuring security price performance which are used in the study. For all of the methodologies, securities are selected as discussed in section 3. For a given security $i$, its monthly arithmetic return, $R_{it}$, is available over a period beginning in the 89th month prior to the event ($t = -89$) and terminating at the end of the tenth month following the event ($t = +10$). There are two observation periods over which return behavior is examined: a single month (month 0) and a series of event-related months (typically, months $-10$ through $+10$). For a particular level of abnormal performance, a given method computes the performance measures for individual securities in each

of the 250 samples and, for each sample, assesses the statistical significance of those measures.[45]

### A.1. Mean adjusted returns

For each security $i$, the mean $K_i$, and standard deviation $\sigma(R_i)$ of its return in months $-89$ through $-11$ are estimated:

$$\hat{K}_i = \frac{1}{79} \sum_{t=-89}^{-11} R_{it}, \tag{A.1}$$

$$\hat{\sigma}(R_i) = \left[ \frac{1}{78} \sum_{t=-89}^{-11} (R_{it} - \hat{K}_i)^2 \right]^{\frac{1}{2}}. \tag{A.2}$$

The measure of abnormal performance for a given security in a given event-related month, $A_{it}$, is the difference between its realized return and the estimate of its mean return in the $(-89, -11)$ period, where this difference is standardized by the estimated standard deviation of the security's return in the $(-89, -11)$ period,[46]

$$A_{it} = (R_{it} - \hat{K}_i)/\hat{\sigma}(R_i). \tag{A.3}$$

[45]For all statistical tests reported in the paper, the $(-10, +10)$ period is ignored in estimating parameters such as the variance of the various performance measures. The simulations presented in tables 1 through 3 have also been replicated using the additional returns from the $(-10, +10)$ period. The additional 21 months of data provided by this period do not appear to have a marked impact on the rejection frequencies or on the distributional properties of the test statistics under the null hypothesis.

However, in an actual event study, the results can be sensitive to the inclusion (or exclusion) of the period surrounding the event. If high levels of abnormal performance are present, then including observations from around the time of the event gives more weight to apparent 'outliers', tending to increase the variance of the security-specific performance measures, and, as borne out by simulations not reported here, lowering the power of the tests. In addition, if there are abnormal returns in the event period, it is difficult to infer 'normal' returns, particularly if the period of the abnormal performance is long and includes an amount of data which is substantial relative to the total available. For a further discussion of reasons to exclude the 'event' period, see Brown, Durbin and Evans (1975). Note that if the event period is excluded in computing parameter estimates which are then used to predict returns into that period, the variance of the performance measure can be adjusted to reflect the predictive nature of the excess returns [see Patell (1976)]. However, event studies typically make no such adjustment; to be consistent with those studies, our simulations, as discussed in this appendix, also make no such adjustment.

[46]The standardization is similar to that which is performed in the Jaffe–Mandelker procedure. Earlier, we noted that the $t$-statistics for the Mean Adjusted Returns method had too much mass in the left-hand tail. That behavior becomes more pronounced without the standardization, making comparisons of power difficult because the level of Type I error is not constant across methodologies.



For month '0', and every month, this procedure yields one performance measure for each of the $N$ securities in the sample.

### The t-tests

The $t$-test for month '0' examines whether or not the average value of the performance measure in month '0' (i.e., the average month '0' standardized difference) is equal to 0. Except when otherwise specified, the $t$-test in the Mean Adjusted Returns method takes into account cross-sectional dependence in the security specific performance measures via a procedure we call Crude Dependence Adjustment.

For all methods using Crude Dependence Adjustment, the standard deviation of the month '0' average performance measure is estimated from the values of the average performance measures in months $-49$ through $-11$. Any cross-sectional dependence in the performance measures is thus taken into account. If the average performance measures for each event-related month are normal, independent,[47] and identically distributed, then under the null hypothesis the ratio of the month '0' average performance measure to the estimated standard deviation is distributed Student-$t$ with 38 degrees of freedom.

With Crude Dependence Adjustment, the test statistic is given by

$$\frac{\frac{1}{N} \sum_{i=1}^{N} A_{i0}}{\left[ \frac{1}{38} \left( \sum_{t=-49}^{-11} \left[ \left( \frac{1}{N} \sum_{i=1}^{N} A_{it} \right) - A^* \right]^2 \right) \right]^{\frac{1}{2}}}, \tag{A.4}$$

where

$$A^* = \left[ \sum_{t=-49}^{-11} \sum_{i=1}^{N} A_{it} \right] \cdot \frac{1}{39N}. \tag{A.5}$$

In the $t$-test for abnormal performance in the $(-10, +10)$ interval, the numerator in (4) becomes

$$\frac{1}{21N} \sum_{t=-10}^{+10} \sum_{i=1}^{N} A_{it}, \tag{A.6}$$

[47]Note that, in general, the average performance measures will not literally be independent, which is one reason we refer to our procedure as a crude one. For example, suppose security A had an event in January and security B had an event in March of the same year. Then the average standardized difference in event months spread two months apart ($-2$ and 0, $-1$ and $+1$, etc.) will be calculated on the basis of observations from the same calendar month, and which are likely to be positively correlated. That the Mean Adjusted Returns method does not appear to reject the null hypothesis 'too often' suggests that the degree of dependence is small with this procedure.

and the denominator is the same as that shown in (4), divided by $\sqrt{21}$. This test statistic is also assumed to be distributed Student-$t$ with 38 degrees of freedom.

*Non-parametric tests*

The test statistic in the sign test for month '0' abnormal performance is given by

$$Z = \frac{|P - 0.5| - 1/2N}{\sqrt{(0.5(0.5)/N)}}, \tag{A.7}$$

where $P$ is the proportion of $A_i$'s in month '0' having positive signs.[48] The test statistic is assumed unit normal under the null hypothesis. The Wilcoxon test is carried out as in Lehmann (1975, pp. 128–129).

### A.2. Market adjusted returns

For month '0', the performance measure for a given sample security is the difference between its return and the corresponding return on the market index,

$$A_{it} = R_{it} - R_{mt}. \tag{A.8}$$

An assumption sufficient for using such a performance measure is that the systematic risk for each sample security is equal to 1. In that case, the expected value of the difference between the return on a security and the return on the market index should, in an asset pricing model framework, be equal to zero.[49] The significance of the month '0' and months $(-10, +10)$ abnormal performance is assessed exactly as in the Mean Adjusted Returns method. For the month '0' $t$-test, the performance measure for a sample is the average difference.

### A.3. Market model residuals

For each security in the sample, we regress its return in months $-89$ through $-11$ against the returns on the market portfolio during the corresponding calendar months. This 'market model' regression yields a 'residual' in each event-related month for each security. For a given security, the market model residual is its measure of abnormal performance. For the

[48] The sign of $Z$ is equal to the sign of the difference between $P$ and 0.5.

[49] For the average difference to be zero, it is not necessary for *all* sample securities to have $\beta = 1$. It is required only that the average $\beta$ be equal to 1.



$t$-test on month '0', the performance measure is the average market model residual. Thus, we examine the significance of

$$\frac{1}{N}\sum_{i=1}^{N} A_{i0}, \tag{A.9}$$

where

$$A_{it} = R_{it} - \hat{\alpha}_i - \hat{\beta}_i R_{mt}. \tag{A.10}$$

Because residual cross-correlation in calendar time is likely to be small (and would generally be even smaller in event time), simulations with Market Model Residuals make no dependence adjustment, unless otherwise stated. For procedures making no dependence adjustment, the significance test on the average residual (or average security specific performance measure) is carried out under the assumption that residuals (or other performance measures) are uncorrelated across securities. The standard deviation of the average performance measure is estimated on the basis of the standard deviation of the performance measure of each sample security in the $(-89, -11)$ period. For month '0', the test statistic is given by

$$\frac{\frac{1}{N}\sum_{i=1}^{N} A_{i0}}{\frac{1}{N}\left(\sum_{i=1}^{N}\left[\frac{1}{77}\sum_{t=-89}^{-11}\left(A_{it} - \left(\sum_{t=-89}^{-11}\frac{A_t}{79}\right)\right)^2\right]\right)^{\frac{1}{2}}}, \tag{A.11}$$

which is distributed Student-$t$ with 78 degrees of freedom for the assumed normal and independent $A_{it}$s.

### A.4. Fama–MacBeth residuals

For each security in the sample, we again use the market model to compute an estimate of its systematic risk over the period from $-89$ through $-11$. Using that estimate, we then compute a 'Fama–MacBeth' residual for each security for each month from $-10$ through $+10$,

$$A_{it} = R_{it} - \hat{\gamma}_1 - \hat{\gamma}_2\hat{\beta}_i. \tag{A.12}$$

For a given month $t$, the Fama–MacBeth residual for security $i$, $A_{it}$, is the return on the security, net of the effect of marketwide factors captured by estimates of $\gamma_1$ and $\gamma_2$. We refer to the $A_{it}$s as 'Fama–MacBeth' residuals because the estimates of $\gamma_1$ and $\gamma_2$ which we use were derived by Fama and

MacBeth (1973).[50] For a given month, these coefficients reflect, respectively, the constant and the slope term in a cross-sectional regression of average portfolio return on average portfolio $\hat{\beta}$. The estimates $\hat{\gamma}_1$ and $\hat{\gamma}_2$ differ from calendar month to calendar month. However, for a given calendar month, they are the same for all securities, and should correspond to the return on the zero beta portfolio and the slope of the market line, respectively.[51]

The month '0' performance measure for a given security is its Fama–MacBeth residual. As with market model residuals, the performance measure for the $t$-test is the average Fama–MacBeth residual, and its statistical significance is assessed exactly as in that method. The test statistic is given in (A.11), unless otherwise stated.

### A.5. Control portfolios

This method forms a portfolio of sample securities where the portfolio has approximately the same estimated systematic risk as the market index. The month '0' performance measure for this method is the difference between the return on that portfolio of sample securities and the average return on the market index in the months when securities experienced events.

The procedure we use to construct portfolios and estimate weights is as follows:

*Portfolio Construction* — Two portfolios are formed from the sample securities. Each portfolio is assigned half of those securities. The first portfolio consists of 'low-$\beta$' sample securities. The second portfolio consists of 'high-$\beta$' sample securities. The composition of each portfolio is determined by a market model regression for each of the securities in the sample for the months $-89$ through $-50$. The securities are ranked according to their estimated $\beta$s and, based on the rankings, securities are assigned to either the high-$\beta$ or low-$\beta$ portfolio.

*Estimation of Weights* — For each month in event-related time, we estimate the returns on each of two portfolios. The first is the equally weighted portfolio of high-$\beta$ securities. The second is the equally weighted portfolio of low-$\beta$ securities. For each equally weighted portfolio, we estimate its $\beta$, based on data from months $-49$ through $-11$. In this way, $\beta$s used for forming portfolios and $\beta$s used for estimating weights will be independent. Given the two estimates of $\beta$, we estimate a unique set of weights

[50]The estimates of $\gamma_1$ and $\gamma_2$ which we use are those reported in Fama (1976, pp. 357–360). The methodology for estimating those coefficients is discussed both there and in Fama and MacBeth (1973). In the original Fama–MacBeth article, $\gamma_1$ and $\gamma_2$ are referred to as $\gamma_0$ and $\gamma_1$, respectively.

[51]See Fama (1976, ch. 9). Brenner (1976) presents evidence that the Fama–MacBeth estimates of $\gamma_1$ are not uncorrelated with the market return.



summing to one which, when applied to the high- and low-$\beta$ portfolios, yields a new portfolio with a $\beta$ of 1 relative to the market index.

Since the $\beta$s of the high- and low-$\beta$ security portfolios are estimates, the two weights derived are only an estimate of the correct weights which would be derived, given the true $\beta$s. The weights are constant over the estimation period, and may imply short-selling. For each event-related month, the return on a risk-adjusted portfolio of sample securities is estimated by applying the calculated weights to the return on the low- and high-$\beta$ portfolios of sample securities. For each event-related month, we then estimate the difference between the return on the risk-adjusted portfolio of sample securities and the average return on the market index. The standard deviation of the difference is calculated on the basis of differences in returns from months $-49$ through $-11$. Thus, the significance tests involve Crude Dependence Adjustment. If the difference in returns on the portfolios is normal, independent, and identically distributed, then the test statistic is distributed Student-$t$ with 38 degrees of freedom, and is given by

$$\frac{D_0}{\left[\frac{1}{38}\sum_{t=-49}^{-11}\left[D_t-\left(\frac{1}{39}\sum_{t=-49}^{-11}D_t\right)\right]^2\right]^{\frac{1}{2}}}, \tag{A.13}$$

where $D_t$ is the difference in returns in event month $t$.

### A.6. The Jaffe–Mandelker methodology[52]

For simulations where the Jaffe–Mandelker method of dependence adjustment is applied to the performance measures (residuals, deviations from mean return, etc.), sample security performance is examined in calendar rather than event time.

#### Measuring performance for a given event-related month

In order to examine security price performance for a given event-related month (say '0') this methodology forms portfolios in *calendar* time. For every month in calendar time, a portfolio is formed. For a given month, the portfolio consists of all securities which experience an event at that time. The portfolio for a given month may contain more than 1 security. This will happen whenever two or more securities have their event in the same calendar month. Conversely, the portfolio for a given month may contain no securities. This will be the case whenever it happens that no firms are experiencing an event in a particular calendar month. Thus, the number of

[52]See Jaffe (1974) and Mandelker (1974).

non-empty portfolios actually formed for investigating performance in event month '0' will be equal to the number of *different* calendar months in which firms experience events.

For each portfolio, the securities included in the portfolio (and their performance measures) are given equal weight, and a portfolio residual is calculated. The portfolio residual for a given calendar month is an unweighted average of the residuals (or other performance measures) of the securities in the portfolio. This 'residual' is then standardized by dividing it by its estimated standard deviation.

The purpose of standardization is to insure that each portfolio residual will have the same variance. Standardization is accomplished by dividing each residual by its estimated standard deviation. In this way, each residual has an estimated variance of 1. If the standardization were not performed, the variance of the residual would not be constant: the variance would be low in months when many securities experienced events and were in the portfolio, and high in calendar months when few securities experienced events. A statistical test which assumed homogeneity of the variance would be inappropriate. Explicitly taking into account changes in the variance should lead to more precise tests; this will be true not only because portfolio size changes, but also because residual variance is not constant across securities.

The standardization procedure yields a vector of residuals, each of which is distributed $t$.[33] There will be one residual for every calendar month in which any sample firm had an event. The test for abnormal performance in month '0' is a test of the hypothesis that the mean standardized portfolio residual is equal to zero. Following Jaffe (1974, p. 418), the mean residual is assumed normally distributed.

## References


Ball, Ray, 1978, Anomalies in relationships between securities' yields and yield-surrogates, Journal of Financial Economics 6, June/Sept., 103–126.

Ball, Ray and Philip Brown, 1968, An empirical evaluation of accounting numbers, Journal of Accounting Research 6, 159–178.

Ball, Ray, Philip Brown and Frank Finn, 1977, Share capitalization changes, information, and the Australian equity market, Australian Journal of Management 2, Oct., 105–117.

Bassett, Gilbert and Roger Koenker, 1978, Asymptotic theory of least absolute error regression, Journal of the American Statistical Association 73, Sept., 618–622.


[33]Details of the standardization procedure are given in Mandelker (1974, pp. 332–333). To compute the standard deviation of, say, the April 1969 residual, that portfolio must be reconstructed. For each month in the 49-month period prior to April 1969 (except the 10 calendar months immediately before April 1969), a residual for the securities which are in the April 1969 portfolio is calculated. The April 1969 portfolio residual is then standardized by dividing it by the standard deviation of that particular portfolio's residual calculated for this 39-month period. If we are dealing with, say, month '0', then for *each* portfolio the entire procedure must be repeated for each calendar month in which an event occurs.




Bawa, Vijay, Stephen Brown and Roger Klein, 1979, Estimation risk and optimal portfolio choice (North-Holland, New York).

Beaver, William H., 1979, Econometric properties of alternative security return metrics, Unpublished manuscript (Stanford University, Stanford, CA).

Black, Fischer, 1972, Capital market equilibrium with restricted borrowing, Journal of Business 45, July, 444–454.

Black, Fischer and Myron Scholes, 1973, The behavior of security returns around ex-dividend days, Unpublished manuscript (Massachusetts Institute of Technology, Cambridge, MA).

Black, Fischer, Michael Jensen and Myron Scholes, 1972, The capital asset pricing model: Some empirical tests, in: M. Jensen, ed., Studies in the theory of capital markets (Praeger, New York).

Brenner, Menachem, 1976, A note on risk, return, and equilibrium: Empirical tests, Journal of Political Economy 84, April, 407–409.

Brenner, Menachem, 1979, The sensitivity of the efficient market hypothesis to alternative specifications of the market model, Journal of Finance 34, Sept., 915–929.

Brown, R.L., J. Durbin and J.M. Evans, 1975, Techniques for testing the constancy of regression relationships over time, Journal of the Royal Statistical Society Series B 37, 149–192.

Charest, Guy, 1978, Split information, stock returns, and market efficiency, Journal of Financial Economics 6, June/Sept., 265–296.

Collins, Daniel W. and Warren T. Dent, 1978, Econometric testing procedures in market based accounting research, Unpublished manuscript (Michigan State University, East Lansing, MI).

Collins, Daniel W. and Warren T. Dent, 1979, The proposed elimination of full cost accounting in the extractive petroleum industry, Journal of Accounting and Economics 1, March, 3–44.

Cornell, Bradford and J. Kimball Dietrich, 1978, Mean-absolute-deviation versus least squares regression estimation of beta coefficients, Journal of Financial and Quantitative Analysis 13, March, 123–131.

Cowles, Alfred, 1933, Can stock market forecasters forecast?, Econometrica 1, 309–324.

Fama, Eugene F., 1976, Foundations of finance (Basic Books, New York).

Fama, Eugene F. and James D. MacBeth, 1973, Risk, return and equilibrium: Empirical tests, Journal of Political Economy 71, May/June, 607–636.

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

Foster, George, 1980, Accounting policy decisions and capital market research, Journal of Accounting and Economics, forthcoming.

Gonedes, Nicholas J., 1978, Corporate signaling, external accounting, and capital market equilibrium: Evidence on dividends, income and extraordinary items, Journal of Accounting Research 16, Spring, 26–79.

Gonedes, Nicholas J., Nicholas Dopuch and Stephen J. Penman, 1976, Disclosure rules, information-production, and capital market equilibrium: The case of forecast disclosure rules, Journal of Accounting Research 14, Spring, 89–137.

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

Kaplan, Robert S. and Richard Roll, 1972, Investor evaluation of accounting information: Some empirical evidence, Journal of Business 45, April, 225–257.

Langetieg, Terence C., 1978, An application of a three factor performance index to measure stockholder gains from merger, Journal of Financial Economics 6, Dec., 365–384.

Latane, Henry A. and Charles P. Jones, 1979, Standardized unexpected earnings — 1971–1977, Journal of Finance 34, 717–724.

Lehmann, E.L., 1975, Nonparametrics: Statistical methods based on ranks (Holden-Day, San Francisco, CA).

Mandelker, Gershon, 1974, Risk and return: The case of merging firms, Journal of Financial Economics 1, Dec., 303–335.

Marsaglia, G., K. Ananthanarayanan and N. Paul, 1973, Random number generator package — 'Super duper', Mimeo. (School of Computer Science, McGill University, Montreal).

Masulis, Ronald W., 1978, The effects of capital structure change on security prices, Unpublished Ph.D. dissertation (University of Chicago, Chicago, IL).

Mayers, David and Edward M. Rice, 1979, Measuring portfolio performance and the empirical content of asset pricing models, Journal of Financial Economics 7, March, 3–28.

Officer, Robert R., 1971, A time series examination of the market factor of the New York stock exchange, Ph.D. dissertation (University of Chicago, Chicago, IL).
Ohlson, James A., 1978, On the theory of residual analyses and abnormal performance metrics, Australian Journal of Management 3, Oct., 175–193.
Ohlson, James A., 1979, Residual (API) analysis and the private value of information, Journal of Accounting Research 17, Autumn, 506–527.
Patell, James M., 1976, Corporate forecasts of earnings per share and stock price behavior: Empirical tests, Journal of Accounting Research 14, Autumn, 246–276.
Patell, James M., 1979, The API and the design of experiments, Journal of Accounting Research 17, Autumn, 528–549.
Roll, Richard, 1977, A critique of the asset pricing theory's tests; Part I: On past and potential testability of the theory, Journal of Financial Economics 4, March 129–176.
Roll, Richard and Stephen A. Ross, 1979, An empirical investigation of the arbitrage pricing theory, Unpublished manuscript (University of California, Los Angeles, CA).
Scholes, Myron and Joseph Williams, 1977, Estimating betas from non-synchronous data, Journal of Financial Economics 5, Dec., 309–328.
Schwert, G. William, 1978, Measuring the effects of regulation: evidence from capital markets, Unpublished manuscript (University of Rochester, Rochester, NY).
Udinsky, Jerald and Daniel Kirshner, 1979, A comparison of relative predictive power for financial models of rates of return, Journal of Financial and Quantitative Analysis 14, June, 293–315.
Warner, Jerold B., 1977, Bankruptcy, absolute priority, and the pricing of risky debt claims, Journal of Financial Economics 4, May, 239–276.
Watts, Ross L., 1978, Systematic 'abnormal' returns after quarterly earnings announcements, Journal of Financial Economics 6, June/Sept., 127–150.

Journal of Financial Economics 8 (1980) 259–282. © North-Holland Publishing Company

# DISCRETELY ADJUSTED OPTION HEDGES*

Phelim P. BOYLE

*University of British Columbia, Vancouver, British Columbia, Canada V6T 1Y8*

David EMANUEL

*University of Texas at Dallas, Richardson, TX 75080, USA*

Received August 1979, revised version received April 1980

This paper analyses the distribution of returns on a hedged portfolio, consisting of a European call option and its associated stock, when the portfolio is rebalanced at discrete time intervals. Under the assumptions of the Black–Scholes model this distribution is particularly skew. In tests of the average return on a hedged portfolio this skewness leads to biased $t$-statistics. The paper explores the nature and extent of this bias and suggests procedures for overcoming it. Other aspects of discrete hedging are also discussed.

## 1. Introduction

The creation and maintenance of a riskless hedge plays an essential role in the derivation of the Black–Scholes option pricing formula. In the case of a European call option, a hedge portfolio is constructed by establishing a long position in the option and a short position in the underlying stock on which the option is written. The relative position in the two securities in the hedge portfolio is determined by the first partial derivative of the option pricing formula with respect to the stock price. [For a more complete description, see either Black and Scholes (1973) or Smith (1976).] Given their assumptions, the effect of diffusion in the stock price is thus eliminated and with continual adjustment of the hedge composition the value of the hedge at maturity becomes riskless. Exploitation of this observation leads to the derivation of the option pricing formulae.

In practice it is clearly not possible to rebalance a portfolio continuously. In their empirical tests, Black and Scholes (1972) assumed that the hedge

*The authors appreciate the assistance of Boon Yong Chew in the preparation of an earlier version of this paper. Earlier versions of this paper were given in seminars at New York University, Yale University, The London Graduate Business School and The Western Finance Association Meetings (1979) in San Francisco. The authors are grateful for the comments received. The authors acknowledge the comments of the referee, Jonathan Ingersoll, on an earlier version of this paper. One of the authors (PPB) acknowledges financial support from a Canada Council Leave Fellowship during the preparation of this paper.