THE JOURNAL OF FINANCE • VOL. LVI, NO. 5 • OCT. 2001

Evaluating Mutual Fund Performance

S. P. KOTHARI and JEROLD B. WARNER*

ABSTRACT

We study standard mutual fund performance measures, using simulated funds whose characteristics mimic actual funds. We find that performance measures used in previous mutual fund research have little ability to detect economically large magnitudes (e.g., three percent per year) of abnormal fund performance, particularly if a fund's style characteristics differ from those of the value-weighted market portfolio. Power can be substantially improved, however, using event-study procedures that analyze a fund's stock trades. These procedures are feasible using time-series data sets on mutual fund portfolio holdings.

THIS PAPER USES simulation procedures to study empirical properties of performance measures for mutual funds (i.e., managed equity portfolios). Recent studies of mutual fund returns (e.g., Carhart (1997), Daniel et al. (1997)) have moved beyond performance measures based on the Capital Asset Pricing Model, such as "Jensen alpha." These studies account for nonmarket factors in the cross section of expected returns, such as size, book-to-market, and momentum.

Applying multifactor benchmarks to performance measurement has been characterized as "simple" and "straightforward" (Fama and French (1993), p. 54). The basis for this view is that multifactor benchmarks are cross-sectionally well specified. However, the power of multifactor benchmarks to detect abnormal performance of a managed portfolio has received little attention. Further, the specific method for implementing a multifactor benchmark could also affect the power of the tests. For example, there is reason to believe that regression-based benchmarks (e.g., four-factor alpha) will have lower power than characteristic-based benchmarks, which form comparison portfolios using information on fund holdings (Daniel et al. (1997)). However, the magnitude of the difference is an unresolved empirical issue.

* Sloan School of Management, Massachusetts Institute of Technology and William E. Simon Graduate School of Business Administration, University of Rochester, respectively. We thank Doug Breeden, Charles Nelson, Wayne Ferson, Bill Schwert, Cliff Smith, René Stulz, two anonymous referees, seminar participants (Rochester, Colorado (Burridge Center Annual Conference) and Duke) for their comments, and Andreas Gintschel, Peter Wysocki, and Tzachi Zach for excellent research assistance. We are grateful to the Research Foundation of the Institute of Chartered Financial Analysts and the Association for Investment Management and Research, the Bradley Policy Research Center at the Simon School, and the John M. Olin Foundation for financial support. S.P. Kothari acknowledges financial support from the New Economy Value Research Lab at the MIT Sloan School of Management.

We provide direct evidence on properties of fund performance measures. Our simulation procedures use random-stratified samples of NYSE/AMEX securities. We form simulated funds and track their performance over time, using a variety of procedures. The simulated funds are designed to mimic the actual characteristics (e.g., size, book-to-market, number of securities, turnover) of funds covered by Morningstar. The simulated funds' performance is ordinary, and well-specified performance measures should not systematically indicate abnormal performance. We explicitly introduce abnormal performance into the portfolios, and focus on the performance measures' power to detect an individual fund's abnormal performance.

We report two main results. First, performance measures typically used in mutual fund research have little ability to detect economically large magnitudes (e.g., three percent per year) of abnormal fund performance, particularly if a fund's style characteristics differ from those of the value-weighted market portfolio. Characteristic-based procedures that compare fund returns to returns on stocks with similar size, book-to-market, and momentum can exhibit modest improvements over regression procedures, but these power comparisons are clouded because style-based funds sometimes exhibit misspecification in both types of measures.

Second, standard event-study procedures that analyze a fund's stock trades can substantially improve power. These procedures are feasible using time-series data on a fund's holdings from CDA (quarterly) or Morningstar (monthly). The trade-based framework extends the characteristic-based approach and exploits information on changes in stock weights, but with the exception of Chen, Jegadeesh, and Wermers (2000), it has generally not been used. These authors find that any abnormal return following mutual funds' aggregate trades is concentrated in the quarters immediately following the trades. Our simulations show that under these conditions (i.e., a stock's abnormal return is somewhat short-lived), the trade-based event-study approach can be quite powerful. We caution, however, that higher power does not occur if abnormal performance is sufficiently long-lived, for example, if abnormal performance lasts four or more quarters. This is a key limitation of the trade-based approach.

Section I outlines the key issues. Section II describes our baseline simulation procedure for regression- and characteristic-based procedures. Section III presents the results. Section IV compares these results to trade-based event-study simulations. We conclude in Section V.

I. Performance Measurement Issues

Our study provides new evidence on empirical properties of performance measures. The underlying multifactor benchmarks are well documented, and the simulations are not a new test of asset pricing models. As discussed in this section, however, the paper's results are not easily inferred from the

asset pricing literature. Applying any benchmark to a managed portfolio involves considerations whose potential consequences cannot easily be studied without using simulated portfolios.

A. Power for Regression-based Performance Measures

Multifactor models are advocated as the basis for performance measurement because they have high explanatory power in asset pricing tests, with three-factor cross-sectional R^2s exceeding 90 percent (Fama and French (1996), p. 57). The standard errors of the performance measures (i.e., the regression intercepts) for actual funds, which are one way of assessing power, are generally not studied. Reported standard errors from the asset pricing regressions are not meaningful for mutual fund performance evaluation, in part because they apply only over observation periods of several decades. In practice, investors are typically interested in performance measures over a three- to five-year period.

The standard errors and hence the power of the tests for mutual funds depend on a number of variables. These include the number and types (market capitalization, book-to-market) of stocks in actual funds, and the covariance structure of excess returns. Our evidence is based on simulated funds whose characteristics mimic Morningstar funds, and we use actual security returns. This experimental design should yield reliable inferences about performance measures' empirical properties, regardless of the true return-generating process. Generally, standard errors of the intercepts (alphas) from excess-return regressions are sufficiently large that there is limited ability to detect abnormal performance.

B. Power Using Characteristics

We also study performance measures that compare fund returns directly to a benchmark portfolio of stocks with similar characteristics (e.g, size, book-to-market, and momentum). The low power of regression-based performance measures could occur because they use no fund-specific information other than returns. The basic regression method simultaneously estimates both relevant characteristics (i.e., factor loadings) and abnormal performance. To improve power, characteristics can be estimated directly from information on each mutual fund's portfolio holdings. As Daniel et al. (1997) emphasize, the use of characteristic-based measures reduces standard errors of abnormal performance measures (see also Daniel and Titman (1997)).

Morningstar now reports the portfolio holdings of each mutual fund on a monthly basis, and CDA/Weisenberger has reported this information on a quarterly basis since 1974. Given this information, it is possible to track changes in a fund's portfolio weights and study the performance of individual stocks subsequent to their purchase (or sale) using event-study procedures.

If trades are information motivated but costly, abnormal performance is more likely to be observed immediately following a decision to trade a stock than following a decision to continue to hold a stock. Power improvements from event-study procedures can occur because greater weight is placed on observations with abnormal performance, and because performance measures' standard errors are for shorter periods and hence lower.

As discussed later, the power gains with event-study procedures should depend critically on how long after the trade any abnormal performance persists. We make a range of assumptions about the timing of abnormal returns, and the assumptions are empirically reasonable given the evidence in Chen et al. ((2000), Tables III and VI). Their results suggest that abnormal performance following aggregate fund trades is concentrated in the quarters immediately following the trades. However, there is cross-sectional variation by fund type, with abnormal performance lasting from one to four quarters.

C. Power, Style, and Specification

Since power will depend on fund style (e.g., large firms have lower return variances than small firms), our simulations form style-based portfolios. Although the paper's focus is power, we also present evidence on how test specification can depend on fund style. Fama and French (1993) argue that their three-factor model "does a good job" on the cross section of average stock returns, but they find misspecification for low book-to-market (i.e., growth) stocks in size quintiles one and five (see Fama and French (1993), Table IXa, and Fama and French (1996), Table I, Panel B). This evidence suggests that style-based funds could be misspecified, at least using regression-based three-factor benchmarks.

II. Baseline Simulation Procedure

A. Actual Fund Characteristics Captured by the Simulations

To capture mutual funds' properties and guide our simulations, we select 50 equity funds at random from Morningstar OnDisc dated January 1996. Table I reports descriptive statistics on the equity funds' asset and portfolio characteristics.

For each fund, Morningstar reports the median market capitalization of the stocks held.[1] From panel A, there is wide variation across the 50 funds. For the median fund, the median market capitalization of the equity holdings is \$6.4 billion, which corresponds to NYSE size decile two. The median fund is tilted toward large stocks (see also Daniel et al. (1997)). To reflect this regularity, the baseline simulations first assume that the probability of a stock's inclusion in the simulated portfolio is equal to a stock's market value weight in the NYSE–AMEX index. We later study many selection

[1] The Morningstar definition of median is that half of the fund's money is invested in stocks of firms with larger than the median market capitalization.

Table I

Descriptive Statistics for a Sample of 50 Randomly-Selected Equity Mutual Funds

In Panel A, fund size is the aggregate net asset value of a mutual fund as of December 31, 1995, or the fund's most recent reporting date before December 31, 1995. Turnover is the percentage of a mutual fund portfolio's holdings that have changed over the past year. NYSE decile rankings are based on the market capitalizations of NYSE stocks as of September 30, 1996, as reported in Stocks, Bonds, Bills, and Inflation, 1997 Yearbook, Ibbotson Associates, Chicago. In Panel B, for each mutual fund, Morningstar reports the percentage of total fund assets invested in each stock. The weights of a fund's assets sum to one. Summary statistics from these weights are reported in Panel B. Using the percentage investments, for each mutual fund we first calculate selected statistics (average weight, minimum, median, maximum, and percentiles of weights). This generates 50 cross-sectional observations for each selected statistic (50 average weights, 50 minimum weights, etc.). The rows of panel B report summary statistics for the 50 cross-sectional observations on each selected statistic; cross-sectional median values of the selected statistics are shown in bold. Data source: Morningstar's Mutual Funds OnDisc, January 1996.

Panel A: General Characteristics

	Fund Size, $million	Number of Stocks Held	Annual Turnover, %	Median Market Capitalization of the Stocks Held by a Mutual Fund, $million[a]	NYSE Decile of the Median Market Capitalization Stock, Decile Ranking as of September 1996
Average	**543.8**	**114**	**58.9**	**8,001.8**	**1(Largest)**
Minimum	26.6	23	4	253	8
10th %	30.1	36	20	1.106.2	5
25th %	51.3	47	28.8	2,632	3
Median	**87.5**	**75**	**47.5**	**6,421.5**	**2**
75th %	271.6	131	76.5	10,912	1
90th %	1,249.3	169	106.7	14,924.4	1
Maximum	10,111.6	892	196	33,685	1

Panel B: Descriptive Statistics for Percentage Portfolio Weights on Individual Assets in Mutual Funds

	Selected Statistics Describing an Individual Fund's Portfolio Weights in Percent							
Cross-sectional (N = 50) Statistic	Average Weight	Minimum Weight	10th %	25th %	Median Weight	75th %	90th %	Maximum Weight
Average	**1.95**	**0.26**	**0.57**	**0.75**	**1.13**	**1.59**	**2.10**	**3.64**
Minimum	0.12	0.01	0.01	0.01	0.03	0.07	0.35	1.05
10th %	0.51	0.01	0.09	0.16	0.29	0.63	0.96	1.93
25th %	0.66	0.03	0.18	0.31	0.53	0.91	1.22	2.11
Median	**1.11**	**0.1**	**0.37**	**0.58**	**0.85**	**1.29**	**1.68**	**2.82**
75th %	1.60	0.48	0.88	1.05	1.47	2.00	2.69	4.05
90th %	2.53	0.71	1.31	1.73	2.38	3.12	3.74	5.47
Maximum	34.9	1.3	1.91	2.24	4.27	5.75	7.40	14.47

[a] The Morningstar definition of median is that half of a mutual fund's money is invested in stocks larger than the median market capitalization.

schemes, including equal probability of each stock's inclusion in a simulated mutual fund. The additional simulations reflect fund style by parameterizing both the market capitalization and the book-to-market ratios of the stocks included.[2]

From panel A, the median number of stocks held is 75. This figure is used as the baseline simulation assumption. Additional results for simulations using 50 and 125 security portfolios yield qualitatively similar inferences (Section III.D). The large number of stocks held suggests that fund managers do not place large bets on any one security. Therefore, the baseline simulations invest equally in the stocks selected. Analysis of actual funds' asset weights (panel B) provides additional evidence that this assumption is reasonable. Across the 50 funds, the typical (i.e., median) maximum asset weight is only 2.82 percent, and the typical median asset weight is 0.85 percent.

Median annual turnover from the Morningstar sample is 47.5 percent. Although this is lower than the 100 percent figure assumed in the simulations, turnover among the actively managed mutual funds is likely to be higher than this median turnover. Since our simulations ignore transaction costs, it is unclear exactly why turnover would affect our results. Nevertheless, we also perform simulations under other assumptions about turnover (not reported), but there is no difference in the conclusions.

Simulations in the paper use gross returns. Consistent with the performance measurement literature, we compare returns to a benchmark that implicitly assumes a buy-and-hold strategy. Implications of transaction costs are already well understood, and simulations that include them are unlikely to produce new insights. Transaction costs (including price concessions) reduce the power of the tests studied in this paper to detect stock-picking ability.

B. Constructing Simulated Funds: Details

We construct a 75-stock mutual fund portfolio each month from January 1966 through December 1994. We then track the 348 simulated mutual fund portfolios' performance over three-year periods (months 1 through 36) using a number of performance measures.[3] As discussed later, these three-year periods are overlapping.

The 75 stocks in each portfolio are selected without replacement from the population of all NYSE–AMEX securities having Center for Research in Security Prices (CRSP) monthly returns. We initially report results using two

[2] We also constructed portfolios by dividend yield, but the paper's conclusions are unchanged and to save space the results are not reported.

[3] We repeated the analysis by constructing 10 portfolios per month, which means tracking 3,480 portfolios over three-year periods. The results on specification and power are virtually identical. This increases our confidence that 348 portfolios is large enough to permit precise inferences. As discussed later, the 348 performance measures are reasonably independent. This is not the case with 10 per month, however, because the average cross-correlation in portfolio raw returns exceeds 0.9 and it is greater than 0.6 in the portfolio performance measures.

stock-selection procedures. First, the probability of selecting a stock is equal to its market value weight at the beginning of the calendar year. Second, each stock has equal probability of being selected, which tilts the portfolio toward small-capitalization stocks relative to the value-weighted index.

Although each portfolio's performance is evaluated over three years, the portfolio composition is changed completely (100 percent turnover) at the beginning of the second and third years (i.e., beginning of months 13 and 25).

Any NYSE–AMEX security with return data available in month 1 is eligible for inclusion in the portfolio formed at the beginning of month 1. Similarly, any security with return data available in month 13 can be included in the portfolio formed at the beginning of month 13. This imposes minimal data-availability requirements, and only securities for which return data become available starting in months 2 through 11 (e.g., initial public offerings) are excluded at the beginning of month 1.

For each of the 348 mutual fund portfolios, we construct a time series of 36 monthly returns starting in month 1. We begin with an equal-weighted portfolio, but the portfolio is not rebalanced at the end of each month. This is consistent with the monthly returns earned on a mutual fund that does not trade any of its stocks in one year. We assume each stock's dividends are reinvested in the stock. Since we reconstruct the mutual fund at the beginning of months 13 and 25, we begin the second and third years with equal-weighted portfolios.

C. Portfolio Performance Measures

The regression-based measures are the estimated intercepts from a regression of 36 monthly portfolio excess returns on one or more factor risk premia. We use three regression-based measures. These are based on the Sharpe-Lintner CAPM, the Fama-French three-factor model, and the Carhart four-factor model. We include the Sharpe-Lintner CAPM measure (Jensen alpha) for illustrative purposes, to permit power comparison with other models, and because, despite its weaknesses, it continues to be popular among practitioners.

The single-beta CAPM Jensen alpha measure (see Jensen (1968)) is the intercept from the regression of portfolio excess returns on the market portfolio excess returns:

$$R_{Pt} - R_{ft} = \alpha_P + \beta_P(R_{Mt} - R_{ft}) + \varepsilon_{Pt} \tag{1}$$

where R_{Pt} is the mutual fund portfolio return in month t, R_{ft} is the risk free return in month t, R_{Mt} is the return on the market portfolio in month t, ε_{Pt} is the white noise error term, and α_P and β_P are the regression's intercept and slope (beta risk) coefficients. We use the CRSP value-weighted index as the market factor.

The Fama-French three-factor model (see Fama and French (1993)) and the Carhart four-factor model regression-based measures (see Carhart (1997), and Daniel et al. (1997)) are estimated from expanded forms of equation (1).

These regressions include the Fama–French book-to-market (HML_t) and size (SMB_t) factor returns and the Carhart momentum factor return. HML_t is the high-minus-low book-to-market portfolio return in month t, SMB_t is the small-minus-big size portfolio return in month t, and the momentum factor is the high-minus-low prior one-year return. We construct the book-to-market and size factors similarly to those in Fama and French (1993) and the momentum factor return as described in Carhart (1997) and Daniel et al. (1997). Details are available on request.

Our characteristic-based measures are a mutual fund's return minus the return on a portfolio of stocks with similar characteristics to those in the fund. We use two such measures. The first is the CRSP value-weighted market adjusted return, and assumes that the funds' stocks are similar to the value-weighted market portfolio. We use this rather naïve performance measure simply because it is common among practitioners to examine whether a fund outperformed the market.

The second characteristic-based measure is the size, book-to-market, and momentum matched return. Specifically, to calculate a mutual fund's performance, we match each stock in the mutual fund to one of the 125 size, book-to-market, and momentum characteristic portfolios. We form the 125 portfolios by triple-sorting all NYSE–AMEX stocks on each firm's size, book-to-market (BM), and past one year's return every July 1. The procedure is detailed in Daniel et al. (1997). A stock's characteristic portfolio-adjusted return for a given month is its return minus the characteristic portfolio return. The fund's characteristic-adjusted monthly return is then calculated by averaging the characteristic-adjusted returns of the stocks in the mutual fund.

D. Distributional Properties of Performance Measures

The procedures just described in Section II.C yield a time series of 348 overlapping performance measures (one set for each simulated portfolio). We first examine the distributional properties of each performance measure. For the null hypothesis that the time series mean of a performance measure is zero the test statistic is

$$t = (1/T) \sum_t \alpha_t / S.E.(\alpha) \tag{2}$$

where $S.E.(\alpha)$ is the standard error of the mean of the estimated performance measures. If the estimated performance measures are independently distributed, then the standard error is given by

$$S.E.(\alpha) = \left[\sum_t (\alpha_t - (1/T) \sum_t \alpha_t)^2 \right]^{1/2} \Big/ (T - 1). \tag{3}$$

Since the alphas are estimated using 36-month overlapping windows, we use a correction for serial dependence in estimating the standard error of the mean (see Newey and West (1987, 1994) and Andrews (1991)) in the calculation of the t-statistic in equation (2).

For each sample, we also examine whether the null hypothesis is rejected, and we report the rejection frequencies across the 348 funds. This is done both before and after abnormal performance is introduced (see Section III.A). Rejection rates after introducing abnormal performance provide direct evidence on power. A regression-based performance measure rejects the null hypothesis if the t-statistic for the estimated alpha from the 36-month regression exceeds the critical value at the one or five percent significance level. For characteristic-based measures, we calculate the time-series mean and standard deviation of the 36 monthly abnormal returns for each fund; the t-statistic is the ratio of the mean to its sample standard error.

III. Simulation Results

A. Summary

Overall, the power of the performance measures is low, particularly for style portfolios, and power improvements from using characteristic-based performance measures rather than regression-based measures seem small. Although the measures are typically reasonably well specified, there is modest misspecification when funds' asset characteristics differ from the value-weighted market portfolio (e.g., style portfolios). These various conclusions are robust to changes in experimental design, such as the number of fund securities or the length of time over which returns are studied.

B. Stock Selection Using Market-value-weight Probabilities

B.1. Specification

Table II, panel A reports distributional properties of the time series of 348 performance measures, with the probability of a stock's selection equal to its market-value weight. The averages of the 348 abnormal performance estimates are economically small in magnitude, ranging from −6 basis points per month for the size, BM, and momentum characteristic-adjusted performance measure to 8 basis points per month for the four-factor alpha. Although the performance measures are typically reasonably well specified for funds that mimic the value-weighted index, the average performance is statistically significant (i.e., mean greater than two standard errors away from zero) in three of the five measures. The Newey–West standard error of the average performance measure ranges from 3 basis points for the market-adjusted abnormal performance measure to 1 basis point for the

Table II

Distributional Properties, Specification, and Power of 348 Characteristic-based and Regression-based Mutual Fund Performance Measures of Portfolios of Securities with Selection Probability Equal to a Security's Market Value Weight

Sample: Each month from January 1966 through December 1994 (348 months), a 75-stock mutual fund portfolio is constructed. Its performance is tracked for a three-year period (months 1 through 36). The portfolio composition is changed 100 percent in months 13 and 25. The 75 stocks are selected without replacement from all NYSE–AMEX stocks with nonmissing return data in month 1, and this procedure is repeated in months 13 and 25 using stocks available in those months. The probability that a stock is included in the portfolio constructed in months 1, 13, or 25 is equal to its market capitalization as a fraction of the aggregate market capitalization of all the stocks eligible for inclusion in months 1, 13, or 25. For each of the 348 portfolios, a time series of monthly returns from month 1 through 36 is constructed. Portfolio returns are equal-weighted at the beginning of months 1, 13, and 25, but they are not rebalanced in the intervening periods. Returns are inclusive of dividends.

Panel A: Descriptive statistics: Market-adjusted return is the mutual fund's return in month t minus the return on the CRSP value-weighted portfolio, averaged over months 1 through 36. Size, BM, and momentum-adjusted return is the mutual fund return in month t minus the return on a size, BM, and prior one-year return quintiles-matched companion portfolio, averaged over months 1 through 36. The three regression-based measures, that is, CAPM, Fama–French three-factor, and Carhart four-factor alphas, are the estimated intercepts of the regressions of the mutual funds' excess returns from months 1 through 36 on (i) the CRSP value-weighted portfolio excess return, (ii) the small-minus-large capitalization stock portfolio return, (iii) the high-minus-low book-to-market portfolio return, and (iv) high-minus-low prior one-year (momentum) portfolio return. See the text for details on variable descriptions and regressions to estimate the regression-based performance measures.

Standard errors, *S.E.*, of the means across the 348 mutual funds' five performance measures are calculated by applying the Newey–West (1987) correction for serial dependence for up to five lags. The average t-statistic for each performance measure is the average of the 348 individual fund t-statistics. Each fund's t-statistic is its performance measure divided by the standard error for the fund.

Panels B and C: Specification and power: Percentage of 348 samples where the null hypothesis of zero abnormal performance is rejected at the one and five percent significance levels using one-sided tests (Panel B) and two-sided tests (Panel C) for various levels of annual portfolio abnormal performance.

Abnormal performance: A given level of annual abnormal performance (e.g., one percent) is induced by adding 1/12 of that amount (e.g., 1/12 percent) per month to the return of each security in each mutual fund.

Panel A: Descriptive Statistics for the Performance Measures

		Characteristic-based Performance Measures		Regression-based Measures		
Summary Statistic	Mutual Fund's Raw Return	Market Adjusted	Size, BM, & Momentum Adjusted	CAPM α	Fama–French 3-factor α	Carhart 4-factor α
Mean	0.94	−0.03	−0.06	−0.00	0.04	0.08
Std. dev., percent	0.61	0.23	0.13	0.23	0.13	0.13
Std. error, percent	0.08	0.03	0.01	0.03	0.01	0.01
Avg. t-statistic	1.45	−0.19	−0.50	−0.25	0.40	0.56
Minimum, percent	−0.95	−0.56	−0.44	−0.56	−0.35	−0.22
Median, percent	0.97	−0.03	−0.05	0.01	0.05	0.07
Maximum, percent	2.7	0.50	0.28	0.51	0.39	0.54

Table II—*Continued*

Annual Abnormal Performance	Significance Level											
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Panel B: Rejection Frequencies Using One-tailed Tests												
0%	17	38	1.7	8.9	0.5	3.4	3	12	2.9	11	5	13
1%	21	44	7	20	3.1	9.8	9	27	11	27	12	31
3%	**28**	**53**	**28**	**42**	**38**	**59**	**38**	**47**	**54**	**75**	**56**	**80**
5%	36	64	48	62	77.6	85.9	52	66	84	94	89	98
7.5%	47	74	73	88	92.8	97.7	78	90	98	100	100	100
10%	56	83	95	100	99.1	100	96	99	100	100	100	100
15%	80	89	100	100	100	100	100	100	100	100	100	100
Panel C: Rejection Frequencies Using Two-tailed Tests												
0%	15	28	6	16	2.6	9.2	7	16	1.4	6.9	3	8
1%	16	31	5.2	19	2.3	7.8	6	21	7.8	20	9	21
3%	**22**	**39**	**22**	**36**	**30**	**49**	**30**	**42**	**46**	**66**	**47**	**69**
5%	27	50	45	53	73.5	82.7	48	58	80	90	85	95
7.5%	38	60	66	81	90.2	95.4	73	84	96	100	99	100
10%	47	73	90	98	97.7	100	93	99	100	100	100	100
15%	72	86	100	100	100	100	100	100	100	100	100	100

multifactor-based performance measures.[4] Serial correlation in the estimated performance measure could arise because we use overlapping measurement windows, or if expected returns change over time. However, the autocorrelation-corrected standard errors are substantially larger than the uncorrected standard errors only in the case of the single-factor performance measures.[5]

The first row in panel B of Table II shows rejection rates under the null hypothesis. There is sometimes a modest degree of misspecification, which is expected based on the means of the abnormal performance estimates reported in Panel A. The Carhart four-factor model rejects the null hypothesis 13 percent of the time with one-tailed t-tests at the 5 percent level of significance. The characteristic-based performance measures are well specified, however.

[4] The Newey–West corrected standard errors reported in this study are based on five lags selected on the basis of sample size. There are alternative lag selection procedures discussed in Andrews (1991) and Newey and West (1987, 1994). These alternative procedures yield 50–100 percent larger standard errors only in the case of the single-factor model abnormal performance estimates. In all other cases, all procedures to implement the Newey–West correction yield virtually identical standard error estimates.

[5] Untabulated results show that only the single-factor model abnormal performance estimates (i.e., the market-adjusted return and Jensen alpha) exhibit large autocorrelations that decline gradually from about 0.8 at the first lag to 0.1 at lag 33. In contrast, the multifactor model abnormal performance estimates exhibit almost no positive autocorrelation. Most of these autocorrelations are not reliably different from zero. The point estimates are generally below 0.1 and several estimates are negative.

B.2. Power

The performance measures' standard deviations in Panel A seem large (e.g., 0.13 to 0.23 percent per month), and suggest the low power of the tests. To provide direct measures of power, we induce a given level of annual abnormal performance (e.g., 1 percent) by adding 1/12 of that amount (e.g., 1/12 percent per month) to the return of each security in each sample. Panel B of Table II reports rejection rates for 1 through 15 percent abnormal performance using one-tailed t-tests for positive abnormal performance.

Overall, the performance measures are only moderately powerful in detecting superior performance, despite the similarity of the funds with the value-weighted portfolio. Consider power when, for example, the induced annual abnormal performance is 3 percent (denoted in bold face in Tables II through VI), which seems quite high and difficult to obtain for mutual funds that look much like the S&P 500 index. The size, BM, and momentum characteristic-adjusted performance measure detects this abnormal performance only 59 percent of the time. The four-factor model alpha rejects the null hypothesis 80 percent of the time, but this does not imply higher power. The comparison is clouded because the four-factor model rejects the null too often (13 percent) when there is no abnormal performance. Rejection rates are higher for the multifactor regression-based and characteristic-based performance measures than for the single-factor-model-based performance measures, but this is consistent with the lower standard deviations of the multifactor measures.

Panel C of Table II shows that the rejection frequencies are still smaller using two-tailed tests, with the corresponding rejection rates with 3 percent abnormal performance dropping to 49 percent and 69 percent. To save space, subsequent tables only report one-tailed results.

C. Stock Selection with Equal Probabilities

Table III reports results when portfolios are formed with every NYSE–AMEX stock having an equal likelihood of being included. By construction, the typical firm selected is of median size (i.e., a mid-cap stock). From Panel A, the average of the 348 market-adjusted abnormal performance estimates is 0.31 percent per month (standard error = 0.07 percent) and the Jensen-alpha estimate is 0.29 percent per month (standard error = 0.07 percent) or about 3.6 percent per year, which is economically large. The observed misspecification is expected because statistically significant firm-size related deviations from the CAPM are well documented.

Panel B of Table III shows the dramatically lower power of the tests than reported in Table II. For example, 3 percent annual abnormal performance is detected only 31 percent of the time using the size, BM, and momentum characteristic-adjusted performance measure. Other measures detect 3 percent abnormal performance less frequently. The fall in power highlights the tests' frailty when applied to funds with asset characteristics that depart from the value-weighted portfolio.

When no abnormal performance is introduced, the three- and four-factor regression models exhibit an average abnormal performance of less than one percent per year, as does the multifactor characteristic-based performance measure (see Panel A of Table III). While these point estimates are economically modest in magnitude, they are statistically significant. Although Fama and French (1993) document misspecification of the three-factor model for small firms, our results suggest that the misspecification can occur even if firms are not, on average, of extreme size. Notwithstanding the significant average abnormal performance, results in the first row of Panel B show that the multifactor performance measures (either characteristic or regression based) are well specified.

D. Style Portfolios

D.1. Book-to-Market

The performance measures' low power is reinforced when style-based portfolios are considered in more detail. Table IV reports results for low (Panels A and B) and high (Panels C and D) book-to-market stock portfolios. All NYSE–AMEX stocks whose book-to-market ratio falls below the median ratio of the stocks ranked at the beginning of each year according to their book-to-market ratios are defined as low book-to-market or growth stocks. The corresponding above-the-median stocks are high book-to-market or value stocks. Book-to-market ratio is calculated using financial data from Compustat. Since financial data on Compustat are not available for every NYSE/AMEX stock, the universe of firms from which the low and high book-to-market stocks are samples is less comprehensive than that used elsewhere in the study. The probability of a stock's selection into the portfolios is equal to the market-value weight of each stock in the above- or below-the-median stocks ranked according to their book-to-market ratios. Firm i's market-value weight is its market capitalization divided by the total market capitalization of all the above- or below-the-median stocks ranked according to their book-to-market ratios.

Panel A of Table IV shows modest misspecification of all five models when applied to low book-to-market stock funds, but the direction of the misspecification is not of a consistent sign. The size, BM, and momentum characteristic-based performance measure is on average -7 basis points per month (standard error is 1 basis point) whereas the four-factor model alpha is 13 basis points per month (standard error is 2 basis points). This misspecification leads the four-factor model alpha measure to incorrectly reject the null hypothesis for 24 percent of the funds at the 5 percent level of significance. The Fama–French three-factor model also rejects the null excessively. Because of these misspecifications, inferences about power must be cautious. Rejection frequencies for 3 percent abnormal performance are 53 and 68 percent using the multifactor characteristic-based and four-factor regression model performance measures, respectively, which is similar to those reported in Table II.

Table III

Distributional Properties, Specification, and Power of 348 Characteristic-based and Regression-based Mutual Fund Performance Measures of Portfolios of Securities with Selection Probability Being Equal Across All Available Firms

Sample: Each month from January 1966 through December 1994 (348 months), a 75-stock mutual fund portfolio is constructed. Its performance is tracked for a three-year period (months 1 through 36). The portfolio composition is changed 100 percent in months 13 and 25. The 75 stocks are selected without replacement from all NYSE–AMEX stocks with nonmissing return data in month 1, and this procedure is repeated in months 13 and 25 using stocks available in those months. The probability that a stock is included in the portfolio constructed in months 1, 13, or 25 is $1/N$, where N is the total number of securities available to be included in a mutual fund at the beginning of months 1, 13, and 25. For each of the 348 portfolios, a time series of monthly returns from months 1 through 36 is constructed. Portfolio returns are equal-weighted at the beginning of months 1, 13, and 25, but they are not rebalanced in the intervening periods. Returns are inclusive of dividends.

Panel A: Descriptive statistics: Market-adjusted return is the mutual fund's return in month t minus the return on the CRSP value-weighted portfolio, averaged over months 1 through 36. Size, BM, and momentum-adjusted return is the mutual fund return in month t minus the return on a size, BM, and prior one-year return quintiles-matched companion portfolio, averaged over months 1 through 36. The three regression-based measures, that is, CAPM, Fama–French three-factor, and Carhart four-factor alphas, are the estimated intercepts of the regressions of the mutual funds' excess returns from months 1 through 36 on (i) the CRSP value-weighted portfolio excess return, (ii) the small-minus-large capitalization stock portfolio return, (iii) the high-minus-low book-to-market portfolio return, and (iv) high-minus-low prior one year (momentum) portfolio return. See the text for details on variable descriptions and regressions to estimate the regression-based performance measures.

Standard errors, *S.E.*, of the means across the 348 mutual funds' five performance measures are calculated by applying the Newey–West (1987) correction for serial dependence for up to five lags. The average t-statistic for each performance measure is the average of the 348 individual fund t-statistics. Each fund's t-statistic is its performance measure divided by the standard error for the fund.

Panel B: Specification and power: Percentage of 348 samples where the null hypothesis of zero abnormal performance is rejected at the one and five percent significance levels using one-sided tests for various levels of annual portfolio abnormal performance.

Abnormal performance: A given level of annual abnormal performance (e.g., one percent) is induced by adding 1/12 of that amount (e.g., 1/12 percent) per month to the return of each security in each mutual fund.

Panel A: Descriptive Statistics for the Performance Measures

		Characteristic-based Performance Measures		Regression-based Measures		
Summary Statistic	Mutual Fund's Raw Return	Market Adjusted	Size, BM, & Momentum Adjusted	CAPM α	Fama–French 3-factor α	Carhart 4-factor α
Mean	1.30	0.31	0.06	0.29	0.03	0.05
Std. dev., percent	0.86	0.59	0.25	0.56	0.30	0.31
Std. err., percent	0.10	0.07	0.01	0.07	0.02	0.02
Avg. t-stat	1.49	0.52	0.19	0.53	0.1	0.16
Minimum, percent	−1.40	−0.92	−0.56	−0.92	−0.71	−0.70
Median, percent	1.30	0.20	0.04	0.22	0.03	0.04
Maximum, percent	3.30	1.90	0.92	1.90	1.10	1.20

Table III—*Continued*

Panel B: Rejection Frequencies Using One-tailed Tests

Annual Abnormal Performance	Significance Level											
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
0%	19	48	7.2	23	1.1	8	7.5	21	2.6	7.2	2	8
1%	23	50	9.5	26	3.2	14	10	27	3.7	12	4.6	12
3%	**31**	**56**	**17**	**36**	**14**	**31**	**17**	**35**	**11**	**28**	**11**	**26**
5%	40	61	26	44	32	60	26	44	27	47	26	47
7.5%	49	66	36	58	63	83	37	60	50	72	48	71
10%	54	75	53	74	86	95	55	75	71	87	70	87
15%	64	83	82	93	98	100	86	96	93	98	93	98

Panels C and D of Table IV report results for high book-to-market stock funds. The results show a marked misspecification of the market-adjusted and CAPM alpha performance measures. In contrast, the multifactor performance measures are well specified, except for the modestly excessive rejection rate (13 percent) using the size, BM, and momentum characteristic-based performance measure. The rejection rates are also the greatest using the multifactor characteristic-based performance measure, whereas they are low using the multifactor regression-based measures. For example, the four-factor regression-based measure detects 3 percent annual abnormal performance in only 27 percent of the funds, whereas the corresponding rejection frequency using the characteristic-based measure is 68 percent. The latter figure overstates power, however, due to the excessive rejection rate under the null.

D.2. Size Portfolios

Table V reports results for large (Panels A and B) and small (Panels C and D) market capitalization stock portfolios. All NYSE–AMEX stocks whose market capitalization falls below (above) the median of the stocks ranked at the beginning of each year according to their market capitalization are defined as small (large) stocks. The results for the large and small market capitalization stocks reinforce the findings discussed above for the high and low book-to-market stock portfolios. Specifically, the performance measures appear slightly misspecified and the power of the tests is higher for large stocks than for small stocks. Both characteristic-based and regression-based multifactor performance measures exhibit low power when applied to small market capitalization stock portfolios. From Panel D, there is only a one-in-five chance of detecting three percent abnormal performance.

E. Sensitivity to Number of Securities and Horizon

Table VI reports results of using 50 and 125 securities in each mutual fund and performance assessment over 5- and 10-year horizons. Stock selection probability is equal to the market value weight of the stocks. Turnover is

Table IV

Distributional Properties, Specification, and Power of 348 Characteristic-based and Regression-based Mutual Fund Performance Measures of Low- and High Book-to-Market Stock Portfolios with the Selection Probability of a Security Equal to Its Market Value Weight within the Low and High Book-to-Market Stocks

Sample: Each month from January 1966 through December 1994 (348 months), a 75-stock mutual fund portfolio is constructed. Its performance is tracked for a three-year period (months 1 through 36). The portfolio composition is changed 100 percent in months 13 and 25. The 75 stocks are selected without replacement from all above-the-median (high book-to-market stocks) or below-the-median (small stocks) of the NYSE–AMEX stocks with nonmissing return data in month 1, and this procedure is repeated in months 13 and 25 using stocks available in those months. For each of the 348 portfolios, a time series of monthly returns from months 1 through 36 is constructed. Portfolio returns are equal-weighted at the beginning of months 1, 13, and 25, but they are not rebalanced in the intervening periods. Returns are inclusive of dividends.

Panels A and C: Descriptive statistics: Market-adjusted return is the mutual fund's return in month t minus the return on the CRSP value-weighted portfolio, averaged over months 1 through 36. Size, BM, and momentum-adjusted return is the mutual fund return in month t minus the return on a size, BM, and prior one-year return quintiles-matched companion portfolio, averaged over months 1 through 36. The three regression-based measures, that is, CAPM, Fama–French three-factor, and Carhart four-factor alphas, are the estimated intercepts of the regressions of the mutual funds' excess returns from months 1 through 36 on (i) the CRSP value-weighted portfolio excess return, (ii) the small-minus-large capitalization stock portfolio return, (iii) the high-minus-low book-to-market portfolio return, and (iv) high-minus-low prior one year (momentum) portfolio return. See the text for details on variable descriptions and regressions to estimate the regression-based performance measures.

Standard errors, *S.E.*, of the means across the 348 mutual funds' five performance measures are calculated by applying the Newey–West (1987) correction for serial dependence for up to five lags. The average t-statistic for each performance measure is the average of the 348 individual fund t-statistics. Each fund's t-statistic is its performance measure divided by the standard error for the fund.

Panels B and D: Specification and power: Percentage of 348 samples where the null hypothesis of zero abnormal performance is rejected at the one and five percent significance levels using one-sided tests for various levels of annual portfolio abnormal performance.

Abnormal performance: A given level of annual abnormal performance (e.g., one percent), is induced by adding 1/12 of that amount (e.g., 1/12 percent) per month to the return of each security in each mutual fund.

Panel A: Descriptive Statistics for Low Book-to-Market Stock Mutual Funds

		Characteristic-based Performance Measures		Regression-based Measures		
Summary Statistic	Mutual Fund's Raw Return	Market Adjusted	Size, BM, & Momentum Adjusted	CAPM α	Fama–French 3-factor α	Carhart 4-factor α
Mean	0.89	−0.09	−0.07	−0.09	0.08	0.13
Std. dev., %	0.64	0.27	0.13	0.26	0.20	0.21
Std. err. %	0.08	0.03	0.01	0.03	0.02	0.02
Avg. t-stat	1.3	−0.45	−0.55	−0.4	0.52	0.73
Minimum, %	−1.00	−0.64	−0.45	−0.73	−0.53	−0.32
Median, %	0.85	−0.07	−0.06	−0.05	0.07	0.11
Maximum, %	2.70	0.58	0.24	0.51	0.67	0.78

Table IV—*Continued*

Panel B: Rejection Frequencies Using One-tailed Tests for Low Book-to-Market Stock Mutual Funds

Annual Abnormal Performance	Significance Level											
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
0%	14	30	0.86	4.9	0	1.1	0.57	3.7	7.2	18	8.9	24
1%	15	36	2.9	13	1.7	6.3	2.6	11	15	31	18	37
3%	**22**	**46**	**18**	**34**	**26**	**53**	**16**	**34**	**43**	**65**	**48**	**68**
5%	28	54	40	53	71	84	39	55	76	86	77	89
7.5%	39	66	60	70	92	96	61	73	91	96	93	99
10%	47	79	78	94	99	100	79	95	96	99	99	100
15%	70	88	100	100	100	100	100	100	100	100	100	100

Panel C: Descriptive Statistics for High Book-to-Market Stock Mutual Funds

Summary Statistic	Mutual Fund's Raw Return	Characteristic-based Performance Measures		Regression-based Measures		
		Market Adjusted	Size, BM, & Momentum Adjusted	CAPM α	Fama–French 3-factor α	Carhart 4-factor α
Mean	1.30	0.28	0.07	0.32	−0.03	−0.05
Std. dev., %	0.55	0.44	0.17	0.41	0.24	0.26
Std. err. %	0.07	0.06	0.02	0.05	0.03	0.03
Avg. t-stat	1.76	0.75	0.39	0.95	−0.29	−0.39
Minimum, %	−0.29	−0.59	−0.29	−0.30	−0.57	−0.60
Median, %	1.30	0.18	0.06	0.23	−0.05	−0.09
Maximum, %	2.80	1.70	0.69	1.80	0.79	0.86

Panel D: Rejection Frequencies Using One-tailed Tests for High Book-to-Market Stock Mutual Funds

Abnormal Performance	Significance Level											
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
0%	23	59	7.5	18	2.9	13	8	22	0.29	4.9	1.1	7.5
1%	27	66	11	26	9.8	27	11	34	2.3	11	3.4	13
3%	**39**	**74**	**25**	**56**	**40**	**68**	**31**	**66**	**13**	**32**	**13**	**27**
5%	51	80	55	82	84	95	64	84	37	65	32	56
7.5%	64	83	85	94	99	100	89	98	78	90	73	86
10%	74	87	95	97	100	100	99	100	95	99	93	98
15%	83	96	100	100	100	100	100	100	100	100	100	100

100 percent per year, but funds are tracked for up to 10 years. We report results using only the multifactor characteristic-based measure and the four-factor regression-based measure.

The results in Table VI suggest that misspecification is quite substantial at 5- and 10-year horizons using 75 securities. This is seen especially from rejection rates under the null hypothesis. For example, the characteristic-based measure's rejection rate under the null hypothesis is zero, compared

Table V

Distributional Properties, Specification, and Power of 348 Characteristic-based and Regression-based Mutual Fund Performance Measures of Large and Small Market Capitalization Stock Portfolios with the Selection Probability of a Security Equal to Its Market Value Weight within the Large and Small Market Capitalization Stocks

Sample: Each month from January 1966 through December 1994 (348 months), a 75-stock mutual fund portfolio is constructed. Its performance is tracked for a three-year period (months 1 through 36). The portfolio composition is changed 100 percent in months 13 and 25. The 75 stocks are selected without replacement from all above-the-median (large stocks) or below-the-median (small stocks) of the NYSE–AMEX stocks with nonmissing return data in month 1, and this procedure is repeated in months 13 and 25 using stocks available in those months. For each of the 348 portfolios, a time series of monthly returns from month 1 through 36 is constructed. Portfolio returns are equal-weighted at the beginning of months 1, 13, and 25, but they are not rebalanced in the intervening periods. Returns are inclusive of dividends.

Panels A and C: Descriptive statistics: Market-adjusted return is the mutual fund's return in month t minus the return on the CRSP value-weighted portfolio, averaged over months 1 through 36. Size, BM, and momentum-adjusted return is the mutual fund return in month t minus the return on a size, BM, and prior one-year return quintiles-matched companion portfolio, averaged over months 1 through 36. The three regression-based measures, that is, CAPM, Fama–French three-factor, and Carhart four-factor alphas, are the estimated intercepts of the regressions of the mutual funds' excess returns from months 1 through 36 on (i) the CRSP value-weighted portfolio excess return, (ii) the small-minus-large capitalization stock portfolio return, (iii) the high-minus-low book-to-market portfolio return, and (iv) high-minus-low prior one year (momentum) portfolio return. See the text for details on variable descriptions and regressions to estimate the regression-based performance measures.

Standard errors, *S.E.*, of the means across the 348 mutual funds' five performance measures are calculated by applying the Newey–West (1987) correction for serial dependence for up to five lags. The average t-statistic for each performance measure is the average of the 348 individual fund t-statistics. Each fund's t-statistic is its performance measure divided by the standard error for the fund.

Panels B and D: Specification and power: Percentage of 348 samples where the null hypothesis of zero abnormal performance is rejected at the one and five percent significance levels using one-sided tests for various levels of annual portfolio abnormal performance.

Abnormal performance: A given level of annual abnormal performance (e.g., one percent) is induced by adding 1/12 of that amount (e.g., 1/12 percent) per month to the return of each security in each mutual fund.

Panel A: Descriptive Statistics for the Performance Measures for Large Stock Mutual Funds

		Characteristic-based Performance Measures		Regression-based Measures		
Summary Statistic	Mutual Fund's Raw Return	Market Adjusted	Size, BM, & Momentum Adjusted	CAPM α	Fama–French 3-factor α	Carhart 4-factor α
Mean	0.93	−0.04	−0.07	−0.01	0.03	0.06
Std. dev., %	0.61	0.24	0.13	0.24	0.13	0.13
Std. err., %	0.08	0.03	0.01	0.03	0.01	0.01
Avg. t-stat	1.44	−0.26	−0.56	−0.07	0.28	0.43
Minimum, %	−0.85	−0.58	−0.52	0.03	−0.51	−0.24
Median, %	0.95	−0.05	−0.06	1.50	0.04	0.06
Maximum, %	2.60	0.49	0.27	3.00	0.39	0.45

Table V—*Continued*

Panel B: Rejection Frequencies Using One-tailed Tests for large Stock Mutual Funds

Annual Abnormal Performance	Significance Level											
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
0%	19	38	2	8	1	2	3	11	2	8	4	10
1%	22	43	5	20	3	9	8	24	8	25	10	27
3%	**28**	**53**	**26**	**43**	**35**	**57**	**33**	**49**	**50**	**74**	**51**	**76**
5%	35	60	48	63	77	86	53	67	85	92	86	96
7.5%	45	74	72	85	91	97	77	88	96	99	98	100
10%	56	81	93	99	98	100	95	99	99	100	100	100
15%	78	89	100	100	100	100	100	100	100	100	100	100

Panel C: Descriptive Statistics for the Performance Measures for Small Stock Mutual Funds

Summary Statistic	Mutual Fund's Raw Return	Characteristic-based Performance Measures		Regression-based Measures		
		Market Adjusted	Size, BM, & Momentum Adjusted	CAPM α	Fama–French 3-factor α	Carhart 4-factor α
Mean	1.30	0.31	0.06	0.29	0.03	0.05
Std. dev., %	0.86	0.59	0.25	0.56	0.30	0.31
Std. err. %	0.10	0.07	0.01	0.07	0.02	0.02
Avg. t-stat	1.49	0.52	0.19	0.53	0.1	0.16
Minimum, %	−1.40	−0.92	−0.56	−0.92	−0.71	−0.70
Median, %	1.30	0.20	0.04	0.22	0.03	0.04
Maximum, %	3.30	1.90	0.92	1.90	1.10	1.20

Panel D: Rejection Frequencies Using One-tailed Tests for Small Stock Mutual Funds

Annual Abnormal Performance	Significance Level											
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
0%	21	47	10	25	1	2	9	22	1	3	0	5
1%	25	50	13	28	1	3	12	26	2	7	2	8
3%	**32**	**53**	**19**	**35**	**3**	**21**	**18**	**36**	**6**	**20**	**9**	**22**
5%	41	55	27	41	24	46	25	43	18	41	20	39
7.5%	46	60	36	49	51	69	36	52	43	63	40	62
10%	51	64	44	58	72	85	47	61	64	83	62	80
15%	59	80	64	83	92	97	65	83	91	97	89	96

to 24 percent for the regression-based measure. This degree of misspecification makes power comparisons for 10-year horizons quite difficult, and similar results apply to 5-year horizons.

The effect of the number of securities in a fund is seen by using a three-year performance assessment period for different size funds and comparing the measure's ability to detect 3 percent abnormal performance to the results reported in Table II using 75-stock funds. From Table VI, the multifactor characteristic-based measure detects 3 percent abnormal performance in

Table VI

Sensitivity of Specification and Power to Sample Size and Horizon: Distributional Properties, Specification, and Power of 348 Characteristic-based and Regression-based Mutual Fund Performance Measures of Portfolios of Securities with Selection Probability Equal to a Security's Market Value Weight

Sample: Each month from January 1966 through December 1994 or December 1992 or December 1987, a 50- or 125-stock mutual fund portfolio is constructed whose performance is tracked for a 3-, 5-, or 10-year period (months 1 through 120), respectively. The portfolio composition is changed 100 percent in months 13 and every 12 months thereafter. The stocks are selected without replacement from all NYSE–AMEX stocks with nonmissing return data in month 1, and this procedure is repeated in months 13 and every 12 months thereafter using stocks available in those months. The probability that a stock is included in the portfolio constructed in any month is equal to its market capitalization as a fraction of the aggregate market capitalization of all the stocks eligible for inclusion in that month. For each time series of portfolios (348 in case of 3-year horizon, 324 in case of 5-year horizon, and 264 in case of 10-year horizon), a time series of monthly returns from months 1 through 36, 60, or 120 is constructed. Portfolio returns are equal-weighted at the beginning of month 1, and every 12 months thereafter, but they are not rebalanced in the intervening periods. Returns are inclusive of dividends.

Size, BM, and momentum-adjusted return is the mutual fund return in month t minus the return on a size, BM, and prior one-year return quintiles-matched companion portfolio, averaged over months 1 through 36, 60, or 120. The Carhart four-factor regression alphas are the estimated intercepts of the regressions of the mutual funds' excess returns from months 1 through 36, 60, or 120 on (i) the CRSP value-weighted portfolio excess return, (ii) the small-minus-large capitalization stock portfolio return, (iii) the high-minus-low book-to-market portfolio return, and (iv) high-minus-low prior one year (momentum) portfolio return. See the text for details on variable descriptions and regressions to estimate the regression-based performance measures.

Standard errors, *S.E.*, of the means across the 348, 324, or 264 mutual funds' performance measures are calculated by applying the Newey–West (1987) correction for serial dependence for up to five lags.

Specification and power: Percentage of 348, 324, or 264 samples where the null hypothesis of zero abnormal performance is rejected at the one and five percent significance levels using one-sided tests for various levels of annual portfolio abnormal performance. A given level of annual abnormal performance (e.g., one percent) is induced by adding 1/12 of that amount (e.g., 1/12 percent) per month to the return of each security in each mutual fund.

	Size, BM, and Momentum Characteristic-adjusted Performance Measure	Four-factor Regression Alpha	Size, BM, and Momentum Characteristic-adjusted Performance Measure	Four-factor Regression Alpha
	Number of Securities = 50, Horizon = 3 Years		Number of Securities = 75, Horizon = 5 Years	
Mean abnormal return, %	−0.06	0.09	−0.06	0.07
Standard error, %	0.02	0.01	0.01	0.01
Rejection frequency for				
Abnormal return = 0%	2.3	12	0.6	15
Abnormal return = 3%	48	72	67	93
Abnormal return = 10%	97	100	100	100

Table VI—*Continued*

	Size, BM, and Momentum Characteristic-adjusted Performance Measure	Four-factor Regression Alpha	Size, BM, and Momentum Characteristic-adjusted Performance Measure	Four-factor Regression Alpha
	Number of Securities = 125, Horizon = 3 Years		Number of Securities = 75, Horizon = 10 Years	
Mean abnormal return, %	−0.03	0.03	−0.07	0.05
Standard error, %	0.03	0.02	0.01	0.01
Rejection frequency				
Abnormal return = 0%	3.4	6.9	0	24
Abnormal return = 3%	76	83	78	98
Abnormal return = 10%	100	100	100	100

48 percent of the funds consisting of 50 stocks, compared to 59 percent when funds consist of 75 stocks (see Table II) and 76 percent when funds consist of 125 stocks. There is also an increase in the rejection frequency using the four-factor regression-based performance measure.

IV. Event-study Simulations and Power Comparisons

The event-study simulations are directly analogous to those in the previous section. Instead of using only fund returns, the event study evaluates performance by exploiting information on when the fund's trades occur.

A. Simulation Design

We again form samples similar to those in Sections II and III, one starting each month from January 1966 until December 1994. We describe the sample construction assuming we are tracking only the performance of a mutual fund's stock purchases. We have also studied power when both purchases and sales are tracked; power is slightly higher for sales than for purchases only, but to save space, the results are not reported.

For each sample, we select six stocks each month for 36 months. The number of stocks selected each month is normally distributed with a mean of six and standard deviation of two. The random number is rounded to be a nonnegative integer. Stocks are selected from the NYSE–AMEX universe, and a stock's selection probability is its market value weight.

The average of six buys per month results in an average of 72 stock purchases per year. Since a typical mutual fund has 75 stocks and the simulations described in Section II assume 100 percent turnover each year, the 72

buys a year for the event-study simulations are roughly equivalent to 100 percent turnover per year. Since the performance assessment in previous sections uses a 36-month evaluation period, the event study also tracks stock purchases in 36 consecutive months.

For each sample, we aggregate all the buys from the 36 months (on average, 216 buys) and evaluate the resulting equal-weighted portfolio's characteristic-adjusted performance over a 1- to 12-month period following each stock's purchase. We report results using size- and book-to-market characteristic-adjusted returns. Experimentation with different characteristic-adjusted performance measures suggests that power is not sensitive to the choice of the characteristic matching. We test the null hypothesis that the T-month abnormal performance of the equal-weighted portfolio of the event-study stocks is zero using a t-statistic (see the notes to Table VII),[6] where T = 1, 3, 6, and 12 months.

B. Introducing Abnormal Performance

In Section III, we added various levels (e.g., 1 percent, 3 percent, etc.) of annual abnormal performance to the returns of the simulated mutual funds. For the event-study samples, we inject comparable levels of fund abnormal performance using the following procedure. For each level of annual abnormal performance, we add the implied average monthly abnormal return to the returns of stocks recently purchased. We assume that abnormal returns are short-lived and that only the recently purchased stocks earn abnormal returns. The implied monthly abnormal return on the recently purchased stocks depends on the period over which we assume the abnormal return is earned. For example, suppose that abnormal performance occurs for three months (the purchase month and two subsequent months). With turnover of 100 percent per year (i.e., 8.33 percent per month), at any point in time, 25 percent of the portfolio has been purchased in the past three months and is experiencing abnormal performance. If the entire portfolio's annual abnormal return is, for example, 1 percent, this 25 percent of the portfolio has 4 percent abnormal return in the three-month holding period, or 1.33 percent per month; we add 1.33 percent to the return on the stocks for each of the first three months they are held.

C. Simulation Results

Table VII reports rejection frequencies for 0 to 15 percent annual abnormal performance and assumed abnormal performance periods of 1 to 12 months. For comparison, we also report rejection rates from Table II for

[6] Since we sample multiple stocks each month in each event-study simulation and because T can exceed a month, there can be both cross-sectional and temporal overlap in excess returns. This very likely violates the independence assumption underlying the test statistic. We attempt to correct for both cross-sectional dependence and dependence due to the use of overlapping return data using the methods described in Chopra, Lakonishok, and Ritter (1992, p. 251) and Newey and West (1987). These corrections yield similar results and are not reported.

regression-based and characteristic-based tests. The first row of Table VII shows that the event-study tests at all horizons (T = 1 to 12 months) are reasonably well specified. This is not surprising given previous evidence on the performance of event-study tests in Brown and Warner (1980).

C.1. Power Comparisons

From Table VII, the size, BM, and momentum-adjusted characteristic-based approach detects 1 percent abnormal performance 9.8 percent of the time. The four-factor model detects the same abnormal performance 31 percent of the time, but the first row indicates that the test is somewhat misspecified. In contrast to these figures, the event-study tests typically have higher (and generally no lower) power. When T = three months, the event-study-based tests reject the null hypothesis of no abnormal performance 88.8 percent of the time.

As T increases, the fraction of the portfolio that was bought within the past T months grows. Therefore, holding the entire portfolio's average annual abnormal performance constant, the average abnormal return on each stock bought is reduced and the standard deviation increases. This naturally reduces the power advantage of the event study. This is seen from the second row of Table VII. For T = six months and when the portfolio's annual abnormal performance is 1 percent, the event study would detect it 22.4 percent of the time compared to 99.7 percent of the time if T = one month.

C.2. Limitations

If abnormal performance is either extremely long-lived (i.e., four or more quarters) or short-lived (weeks), the one- through six-month results in Table VII overstate the gains from trade-based procedures, although the empirical basis for these alternative scenarios about abnormal performance is unclear. At one-year horizons, similar rejection frequencies using the regression-, characteristic-based, and event-study-based approaches are expected because the assumption about the horizon over which abnormal returns are earned (i.e., one year) is identical across the three approaches. From Table VII, differences in rejection rates at one-year horizons do not seem dramatic, although the event-study procedures sometimes have lower rejection rates than the characteristic based procedures.[7] At the other extreme, the simulations where the abnormal performance period is only one month are somewhat artificial, reflecting an implicit assumption that trades take place on the first of the month. If abnormal performance only occurs between the time of the trade and the end of the month, the event-study approach will miss it. In addition, the one-month abnormal performance period

[7] The event study does not include the momentum factor in constructing companion portfolios (see Lyon, Barber, and Tsai (1999)). Higher rejection rates and more powerful tests would be expected by including a momentum factor, but this would only reinforce the paper's conclusions about power gains using a trade-based approach.

Table VII

Specification and Power of Regression-Based and Trade-Based Event-Study Tests of Performance

Table VII provides percentage of samples where the null hypothesis of zero abnormal performance is rejected at the five percent significance level using one-sided tests for various levels of annual portfolio abnormal performance. The multifactor characteristic-based (i.e., size, BM, and momentum-adjusted) approach and the four-factor regression-based approach use fund returns. The event-study approach only studies returns to securities recently purchased. Abnormal annual fund performance introduced is the same under both approaches, but in the event-study approach, abnormal performance occurs only for the recently purchased securities.

Samples for the multifactor characteristic-adjusted and four-factor-model-based tests: Each month from January 1966 through December 1994, a 75-stock mutual fund portfolio is constructed. Its performance is tracked for a three-year period (months 1 through 36). The portfolio composition is changed 100 percent in months 13 and 25. The 75 stocks are selected without replacement from all NYSE/AMEX stocks with nonmissing return data in month 1, and this procedure is repeated in months 13 and 25 using stocks available in those months. The probability that a stock gets included in the portfolio constructed in month 1, 13, or 25 is equal to its market capitalization as a fraction of the aggregate market capitalization of all the stocks eligible for inclusion in months 1, 13, or 25. For each of the 348 portfolios, a time series of monthly returns from month 1 through 36 is constructed. Portfolio returns are equal-weighted at the beginning of months 1, 13, and 25, but they are not rebalanced in the intervening periods. Returns are inclusive of dividends. The characteristic-based and four-factor-based performance measures are as defined in Table II.

Samples for the event-study tests: Each month from January 1964 through December 1991, a sample of stock purchases is constructed. Each sample consists of an average of six securities purchased each month for 36 consecutive months for a total of 216 purchases on average. The six security purchases per month represents approximately 100 percent annual turnover in a fund consisting of 75 securities. The assumed abnormal return period for each newly purchased security is 1, 3, 6, or 12 months. Abnormal return is defined as the stock return minus the return on a size- and book-to-market matched companion portfolio. The test statistic is

$$t = (1/N) \sum_{t} AR_{iT}/S.E.(AR)$$

where AR_{iT} is security i's T-month abnormal return calculated by compounding the stock's monthly abnormal returns over T months; N is the number of stocks in the event-study portfolio, i varies from 1 to N, and $S.E.(AR)$ is the standard error of the mean of the T-month abnormal returns. The standard error is

$$S.E.(AR) = \left[\sum_{t} (AR_{iT} - (1/N)\Sigma_i AR_{iT})^2 \right]^{1/2} \Big/ (N-1).$$

Abnormal performance: In the multifactor characteristic-based and the regression-based approaches, a given level of annual abnormal performance (e.g., one percent) is induced by adding 1/12 of that amount (e.g., 1/12 percent) per month to the return of each security in each sample. For the event-study samples, for each level of annual abnormal performance, we add the implied average monthly abnormal return on the stocks purchased. The implied monthly abnormal return on the recently purchased stocks depends on the period over which the abnormal return is assumed to be earned. For example, suppose that abnormal performance occurs for three months (the purchase month and two subsequent months). With turnover of 100 percent per year (i.e., 8.33 percent per month), at any point in time, 25 percent of the portfolio has been purchased in the past three months and is experiencing abnormal performance. If the entire portfolio's annual abnormal return is, for example, 1 percent, this 25 percent of the portfolio has 4 percent abnormal return in the three-month holding period, or 1.33 percent per month; we add 1.33 percent to the return on the stocks for each of the first three months they are held.

Table VII—*Continued*

Annual Abnormal Portfolio Return	Rejection Rates Using					
	Multifactor Characteristic-adjusted	Carhart 4-factor Alpha	Event-Study Approach: Assumed Abnormal Return Period for the Fund's Newly Purchased Securities			
			1 Month	3 Months	6 Months	1 Year
None	3.4	13	5.2%	4.9%	6.6%	6.3%
1%	10	31	100	89	22	12
3	59	80	100	100	85	26
5	86	98	100	100	100	48
7.5	98	100	100	100	100	75
10	100	100	100	100	100	93
15	100	100	100	100	100	100

simulations implicitly assume that monthly (Morningstar) holdings data are used; with only quarterly (CDA) data, trades cannot be pinpointed to the month and power will be lower than suggested by the one-month results.

V. Summary and Conclusions

Although there is a large literature on mutual fund performance measures, the ability to detect abnormal performance for an individual fund has received little attention. Our main message is that standard mutual fund performance measures are unreliable and can result in false inferences. It is hard to detect abnormal performance when it exists, particularly for a fund whose style characteristics differ from those of the value-weighted market portfolio.

Power improvements from analyzing a fund's stock trades can be substantial, but this is subject to an important caveat. The improvements occur under the presumption that fund managers' profit opportunities are somewhat short-lived and are concentrated in a few quarters. Although consistent with the empirical evidence (Chen et al. (2000)), this is, nevertheless, a critical presumption. Further, all procedures' power will be a decreasing function of the amount of a fund's liquidity (i.e., non-information-based) trading and its trading costs. Whether substantial abnormal performance for an individual fund can, in fact, be detected using trade-based procedures remains unanswered.

REFERENCES

Andrews, Donald W. K., 1991, Heteroskedasticity and autocorrelation consistent covariance matrix estimation, *Econometrica* 59, 817–858.

Brown, Stephen J., and Jerold B. Warner, 1980, Measuring security price performance, *Journal of Financial Economics* 8, 205–258.

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

Chen, Hsiu-Lang, Narasimhan Jegadeesh, and Russ Wermers, 2000, The value of active mutual fund management: An examination of the stockholdings and trades of fund managers, *Journal of Financial and Quantitative Analysis* 35, 343–368.

Chopra, Navin, Josef Lakonishok, and Jay R. Ritter, 1992, Measuring abnormal performance: Do stocks overreact? *Journal of Financial Economics* 31, 235–268.

Daniel, Kent, Mark Grinblatt, Sheridan Titman, and Russ Wermers, 1997, Measuring mutual fund performance with characteristic-based benchmarks, *Journal of Finance* 52, 1035–1058.

Daniel, Kent, and Sheridan Titman, 1997, Evidence on the characteristics of cross-sectional variation in stock returns, *Journal of Finance* 52, 1–33.

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

Fama, Eugene F., and Kenneth R. French, 1996, Multifactor explanations of asset pricing anomalies, *Journal of Finance* 51, 55–84.

Jensen, Michael C., 1968, The performance of mutual funds in the period 1945–1964, *Journal of Finance* 23, 389–416.

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

Newey, Whitney D., and Kenneth D. West, 1987, A simple, positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55, 703–708.

Newey, Whitney D., and Kenneth D. West, 1994, Automatic lag selection in covariance matrix estimation, *Review of Economic Studies* 61, 631–653.