THE JOURNAL OF FINANCE • VOL. XLII, NO. 2 • JUNE 1987

Mutual Fund Performance Evaluation: A Comparison of Benchmarks and Benchmark Comparisons

BRUCE N. LEHMANN and DAVID M. MODEST*

ABSTRACT

The authors' main goal in this paper is to ascertain whether conventional measures of abnormal mutual fund performance are sensitive to the benchmark chosen to measure normal performance. They employ the standard CAPM benchmarks and a variety of APT benchmarks to investigate this question. They find little similarity between the absolute and relative mutual fund rankings obtained from these alternative benchmarks, which suggests the importance of knowing the appropriate model for risk and return in this context. In addition, the rankings are not insensitive to the method used to construct the APT benchmark. Finally, they find statistically significant measured abnormal performance using all the benchmarks. The economic explanation for this phenomenon appears to be an open question.

THE PROBLEM OF HOW to properly evaluate portfolio performance remains largely unresolved despite broad agreement on an intuitive level that an actively managed portfolio with superior performance should exhibit higher average returns than a passively managed portfolio with the same amount of risk. Unfortunately, two obstacles prevent the easy implementation of this intuitive notion of superior performance to evaluate the track record of managed funds. The first stems from disagreement on the appropriate way to quantify risk and, hence, on what constitutes normal performance. The second problem concerns errors in inference that may arise when portfolio managers can, in fact, outperform the market. In this paper, we provide empirical evidence on the nature of these problems by examining (a) the extent to which alternative benchmarks for normal performance alter the usual performance measures of mutual funds and (b) the efficacy of standard security market line analysis given the shifting composition of managed portfolios.

In order to measure abnormal performance by mutual funds, it is necessary to have a benchmark for normal performance. Numerous investigators have employed the usual proxies for the market portfolio as Capital Asset Pricing Model (CAPM) benchmarks to evaluate the performance of mutual funds.[1] Roll [32,

* Lehmann is from the Department of Economics and the Graduate School of Business, Columbia University, and the National Bureau of Economic Research. Modest is from the Graduate School of Business, Columbia University, and the School of Business Administration, University of California, Berkeley. We would like to express our gratitude to Roy Henriksson for providing us with his mutual fund data base, Karen Bettauer for assistance in updating it, and the Faculty Research Fund of the Columbia Business School and the Institute for Quantitative Research in Finance for their support. We are also grateful for comments from Allan Kleidon, Robert Korajczyk, Cheng-Few Lee, Arthur Warga, seminar participants at Stanford University and the Universities of California, Berkeley and Los Angeles, and an anonymous referee. We alone remain responsible for any remaining errors.

[1] These include Treynor [42], Sharpe [40], and Jensen [20, 21].

33], however, has argued that the use of the CAPM as a benchmark is logically inconsistent since the model assumes that all investors have common beliefs and information and, hence, that any measured abnormal performance can only occur when the market proxy is inefficient. In the absence of any systematic evidence of abnormal performance by mutual funds, Roll's critique would appear to be an academic point. Yet there is plenty of ancillary empirical evidence indicating the mean-variance inefficiency of the usual indices, including the anomalies involving dividend yield, firm size, and price/earnings ratios, which leads one to question the use of the usual CAPM market proxies as performance benchmarks.

The apparent inefficiency of the usual market proxies, coupled with concern over the testability of the CAPM, has led researchers to explore alternative asset-pricing theories. One theory that has stimulated much recent research is the Arbitrage Pricing Theory (APT) developed by Ross [36, 37]. In that seminal work, Ross contended that systematic risk need not be adequately represented by a single common factor such as the return on the market and instead presumed that there are K common sources of covariation (risk) affecting security returns. These K factors constitute another potential benchmark with which to measure normal performance.

Previous research would suggest that alternative risk-adjustment procedures should lead to few substantive differences in performance measures. For example, Stambaugh [41] found that the choice of a market proxy made little difference in CAPM tests. Moreover, Roll [34] found that three market proxies provided nearly identical performance measures for *randomly* selected portfolios and that these risk-adjustment methods produced almost the same rankings as no adjustment at all. Similarly, Copeland and Mayers [9] and Chen, Copeland, and Mayers [7] found that the choice of a performance benchmark did not affect inferences regarding the Value Line enigma. As yet, there is no direct evidence on the sensitivity of commonly used mutual fund performance measures to the choice of benchmarks. The main goal of this paper is to provide such evidence.

Even if there were no question about the appropriate benchmark, it is still difficult to measure managerial performance when mutual fund managers are superior investors. This second difficulty arises from problems associated with measuring portfolio risk when managers act on private information and revise the composition of their portfolios and the risk level of their funds. Mayers and Rice [30] provided sufficient conditions under which the standard security market line analysis is a valid measure of portfolio performance ability. Unfortunately, as shown by Dybvig and Ross [12], their sanguine conclusion rests on the assumption that managers possess no market-timing ability and that any abnormal performance is due to stock selection. This second difficulty occurs because uninformed investors, unable to observe managers' private information signals or actual portfolio choices, may perceive implicit changes in the distribution of an actively managed portfolio's returns due to market timing as needless additions to variance when they are forced to draw inferences solely on the basis of the realized returns of the portfolio.[2]

The next section examines the ability of security market line regression (as in Jensen [20, 21]) and Jensen [22] and Treynor and Mazuy [44] quadratic regres-

[2] See Admati and Ross [2] and Grinblatt and Titman [14] for further details and references.

sions to detect abnormal performance and market-timing ability when there are K sources of systematic risk underlying security returns. In Section II, we discuss the construction of the reference portfolios employed as proxies for the K common factors. Section III describes the mutual fund data set and discusses other data-analytic issues.

The remainder of the paper provides empirical evidence on both the sensitivity of mutual fund performance measures to alternative benchmarks and the adequacy of security market line analysis. Section IV begins by contrasting the inferences yielded by alternative APT benchmarks constructed from differently sized cross-sections of security returns, by alternative estimation methods, and with different numbers of common factors. We then compare standard CAPM performance measures with those of the APT benchmarks. Finally, we use quadratic regressions to examine the problems associated with the shifting composition and risk of managed portfolios. The final section provides concluding remarks.

In brief, our investigation yielded several noteworthy conclusions. We found mutual fund rankings to be very sensitive to the asset-pricing model chosen to measure normal performance. Similarly, alternative APT implementations often suggested substantially different absolute and relative mutual fund rankings. All methods give the appearance of widespread abnormal performance by mutual funds in two of the three sample periods. Unfortunately, the evidence from the quadratic regressions is inconclusive regarding the role of actual or spurious market timing in this finding.

I. On the Detection of Abnormal Performance

In this section, we present the general framework used below for evaluating the stock-selection and market-timing ability of a sample of 130 mutual funds. This basic framework was initially employed by Jensen [20, 21] and has recently been re-examined by Admati et al. [1].

The main assumption underlying the analysis is that the returns on individual securities are generated by a K-factor linear model:

$$\tilde{R}_{it} = E_i + \sum_{k=1}^{K} b_{ik}\tilde{\delta}_{kt} + \tilde{\varepsilon}_{it}, \tag{1}$$

where

$\tilde{R}_{it} \equiv$ return on security i between time $t-1$ and time t for $i = 1, \cdots, N$;
$E_i \equiv$ expected return on security i;
$\tilde{\delta}_{kt} \equiv$ value taken by the kth common factor between time $t-1$ and t (normalized to have zero mean);
$b_{ik} \equiv$ sensitivity of the return of security i to the kth common factor; and
$\tilde{\varepsilon}_{it} \equiv$ the idiosyncratic risk (return) of the ith security between time $t-1$ and time t that is assumed to have a zero mean and finite variance conditional on the realization of the factors.

Note that, under an appropriate choice of *the factors*, equation (1) is consistent, in principle, with a number of alternative asset-pricing models including the traditional CAPM of Sharpe [39] and Lintner [27], the zero-beta model of Black

[3], the intertemporal models of Merton [31], Long [28], Breeden [4], and Cox, Ingersoll, and Ross [10], the skewness model of Kraus and Litzenberger [24], and the arbitrage pricing theory of Ross [36].

We presume that one of these theories accounts for expected returns from the perspective of uninformed investors and that there are a set of K basis or reference portfolios with returns that are perfectly correlated with the realizations of the common factors.[3] In this circumstance, we can compactly write the return-generating process for N individual securities as

$$\tilde{\underline{R}}_t = B\tilde{\underline{R}}_{mt} + \tilde{\underline{\varepsilon}}_t, \tag{2}$$

where $\tilde{\underline{R}}_{mt}$ is a $K \times 1$ vector of returns on the reference portfolios and B is the $N \times K$ matrix of factor sensitivities. Here we let $\tilde{\underline{R}}_t$ and $\tilde{\underline{R}}_{mt}$ denote excess returns above the riskless-rate or zero-beta return where appropriate.

Given the return-generating process for individual securities (2), the return on any mutual fund portfolio can be decomposed as

$$\tilde{R}_{pt} = \sum_{i=1}^{N} \omega_i(\underline{s}_t)\tilde{R}_{it} = \sum_{i=1}^{N} [\omega_i(\underline{s}_t)\underline{b}_i'\tilde{\underline{R}}_{mt} + \omega_i(\underline{s}_t)\tilde{\varepsilon}_{it}], \tag{3}$$

where $\omega_i(\underline{s}_t)$ is the weight of security i in the portfolio at date t, $\underline{b}_i'$ is a $1 \times K$ vector consisting of the ith row of B, and $\underline{s}_t$ is a vector of signals received by the mutual fund manager for predicting $\tilde{\underline{R}}_{mt}$ and $\tilde{\underline{\varepsilon}}_t$. It will prove useful to rewrite (3) as

$$\tilde{R}_{pt} = \underline{\beta}_{pt}'\tilde{\underline{R}}_{mt} + \tilde{\varepsilon}_{pt}, \tag{4}$$

where

$$\underline{\beta}_{pt}' = \sum_{i=1}^{N} \omega_i(\underline{s}_t)\underline{b}_i' = \overline{\underline{\beta}}_p' + \underline{x}(\underline{s}_t)',$$

$$\tilde{\varepsilon}_{pt} = \sum_{i=1}^{N} \omega_i(\underline{s}_t)\tilde{\varepsilon}_{it}. \tag{5}$$

In (4), market-timing attempts are reflected in movements in the fund sensitivities, $\underline{\beta}_{pt}$, while stock-selection ability is embedded in the residual disturbance, a reasonable definition given the assumptions underlying (2).[4] Hence, the elements of $\overline{\underline{\beta}}_p$ are the target or average sensitivities of the fund to the K common factors, and $\underline{x}(\underline{s}_t)$ are the time-t deviations from $\overline{\underline{\beta}}_p$ selected by the manager in attempts to time factor movements (which have zero mean by definition). Similarly, if the manager possesses stock-selection ability, $\tilde{\varepsilon}_{pt}$ will not have a zero population mean.

The usual security market line procedure involves regressing the excess returns of the mutual fund on the excess returns of the reference portfolios. Letting $E^*[X \mid Y]$ denote the minimum-variance linear estimator of X given Y (i.e., the

[3] If the underlying asset-pricing theory is the static CAPM, the two reference portfolios are the true market portfolio and its orthogonal partner or the riskless asset, while the intertemporal CAPM adds the relevant state-variable hedge portfolios. The three-moment CAPM would involve the true market portfolio, the hedge portfolio with maximal correlation with squared excess returns on the market portfolio, and a zero-beta portfolio or riskless asset. In an APT framework, we assume that exact factor pricing obtains.

[4] This is a version of the portfolio approach discussed in Admati et al. [1].

regression function), the regression of $\tilde{R}_{pt}$ on $\underline{\tilde{R}}_{mt}$ results in

$$\mathrm{E}^*[\tilde{R}_{pt} \mid \underline{\tilde{R}}_{mt}] = \hat{\alpha}_p + \underline{\hat{\beta}}_p' \underline{\tilde{R}}_{mt}, \tag{6}$$

where

$$\begin{aligned}
\hat{\alpha}_p &= [\bar{\varepsilon}_p - \mathrm{Cov}\{\underline{x}_t' \underline{\tilde{R}}_{mt}, \underline{\tilde{R}}_{mt}\}' \textstyle\sum_m^{-1} \underline{\bar{R}}_m + \mathrm{E}\{\underline{x}_t' \underline{\tilde{R}}_{mt}\}], \\
\underline{\hat{\beta}}_p &= [\underline{\bar{\beta}}_p + \textstyle\sum_m^{-1} \mathrm{Cov}\{\underline{x}_t' \underline{\tilde{R}}_{mt}, \underline{\tilde{R}}_{mt}\}], \\
\bar{\varepsilon}_p &= \textstyle\sum_{i=1}^{N} \mathrm{Cov}\{\omega_i(\underline{s}_t), \tilde{\varepsilon}_{it}\}, \\
\textstyle\sum_m &= \mathrm{E}[\{\underline{\tilde{R}}_{mt} - \underline{\bar{R}}_m\}\{\underline{\tilde{R}}_{mt} - \underline{\bar{R}}_m\}'], \\
\underline{\bar{R}}_m &= \mathrm{E}[\underline{\tilde{R}}_{mt}],
\end{aligned} \tag{7}$$

$\underline{x}_t'$ is used as shorthand notation for $\underline{x}(\underline{s}_t)'$, and $\mathrm{Cov}\{\underline{x}_t' \underline{\tilde{R}}_{mt}, \underline{\tilde{R}}_{mt}\}$ is a $K \times 1$ vector of the covariances between $\underline{x}_t' \underline{\tilde{R}}_{mt}$ and the K elements of $\underline{\tilde{R}}_{mt}$. The coefficient $\hat{\alpha}_p$ is the usual Jensen performance measure.

In the absence of an ability to pick stocks (i.e., $\bar{\varepsilon}_p = 0$) and to time the market (i.e., $\mathrm{E}\{\underline{x}_t' \underline{\tilde{R}}_{mt}\} = \mathrm{Cov}\{\underline{x}_t' \underline{\tilde{R}}_{mt}, \tilde{R}_{jt}\} = 0$ for all $j = 1, \cdots, K$), the regression equation (6) will indicate no abnormal performance since, in this instance,

$$\mathrm{E}^*[\tilde{R}_{pt} \mid \underline{\tilde{R}}_{mt}] = \underline{\bar{\beta}}_p' \underline{\tilde{R}}_{mt}. \tag{8}$$

If the mutual fund manager possesses stock-selection ability but no market-timing ability, the regression *will* indicate superior performance since

$$\mathrm{E}^*[\tilde{R}_{pt} \mid \underline{\tilde{R}}_{mt}] = \bar{\varepsilon}_p + \underline{\bar{\beta}}_p' \underline{\tilde{R}}_{mt}, \tag{9}$$

where $\bar{\varepsilon}_p > 0$ under mild restrictions.[5] Finally, if portfolio managers possess market-timing ability as well, the Jensen measure may be positive or negative depending on the terms in brackets on the first line of (7). Hence, the Jensen measure will (in the population) indicate abnormal performance,[6] but it cannot be used to evaluate managers since $\hat{\alpha}_p$ could be positive even if the manager were an unsuccessful stock picker and a perverse market timer and conversely could be negative if the manager were both a successful stock picker and a successful market timer. Yet there is a hint in (6) and (7) of the possibility of detecting the presence of market-timing ability due to the terms involving $\mathrm{Cov}\{\underline{x}_t' \underline{\tilde{R}}_{mt}, \underline{\tilde{R}}_{mt}\}$ and $\mathrm{E}\{\underline{x}_t' \underline{\tilde{R}}_{mt}\}$. These terms suggest that perhaps a quadratic regression could detect market-timing ability when individual returns unconditionally follow a linear factor structure as in (2).

[5] This is merely a restatement of the Mayers and Rice [30] proposition, as simplified and extended by Dybvig and Ross [12], that the Jensen measure will correctly indicate superior performance when managers possess security-selection ability but are unable to time the market. It has also been recently (and independently) noted in a multifactor setting by Connor and Korajczyk [8], who ignore any potential market timing by managers. While $\hat{\alpha}_p > 0$ will indicate superior performance in this circumstance, the funds cannot be relatively ranked on the basis of the alphas without further assumptions about investor preferences. For example, constant-absolute-risk-aversion investors will rank funds on the basis of their Treynor-Black [43] appraisal ratios (the alphas divided by the residual standard deviations of the funds) given normally distributed returns.

[6] Grinblatt and Titman [14] argue that the Jensen measure will be reliably non-negative under reasonable assumptions.

The quadratic-regression framework originally was examined by Treynor and Mazuy [44]. Their basic idea was quite simple; market timers should make money when the market rises or falls dramatically, that is, when the squared return on the market is large. Its possibilities as a framework for separating market-timing ability from stock-selection ability were studied by Jensen [22], an analysis that was corrected and extended in Admati et al. [1].

Consider (for the sake of notational simplicity) the one-factor version of (4):

$$\tilde{R}_{pt} = \beta_{pt}\tilde{R}_{mt} + \tilde{\varepsilon}_{pt}, \tag{10}$$

and the associated quadratic regression:[7]

$$\mathrm{E}^*[\tilde{R}_{pt} \mid \tilde{R}_{mt}, \tilde{R}^2_{mt}] = \alpha^*_p + b^*_{1p}\tilde{R}_{mt} + b^*_{2p}\tilde{R}^2_{mt}. \tag{11}$$

The regression slope coefficients are given by

$$\begin{aligned}
\begin{bmatrix} b^*_{1p} \\ b^*_{2p} \end{bmatrix} &= \left(\mathrm{Var}\begin{bmatrix} \tilde{R}_{mt} \\ \tilde{R}^2_{mt} \end{bmatrix}\right)^{-1} \mathrm{Cov}\left[\tilde{R}_{pt}, \begin{pmatrix} \tilde{R}_{mt} \\ \tilde{R}^2_{mt} \end{pmatrix}\right] \\
&= \begin{pmatrix} \sigma^2_m & \sigma_{3m} \\ \sigma_{3m} & \sigma_{4m} \end{pmatrix}^{-1} \begin{bmatrix} \bar{\beta}_p\sigma^2_m + \mathrm{Cov}(x_t, \tilde{R}^2_{mt}) \\ \bar{\beta}_p\sigma_{3m} + \mathrm{Cov}(x_t, \tilde{R}^3_{mt}) \end{bmatrix} \\
&= \begin{bmatrix} \bar{\beta}_p \\ 0 \end{bmatrix} + \frac{1}{\sigma^2_m\sigma_{4m} - \sigma^2_{3m}} \begin{bmatrix} \sigma_{4m} & -\sigma_{3m} \\ -\sigma_{3m} & \sigma^2_m \end{bmatrix} \begin{bmatrix} \mathrm{Cov}(x_t, \tilde{R}^2_{mt}) \\ \mathrm{Cov}(x_t, \tilde{R}^3_{mt}) \end{bmatrix} \\
&\equiv \begin{bmatrix} \bar{\beta}_p \\ 0 \end{bmatrix} + \begin{bmatrix} \gamma_{1p} \\ \gamma_{2p} \end{bmatrix},
\end{aligned} \tag{12}$$

where σ_{3m} and σ_{4m} are the skewness and kurtosis of $\tilde{R}_{mt}$, and $\bar{\beta}_p$ is the target β of the mutual fund. Similarly, the intercept of the quadratic regression is

$$\begin{aligned}
\alpha^*_p &= \bar{\varepsilon}_p + \bar{\beta}_p\bar{R}_m + \mathrm{Cov}(x_t, \tilde{R}_{mt}) - b^*_{1p}\bar{R}_m - b^*_{2p}\bar{R}^2_m \\
&= \bar{\varepsilon}_p + \mathrm{Cov}(x_t, \tilde{R}_{mt}) - \gamma_{1p}\bar{R}_m - \gamma_{2p}\bar{R}^2_m.
\end{aligned} \tag{13}$$

In the absence of market-timing ability, $\mathrm{Cov}\{x_t, \tilde{R}^2_{mt}\}$ and $\mathrm{Cov}\{x_t, \tilde{R}^3_{mt}\}$ are both zero so that the coefficient on $\tilde{R}_{mt}$ will be the target beta of the fund and the coefficient on $\tilde{R}^2_{mt}$ will be zero. However, if the manager has timing ability, then nonzero values of x_t will, in general, be correlated with $\tilde{R}^2_{mt}$ and $\tilde{R}^3_{mt}$, in which case b^*_{2p} will be nonzero in the population,[8] indicating the presence of market-timing ability.[9]

In general, without further restriction on distributions and preferences, it is not possible to measure the magnitudes of market-timing and security-selection

[7] In the multifactor case, all second moments would be included in the regression—both the variances of and covariances among the basis portfolio returns—similar in spirit to the heteroscedasticity tests of White [45].

[8] To be precise, this will apply as long as the market-timing information, the preferences of the manager, and the distribution of $\tilde{R}_{mt}$ are such that the appropriate combination of $\mathrm{Cov}\{x_t, \tilde{R}^2_{mt}\}$, $\mathrm{Cov}\{x_t, \tilde{R}^3_{mt}\}$, σ_{3m}, and σ_{4m} in (12) are nonzero.

[9] This also presumes that managers do not revise their portfolios more frequently than portfolio returns are observed and that all systematic changes in the funds' betas are the result of conscious decisions by management and not the result of inadvertent shifts, such as Jagannathan and Korajczyk [19] suggest might occur from changes in firms' debt/equity ratios.

abilities. As is apparent from equation (12), if there is no co-skewness between the fluctuations in the fund beta and the return on the factor (i.e., $\text{Cov}\{x_t, \tilde{R}_{mt}^2\}$ is zero), then it will be possible to estimate the target beta of the fund and $\text{Cov}\{x_t, \tilde{R}_{mt}^3\}$, but it will not be possible to separate the two sources of abnormal performance in (13).[10] Still, the potential capability to detect the presence of market-timing ability with the simple quadratic-regression procedure represents a promising advance that we will begin to examine below.

II. The Construction of Reference Portfolios

The analysis in Section I presupposes the existence of reference portfolios that are perfectly correlated with the common factors underlying security returns. In this section, we discuss the construction of reference portfolios designed to mimic the realization of the common factors. Two broad classes of portfolios are used in the empirical tests: those associated with the CAPM and those associated with the APT. In implementing the CAPM to obtain risk-adjusted excess returns, we use the standard market proxies (the CRSP equally weighted and value-weighted indices of NYSE stocks taken from the CRSP monthly index file) and ignore the unobservability of the true market portfolio emphasized by Roll [32]. The construction of reference portfolios for the APT benchmarks, however, is not so straightforward.

The construction of APT reference portfolios is a two-step procedure. First, the sensitivities to the common factors are estimated for a collection of individual securities. In the second step, these estimated factor loadings are used to construct basis portfolios to mimic the common factors. Unfortunately, this description overlooks a bewildering number of complications. There are two versions of the APT (the riskless-rate and zero-beta models), a variety of methods for estimating the factor sensitivities of individual securities, and several possible portfolio-formation procedures that use the estimated factor loadings and idiosyncratic variances. In addition, there are important data-analytic choices including the number of securities to include in the first-stage estimation as well as the periodicity of data appropriate for estimating the factor loadings.

The determination of what combination of these choices is best along with the comparative merits of the CAPM or APT lies well beyond the scope of this paper. Instead, we examine the more modest question of whether different methods for constructing reference portfolios lead to different conclusions about the relative performance of mutual funds. Note that our focus is on this question and that we make no effort to distinguish *true* benchmark error (i.e., the errors implied by alternative theoretical models) from that induced by measurement error in the construction of benchmarks.[11] Accordingly, we consider many permutations

[10] This can be accomplished by placing sufficient structure on the problem so that measurement of $\text{Cov}\{x_t, \tilde{R}_{mt}^3\}$ leads to estimation of $\text{Cov}\{x_t, \tilde{R}_{mt}\}$, which permits the estimation of $\tilde{\varepsilon}_p$ using equation (13). For example, Admati et al. [1] assume the joint normality of $\tilde{R}_{mt}$, $\underset{\sim}{g}_t$, and $\tilde{\varepsilon}_{pt}$ and the linearity of $\underset{\sim}{x}_t$ in $\underset{\sim}{g}_t$.

[11] For example, the differences in performance measures yielded by alternative five-factor APT benchmarks can be attributed only to measurement error (when a five-factor model is the appropriate economic model) since we are unable to observe or identify *a priori* the "true" factors. A similar problem arises in interpreting the results using the CAPM benchmarks due to the unobservability of the *true* market portfolio and the consequent measurement error in any proxy.

and combinations of these choices below. If different methods yield similar conclusions, researchers would perhaps be best advised to choose the least computationally intensive procedures. However, if the methods yield different performance measures, then presumably care must be taken in choosing a benchmark for normal preference, and further research is needed to determine which procedures provide better estimates of the common factors underlying security returns. The remainder of this section is devoted to a brief description of the elements of the different basis-portfolio-construction strategies examined below. In Section III, we discuss some of the data-analytic alternatives involved in basis-portfolio construction.

A. *Riskless-Rate and Zero-Beta Versions of the APT*

Like the CAPM, the APT has both a riskless-rate and a zero-beta formulation. The riskless-rate version is appropriate when it is possible to form a positive investment portfolio of risky assets with a return variance that goes to zero as the number of assets that satisfy the factor structure (1) grows large. The zero-beta formulation arises when it is not possible to form a limiting riskless portfolio of risky assets. This will occur if and only if there exists a factor to which all securities have equal sensitivity under an appropriate transformation of the factor space.

If the riskless-rate formulation of the APT is true and there is exact factor pricing, then security returns satisfy

$$\underset{\sim}{\tilde{R}}_t - \underset{\sim}{\iota} R_f = B(\underset{\sim}{\tilde{R}}_{mt} - \underset{\sim}{\iota} R_f) + \underset{\sim}{\tilde{\varepsilon}}_t, \tag{14}$$

where R_f is the return on the limiting riskless portfolio of risky assets and $\underset{\sim}{\iota}$ is a vector of ones. Similarly, the exact factor-pricing version of the zero-beta APT yields

$$\underset{\sim}{\tilde{R}}_t = B\underset{\sim}{\tilde{R}}_{mt} + \underset{\sim}{\tilde{\varepsilon}}_t \tag{15}$$

since the zero-beta portfolio corresponds to one of the common factors underlying security returns. In the empirical tests, we report on both the riskless-rate and zero-beta formulations of the APT to indicate any sensitivity of mutual fund performance measures to alternative versions of the APT. A more complete discussion of the two models and their empirical validity is contained in Lehmann and Modest [26].

B. *Estimation Methods*

Four different methods for estimating the factor loadings and idiosyncratic variances underlying the APT are briefly described in this section: two maximum-likelihood factor-analysis procedures, a principal-components procedure, and an instrumental-variables estimator. In theory, as the number of securities grows large, all four methods provide consistent estimates of the factors and, as the number of observations grows large, consistent estimates of the factor loadings and idiosyncratic variances as well. However, it is obviously of greater than academic interest to know whether they provide substantively different answers in actual practice.

The principal assumption of the APT is that security returns are generated by a K-factor linear structure. Given the structure in (1) in conjunction with the assumptions $E[\tilde{\underline{\varepsilon}}_t \mid \tilde{\underline{\delta}}_t] = \underline{0}$ and $E[\tilde{\underline{\varepsilon}}_t \tilde{\underline{\varepsilon}}_t' \mid \tilde{\underline{\delta}}_t] = \Omega$ (Ω a positive-definite symmetric matrix) and the normalization of the factors $\tilde{\underline{\delta}}_t$ such that $E[\tilde{\underline{\delta}}_t] = \underline{0}$ and $E[\tilde{\underline{\delta}}_t \tilde{\underline{\delta}}_t'] = I$, the covariance matrix of security returns, Σ, can be written as $BB' + \Omega$. Theoretically, the APT requires only that the off-diagonal elements of Ω are sufficiently sparse so that the residual risks are diversifiable (in the limit) and, hence, that security returns satisfy an approximate factor structure.[12]

Given this factor structure, Chamberlain and Rothschild [5] have shown that consistent estimates of the factors can be obtained from the eigenvectors associated with the K largest eigenvalues of the matrix $\Upsilon^{-1}\Sigma$, where Υ is *any* arbitrary positive-definite matrix with eigenvalues bounded away from zero and infinity. Standard maximum-likelihood factor analysis (under the normality assumption) is numerically equivalent to calculating the largest K eigenvectors of the matrix $\Upsilon^{-1}\Sigma$, where Υ is set equal to $\hat{D}$, which is a diagonal matrix with the estimated residual variances.[13] This is the procedure that we refer to as unrestricted maximum-likelihood factor analysis.[14] The procedure we refer to as restricted maximum-likelihood factor analysis involves maximizing the log-likelihood function of Σ conditional on the factor structure of returns and the requirement that mean security returns are spanned by their factor loadings and the factor risk premia.[15]

The principal-components method, on the other hand, involves the corresponding eigenvectors of the matrix $\Upsilon^{-1}\Sigma$, with Υ set equal to an identity matrix. We employed the singular-value decomposition algorithm included in the NAG Subroutine Library to obtain the required eigenvalues and eigenvectors. Each column of the K eigenvectors was multiplied by the square root of the corresponding eigenvalue in order to scale the factors to have unit variances. Estimates of the idiosyncratic variances were then obtained from the diagonal elements of the matrix $\Sigma - BB'$.

Another relatively inexpensive alternative to maximum-likelihood factor analysis is instrumental-variables factor analysis along the lines suggested by Madansky [29] and Hagglund [15]. The basic idea of these instrumental-variables estimators is to substitute consistent estimates of the factors for the factors

[12] The formal requirement is that, as $N \to \infty$, the eigenvalues of Ω remain bounded.

[13] Formally, this involves maximizing the likelihood function: $\mathscr{L}(\sum \mid S) = \frac{-NT}{2} \ln(2\pi) - \frac{T}{2} \ln|\sum| - \frac{1}{2} \sum_{t=1}^{T} (\tilde{\underline{R}}_t - \bar{\underline{R}})' \sum^{-1} (\tilde{\underline{R}}_t - \bar{\underline{R}})$. Connor and Korajczyk [8] prove that the consistency of the factor estimates also holds when the population covariance matrix $\sum$ is replaced by the sample covariance matrix with Υ set equal to the identity matrix. An analogous proof can be used to show the consistency of maximum-likelihood factor analysis with Υ set equal to the diagonal matrix with elements bounded away from zero and infinity.

[14] While this is a conceptually simple exercise, it is computationally infeasible to obtain these estimates by iteratively solving the first-order conditions when the number of securities being analyzed is substantial. We therefore employed a significantly cheaper alternative: the EM algorithm of Dempster, Laird, and Rubin [11] as applied to factor analysis by Rubin and Thayer [38].

[15] This involves maximizing the function: $\mathscr{L}(\sum \mid S) = \frac{-NT}{2} \ln(2\pi) - \frac{T}{2} \ln|\sum| - \frac{1}{2} \sum_{t=1}^{T} (\tilde{\underline{R}}_t - [\underline{\iota}\lambda_0 + B\underline{\lambda}])' \sum^{-1} (\tilde{\underline{R}}_t - [\underline{\iota}\lambda_0 + B\underline{\lambda}])$, where $\underline{\iota}$ is a vector of ones, λ_0 is the riskless rate if the riskless-rate version of the APT is appropriate and zero otherwise, and $\underline{\lambda}$ is the vector of factor risk premia.

themselves in equation (1) and then to estimate the factor loadings B by ordinary least squares (OLS). The formulation employed here involves normalizing the factor space so that the factor-loading matrix on the first K security returns is an identity matrix and using all but the ith security return as instruments for the first K security returns to provide consistent estimates of the factor loadings on the ith security returns. The procedure we employ closely follows Hagglund [15] and is described in detail in Lehmann and Modest [25].

C. Portfolio-Formation Procedures

The most commonly used procedure for constructing portfolios to mimic the common factors involves treating the returns on the N securities as the dependent variable, the factor loadings as explanatory variables, and the factor realizations as parameters to be estimated from cross-sectional regressions along the lines of Fama and MacBeth [13].[16] For the zero-beta model (15), the generalized least-squares (GLS) version of this estimator (i.e., portfolio weights given by $(B'\Omega^{-1}B)^{-1}B'\Omega^{-1}$) provides the minimum-variance linear unbiased estimate of the factors given the population factor loadings (B) and idiosyncratic covariance matrix (Ω). Similarly, the factors associated with the riskless-rate version can be estimated by replacing the factor-loading matrix B with the augmented factor-loading matrix B^* ($B^* = [\iota : B]$). In practice, all APT applications that we are aware of replace Ω with a diagonal matrix D consisting of the idiosyncratic variances, thereby ignoring the off-diagonal elements.

Since the population values of the factor loadings B and the idiosyncratic variances D are unobservable, most investigators employ the corresponding parameter estimates in the construction of the weighted least-squares (WLS) basis portfolios. This presents empirical problems for the zero-beta model (15) and the riskless-rate model (14) estimated relative to a measured riskless rate because the sample estimates of the WLS portfolio weights $(B'D^{-1}B)^{-1}B'D^{-1}$ are extremely poorly diversified, with weights typically in excess of one hundred percent in absolute value using the conventional normalization of the common factors, which suggests that the WLS portfolios will not be rid of idiosyncratic risk. This occurs because the usual normalization requires the factors to have unit variance (compared with a typical daily-return variance on the order of 10^{-5}) and because of sampling error in the parameter estimates.

In what follows, we employ a method we refer to as the *minimum-idiosyncratic-risk procedure* as an alternative to the WLS portfolio-formation procedure. The WLS procedure for mimicking the kth common factor produces the minimum-idiosyncratic-variance portfolio with a sample loading of one on the kth common factor and loadings of zero on the other factors. The corresponding minimum-idiosyncratic-risk portfolio is the minimum-idiosyncratic-variance portfolio that costs a dollar and has sample loadings of zero on the other common factors.[17] The difference between the two procedures lies in the requirement that the WLS

[16] This section is an abbreviated version of Section III of Lehmann and Modest [25], which contains a detailed discussion of these issues as well as documentation of the claims made in the text.

[17] This estimator can be computed as follows. Let $B = (b_1 b_2 \cdots b_k)$ and suppose that we are interested in mimicking the jth factor. The minimum-idiosyncratic-risk estimator is $D^{-1}B^*[B^{*\prime}D^{-1}B^*]^{-1}e_j$, where $B^* = (b_1 b_2 \cdots \iota \cdots b_k)$ and ι is a vector of ones in the jth column.

portfolios have a sample loading of unity on the factor being mimicked prior to normalization in order to cost a dollar, while the minimum-idiosyncratic-risk portfolios must simply cost a dollar. The minimum-idiosyncratic-risk procedure always produces well-diversified portfolios since it is not sensitive to the normalization of the common factors. The evidence presented in Lehmann and Modest [25] suggests that the minimum-idiosyncratic-risk procedure performed at least as well as (and usually better than) its competitors, so we employ it in this investigation.

When implementing the riskless-rate version, excess returns are constructed relative to those of the orthogonal minimum-idiosyncratic-variance portfolio that costs a dollar and has sample loadings of zero on all factors. The returns on this orthogonal portfolio are identical to the intercepts obtained from the cross-sectional regressions of individual security returns for each month on a constant and the factor loadings (i.e., with sample estimates of the portfolio weights $(B^{*\prime}D^{-1}B^{*})^{-1}B^{*\prime}D^{-1}$). Hence, the results obtained from the minimum-idiosyncratic-risk procedure are identical to those from the conventional WLS Fama-MacBeth strategy for the riskless-rate model.[18]

III. The Data

Our mutual fund data base consists of monthly returns on 130 mutual funds over the fifteen-year period January 1968 through December 1982. We are grateful to Roy Henriksson for graciously supplying us with the bulk of these data. The monthly returns are calculated from the end-of-month bid prices and monthly dividends obtained from Standard and Poor's *Over-the-Counter Daily Stock Price Records*, Weisenberger's *Investment Companies* annual compendium, and Moody's *Annual Dividend Record*. The *Over-the-Counter Daily Stock Price Record* omits a significant fraction of the dividends paid, and the other two sources are required to obtain accurate dividend information. The sample was chosen to include a variety of funds with differing risk postures but omitted all municipal bond and option funds. It contains only funds that survived from January 1968 to June 1980, which raises the possibility of survivorship bias, although the results obtained after June 1980 do not suggest that this is a serious problem. Due to our concern that the risk levels of the funds were not constant over the fifteen-year period, we restricted our attention to examining the behavior of the funds over three five-year subperiods: January 1968 through December 1972, January 1973 through December 1977, and January 1978 through December 1982.

One choice facing researchers in constructing APT basis portfolios is the appropriate frequency of observation for estimating the factor models of security returns. Following Roll and Ross [35], we opted for the putative benefits of a large sample and used daily data to estimate the factor models. The primary advantage of daily data is the potential increase in precision of the estimated variances and covariances (the inputs to the factor-analysis model) that comes with sampling the data more often. The main disadvantages are the persistent

[18] We also constructed excess mutual fund and basis-portfolio returns relative to the one-month Treasury bill rate, but this did not change the qualitative conclusions reached in the text.

incidence of nontrading and thin trading, which bias the estimates of second-order moments, and the biases in mean returns associated with bid-ask spreads. Portfolio weights constructed from daily data based on the minimum-idiosyncratic-risk procedure outlined above were then multiplied by monthly security returns to construct monthly returns on our basis portfolios.[19]

Another decision involves the number of securities to be employed in the analysis. Large cross-sections lead to more precise estimates of the factors in the absence of measurement error in the factor-model estimates. Of course, the question of how large a "large" cross-section is is an empirical one, so we report below on the effects of different numbers of securities (between thirty and 750) on mutual fund rankings. Computational considerations required the analysis of no more than 750 securities simultaneously.[20] The CRSP daily file contains 1359, 1346, and 1281 continuously listed securities with no missing observations during the three five-year periods covered by our mutual fund data. We confined our attention to these firms in order to have the same number of observations for each security and ignored any potential selection bias associated with this choice. The CRSP daily file lists securities in alphabetical order by most recent name. To guard against any biases induced by the natural progression of letters (e.g., General Electric, General Motors, $\cdots$), we randomly reordered the firms. The number of daily observations in these samples was 1234, 1263, and 1264, respectively. The usual sample covariance matrix of these security returns provided the basic input to our subsequent analysis.

IV. Empirical Results

In this section, we provide the promised comparison of benchmarks using the monthly-returns data base described above. Tables I through VI compare performance measures yielded by alternative APT benchmarks. The tables summarize and contrast the behavior of the intercepts from simple Jensen-style ordinary least-squares regressions of mutual fund returns on the APT basis portfolios as given by equation (6). The basic questions here are whether different basis-portfolio-construction procedures lead to different conclusions and whether performance measures are sensitive to the number of factors assumed to underlie security returns. Tables VII and VIII provide the corresponding information comparing APT and CAPM benchmarks in order to highlight the contrasts across asset-pricing models. Table IX presents evidence on the intertemporal relations among the Jensen measures produced by each method across sample periods. Finally, Tables X and XI summarize the information from quadratic regressions along the lines of equation (11) using both APT and CAPM benchmarks in order to shed some light on one possible cause of the anomalous behavior of the Jensen measures.

The first nine tables provide evidence on the similarities and differences in the intercepts across benchmarks. The first eight tables come in pairs, i.e., four sets

[19] We discarded the alternative daily-rebalancing strategy, which would involve multiplying the portfolio weights by daily security returns and then aggregating these daily returns to obtain monthly portfolio returns, due to concern over bid-ask spread bias.

[20] We have carried out runs using as many as one thousand securities. However, the larger number of securities yielded a minimal improvement over the performance of reference portfolios based on 750 securities and proved to be disproportionately expensive in terms of computational time.

Table I

Statistics of Intercepts across Estimation Methods of the APT[a]

Sample Period	Statistic	Maximum Likelihood		Restricted Maximum Likelihood		Instrumental Variables		Principal Components	
		Zero Beta	Riskless Rate	Zero Beta	Riskless Rate	Zero Beta	Riskless Rate	Zero Beta	Riskless Rate
January 1968 through December 1972	Mean intercept	−3.22	−4.85	−3.53	−5.22	−4.77	−5.13	−1.47	−3.47
		(4.79)	(3.86)	(4.98)	(3.93)	(5.42)	(4.08)	(4.04)	(3.61)
	Average absolute intercept	4.21	5.27	4.48	5.62	5.54	5.60	3.17	4.16
	Average absolute *t*-statistic	1.34	2.05	1.38	2.15	1.67	1.98	1.11	1.53
		(1.04)	(1.16)	(1.06)	(1.19)	(1.13)	(1.05)	(.85)	(.96)
	Average absolute *t*-adjusted	1.51	2.30	1.56	2.43	1.79	2.20	1.25	1.71
		(1.13)	(1.33)	(1.16)	(1.36)	(1.17)	(1.22)	(.95)	(1.09)
January 1973 through December 1977	Mean intercept	−2.67	−5.45	−2.58	−5.35	−4.05	−6.98	−3.97	−6.52
		(3.76)	(3.60)	(3.73)	(3.57)	(4.20)	(3.89)	(4.42)	(4.12)
	Average absolute intercept	3.58	5.80	3.49	5.70	4.75	7.21	4.88	6.84
	Average absolute *t*-statistic	1.40	2.36	1.36	2.29	1.71	2.72	1.76	2.37
		(1.02)	(1.17)	(.99)	(1.13)	(1.07)	(1.19)	(1.10)	(1.19)
	Average absolute *t*-adjusted	1.59	2.68	1.54	2.60	1.95	3.04	1.99	2.63
		(1.17)	(1.35)	(1.15)	(1.31)	(1.24)	(1.38)	(1.24)	(1.33)
January 1978 through December 1982	Mean intercept	−3.25	−3.85	−3.36	−4.14	−1.48	−1.96	−.57	−1.78
		(3.35)	(3.30)	(3.46)	(3.53)	(3.17)	(3.14)	(3.20)	(2.83)
	Average absolute intercept	3.67	4.12	3.77	4.39	2.68	2.81	2.41	2.64
	Average absolute *t*-statistic	1.26	1.45	1.26	1.50	0.92	0.99	0.83	0.91
		(.87)	(.92)	(.88)	(.95)	(.73)	(.77)	(.67)	(.69)
	Average absolute *t*-adjusted	1.51	1.72	1.51	1.79	1.07	1.16	0.94	1.04
		(1.05)	(1.09)	(1.07)	(1.12)	(.86)	(.95)	(.76)	(.78)

[a] Intercepts are measured in percent per annum. Standard deviations are in parentheses. Number of factors: 10; number of funds: 130; number of securities used in estimation: 750.

Table II

Comparisons of Treynor-Black Appraisal Ratios across Estimation Methods of the APT (Riskless-Rate Version of the APT)[a]

Sample Period		Simple Correlations			Fractions of Funds with Significant Abnormal Performance at the Five Percent Level for Both/Only One Benchmark(s)		
		Restricted Maximum Likelihood	Instrumental Variables	Principal Components	Restricted Maximum Likelihood	Instrumental Variables	Principal Components
January 1968 through December 1972	Maximum likelihood	1.00	.92	.91	.50/.05	.44/.12	.28/.25
	Restricted maximum likelihood		.91	.89		.46/.12	.28/.28
	Instrumental variables			.83			.27/.26
January 1973 through December 1977	Maximum likelihood	1.00	.97	.92	.63/.00	.61/.13	.58/.10
	Restricted maximum likelihood		.97	.91		.61/.13	.58/.10
	Instrumental variables			.92			.63/.09
January 1978 through December 1982	Maximum likelihood	.99	.86	.76	.32/.02	.10/.25	.08/.28
	Restricted maximum likelihood		.81	.69		.09/.26	.07/.30
	Instrumental variables			.84			.06/.10

[a] Number of factors: 10; number of funds: 130; number of securities used in estimation: 750.

Table III

Statistics of Intercepts across Number of Securities Used in Estimating the APT[a]

Sample Period	Statistic	30		250		750	
		Zero Beta	Riskless Rate	Zero Beta	Riskless Rate	Zero Beta	Riskless Rate
January 1968 through December 1972	Mean intercept	2.22	−.30	−1.83	−4.81	−3.22	−4.85
		(4.65)	(3.98)	(4.75)	(4.07)	(4.79)	(3.86)
	Average absolute intercept	3.82	2.89	3.41	4.13	4.21	5.36
	Average absolute *t*-statistic	1.09	0.68	1.02	1.71	1.34	2.05
		(.83)	(.59)	(.85)	(1.01)	(1.04)	(1.16)
	Average absolute *t*-adjusted	1.23	0.73	1.15	1.84	1.51	2.30
		(.96)	(.64)	(.90)	(1.08)	(1.13)	(1.33)
January 1973 through December 1977	Mean intercept	−4.09	−7.36	−3.81	−7.82	−2.67	−5.45
		(5.38)	(4.51)	(4.25)	(4.16)	(3.76)	(3.60)
	Average absolute intercept	5.56	7.67	4.63	8.00	3.58	5.80
	Average absolute *t*-statistic	1.62	1.87	1.63	2.76	1.40	2.36
		(1.00)	(1.07)	(1.02)	(1.17)	(1.02)	(1.17)
	Average absolute *t*-adjusted	1.82	2.04	1.89	2.91	1.59	2.68
		(1.15)	(1.19)	(1.21)	(1.24)	(1.17)	(1.35)
January 1978 through December 1982	Mean intercept	2.18	1.05	−2.00	−3.22	−3.25	−3.85
		(3.98)	(3.85)	(3.28)	(3.12)	(3.35)	(3.30)
	Average absolute intercept	3.59	2.95	2.94	3.62	3.67	4.12
	Average absolute *t*-statistic	0.95	0.70	0.89	1.12	1.26	1.45
		(.70)	(.55)	(.69)	(.75)	(.87)	(.92)
	Average absolute *t*-adjusted	1.08	0.81	1.02	1.31	1.51	1.72
		(.78)	(.65)	(.83)	(.89)	(1.05)	(1.09)

[a] Intercepts are measured in percent per annum. Standard deviations are in parentheses. Number of factors: 10; estimation method: maximum likelihood; number of funds: 130.

Table IV

Comparisons of Treynor-Black Appraisal Ratios across Number of Securities Used in Estimating the APT (Riskless-Rate Version of the APT)[a]

Sample Period		Simple Correlations		Fractions of Funds with Significant Abnormal Performance at the Five Percent Level for Both/Only One Benchmark(s)	
		250	750	250	750
January 1968 through	30	.79	.76	.03/.32	.04/.48
December 1972	250		.91		.31/.22
January 1973 through	30	.89	.87	.43/.29	.42/.22
December 1977	250		.96		.61/.14
January 1978 through	30	.73	.56	.01/.15	.01/.35
December 1982	250		.92		.13/.21

[a] Number of factors: 10; estimation method: maximum likelihood; number of funds: 130.

of two tables each. The first table in each pair provides four summary measures describing the typical behavior of Jensen measures from alternative benchmarks for each of our three sample periods. The first three statistics are the mean intercept, the mean absolute intercept, and the average absolute t-statistic.[21] These t-statistics are simply the estimated intercepts divided by the usual ordinary least-squares standard errors calculated under the assumption that the residuals are independent and have constant variance over time. Unfortunately, as long as managers vary the composition of their portfolios in attempts to outperform the market, mutual fund returns will have nonstationary variances even if the return-generating process of individual securities is stationary. To guard against this possibility, we also present adjusted t-statistics using estimated standard errors that are consistent in the presence of arbitrary forms of conditional heteroscedasticity.[22]

These summary statistics merely serve to characterize the behavior of the *sample* intercepts and cannot be used to draw inferences about the typical behavior of the *true* intercepts without further assumptions that facilitate statistical inference.[23] Unfortunately, simultaneous inference procedures based on the F or *chi-squared* distribution require the inverse of the sample covariance matrix of the residuals from the regression given by (6), which is singular because the number of funds (130) is greater than the number of time-series degrees of

[21] In parentheses below the sample averages, we also present the sample standard deviations of the intercepts and the t-statistics across funds. The mean intercepts are measured in percent per annum (i.e., monthly intercepts multiplied by twelve).

[22] See Hansen [16], White [45], and Hsieh [18] for further details.

[23] Note that by *true* intercept we mean the population (i.e., expected) intercept associated with a given benchmark. For our APT benchmarks, this means that both *true* benchmark error and the effects of estimation error in the benchmark-construction process can cause differences in performance measures and affect the sample standard errors of the Jensen measures as well. Hence, as noted above, our analysis records differences in *measured* performance without attempting to distinguish among various sources of these differences.

Table V

Statistics of Intercepts across Number of Factors[a]

Sample Period	Statistic	5		10		15	
		Zero Beta	Riskless Rate	Zero Beta	Riskless Rate	Zero Beta	Riskless Rate
January 1968 through December 1972	Mean intercept	−3.76	−4.87	−3.22	−4.85	−4.28	−4.74
		(5.18)	(4.15)	(4.79)	(3.86)	(5.86)	(4.52)
	Mean absolute intercept	4.73	5.38	4.21	5.27	5.10	5.26
	Average absolute *t*-statistic	1.55	1.98	1.34	2.05	1.45	1.88
		(1.21)	(1.12)	(1.04)	(1.16)	(1.14)	(1.17)
	Average absolute *t*-adjusted	1.53	2.05	1.51	2.30	1.74	2.22
		(1.22)	(1.15)	(1.13)	(1.33)	(1.34)	(1.37)
January 1973 through December 1977	Mean intercept	−3.59	−6.32	−2.67	−5.45	−3.00	−5.77
		(4.33)	(4.06)	(3.76)	(3.60)	(3.80)	(3.81)
	Mean absolute intercept	4.55	6.66	3.58	5.80	3.79	6.13
	Average absolute *t*-statistic	1.70	2.60	1.40	2.36	1.46	2.52
		(1.13)	(1.31)	(1.02)	(1.17)	(1.02)	(1.21)
	Average absolute *t*-adjusted	1.82	2.78	1.59	2.68	1.68	2.89
		(1.21)	(1.44)	(1.17)	(1.35)	(1.19)	(1.41)
January 1978 through December 1982	Mean intercept	−.89	−1.52	−3.25	−3.85	−3.51	−3.93
		(3.22)	(2.98)	(3.35)	(3.30)	(3.51)	(3.37)
	Mean absolute intercept	2.52	2.54	3.67	4.10	3.95	4.22
	Average absolute *t*-statistic	0.89	0.94	1.26	1.45	1.38	1.48
		(.75)	(.79)	(.87)	(.92)	(.93)	(.94)
	Average absolute *t*-adjusted	0.96	1.02	1.51	1.72	1.73	1.86
		(.82)	(.86)	(1.05)	(1.09)	(1.22)	(1.19)

[a] Intercepts are measured in percent per annum. Standard deviations are in parentheses. Estimation method: maximum likelihood; number of securities used in estimation: 750; number of funds: 130.

Table VI

Comparisons of Treynor-Black Appraisal Ratios across Number of Factors (Riskless-Rate Version of the APT)[a]

Sample Period		Simple Correlations		Fractions of Funds with Significant Abnormal Performance at the Five Percent Level for Both/Only One Benchmark(s)	
		10	15	10	15
January 1968 through December 1972	5	.97	.92	.44/.09	.35/.16
	10		.93		.36/.19
January 1973 through December 1977	5	.96	.92	.59/.10	.62/.09
	10		.96		.62/.05
January 1978 through December 1982	5	.69	.59	.09/.28	.07/.30
	10		.95		.27/.09

[a] Estimation method: maximum likelihood; number of securities used in estimation: 750; number of funds: 130.

freedom (sixty).[24] Since this is a common problem, it is customary in the literature to ignore inference procedures and instead to report the implied rankings of mutual funds and the "big winners" among them. We will similarly confine our attention to measures of ordinal and cardinal rankings of mutual funds in order to examine the sensitivity of this common practice to the choice of benchmark. Note that we ascribe no particular economic significance to such rankings; they can reflect benchmark error as well as true abnormal performance.

The second table in each pair describes the degree of association among the inferences produced by different performance benchmarks. Each table examines the relations among the Treynor-Black appraisal ratios—the estimated intercepts (i.e., Jensen measures) divided by their idiosyncratic standard deviations—yielded by different risk-adjustment procedures. The Treynor-Black appraisal ratio is, of course, proportional to the usual OLS t-statistic for the intercept. The tables report two measures to characterize relations among them: simple (i.e., Pearson product-moment) correlations and relations among "significant" Treynor-Black measures. The second measure consists of two statistics: the fraction of intercepts significant at the five percent level for *both* benchmarks and the fraction significant at the five percent level for *only one* benchmark.[25] These

[24] There are conservative testing procedures for the presence of abnormal performance by *any* mutual fund in the sample. The Bonferroni inequality states that, if we examine N possibly dependent t-statistics at the critical value associated with α/N, then we are sure that we have at most a joint test at the significance level α. Any individual t-statistic in our sample that is greater than four is large enough to ensure that the joint F-statistic is significant at better than the one percent level. For each period and method, there are a number of t-values greater than four, indicating the presence of abnormal performance. Unfortunately, this procedure does not lend itself to comparisons of abnormal performance measures *across* benchmarks.

[25] These numbers are obtained from the usual 2 × 2 contingency table for measuring association among discrete variates. The first is the upper left-hand box (significant intercepts with both benchmarks), while the second is the sum of the off-diagonal boxes. The lower right-hand box (insignificant intercepts with both benchmarks) is simply one minus the sum of these two numbers.

Table VII

Statistics of Intercepts across Benchmarks[a]

Sample Period	Statistic	APT		CRSP		
		Zero Beta	Riskless Rate	Value Weighted: Excess Returns	Equally Weighted: Excess Returns	No Risk Adjustment: Excess Returns
January 1968 through December 1972	Mean intercept	−3.22	−4.85	−1.41	−.15	6.53
		(4.79)	(3.86)	(4.37)	(4.23)	(4.07)
	Mean absolute intercept	4.21	5.27	3.12	2.81	6.91
	Average absolute *t*-statistic	1.34	2.05	0.99	0.81	1.03
		(1.04)	(1.16)	(.81)	(.71)	(.58)
	Average absolute *t*-adjusted	1.51	2.30	1.01	0.83	1.04
		(1.13)	(1.33)	(.82)	(.72)	(.59)
January 1973 through December 1977	Mean intercept	−2.67	−5.45	−.79	−6.32	1.33
		(3.76)	(3.60)	(4.54)	(4.91)	(4.83)
	Mean absolute intercept	3.58	5.80	3.62	6.83	3.82
	Average absolute *t*-statistic	1.40	2.36	1.31	1.52	0.53
		(1.02)	(1.17)	(.86)	(.82)	(.53)
	Average absolute *t*-adjusted	1.59	2.68	1.31	1.58	0.54
		(1.17)	(1.35)	(.86)	(.86)	(.53)
January 1978 through December 1982	Mean intercept	−3.25	−3.85	1.40	−3.19	16.36
		(3.35)	(3.30)	(3.98)	(3.27)	(4.82)
	Mean absolute intercept	3.67	4.12	2.94	3.95	16.36
	Average absolute *t*-statistic	1.26	1.45	0.91	1.14	2.11
		(.87)	(.92)	(.69)	(.68)	(.44)
	Average absolute *t*-adjusted	1.51	1.72	0.94	1.18	2.12
		(1.05)	(1.09)	(.71)	(.71)	(.45)

[a] Intercepts are measured in percent per annum. Standard deviations are in parentheses. APT: number of factors: 10; estimation method: maximum likelihood; number of securities used in estimation: 750. Number of funds: 130.

Table VIII

Comparisons of Treynor-Black Appraisal Ratios Across Benchmarks (Riskless-Rate Version of the APT)[a]

Sample Period		Simple Correlations			Fractions of Funds with Significant Abnormal Performance at the Five Percent Level for Both/Only One Benchmark(s)		
		CRSP			CRSP		
		Value Weighted	Equally Weighted	No Risk Adjustment	Value Weighted	Equally Weighted	No Risk Adjustment
January 1968 through December 1972	APT	.80	.70	.63	.13/.39	.04/.48	.01/.53
	CRSP (value weighted)		.96	.89		.05/.09	.02/.14
	CRSP (equally weighted)			.92			.02/.05
January 1973 through December 1977	APT	.83	.87	.75	.18/.52	.31/.32	.02/.62
	CRSP (value weighted)		.92	.88		.17/.22	.02/.24
	CRSP (equally weighted)			.91			.01/.32
January 1978 through December 1982	APT	.50	.57	.52	.00/.41	.09/.29	.13/.68
	CRSP (value weighted)		.89	.68		.00/.22	.08/.54
	CRSP (equally weighted)			.61			.03/.70

[a] APT: number of factors: 10; estimation method: maximum likelihood; number of securities used in estimation: 750. Number of funds: 130.

Table IX

Intertemporal Persistence of Treynor-Black Appraisal Ratios

Benchmark	Simple Correlations			Fractions of Funds with Significant Abnormal Performance at the Five Percent Level for Both/Only One Period(s)		
	1968–1972/ 1973–1977	1973–1977/ 1978–1982	1968–1972/ 1978–1982	1968–1972/ 1973–1977	1973–1977/ 1978–1982	1968–1972/ 1978–1982
APT[a]						
Maximum likelihood	.39	.41	.17	.37/.40	.25/.45	.22/.41
Restricted maximum likelihood	.40	.42	.19	.40/.38	.26/.44	.22/.45
Instrumental variables	.45	.30	.08	.42/.38	.09/.64	.06/.49
Principal components	.26	.32	.11	.23/.48	.08/.59	.03/.35
Maximum likelihood: 30 securities	.17	.39	.04	.02/.45	.01/.45	.00/.08
Maximum likelihood: 250 securities	.39	.39	.18	.29/.47	.14/.58	.05/.36
Maximum likelihood: 5 factors	.40	.27	.05	.37/.38	.11/.58	.07/.46
Maximum likelihood: 15 factors	.44	.44	.22	.36/.35	.25/.46	.15/.42
CRSP						
Value weighted	.19	.17	−.09	.03/.33	.02/.29	.01/.21
Equally weighted	.25	.14	−.15	.02/.33	.07/.32	.00/.20
No risk adjustment	.39	.30	−.01	.02/.03	.02/.61	.02/.62

[a] APT benchmarks were constructed using 750 securities under the assumption that there are ten common factors, unless otherwise indicated.

Table X

Results from APT Quadratic-Regression Tests for Market Timing[a]

	Percentage of Funds with Quadratic Terms Statistically Different from Zero at the Following Significance Levels									
	F-Test					Chi-Square Test				
Sample Period	≤1%	≤5%	≤10%	≤15%	>15%	≤1%	≤5%	≤10%	≤15%	>15%
January 1968 through December 1972	2.3	5.4	13.1	20.0	80.0	14.6	23.8	35.4	42.3	57.7
January 1973 through December 1977	21.5	38.5	48.5	53.8	46.2	61.5	72.3	76.9	79.2	20.8
January 1978 through December 1982	1.5	10.0	14.6	21.5	78.5	21.5	36.9	41.5	45.4	54.6

[a] APT: number of factors: 5; estimation method: maximum likelihood; number of securities used in estimation: 750. Number of funds: 130.

Table XI

Results from CRSP Quadratic-Regression Tests for Market Timing[a]

Sample Period	Percentage of Funds with Quadratic Terms Statistically Different from Zero at the Following Significance Levels									
	t-Test					Adjusted *t*-Test				
	≤1%	≤5%	≤10%	≤15%	>15%	≤1%	≤5%	≤10%	≤15%	>15%
Value-Weighted CRSP Index as Market Proxy										
January 1968 through December 1972	2.3	9.2	13.8	20.0	80.0	3.1	10.0	20.0	24.6	75.4
January 1973 through December 1977	2.3	5.4	10.8	17.7	82.3	2.3	6.9	10.8	14.6	85.4
January 1978 through December 1982	16.9	30.0	34.6	41.5	58.5	16.1	25.4	31.5	36.9	63.1
Equally Weighted CRSP Index as Market Proxy										
January 1968 through December 1972	4.6	12.3	20.0	32.3	67.7	7.7	16.2	28.5	36.9	63.1
January 1973 through December 1977	60.0	80.8	87.7	90.0	10.0	60.7	83.8	89.2	91.5	8.5
January 1978 through December 1982	20.0	42.3	49.2	53.8	46.1	33.1	44.6	54.6	60.8	39.2

[a] Number of funds: 130.

numbers indicate whether changing the benchmark changes the percentage or composition of funds that display "significant" abnormal performance and, hence, provide a measure of the sensitivity of cardinal rankings to the benchmark chosen. The correlations among Treynor-Black measures were qualitatively similar to those of the Jensen intercepts, so reference will be made to the latter in the text only where substantive differences were found. Similarly, the implications of the simple correlations will occasionally be highlighted by reference to the typical difference in the ranks of the Jensen measures yielded by different benchmarks.[26] This is often an apt summary statistic since one application of mutual fund performance evaluation is the provision of such ordinal rankings. Once again, we cannot report standard errors and confidence intervals for our summary measures due to the likely presence of correlation among the intercepts across funds.

Table IX merits separate comment. It is commonplace in the mutual fund literature for investigators to report the correlations among Jensen measures for individual mutual funds across sample periods. Accordingly, Table IX reports the simple correlations and relations among "significant" Treynor-Black measures produced by each benchmark over time. For each benchmark, this involves comparing the Treynor-Black measures for each fund yielded by the 1968 through 1972 sample with those produced in both the 1973 through 1977 and 1978 through 1982 periods, as well as relating those from the latter two periods. Of course, little can be inferred about abnormal performance from these statistics; high serial correlation among Treynor-Black measures could simply reflect consistent abnormal performance or benchmark error, while low serial correlation could result from secular changes in their risk postures. Table IX can only indicate whether alternative risk-adjustment procedures imply different degrees of intertemporal persistence in the performance measures of individual mutual funds.

One general observation about the Jensen measures described in the tables is the persistent incidence of negative intercepts across all three sample periods, especially with the APT benchmarks. Consider, for instance, the average estimated alphas from the regressions run in excess-return form using the unrestricted maximum-likelihood estimation procedure presented in Table I (which we simply refer to as maximum likelihood in the tables). The average annual excess return of the funds implied by this benchmark is −4.85 percent for the first five-year period, −5.45 percent for the second period, and −3.85 percent per year for the third period.[27] The Jensen measures of the funds (not presented in the tables) are almost uniformly negative, and, hence, the mean is not being

[26] The statistic we employ is the standard deviation (or square root) of the average squared difference in ranks of the Jensen measures. This is obtained from the Spearman rank correlation:

$$\rho_{jk}^{\text{rank}} = 1 - \frac{6 \sum_{i=1}^{N} (y_{ij} - y_{ik})^2}{(N(N^2 - 1))} = 1 - \frac{6 \times \text{average squared difference of ranks}}{(N^2 - 1)},$$

where y_{ij} is the rank of the ith firm using the jth benchmark.

[27] Since the conclusions one would reach about the inferences implied by alternative APT benchmarks turn out to be independent of whether the zero-beta or riskless-rate models are employed, we limit our discussion in the text to the latter. We emphasize these results due to evidence presented in Lehmann and Modest [26] that suggests the preferability of this form of the APT.

pulled down by a few funds with "exceptionally poor" performance. This also can be seen by noting the relatively small difference between the absolute values of the average intercepts and the sample means of the absolute values of the individual intercepts. Not only are the absolute magnitudes of the intercepts large, but, in the first two five-year periods, the *average* values of the adjusted and unadjusted t-statistics are sufficiently large to suggest that many of the intercepts are significantly different from zero. In short, our APT benchmarks suggest the presence of widespread abnormal performance by mutual funds across all sample periods; we discuss the economic significance of this finding at the end of the section.

Tables I and II examine the impact of alternative methods of estimating the factor model for security returns underlying the APT. Four different estimation methods were compared using samples of 750 securities: two maximum-likelihood procedures, an instrumental-variables estimator, and the method of principal components. These procedures were discussed in Section II, Subsection B. Tables I and II compare these APT benchmarks assuming that there are ten common sources of systematic risk. For each estimation method, statistics are presented for regressions run in raw-return form, which corresponds to the zero-beta version of the APT (15), and for regressions run in excess-return form, which corresponds to the riskless-rate model (14).

Examination of Table I reveals that there is considerable variation in the mean intercept across estimation methods. The differences in the mean alphas from the restricted and unrestricted maximum-likelihood procedures were minor; the restricted procedure led to mean Jensen measures that were thirty-seven basis points per annum lower in the first five-year period, ten basis points higher in the second period, and twenty-nine basis points per year lower in the third period. In contrast, the average intercepts using the instrumental-variables estimation procedure were twenty-eight basis points lower, 153 basis points lower, and 189 basis points higher in periods one through three, respectively, than those produced by the unrestricted maximum-likelihood procedure. The corresponding numbers using the principal-components procedure were +138 basis points, −107 basis points, and +207 basis points.

Table II also suggests that the principal-components and instrumental-variables procedures can yield very different inferences from those produced by the maximum-likelihood procedures. The correlations between the Treynor-Black measures from the unrestricted and restricted maximum-likelihood procedures are in excess of .99, and the methods differed on fewer than 5.4 percent of the "significant" t-values in all three periods. There is less similarity between the measures produced by the unrestricted maximum-likelihood procedure and the instrumental-variables or principal-components methods. The correlations between the Treynor-Black measures yielded by the unrestricted maximum-likelihood method and the instrumental-variables procedure are .9179, .9661, and .8587, and they differed on 12.31 percent, 13.08 percent, and 24.62 percent of the "significant" ones in the three five-year periods. The relations between the inferences suggested by the unrestricted maximum-likelihood procedure and the method of principal components are weaker with correlations of .9078, .9229, and .7572 and differences on 24.62 percent, 10.0 percent, and 28.46 percent of the

"significant" t-ratios in the three five-year periods. Note that the correlations reported here are associated with typical risk differences on the order of sixteen positions for the instrumental-variables procedure and twenty positions for principal components compared with the unrestricted maximum-likelihood procedure. Put differently, the relevant statistics (not reported here) indicate that the Treynor-Black measures yielded by the instrumental-variables and principal-components procedures are more similar to those produced by the unrestricted maximum-likelihood procedure with 250 (as opposed to 750) securities in the latter two periods.

The statistics reported in Table IX reveal a similar pattern. To be sure, the correlations of the Treynor-Black measures are much smaller and the proportion of "significant" disagreements much larger across sample periods for each benchmark. Nevertheless, there are some interesting differences across benchmarks. The restricted and unrestricted maximum-likelihood procedures yield Treynor-Black measures with similar degrees of intertemporal persistence. The same cannot be said of the principal-components and instrumental-variables methods; both the simple correlations among Treynor-Black measures and the proportion of agreements on "significant" ones are much larger for the maximum-likelihood methods for two of the three intertemporal comparisons. The numbers reported in the last column of Table IX are particularly striking; 21.54 percent of the funds with "significant" maximum-likelihood Treynor-Black measures in the 1968 through 1972 period had "significant" ones in the 1978 through 1982 period, while the proportions for the instrumental-variables and principal-components procedures were only 6.15 percent and 3.08 percent, respectively.

In order to study whether the number of securities used in factor-model estimation has an economically significant impact on performance evaluation, we performed unrestricted maximum-likelihood factor analysis on the first thirty, 250, and 750 securities in our randomly sampled data file and used these subsets of securities to construct the reference portfolios. The results are presented in Tables III and IV for the ten-factor model. The statistics reported in Table III indicate considerable dependence of the mean alphas on the number of securities used in the estimation. Previous authors, such as Roll and Ross [35], have based their inferences concerning the APT on maximum-likelihood estimation of factor models involving thirty to sixty securities. As is evident, this leads to very different conclusions about mutual fund performance than those yielded by the 250- or 750-security benchmarks. The difference between the mean alphas using thirty securities and those based on 750 securities is, on an annual basis, +455 basis points, −191 basis points, and +490 basis points for the three five-year periods, respectively. The corresponding differences with the 250-security benchmark are the more modest values of +4, −237, and +63 basis points.

Tables IV and IX indicate that the relations among performance measures are sensitive to this choice as well. The simple correlations between the Treynor-Black appraisal ratios from the thirty- and 750-security benchmarks range from .5611 to .8688, while the differences in "significant" ones range from 21.54 percent to 47.69 percent. They also yield very different degrees of intertemporal persistence; the correlations of the 750-security appraisal ratios are considerably larger than those of the thirty-security benchmark across two of the three intertemporal

comparisons, while there is negligible (less than 1.54 percent) intertemporal agreement on "significant" ones for the thirty-security benchmark. The corresponding ranges for the Treynor-Black measures given by the 250- and 750-security benchmarks are narrower: .9148 to .9641 for the simple correlations and 13.85 percent and 22.31 percent for disagreements on "significant" ones within each period. Similarly, the correlations of the 250-security appraisal ratios across periods are comparable to those yielded by the 750-security benchmark, but there is considerably less agreement on "significant" ones in two of the three comparisons, 13.85 percent vs. 25.38 percent across the 1973 through 1977 and 1978 through 1982 sample periods and 5.38 percent vs. 21.54 percent for the 1968 through 1972 and 1978 through 1982 comparison. In short, Tables I to IV and IX suggest that alternative basis-portfolio-construction procedures can lead to substantially different conclusions.

In our final comparison of APT benchmarks, we examined the sensitivity of mutual fund performance measures to the number of common factors assumed to affect security returns. In particular, we contrasted the performance measures yielded by basis portfolios constructed under the alternative assumptions that there are five, ten, and fifteen common factors using the estimated factor loadings and idiosyncratic variances from maximum-likelihood factor analysis of 750 securities.[28] Tables V and VI present this evidence. The difference between the mean intercepts from estimating five factors relative to ten factors is (on an annual basis) −2 basis points, −87 basis points, and +233 basis points in the three five-year periods, respectively. The corresponding differences between the mean intercepts using ten and fifteen factors are +11, −32, and −8 basis points, respectively. Except for the difference between the mean intercepts using five and ten factors in the third and, perhaps, the second five-year periods, these differences are all quite small.

A similar picture arises from an examination of the statistics in Tables VI and IX: the ten- and fifteen-factor models yielded similar performance measures in all comparisons, while the five-factor model produced substantially different inferences in the final period. The simple correlations between the t-ratios from the five- and ten-factor models range from .6911 to .9666, and the differences in "significant" ones are on the order of ten percent for the first two periods and 28.46 percent for the third period. Similarly, the intertemporal persistence of the ten-factor Treynor-Black measures is greater than that produced by the five-factor model for comparisons involving the third sample period. The corresponding relations between the ten- and fifteen-factor Treynor-Black measures involve simple correlations ranging from .9264 to .9641 and "significant mismatches" ranging from 5.38 percent to 19.23 percent within the three sample periods, as well as similar intertemporal patterns in all but the last column of Table IX. Thus, the choice of the number of factors does not appear to be as important as the other ones for evaluating the performance of mutual funds although, in the

[28] This arbitrary choice (of the number of factors to examine) was based on the limitations of our computer resources and the needs of our related research. The differences between the results obtained below for one (i.e., the CRSP indices), five, and ten factors and the similarities between the ten- and fifteen-factor performance measures suggest that it might be interesting to determine "where" the changes occur between one and ten factors. We leave this for future research.

third sample period, the five-factor performance measures behave differently from those yielded by the ten- and fifteen-factor models.

Having compared alternative APT benchmarks, it is natural to ask whether the APT has anything different to say about performance evaluation than do the standard implementations of the CAPM. Tables VII and VIII present summary statistics that shed light on this question. As a point of reference, we also present summary statistics based on no risk adjustment as well. The difference between the mean intercepts using the APT benchmark and the mean intercepts using either of the CAPM benchmarks is striking. While the alphas from the APT benchmarks are markedly negative in all three periods, the CAPM alphas are much less negative and less statistically significant. The means of the CAPM alphas using the value-weighted index are (on an annual basis) 344, 466, and 525 basis points higher in the three five-year periods than the mean value of the APT alphas using the unrestricted maximum-likelihood estimation procedure with 750 securities to construct the APT benchmark. The corresponding differences between the CAPM alphas using the equally weighted index are +470, −87, and +66 basis points.

These sharp differences are also reflected in the dissimilarities among the Treynor-Black measures. The simple correlations between the APT and CAPM t-ratios using the value-weighted index are .8046, .8337, and .4979, while they disagree on 39.23 percent, 52.31 percent, and 40.77 percent of the "significant" ones. The corresponding statistics relating the equally weighted CAPM and APT benchmarks are .7037, .8667, .5678, and 48.46 percent, 32.31 percent, and 29.23 percent, respectively. Similarly, the correlations over time of the Treynor-Black measures produced by the two CRSP indices are much smaller than those associated with the APT model in all three comparisons, and there is negligible intertemporal persistence of "significant" ones as well.

These numbers are striking for two reasons. First, the APT and CAPM benchmarks differ more than they agree on the "significant" Treynor-Black measures in all three periods. Second, the Jensen measures produced by the CAPM benchmarks are more similar to no risk adjustment than they are to the APT benchmarks! The typical rank differences between the APT Jensen measures and no risk adjustment are twenty-two, nineteen, and forty-seven positions in the three five-year periods. In contrast, the CAPM benchmarks produce rankings quite similar to no risk adjustment: the value-weighted index yields typical rank differences of seven, seven, and twelve positions, while the equally weighted index produces typical rank differences of only four, three, and twenty-three positions in the three periods.[29] Similarly, the relevant statistics (not reported here) indicate that the Treynor-Black measures yielded by the two CRSP indices are more highly correlated with subsequent sample-mean excess mutual fund returns (not risk adjusted) than they are with the corresponding

[29] The correlations among the Jensen measures themselves are similarly revealing. The correlations between the APT intercepts and the sample mean (i.e., no risk adjustment) excess mutual fund returns above the risk-free rate are .8952, .8887, and .1793 in the three five-year periods, compared with .9907, .9835, and .9651 for the value-weighted index and .9970, .9970, and .8767 for the equally weighted index. This complements Roll's [34] observation that different indices yield inferences similar to no risk adjustment in a study involving simulated passively managed funds.

appraisal ratios. Obviously, inferences about mutual fund performance are dramatically affected by the choice between an APT benchmark and a CAPM benchmark.

What accounts for the sharply negative intercepts? We offer two potential explanations. The first possibility is that the measured abnormal performance simply reflects the mean-variance inefficiency of our benchmarks. For example, in Lehmann and Modest [25], we found that the APT could explain the empirical anomalies involving dividend yield and own variance but could not account for size-related anomalies. In particular, the value-weighted CRSP index has significant negative Jensen measures relative to the APT benchmark in each of the five-year periods covered by the mutual fund data. Hence, mutual funds would tend to have negative Jensen measures when stocks with large market capitalizations constitute a large fraction of their portfolios although it is worth noting that the intercepts from the value-weighted regressions are not as large and negative as the mean intercepts from the mutual fund regressions. The second explanation involves true or spurious market timing by mutual fund managers. As discussed in Section I, the fund's alpha can be a reasonable measure of its stock-selection ability (in the absence of benchmark error) if its risk level is constant. However, if the fund's risk level is not constant (perhaps due to shifts associated with market-timing attempts or because of the option nature of levered securities), its Jensen measure may be arbitrarily positive or negative depending on the covariance between changes in its risk posture and the returns on the factors.

We ran quadratic regressions of the mutual funds' returns on the factors and the factors squared, as outlined in equation (11) for the single-factor case, to see whether real or artificial market timing accounts for the incidence of persistently negative intercepts.[30] The null hypothesis that the coefficients on the quadratic terms are zero reflects the joint hypothesis that the risk of the funds is constant and that the returns of the individual securities are generated by a K-factor linear structure.[31] The usual F-statistic is appropriate for testing this hypothesis when the residuals from the quadratic regressions are homoscedastic but is inappropriate in the presence of residual heteroscedasticity. An asymptotically valid test can be conducted using the procedures of Hansen [16], White [45], and Hsieh [18] to construct heteroscedastic-consistent covariance matrices along the lines of the adjusted t-statistics discussed above. This test has the shortcoming that its small sample distribution is not known; reliance must be placed on its asymptotic chi-squared distribution.

In Table X, we present summary statistics for tests that the risk levels of the funds are constant under the assumption that there are five common factors affecting security returns. For each sample period, we report the fraction of the funds for which we could reject the hypothesis that the quadratic terms were

[30] Recent empirical papers that have examined the ability of managed funds to time the market include Kon [23], Chang and Lewellen [6], and Henriksson [17]. None have employed the quadratic-regression approach.

[31] This assumes, of course, that we know K and that we have basis portfolios that are measured without error. The results obtained below could merely reflect particular sorts of errors in our benchmarks even if the risk levels of funds were constant and returns came from a stationary distribution.

zero at the one percent, five percent, ten percent, and fifteen percent significance levels. The quadratic regressions include only the own-squared terms and none of the cross-product terms (which would have raised the number of regressors to fifteen) and are limited to the five-factor model (since the ten-factor model would have involved twenty-one regressors) due to the limited number of time-series observations (sixty). We report both F-tests, which are valid when the residuals are homoscedastic, and chi-squared statistics, which are asymptotically valid under arbitrary forms of conditional heteroscedasticity.

An examination of Table X reveals that the chi-squared tests lead to a greater number of rejections of the null hypothesis of constant risk levels relative to the five-factor model in all three periods than would have been expected *a priori* under the assumption of independence. The F-tests, on the other hand, lead to approximately the number of rejections that would have been expected, except in the second five-year period. For instance, in the first five-year period, the chi-squared test rejects the null hypothesis that the quadratic terms are zero for 14.6 percent of the firms at the one percent significance level, 23.8 percent at the five percent level, 35.4 percent at the ten percent level, and 42.3 percent at the fifteen percent level. The F-test, however, leads to rejection of the null for only 2.3 percent of the firms at the one percent level, 5.4 percent at the five percent level, 13.1 percent at the ten percent level, and 20.0 percent at the fifteen percent level. It is difficult to reconcile these differences, as the F-test is valid only under the potentially dubious assumption of homoscedasticity while the small sample properties of the chi-squared test are unknown.[32]

Finally, in Table XI, we present summary statistics for quadratic-regression tests under the assumption that the market model (using either the CRSP equally or value-weighted index) describes the return-generating process of individual securities. For each sample period, we report the fraction of the funds for which we could reject the hypothesis that the quadratic term was zero at the one percent, five percent, ten percent, and fifteen percent significance levels using both CRSP indices. We report both the ordinary t-tests, which are valid under residual homoscedasticity, and adjusted t-tests, which are constructed using heteroscedastic-consistent standard errors. Both kinds of t-tests yielded by the equally weighted index lead to a greater number of rejections of the null hypothesis of constant risk levels in all three periods than would be expected under the assumption of independence across funds.[33] For example, in the first five-year period (where the smallest number of rejections occurs), the adjusted t-test rejects the null hypothesis that the quadratic term is zero for 7.7 percent of the firms at

[32] One natural way to confront this problem is to carry out a direct test for heteroscedasticity. Monte Carlo evidence presented in Hsieh [18], however, suggests that such tests have little power to discriminate between homoscedasticity and heteroscedasticity. Hsieh suggests that, even in small samples, there is little harm (and potentially a significant gain) to always using the heteroscedastic-consistent standard errors. Unfortunately, this does not speak to the problem here, which is the rate of convergence of the test statistic to its asymptotic distribution.

[33] See Jagannathan and Korajczyk [19] for an argument that such results could arise from artificial market timing due to the differential leverage of the firms in the CRSP indices and those invested in by mutual funds. Like many others, we found predominantly negative quadratic terms. Unlike the predictions in Jagannathan and Korajczyk, however, we found no systematic evidence that firms with large negative quadratic terms have large positive intercepts. In particular, we have been unable to document any substantive correlation between the alphas (or the t-statistics on the alphas) and the coefficients on the squared terms (or the t-statistics on the squared terms).

the one percent significance level, 16.2 percent at the five percent level, 28.5 percent at the ten percent level, and 36.9 percent at the fifteen percent level. In the second five-year period (where the largest number of rejections occurs), the adjusted t-tests lead to rejection of the null for 60.7 percent of the firms at the one percent level, 83.8 percent at the five percent level, 89.2 percent at the ten percent level, and 91.5 percent at the fifteen percent level. Using the value-weighted CRSP index as a market proxy, an abnormal number of rejections occurs only in the third five-year period.[34]

V. Conclusion

In this paper, we have examined the returns of 130 mutual funds over the period January 1968 through December 1982 to find out whether inferences about their performance are sensitive to the benchmark chosen to measure normal performance. In this pursuit, we studied the behavior of the intercepts from Jensen-style mutual fund regressions that used different risk-adjustment procedures, including alternative APT and CAPM benchmarks. We provided summary statistics describing the sensitivity of mutual fund performance measures to the chosen benchmark.

Three conclusions emerged from this comparison. First, the Jensen measures and Treynor-Black appraisal ratios of individual mutual funds are quite sensitive to the method used to construct the APT benchmark. One would reach very different conclusions about the funds' performance using smaller numbers of securities in the analysis or the instrumental-variables or principal-components estimation methods than one would arrive at using the maximum-likelihood procedures with 750 securities. Second, the rankings of the funds are less sensitive to the exact number of common sources of systematic risk that are assumed to impinge on security returns. There were negligible differences in the inferences yielded by the ten- and fifteen-factor models and only small differences with the five-factor benchmark, except for the final five-year period. Third, there are considerable differences between the performance measures yielded by the standard CAPM benchmarks and those produced with the APT benchmarks, which suggests the importance of knowing the appropriate model for risk and expected return in this context.

In short, the one firm conclusion that can be reached from our analysis is that the choice of what constitutes normal performance is important for evaluating the performance of managed portfolios. It is also worth stressing that these findings are in no way compromised by the potential problems associated with the shifting risk levels of managed funds. These problems affect only the *interpretation* of the Jensen and Treynor-Black measures. If the choice of a benchmark were an unimportant one, different benchmarks should have yielded similar results; the overwhelming fact is that they did not.

These findings stand in sharp contrast to much of the conventional wisdom in the literature. We conjecture that many investigators would not have expected

[34] Note that the statistics reported in Tables X and XI often vary as much across time periods for a given benchmark as they do across benchmarks for a given time period. This is doubtless a consequence of the nature of market fluctuations in our sample periods. For example, the largest decline in equity market values (and in squared returns on well-diversified portfolios) in forty-five years occurred in 1974, and the 1978 through 1982 period witnessed an extremely strong small-firm effect.

substantive differences across the different APT benchmarks. Conversely, some would doubtless have predicted large differences in the inferences produced by APT benchmarks with different numbers of factors. Finally, previous evidence suggests that alternative risk-adjustment procedures lead to similar inferences in settings other than the present one. Our comprehensive examination of mutual fund performance suggests that each of these intuitions is unreliable in this context.

Along with the three conclusions that have emerged from our analysis, one puzzle has also arisen—the persistent incidence of large and negative Jensen measures. While theoretically it is possible that this measured abnormal performance can be attributed to real or artificial market timing or to a value-weighted bias in our constructed benchmarks, the preliminary evidence is not conclusive. We are currently pursuing this line of research.

REFERENCES

1. Anat Admati, Sudipto Bhattacharya, Paul C. Pfleiderer, and Stephen A. Ross. "On Timing and Selectivity." *Journal of Finance* 41 (July 1986), 715–30.
2. Anat Admati and Stephen A. Ross. "Measuring Investment Performance in a Rational Expectations Model." *Journal of Business* 58 (Jaunary 1985), 1–26.
3. Fischer Black. "Capital Market Equilibrium with Restricted Borrowing." *Journal of Business* 45 (July 1972), 444–55.
4. Douglas T. Breeden. "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities.." *Journal of Financial Economics* 7 (September 1979), 265–96.
5. Gary Chamberlain and Michael Rothschild. "Arbitrage and Mean/Variance Analysis on Large Markets." *Econometrica* 51 (September 1983), 1281–1304.
6. Eric C. Chang and Wilbur G. Lewellen. "Market Timing and Mutual Fund Investment Performance." *Journal of Business* 57 (January 1984), 57–72.
7. Nai-fu Chen, Thomas Copeland, and David Mayers. "A Comparison of APM, CAPM, and Market-Model Portfolio Performance Methodologies: The Value Line Case (1965–1978)." Working Paper, UCLA, 1983.
8. Gregory Connor and Robert A. Korajczyk. "Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis." *Journal of Financial Economics* 15 (March 1986), 373–94,
9. Thomas Copeland and David Mayers. "The Value Line Enigma (1965–1978): A Case Study of Performance Evaluation Issues." *Journal of Financial Economics* 10 (November 1982), 289–322.
10. John C. Cox, Jonathan E. Ingersoll, Jr., and Stephen A. Ross. "An Intertemporal General Equilibrium Model of Prices." *Econometrica* 53 (March 1985), 363–84.
11. Arthur P. Dempster, N. M. Laird, and Donald B. Rubin. "Maximum Likelihood from Incomplete Data via the E-M Algorithm." *Journal of the Royal Statistical Society Series B* 39 (1977), 1–38.
12. Philip H. Dybvig and Stephen A. Ross. "Differential Information and Performance Measurement Using a Security Market Line." *Journal of Finance* 40 (June 1985), 383–99.
13. Eugene F. Fama and James D. MacBeth. "Risk, Return, and Equilibrium: Some Empirical Tests." *Journal of Political Economy* 81 (May 1973), 607–36.
14. Mark Grinblatt and Sheridan Titman. "Portfolio Performance Evaluation: Old Issues and New Insights." Technical Report 16-84, Graduate School of Management, UCLA, 1985
15. Gosta Hagglund. "Factor Analysis by Instrumental Variables Methods." *Psychometrika* 47 (June 1982), 209–21.
16. Lars P. Hansen. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50 (June 1982), 1029–54.
17. Roy D. Henriksson. "Market Timing and Mutual Fund Performance: An Empirical Investigation." *Journal of Business* 57 (January 1984), 73–96.
18. David A. Hsieh. "A Heteroscedasticity-Consistent Covariance Matrix Estimator for Time Series Regressions." *Journal of Econometrics* 22 (August 1983), 281–90.
19. Ravi Jagannathan and Robert A. Korajczyk. "Assessing the Market Timing Performance of

Managed Portfolios." *Journal of Business* 59 (April 1986), 217–35.
20. Michael C. Jensen. "The Performance of Mutual Funds in the Period 1945–1964." *Journal of Finance* 23 (May 1968), 389–416.
21. ———. "Risk, the Pricing of Capital Assets, and the Evaluation of Investment Portfolios." *Journal of Business* 42 (April 1969), 167–247.
22. ———. "Optimal Utilization of Market Forecasts and the Evaluation of Investment Portfolio Performance." In G. P. Szego and Karl Shell (eds.), *Mathematical Methods in Investment and Finance.* Amsterdam: North-Holland, 1972, 310–35.
23. Stanley J. Kon. "The Market Timing of Mutual Fund Managers." *Journal of Business* 56 (July 1983), 323–47.
24. Alan Kraus and Robert H. Litzenberger. "Skewness Preference and the Valuation of Risky Assets." *Journal of Finance* 31 (September 1976), 1085–1100.
25. Bruce N. Lehmann and David M. Modest. "The Empirical Foundations of the Arbitrage Pricing Theory II: The Optimal Construction of Basis Portfolios." Working Paper, Graduate School of Business, Columbia University, August 1985.
26. ———. "The Empirical Foundations of the Arbitrage Pricing Theory I: The Empirical Tests." Working Paper, Graduate School of Business, Columbia University, April 1987.
27. John Lintner. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *Review of Economics and Statistics* 47 (February 1965), 13–37.
28. John B. Long, Jr. "Stock Prices, Inflation and the Term Structure of Interest Rates." *Journal of Financial Economics* 1 (July 1974), 131–70.
29. Albert Madansky. "Instrumental Variables in Factor Analysis." *Psychometrika* 29 (June 1964), 105–13.
30. David Mayers and Edward Rice. "Measuring Portfolio Performance and the Empirical Content of Asset Pricing Models." *Journal of Financial Economics* 7 (March 1979), 3–29.
31. Robert C. Merton. "An Intertemporal Capital Asset Pricing Model." *Econometrica* 41 (September 1973), 867–87.
32. Richard W. Roll. "A Critique of the Asset Pricing Theory's Tests—Part I: On Past and Potential Testability of the Theory." *Journal of Financial Economics* 4 (May 1977), 129–76.
33. ———. "Ambiguity When Performance is Measured by the Securities Market Line." *Journal of Finance* 33 (September 1978), 1051–69.
34. ———. "Sensitivity of Performance Measurement to Index Choice: Commonly-Used Indices." Working Paper, Graduate School of Management, UCLA, March 1979.
35. ——— and Stephen A. Ross. "An Empirical Investigation of the Arbitrage Pricing Theory." *Journal of Finance* 35 (December 1980), 1073–1103.
36. Stephen A. Ross. "The Arbitrage Theory of Capital Asset Pricing." *Journal of Economic Theory* 13 (December 1976), 341–60.
37. ———. "Risk, Return, and Arbitrage." In Irwin Friend and James L. Bicksler (eds.), *Risk and Return in Finance: Volume 1.* Cambridge, MA: Ballinger, 1977, 189–218.
38. Donald B. Rubin and Dorothy T. Thayer. "EM Algorithms for ML Factor Analysis." *Psychometrika* 57 (March 1982), 69–76.
39. William F. Sharpe. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance* 19 (September 1964), 425–42.
40. ———. "Mutual Fund Performance." *Journal of Business* 39 (January 1966 supplement), 119–38.
41. Robert F. Stambaugh. "On the Exclusion of Assets from Tests of the Two Parameter Model." *Journal of Financial Economics* 10 (November 1982), 235–68.
42. Jack L. Treynor. "How to Rate Management of Investment Funds." *Harvard Business Review* 43 (January–February 1965), 63–75.
43. ——— and Fischer Black. "Portfolio Selection Using Special Information, under the Assumptions of the Diagonal Model, with Mean-Variance Portfolio Objectives, and without Constraints." In G. P. Szego and Karl Shell (eds.), *Mathematical Models in Investment and Finance.* Amsterdam: North-Holland, 1972, 367–84.
44. Jack L. Treynor and F. Mazuy. "Can Mutual Funds Outguess the Market?" *Harvard Business Review* 44 (July–August 1966), 131–36.
45. Halbert White. "A Heteroscedasticity-Consistent Covariance Estimator and a Direct Test for Heteroscedasticity." *Econometrica* 48 (May 1980), 817–38.