R. Jarrow et al., Eds., *Handbooks in OR & MS, Vol. 9*

Chapter 19

Performance Evaluation

Mark Grinblatt
Anderson Graduate School of Management, University of California, Los Angeles, CA 90095-1481, U.S.A.

Sheridan Titman
W.E. Carroll School of Management, Boston College, Chestnut Hills, MA 02167, U.S.A.

1. Introduction

Trillions of dollars are invested in stocks worldwide by institutional portfolio managers. From a social perspective it is important to know whether these investors as a group add value to the portfolios they manage or whether they merely generate wasteful transaction costs through their active management. At the micro level it is important to know how to select a portfolio manager with the ability to add value to the portfolio he manages. Performance evaluation is a topic in financial economics that seeks to address both of these issues. In particular, it studies whether superior returns can be generated by active managers who are better able to collect and interpret information that helps forecast securities returns.[1]

To evaluate whether a manager has generated superior returns we need to adjust his portfolio return for risk. Since the mean returns of securities are positively related to their risk, performance measures are based on techniques that adjust for priced risk. Some performance measures use the diversification of a portfolio as an additional criterion for evaluation. This requires an adjustment for both priced risk and unpriced risk, typified by the performance measure known as Sharpe's ratio [Sharpe, 1966]. Sharpe's ratio is the excess return of the portfolio (above the risk-free return) divided by the standard deviation of the return of the portfolio.

Adjusting for performance based on the total risk of a portfolio rather than the priced risk of a portfolio is no longer popular and we think inappropriate. This is

[1] Many investors assert that they make money by 'arbitraging' mispriced derivatives. For example, they may buy a European call option that is underpriced relative to the Black-Scholes model and short the underlying stock in appropriate amounts so as to achieve a riskless return that exceeds the current riskless rate available in the fixed income markets. This type of performance is based on model failure, rather than asymmetric information. Indeed, in the effectively complete markets world of derivatives pricing, no asymmetric information about the mean returns of securities is permitted. Since there is no available performance methodology that addresses the issue of performance based on model failure, we do not discuss the issue here.

because the managers whose performance is typically evaluated rarely manage the entire savings of an investor. Investors in mutual funds, for example, typically hold a number of funds and may personally manage a large portion of their wealth. They may also hold a substantial fraction of their wealth in the home they own or the human capital they possess. Even if we argue that for some individuals, most of their wealth is held in their pension fund, most pension funds farm out the management of their assets to a number of different firms. It therefore seems more important to focus on the marginal contributions of a managed portfolio to the risk and expected return of an investor. This necessarily involves adjusting for risk with a marginal risk measure, like beta.

There are two basic classes of performance measures analyzed in this chapter. An intuitive way to think about both classes of performance measures is that they compare the returns of the actively managed portfolio that is being evaluated with a passive (i.e. buy and hold) portfolio with the same level of risk. The first class requires the observation of the returns of the evaluated portfolio as well as the returns of a benchmark that consists of one or more portfolios along with a risk-free asset. The second class utilizes information about the composition of the evaluated portfolio but does not necessarily require a benchmark portfolio(s). In most cases the first class of measures assumes that stock returns are normally distributed. This assumption is not needed for the second class of measures. However, both classes of measures require that stock returns be drawn from a stationary distribution.

The stationarity requirement is considered by some to be a serious weakness of the performance evaluation literature. Given the recent literature on the nonstationarity of expected stock returns, as found, for example, in Ferson [1995, chapter 5 in this volume] and Hawawini & Keim [1995, chapter 17 in this volume], this concern seems particularly valid. However, this assumption is needed because it is generally impossible for an observer to empirically distinguish between the performance of informed investors and the 'performance' of uninformed investors who optimally respond to changes in the parameters of the return generating process. Indeed, a fair 'philosophical' distinction between an economy with informed investors and an economy with changing parameters is one of magnitude. In the former, only a few investors observe the nonstationarities, whereas in the latter, virtually all investors observe the nonstationarities (and only the evaluator is naive). A more sophisticated evaluation technique that models the nonstationarities known to uninformed market participants can in principle avoid this problem, but modelling what is known by 'the market' is speculative at best.

2. Measures based solely on returns

2.1. Treynor's ratio, Jensen's alpha, and the Treynor–Black appraisal ratio

A number of measures of performance are based on the capital asset pricing model, a theory relating expected return to a measure of risk known as 'beta'.

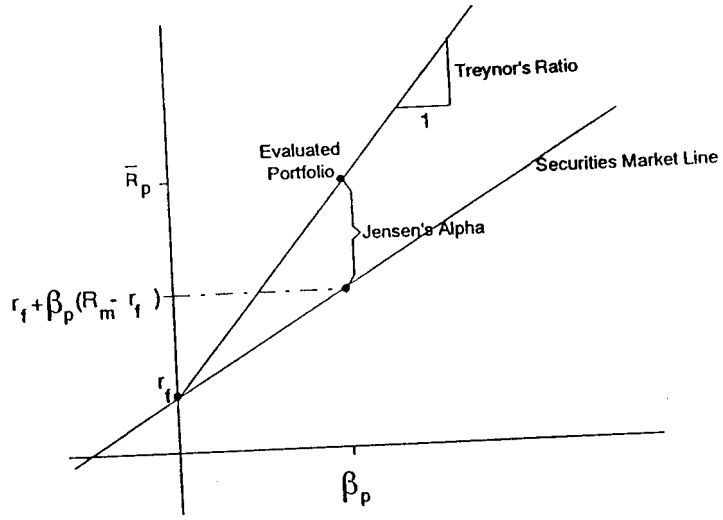

Fig. 1. The distinction between Treynor's ratio and Jensen's alpha.

Beta is the slope coefficient in a regression of the return of the portfolio being evaluated against the return of a proxy for the market portfolio. The relation between between beta and expected return is graphically represented by the securities market line, observed in Figure 1 (and discussed in Ferson [1995, chapter 5 in this volume]). The true mean returns and betas (as opposed to their estimated values) of all securities and all passive portfolios of securities lie on this line if the Capital Asset Pricing Model (CAPM) is true.

2.1.1. Treynor's ratio

Treynor's ratio [Treynor, 1965] was the first academic attempt to adjust returns with betas in order to measure performance. It is computed as

$$\frac{\bar{R}_p - r_f}{\beta_p},$$

the average return of the portfolio in excess of the risk-free return divided by the portfolio's beta. It is thus the slope of a line connecting the risk-free return to the evaluated portfolio in mean–beta space, as illustrated in Figure 1.

2.1.2. Jensen's alpha

A variation of the Treynor approach, known as Jensen's alpha, is the arithmetic difference of the portfolio's return from the return of a portfolio on the securities-market line with the same beta, as illustrated in Figure 1. Since all securities are expected to lie on the securities market line if the CAPM holds, the alphas of passively managed portfolios (with returns measured before transaction costs,

fees, and expenses), are expected to be zero. An actively managed portfolio with a significantly positive Jensen's alpha is therefore interpreted as a portfolio that is managed with superior forecasts. We will later examine the conditions under which this interpretation is correct.

Jensen's alpha is a more commonly used measure of performance than Treynor's ratio. One reason for its popularity is that it is easily computed by finding the intercept in the regression,

$$\tilde{R}_p - r_f = \alpha_p + \beta_p(\tilde{R}_M - r_f) + \tilde{\epsilon}_p \tag{1}$$

The intercept, α_p, computed as

$$\alpha_p = \bar{R}_p - r_f - \beta_p(\bar{R}_M - r_f), \tag{2}$$

is the average excess return of the portfolio less the product of the portfolio's beta and the average excess return of the market portfolio. Intuitively, Jensen's alpha is the difference between the return of the evaluated portfolio and the return of the passive portfolio consisting of beta units of the benchmark portfolio and one minus beta units of the risk-free asset.

2.1.3. Ranking of forecasting ability and the Treynor–Black appraisal ratio

Treynor's ratio was designed to rank portfolios and does not determine whether a particular portfolio was managed by someone with superior abilities. However, if 'uninformed managers' have portfolios (absent fees, expenses, and transaction costs) that plot on the securities market line, any manager with forecasting ability is expected to have a Treynor ratio (absent these same frictions) that is significantly in excess of the slope of the securities market line, $\bar{R}_M - r_f$, where $\bar{R}_M$ is the average return of the proxy for the market portfolio.

A manager's aggressiveness in using information will alter the expected return of a portfolio. Hence a manager with low risk aversion and good information may outperform a manager with a high degree of risk aversion and great information. Measures like Treynor's, that involve a ratio of return to risk mitigate this problem. However, there is no a priori reason to believe that the beta adjustment suggested by Treynor is the correct one for a cardinal ranking of managerial information precision. For example, a manager with the ability to forecast the epsilon in a market model regression

$$\tilde{r}_i - r_f = \alpha_i + \beta_i(\tilde{R}_M - r_f) + \tilde{\epsilon}_i \tag{3}$$

might hedge out the market risk associated with the acquisition of large positions in the security. This requires shorting β_i dollars of the market portfolio for each \$1 invested in the ith security. If these forecasts are imperfect, so that $\tilde{\epsilon}_i$ conditioned on the manager's information still has some variability, a less risk averse manager will generally take a larger position in security i. In this case, simple division by the portfolio beta will not capture the relatively higher unsystematic risk of the less risk averse manager.

A measure derived from Jensen's alpha can rank managers according to information precision. Connor & Koraczyk [1986] describe a case in which the Treynor & Black [1973] appraisal ratio,

$$\frac{\alpha_p}{s_p},$$

which is Jensen's alpha divided by the standard deviation of the error term in the regression used to obtain alpha, properly ranks managers according to their forecasting abilities. However, this result requires a number of assumptions before it is valid, including: no ability to forecast the market, multivariate normal returns, exponential utility as the criterion for investment for all managers, and the tradability of all assets for all managers. These restrictions appear to be stringent enough to preclude the usefulness of this ratio as a tool for ranking

While the ability to rank managers according to the precision of their forecasting ability is an ideal, we must generally content ourselves with the separation of portfolio managers into two classes: those with superior forecasting ability and those without it. One cannot be greatly disappointed that ranking is probably impossible because of differences in managerial (and ultimately client) risk aversion. However, the ability to forecast securities returns is rare if one largely accepts the common academic view of the efficient markets hypothesis. A measure of performance that merely identifies the few managers with forecasting ability would then be quite useful.

2.2. Asset pricing, performance measurement, and Roll's critique

The measures of performance that we discussed in the last section used the Capital Asset Pricing Model as the theoretical basis for their construction. As a result, the benchmarks originally used to compute these measures are proxies for the value-weighted market portfolio. However, other benchmarks have been used to estimate Jensen's alpha, Treynor's ratio, and the Treynor–Black ratio. One can also use a multiple portfolio benchmark with Jensen's alpha. The summed product of the multiple regression betas and the average excess returns of the benchmark portfolios is then subtracted from the average excess return of the portfolio being evaluated. The alphas obtained from a multiple portfolio benchmark are equivalent to the alphas that would be obtained from using the ex-post efficient combination of the portfolios in the benchmark.[2]

The choice of a benchmark portfolio is probably the most controversial issue in performance evaluation. The debate about benchmarks was initiated by Roll [1978] who noted that different benchmark portfolios provide different risk adjustments and hence different assessments of abnormal performance. He showed that two benchmark portfolios lying inside the mean–variance efficient frontier

[2] Grinblatt & Titman [1987] show that the mean-variance efficient combination of a set of multiple benchmarks will be itself mean-variance efficient if and only if the Jensen's alphas derived from these multiple portfolio benchmarks are all zero.

could reverse the rankings of a group of passive portfolios. Portfolios lying above the securities market line with one benchmark lie below the securities market line with the other benchmark and vice versa. On the other hand, a benchmark portfolio that is mean–variance efficient cannot distinguish between passive portfolios. Passive portfolios, like all securities, lie on the securities market line in this case.

The reason for this is a mathematical relation. The equation of the securities market line,

$$\bar{r}_i = r_f + \beta_i(\bar{R}_E - r_f), \qquad \text{for all } i$$

where

$$\beta_i = \frac{\text{cov}(\tilde{r}_i, \tilde{R}_E)}{\text{var}(\tilde{R}_E)},$$

is merely the (necessary and sufficient) first order condition for the mean–variance efficiency of the benchmark (portfolio E) used to compute beta. Because of this mathematical property of mean–variance efficiency, it would seem that a proper benchmark portfolio needs to be both mean–variance efficient and mean–variance inefficient at the same time. It needs to be mean–variance efficient so that the portfolios of uninformed managers and all passive portfolios will have Jensen's alphas of zero. It needs to be mean–variance inefficient for the portfolios of managers with forecasting ability so that these portfolios can have nonzero alphas.

This would seem to be an impossible task for a portfolio, but for the fact that the information sets of managers with forecasting ability differ from those without forecasting ability. Two managers with different information sets would necessarily draw different mean–standard deviation diagrams. In particular, the manager with forecasting ability would have mean–variance frontiers that are improved by dynamic portfolio strategies — strategies that weight more heavily those securities that are forecasted to have unusually high returns in a period. Managers lacking this ability cannot achieve a better mean–variance tradeoff by dynamically changing their portfolio weights. Hence, their efficient frontier plots inside the efficient frontier of informed managers, as in Figure 2. This insight implies that portfolio performance evaluation may escape Roll's critique if we use a benchmark that lies on the line connecting points A and B in Figure 2. This benchmark is mean–variance efficient with respect to passive portfolios but not with respect to the dynamic portfolios chosen by managers with forecasting ability.

The performance obtained with a benchmark having this property is analyzed in models developed by Mayers & Rice [1979], Dybvig & Ross [1985a, b], and Grinblatt & Titman [1989b]. In these models, investors with superior information about individual securities returns (i.e. selectivity information) but with no information about the return on the benchmark (i.e. timing information) achieve positive alphas if the benchmark is mean–variance efficient from the perspective of

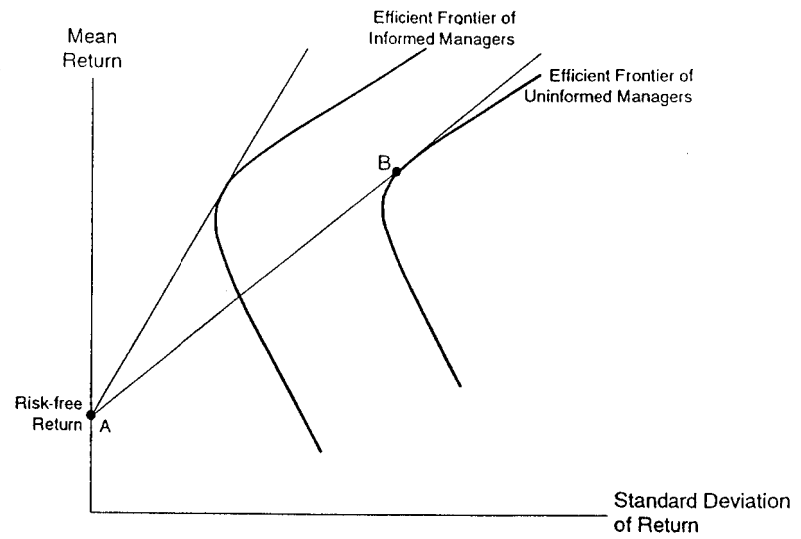

Fig. 2. Differences between the efficient frontiers of informed and uninformed portfolio managers.

an investor without forecasting ability. For investors with timing information, this result does not necessarily hold. We will defer discussion of this timing problem until later.

A practical complication still remains. Figure 2 is based on the ex-ante means and standard deviations of portfolios rather than their estimated means and standard deviations. It suggests that we employ a benchmark portfolio on the ex-ante efficient frontier, but finding such a portfolio is not an easy task. Motivated by the CAPM, early empirical studies of mutual fund performance made use of value-weighted portfolios as benchmarks. These studies were undertaken prior to the late 1970s, when a number of CAPM anomalies were discovered.[3] Given the more recent empirical evidence, the benchmarks used in these studies appear to be inappropriate for studying portfolio performance, since they can be gamed by managers aware of CAPM-related anomalies, like the well-known firm size or dividend yield effects. In the next section, we outline the results from empirical studies of fund performance with techniques that are based on securities market line analysis. We can see how benchmarks have evolved as we analyze these studies in chronological order.

[3] While there was evidence in the early 1970s of a beta-related anomaly, this anomaly was small for stocks with betas close to 1 (e.g. portfolios with betas from 0.85 to 1.1 deviated by at most 0.5% per year from the securities market line, which is well within levels of statistical tolerance). Since most mutual funds, and certainly the average mutual fund, have betas close to 1, early empirical studies of mutual fund performance felt that CAPM proxies were appropriate benchmarks.

2.3. Empirical studies that employ securities market line analysis

2.3.1. Studies that employ a single portfolio benchmark

Jensen's doctoral dissertation, summarized in articles in the *Journal of Finance* [1968] and the *Journal of Business* [1969], was one of the earliest and most influential studies of mutual fund performance. The study used Jensen's alpha to evaluate the yearly returns of 115 mutual funds over the 1945–1964 period. The sample included all of the funds that were followed by Wiesenberger's *Investment Companies* over the entire 1955–1964 period. The benchmark used to evaluate these funds was the S&P composite index which is a value-weighted portfolio (consisting of 500 stocks after March 1, 1957 and 90 stocks prior to this date).

Jensen concluded that over the 1955–1964 period, mutual funds, on average, achieved a risk-adjusted performance of about −0.9% per year.[4] When the various commissions and expenses of the funds were added back to the funds' returns, the risk-adjusted performance was virtually zero. Thus, the average returns of the funds were consistent with what would be expected in an efficient market. Similar results were found by McDonald [1974] using the equally-weighted NYSE index as his benchmark.

For a smaller sample of 56 funds, Jensen also examined whether funds that did relatively well (poorly) from 1945–1954 also did well (poorly) from 1955–1964. The correlation between the first half returns and second half returns was 0.64 indicating that the performance of some mutual funds persists over time. This would be evidence against market efficiency if it were the positive performers that did well persistently. However, the evidence suggests that it is primarily the bad performers that exhibit persistent performance. These persistently bad performance numbers could have been generated by funds that generated very high commissions and other expenses. The study failed to find evidence of persistently positive performance.

In a comment on Jensen's study, Mains [1977] argued that the mutual fund returns used by Jensen were biased downwards. Jensen assumed that the dividends of the fund were paid out at the end of the year which was in fact not true. As a result, the interest income on the dividend payments was ignored in Jensen's analysis. The way in which Jensen added back expenses and commissions produced a similar bias in his estimates of gross returns. Mains also questioned Jensen's estimates of systematic risk.

To reassess the Jensen results, Mains analyzed the monthly returns of 70 of the 115 funds examined by Jensen over the same 1955 to 1964 time period. The use of monthly returns, rather than the yearly returns used by Jensen, eliminates the bias associated with the assumption that the dividends were paid at the end of the year and provides superior estimates of the betas and alphas of the funds.

[4] While some funds achieved positive abnormal returns, it is difficult to ascertain the implications of this for the efficient markets hypothesis because of the multiple comparison being made. That is, even if no superior fund management ability existed, we would expect some funds to achiever superior risk-adjusted returns by chance.

Mains concluded that, on average, the net risk-adjusted returns of mutual funds were about zero. However, after adding back expenses, their gross risk-adjusted returns were about 1% per year. Although the study did not report the statistical significance of that 1% abnormal return, it is unlikely that a 1% return significantly differs from zero.

A more recent application of the securities market line methodology was undertaken by Ippolito [1989]. Using the S&P 500 benchmark, Ippolito found slightly positive average alphas (0.83% per year) for his sample of mutual funds in a later period (1965–1984). Ippolito interprets his findings as indicating that mutual funds are able to use their superior information to generate abnormal returns. However, as we discuss in the next subsection, such conclusions are highly dependent on the choice of benchmark portfolios.

2.3.2. Studies with multiple portfolio benchmarks

More recent studies have examined the sensitivity of performance inferences to the choice of the benchmark portfolio, examining multiple as well as single portfolio benchmarks. There are a number of advantages to using multiple portfolio benchmarks. First, unless stock returns are generated by only one common factor, it is unlikely that an arbitrarily chosen diversified portfolio will be mean–variance efficient. However, as Grinblatt & Titman [1987] emphasize, if securities returns are generated by at most k factors, then in the absence of arbitrage, any k well-diversified portfolios will sum to the mean–variance efficient frontier. For this reason we generally feel more comfortable with multiple portfolio benchmarks than with single portfolio benchmarks. Moreover, there is an alternative asset pricing theory, the Arbitrage Pricing Theory (APT), summarized by Connor & Korajczyk [1995, chapter 4 of this volume], that makes use of a multiple portfolio benchmark and which provides guidance on how to select a multiple portfolio benchmark.

Empirical investigations of mutual fund performance with multiple portfolio benchmarks can be found in papers by Lehmann & Modest [1987], Grinblatt & Titman [1989a, 1994], Connor & Korajczyk [1991], and Elton, Gruber, Das & Hlavka [1993]. The study by Lehmann and Modest was the first to adapt the APT to performance evaluation. The benchmarks considered by Lehmann and Modest included the CRSP value-weighted index of all NYSE and AMEX listed stocks (VW), the CRSP equally-weighted index of all NYSE and AMEX listed stocks (EW), and 5, 10, and 15 portfolio benchmarks formed using a variety of factor analysis methods. Over the 1968 to 1982 time period they found that benchmarks formed with the various factor analysis methods produced similar performance numbers. However, the performance numbers with the EW and VW benchmarks differed from each other as well as from the numbers generated with the factor analysis benchmarks. The factor analysis benchmark that they examined in the greatest detail generated very negative performance numbers on average, approximately −4% per year. The EW index generated performance numbers of similar magnitude while the VW index generated performance numbers that were close to zero on average.

Although the factor analysis benchmark has strong theoretical motivation, the extreme negative average performance numbers it generates indicate that the benchmark may not be appropriate. Given that the 4% negative performance greatly exceeds the level of expenses and commissions, the performance numbers suggest that the funds must, on average, be systematically selecting bad stocks, which does not seem plausible. Lehmann and Modest provide two possible explanations for this perverse finding: one is that the mutual funds exhibit timing ability, which biases the beta estimates upwards and the intercepts downwards (see Section 2.4); the second is that the benchmark portfolios cannot be combined to form a point on the mean–variance efficient frontier.

Lehmann and Modest's earlier work[5] suggests that the second explanation is very plausible. The factor analysis benchmark, like the EW benchmark, exhibits a strong size-related bias. In particular, the stocks of large firms exhibit negative alphas when evaluated by this benchmark. Since mutual funds generally select larger than average stocks, it follows that they are likely to exhibit negative performance when measured against this benchmark. Lehmann & Modest [1987] also examined the possibility that the negative performance estimates result from the mutual funds timing the market. They did this by employing a Treynor & Mazuy [1966] quadratic regression, but failed to find evidence of pervasive timing behavior.[6]

Connor & Korajczyk [1991], using a five factor model, analyzed the same sample of mutual funds over the same time period as did Lehmann and Modest. To derive their five portfolio benchmark they first constructed five portfolios using principal components analysis. Four linear combinations of these five portfolios that best mimic a set of four prespecified macroeconomic factors are then formed to yield four new factor portfolios that correspond to the four macro-factors. A fifth factor, the residual of the regression of the value-weighted index on the previously described four macro-factor portfolios, was also used. (The performance results generated with this residual are the same as those that would have been generated by including the value-weighted index itself as the fifth factor.) Because of the inclusion of this fifth factor, Connor and Koraczyk did not find the same evidence of negative performance found by Lehmann and Modest.

Grinblatt & Titman [1989a] provided further evidence that the negative abnormal return generated with Lehmann and Modest's factor analysis benchmark is the result of the inefficiency of the benchmark. In addition to examining performance with the equally-weighted and value-weighted indices and the Lehmann and Modest 10 factor benchmark, they developed a second multiple portfolio benchmark, referred to as 'P8', that is formed on the basis of securities characteristics. This eight portfolio benchmark consisted of four size-based portfolios, three dividend-yield-based portfolios, and the lowest past returns portfolio. The rationale for forming benchmark portfolios based on securities characteristics is that these characteristics may be better proxies for the true factors than factors formed with statistical factor analysis. In their sample period, the P8 benchmark could not be gamed by simple strategies based on well-known CAPM and APT anomalies, such as firm size, dividend yield, beta, skewness, interest rate sensitivity, or past performance.[7] In addition, the benchmark did not generate significantly different alphas for portfolios grouped by industry.[8]

[5] Lehmann & Modest [1989], which developed the factor analysis technique used in Lehmann & Modest [1987], was published later than their mutual fund study.

[6] Even if timing ability were pervasive, it is unlikely that it would result in negative intercepts. Under reasonable parameter values, the intercept is likely to be positive for positive timers. It is more plausible that the negative intercepts are caused by perverse timers. For example, mutual funds might choose to be very conservative when return variances and expected returns are unusually high.

In addition to examining the actual returns of the mutual funds, Grinblatt & Titman [1989a] analyzed what they called 'hypothetical portfolios', formed from the quarterly holdings of the mutual funds. In contrast to the actual portfolios of the funds, the hypothetical portfolios consisted entirely of equity. Since a mutual fund manager's decisions to allocate assets between cash, bonds and stocks do not affect these hypothetical returns, their betas are likely to vary much less than the betas of the actual fund returns. Hence, performance measurement biases arising because of timing the market are substantially lower with these hypothetical returns. In addition, the hypothetical portfolio returns include no expenses or transaction costs and thus should generate zero performance under the null hypothesis that fund managers have no special information.

The Grinblatt & Titman [1989a] findings confirmed the Lehmann and Modest conclusion that benchmark choice does matter. For the actual mutual fund returns, the EW index and the factor analysis benchmark generated very negative performance numbers on average. In contrast, the performance numbers generated by the VW index and the P8 benchmark yielded performance numbers that were close to zero on average. The various benchmarks also ranked the mutual funds differently, [Grinblatt & Titman, 1988]. The P8 benchmark ranked funds very differently than the EW index and the 10 factor benchmark. However, the 10 factor benchmark provided performance scores that were similar to the EW index on a fund by fund basis. The cross-sectional correlation coefficient [from Grinblatt & Titman, 1994] was 0.86.[9]

An evaluation of the hypothetical returns of the portfolios formed from the quarterly holdings revealed slightly positive performance with the VW index (1.9% per year) and the P8 benchmark (1.1% per year) but negative average

[7] See Hawawini & Keim [1995, chapter 17 of this volume) and De Bondt & Thaler [1995, chapter 13 of this volume] for more detail.

[8] A recent paper by Sharpe [1992] forms benchmarks in a similar manner. Sharpe postulates that the 'style' of each managed portfolio can be characterized as a linear function of 12 prespecified passive portfolios that represent the various dimensions of investment style. To find the combination of these 12 portfolios that has the same style as the managed portfolio that is being evaluated, he regresses the managed portfolio's return on the returns of the 12 style portfolios. The estimated betas from this regression are then used as portfolios weights to construct a passive portfolio with the same style as the managed portfolio.

[9] This distinction from the Lehmann and Modest results can be explained by differences in the sample period.

performance with both the EW index (−3.0% per year) and the factor analysis benchmark (−2.4% per year). The positive performance was statistically significant with the VW index but not with the P8 benchmark. The negative performance was statistically significant with the factor analysis benchmark but not with the EW index. Since we can rule out timing and expenses as an explanation for the negative performance numbers generated with the equally-weighted index and the factor analysis benchmark, we must conclude that these benchmarks are not mean–variance efficient and are thus inappropriate for the evaluation of fund performance.

Grinblatt & Titman's [1989a] analysis of the aggressive growth funds is especially useful for understanding the suitability of the different benchmarks. The hypothetical returns of these funds generated significant positive performance (3.3% per year) when measured relative to the P8 benchmark. However, the performance is substantially negative (−3.7%) when measured relative to the factor analysis benchmark. To understand this, consider the fact that aggressive growth funds invest heavily in relatively large firms with low dividend yields. Since the returns of large firms with low dividend yields do poorly relative to the factor analysis benchmark (as shown in Grinblatt & Titman [1988]), mutual funds that follow such a strategy will also exhibit poor performance when measured with respect to this benchmark even if some of the funds really do have superior abilities. However, when these biases are eliminated with the P8 benchmark, the hypothetical returns of the aggressive growth funds exhibit positive performance on average.

The positive performance of the hypothetical returns of the aggressive growth funds does not imply that investors can realize abnormal returns by holding shares directly in these funds. The abnormal performance of the actual fund returns is close to zero on average. The difference in the performance of the hypothetical returns and the actual returns can be attributed to the expenses of the funds and the costs of trading (e.g. brokerage commisions and the bid–ask spread). The estimates of these costs that are derived from this difference is about 2.5% per year for the average mutual fund in the sample (and about 3% for the average aggressive growth fund).

The work of Elton, Gruber, Das & Hlavka [1993], punctuates this evidence on the importance of the benchmark portfolio in drawing performance evaluation conclusions. They propose a 3-index benchmark that includes bond portfolios as well as stock portfolios. Specifically, their benchmark is comprised of the S&P 500 index, a small stock index, and a bond index. They used this benchmark to reevaluate Ippolito's conclusions about mutual funds generating abnormal returns. They found that with their benchmark, Ippolito's sample of mutual funds generated insignificant negative abnormal performance, attributing Ippolito's 'performance' to fund holdings of non-S&P 500 stocks and bonds.

The importance of the choice of a benchmark portfolio is nicely illustrated by the contradictory results on average fund performance between single portfolio and multiple portfolio benchmarks and between different multiple portfolio benchmarks. The lesson of Roll [1978] could not be more clear: When evaluating the performance of a fund manager, the benchmark must be ex-ante efficient with respect to the mean–variance set generated by the passive investment strategies that the fund manager considers feasible.

2.3.3. Is there differential performance?

The positive abnormal performance of the hypothetical mutual fund returns in the Grinblatt & Titman [1989a] study suggests that at least some funds have superior selection ability. Given this, it is natural to ask whether some funds realize better performance than do other funds. We will discuss two ways to test this proposition: The most general test is to simultaneously estimate the market model regressions for each of the mutual funds and to jointly test the restriction that their Jensen's alphas are equal to each other. This test does not quantify differential performance. It merely rejects or fails to reject the hypothesis that all funds have the same risk-adjusted returns. A second test, which analyzes whether the past performance of a mutual fund is a good indicator of its future performance, has the ability to quantify differences in performance.

Within the first class of tests, Grinblatt & Titman [1989a] estimated a series of joint F-tests to determine whether mutual funds with the same investment objectives all generate the same performance.[10] These tests revealed evidence of differential performance for the actual as well as the hypothetical returns of the aggressive growth and growth funds. There was no evidence of differential performance among the funds with other investment objectives.

Tests of the persistence of mutual fund performance may be somewhat less general, but they more directly address the question that is of most interest to mutual fund investors. Is the past performance of a fund a good indicator of its future performance?

Direct tests of this proposition are found in Grinblatt & Titman [1992] and Hendricks, Patel & Zeckhauser [1993]. Grinblatt and Titman examined the actual returns of a sample of 279 funds over the 1975 to 1984 time period using their P8 benchmark. They divided the sample into 1975–1979 and 1980–1984 subperiods and examined whether better than average performance in the earlier half is indicative of better than average performance in the later half. Their results provide weak support for the hypothesis that better than average performance persists over time. For example, the subsample of funds that achieved performance in the top decile in the first subperiod did not realize abnormal performance on average in the second subperiod (although the performance would be positive if expenses were added to the returns). In contrast, funds that performed in the bottom decile in the first subperiod realized abnormal performance of −3.6% per year. The difference between the performance of these groups of funds is statistically significant at the 5% level.

The Hendricks, Patel, and Zeckhauser study, which looks at no-load growth-oriented mutual funds from 1974–1988, provides stronger evidence that funds

[10] Because their sample included 157 mutual funds, but only 120 monthly observations, they did not have enough degrees of freedom to jointly test the equality of all of the intercepts.

that do well in the past do well in the future. In their study, funds in the top octile of past performers over the past year (as measured with raw returns), outperformed the lowest octile past performers in the following year, by 10–16% per year (as measured by risk-adjusted returns, with the variation depending on the benchmark). Their analysis suggests that the best way to profit from this persistence is to focus on the raw returns of funds in the prior four quarters.[11] Information about performance beyond the previous four quarters does not seem to predict future performance. In contrast to Grinblatt & Titman [1992], they find profits from buying the winners as well as from selling the losers. While the profits from buying the past performers in the top octile are large for some benchmarks, they generally are not statistically significant.[12]

2.4. *Timing, selectivity and biases in Jensen's alpha*

We now analyze how timing and selectivity ability affect performance and discuss methods to separate these two types of performance ability. The distinction between timing and selectivity is important in that it is generally more difficult to evaluate performance when there is any timing ability contributing to it.

[11] The persistence of abnormal performance for only about four quarters is not consistent with the idea that some fund managers have superior ability. Ability should last more than four quarters. In addition, four quarters of historical returns is not a sufficiently long time series to draw proper statistical inferences about ability, given the volatility of mutual fund returns. Hence, the fund rankings of Hendricks, Patel & Zeckhauser [1993] are largely due to noise in stock returns. Their evidence on persistence may be related to recent evidence of persistence in individual stock returns [e.g. Jegadeesh & Titman, 1993]. Funds that happen to hold stocks that do well, and continue to hold those stocks, will continue to do well because of the persistence in the individual stock returns. While Hendricks, Patel & Zeckhauser [1993] employ simulations to conclude that persistence in stock returns is not driving their hot-hands effect, their simulations are based on randomly held equally-weighted portfolios of 100 stocks. Relative to actual fund strategies over this time period, this approach may be biased against finding a stock persistence effect.

[12] A recent paper by Brown, Goetzmann, Ibbotson & Ross [1992] argues that results of persistence will appear spuriously in samples limited to surviving mutual funds. Their argument is that funds that choose high risk strategies and survive in the first half of the sample period are likely to have above average returns. If these funds continue their high risk strategy and continue to survive, they are also likely to achieve above normal returns in the second half of the sample. This bias in favor of finding persistence is offset somewhat by the fact that funds which do poorly in the first half of the sample are more likely to exit the sample in the test period (because of poor performance in the test period) than those funds that did well in the first half. Survivorship bias of this type is therefore more severe for the past losers than for the past winners, biasing our tests against finding persistence. Apparently, these two effects are either unimportant in practice, or alternatively, they cancel each other out. Tests of the persistence of mutual fund performance on samples that are not subject to a survival requirement [e.g. Grinblatt & Titman, 1993, footnote 13], provide evidence that is similar to tests on samples that require survival for the entire sample period. Malkiel's [1995] work, however, suggests that conclusions about the unimportance of survivorship may be sensitive to the time period studied. In particular, he finds that there is no evidence of persistence among a sample of surviving and non-surviving funds in the 1980s (but there is persistence in the 1970s). Brown & Goetzmann [1995], using data on non-surviving funds, similarly find that survivorship bias plays some role (albeit a modest one) in the hot hands persistence findings.

We can more formally classify these two types of ability with a simple regression. Let

$\tilde{r}_j$ = excess return of asset j

$\tilde{x}_j$ = the investor's portfolio weight on asset j, which is random, since the investor may alter his portfolio in response to (real or imagined) information.

$\tilde{R}_p$ = the summed product of $\tilde{r}_j$ and $\tilde{x}_j$
= excess return of the investor's portfolio of the N risky assets

$\tilde{R}_E$ = the excess return of the portfolio of risky assets that is mean–variance efficient from the perspective of an uninformed observer.

A regression of the excess return of security i on the excess return of a mean–variance efficient benchmark portfolio implies that the excess return of each asset is

$$\tilde{r}_j = \beta_j \tilde{R}_E + \tilde{\epsilon}_j \tag{4}$$

where

$$\beta_j = \frac{\operatorname{cov}(\tilde{r}_j, \tilde{R}_E)}{\operatorname{var}(\tilde{R}_E)}$$

and, given the efficiency of the benchmark, the mean of $\tilde{\epsilon}_j$ is zero. It follows that the excess return of the investor's portfolio is

$$\tilde{R}_p = \tilde{\beta}_p \tilde{R}_E + \tilde{\epsilon}_p \tag{5}$$

where

$$\tilde{\beta}_p = \sum_{j=1}^{N} \tilde{x}_j \beta_j \quad \text{and} \quad \tilde{\epsilon}_p = \sum_{j=1}^{N} \tilde{x}_j \tilde{\epsilon}_j.$$

Taking the expected value of both sides of equation (5) yields

$$E(\tilde{R}_p) = E(\tilde{\beta}_p) E(\tilde{R}_E) + \operatorname{cov}(\tilde{\beta}_p, \tilde{R}_E) + E(\tilde{\epsilon}_p) \tag{6}$$

The first term on the right hand side of equation (6) is the expected excess return of the portfolio conditional on knowing the portfolio's target risk level. The second term, the covariance between the portfolio beta and the return of the benchmark, is the contribution of timing to the excess return. The third term, which is the sum of the covariances between the portfolio holdings and the residuals, is the contribution of selectivity to the excess return. Total abnormal performance, the sum of the latter two terms, can be shown to be the sum of the covariances between the portfolio weights and the returns,

$$\sum_{j=1}^{N} \operatorname{cov}(\tilde{x}_j, \tilde{r}_j). \tag{7}$$

Grinblatt & Titman [1989b] have shown that this covariance should be positive for investors with the ability to forecast returns.[13]

It is often difficult to properly capture performance with Jensen's alpha, Treynor's ratio, or the Treynor–Black appraisal ratio if betas change. For example, if the beta of the portfolio increases as the forecasted benchmark return increases, the single portfolio beta estimated with Jensen's regression will overestimate the average beta, $E(\tilde{\beta}_p)$. The resulting intercept is then underestimated and under certain conditions can be negative for investors with timing information.[14] Figure 3, drawn from Grinblatt & Titman [1989b], provides a binomial illustration of this phenomenon. Popular examples in the literature illustrate that this phenomenon can also occur for certain parameter values when the portfolio beta is a linear function of a normally distributed timing signal.[15]

The large sample value (or probability limit) of Jensen's alpha can be decomposed to better analyze this phenomenon and to point out the relation of this performance measure to timing and selectivity. Grinblatt & Titman [1989b] showed that Jensen's alpha can be decomposed into the sum of three terms:

$$\alpha = \operatorname{plim}\left[\frac{1}{T}\sum_{t=1}^{T}(\tilde{\beta}_{pt} - b_p)\right]\hat{R}_E + \operatorname{plim}\left[\frac{1}{T}\sum_{t=1}^{T}\tilde{\beta}_{pt}(\tilde{R}_{Et} - \hat{R}_E)\right] + \operatorname{plim}\left[\frac{1}{T}\sum_{t=1}^{T}\tilde{\epsilon}_{pt}\right], \tag{8}$$

where

b_p = the probability limit of the least squares slope coefficient from the time-series regression of excess returns of the evaluated portfolio against the excess returns of the efficient benchmark portfolio and

$\hat{R}_E$ = the probability limit of the sample mean of the excess returns of the benchmark portfolio.

[13] Verrecchia [1980] devised an example where betas were properly measured and where performance was negative for an investor with timing information. Grinblatt & Titman [1989b] showed that this was due to a wealth effect. Investors with positive information about the market reduced their betas because their information made them wealthier. The particular utility function used by Verrecchia, quadratic utility, is one where increased wealth make the investor's risk aversion arbitrarily large at large wealth levels. Hence, this was an risky assets were Giffen goods–wealth effects dominated substitution effects.

[14] See Jensen [1972], Dybvig & Ross [1985a], Admati & Ross [1985], Admati, Bhattacharya, Pfleiderer & Ross [1986], and Grinblatt & Titman [1989b] for other conditions that lead to this result.

[15] See, for example, Grant [1977], Dybvig & Ross [1985a], and Grinblatt & Titman [1989b].

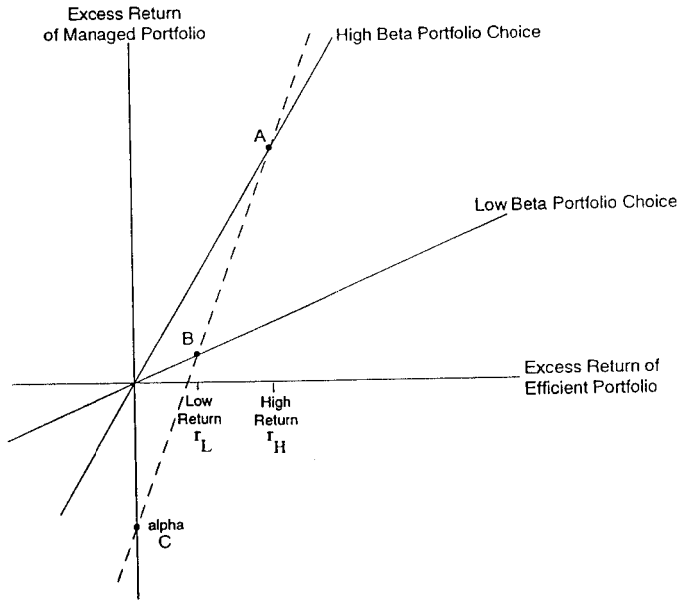

Fig. 3. An example of a negative Jensen measure for a market timer. The two solid lines plot the excess return of a managed portfolio of a risk-free investment and an investment in the risky efficient portfolio against the latter's excess return for two different choices of beta. A market-timing strategy, constrained to choose between the two betas, would plot at point A (point B) if information indicated that the excess return of the efficient portfolio was expected to be r_H (r_L). The slope of the dotted line is the estimated beta in the Jensen regression, and the intercept is the Jensen measure.

The three terms in equation (8) are respectively the component of performance that results from large sample biases in estimated beta, the component that results from timing, and the component that results from selectivity. If the first term is zero, Jensen's alpha aggregrates the sum of the timing and selectivity components. However, this is only the case in the absence of timing information [Grinblatt & Titman, 1989b, Lemma 1].

2.5. *Regression-based timing measures*

There are several procedures that have been proposed to correct for the bias in Jensen's alpha induced by the effect of timing ability on the estimate of beta. The first is a quadratic regression, proposed by Treynor & Mazuy [1966]. This regression,

$$\tilde{R}_p - r_f = \alpha_p + \beta_{0p}(\tilde{R}_E - r_f) + \beta_{1p}(\tilde{R}_E - r_f)^2 + \tilde{\epsilon}_p, \tag{9}$$

is identical to the regression used to compute Jensen's alpha, except that an

additional term, the squared excess return of the benchmark portfolio is included. Admati, Bhattacharya, Pfleiderer & Ross [1986] developed conditions under which the coefficient on the quadratic term can be used to detect the precision of the manager's timing forecasts. Like Connor & Koraczyk, Admati, Bhattacharya, Pfleiderer & Ross assume exponential utility and multivariate normality. It is easy to show that these assumptions imply that the portfolio beta is a linear function of the timing signal.

Under the Admati, Bhattacharya, Pfleiderer & Ross conditions, one can relate the parameters of the Treynor–Mazuy regression to the timing and selectivity components discussed above. In this case, $\beta_{1p} var(\tilde{R}_E)$ is the timing component of performance and α_p is the selectivity component of performance. Total performance is the sum of the two terms. There is no mismeasurement of beta in this case. However, the moment we depart from beta linearity, we can no longer assert this. This makes the application of this regression somewhat narrow.

To see that timing and selectivity are represented by these terms under the narrow set of assumptions, imagine that the manager observes the excess return of the benchmark plus noise, $\tilde{r}_E + \tilde{\delta}$. His forecast of the benchmark return is a linear function of this signal and his portfolio beta is a linear function of the optimal forecast. This implies that

$$\tilde{\beta}_p = \gamma_0 + \gamma_1(\tilde{R}_E + \tilde{\delta})$$

and the excess return of the portfolio is

$$[\gamma_0 + \gamma_1(\tilde{R}_E + \tilde{\delta})]\tilde{R}_E + \tilde{\epsilon}_p.$$

From this expression, let us regard

$$\gamma_1 \tilde{\delta} \tilde{R}_E + (\tilde{\epsilon}_p - E(\tilde{\epsilon}_p))$$

as the regression residual in the Treynor–Mazuy regression. Given exponential utility and the independence of the signals, we know that the portfolio weights that determine $\tilde{\epsilon}_p$ are independent of $\tilde{R}_E$ because of their independence from both the forecast of the benchmark return and of the expected wealth that is tied to this forecast. In addition, $\tilde{\delta}$, being noise, is independent of $\tilde{R}_E$, making the regression residual, $\gamma_1 \tilde{\delta} \tilde{R}_E + (\tilde{\epsilon}_p - E(\tilde{\epsilon}_p))$, uncorrelated with both of the independent variables in the quadratic regression [equation (9)]. γ_0, which multiplies the first variable in the Treynor–Mazuy regression, must therefore equal the asymptotic value of β_1, while γ_1, which multiplies the quadratic term, must equal the asymptotic value of β_2. $E(\tilde{\epsilon}_p)$, which is the large sample expectation of what is left, must equal the asymptotic value of α_p in the quadratic regression.

Empirical work with quadratic regressions has been limited and somewhat disappointing. Work by Grinblatt & Titman [1988], and Cumby & Glen [1990] finds that a large proportion of mutual funds have negative coefficients on the quadratic term. Lehmann & Modest [1987] and Lee & Rahman [1990] also examine the Treynor–Mazuy regression, however neither of these papers report whether the significance of the coefficients is due to their being positive or negative.

Another returns-based approach for estimating timing performance is the option approach developed by Merton [1981] and Henriksson & Merton [1981]. The regression used is similar to the Treynor–Mazuy regression, except that $\max(0, r_f - \tilde{R}_E)$ is used in place of the quadratic term on the right hand side of the regression. This term is the end-of-period value of a put option on the benchmark portfolio with a strike price equal to the risk-free return. Like the Treynor–Mazuy regression, the model used to develop this regression is based on a narrow set of behavioral assumptions. In contrast to the linear beta adjustment of the Treynor–Mazuy framework, the portfolio beta in the Henriksson and Merton study is assumed to switch between two betas: a high beta corresponding to a large forecasted benchmark return and a beta of zero, corresponding to a forecasted benchmark return that is less than the risk-free return.

Chang & Lewellen [1984] and Henriksson [1984] applied the Henriksson & Merton technique to samples of mutual funds and did not find evidence that funds were sytematically timing the market. If anything, there seems to be evidence of negative timing. The application of this technique to a multi-portfolio benchmark in Connor & Korajczyk [1991] reveals similar results.

In a similar spirit, Kon [1983] and Kon & Jen [1979] have used switching regression techniques to estimate performance. Rather than forcing one of the betas to be zero, the Kon & Jen and Kon approaches assume that one of two (or more) unknown betas is selected and use econometric techniques to infer estimates of them and of their contribution to performance. The more sophisticated of the two papers, Kon [1983], concludes that there is no evidence of timing performance within funds as a group.

Although the adjustment for performance developed with either the option approach or quadratic regression approach provide a reasonable estimate of whether timing exists or not, the actual contribution of timing ability to the portfolio return as well as the contribution of selectivity ability will generally be estimated with a bias. This is because investment behavior is unlikely to conform to the rather narrow behavioral assumptions used in these models.

In addition to having restrictive behavioral assumptions, the Treynor–Mazuy and Henriksson–Merton approaches require that stock returns not be co-skewed with the benchmark return. We know, however, from Kraus & Litzenberger [1976] that many stocks are co-skewed with the market return. They illustrated this by examining slope coefficients in quadratic regressions that are identical to the regression specified in equation (9). Since individual stocks exhibit 'timing' performance with these regressions, the Treynor–Mazuy approach will falsely classify passive investors who select stocks with returns that are co-skewed with the benchmark return as successful timers. For similar reasons, the Henrikkson & Merton and Kon & Jen approaches will lead to misclassifications when stock returns are co-skewed with the benchmark. Moreover, even if stock returns are multivariate normally distributed, dynamic portfolio strategies (e.g. synthetic put options or portfolio insurance on the market) generate co-skewness.[16] Hence, the

[16] See Jagannathan & Korajczyk [1986].

coefficient, β_{1p}, in the quadratic regression, as well as the binomial estimates of timing, will be positive for managers who follow such strategies even though they do not possess any real timing ability.

2.6. The positive period weighting measure

Grinblatt & Titman [1989b] developed an alternative performance measure that they describe as a weighting of the evaluated portfolio's time series of excess returns. They showed that if these period weights are nonnegative, and if the weighted average of the time series of excess returns of the benchmark portfolio is zero when using these weights, then the weighted-average of the excess returns of an evaluated portfolio will be positive if and only if it is managed by an investor with forecasting ability that enables him to time the market or select individual stocks to achieve a superior risk-return tradeoff.

Algebraically, the positive period weighting measure is

$$\sum_{t=1}^{T} w_t R_{pt}, \tag{10}$$

where

$$w_t \geq 0, \quad \sum_{t=1}^{T} w_t = 1, \quad \text{and} \quad \sum_{t=1}^{T} w_t R_{Et} = 0.$$

To understand the measure, consider the case where the period weights equal he marginal utilities of an investor who holds the benchmark portfolio. The ieasure then represents the marginal amount that the investor's expected utility ill increase from adding a small amount of the evaluated portfolio to his existing ortfolio. If the evaluated portfolio is managed with no special information, then lding it to the mean–variance efficient portfolio creates no improvement so e measure would be zero in that case. However, if the evaluated portfolio is anaged with special information, its addition does lead to an increase in utility. Jensen's alpha has implicit weights that satisfy two of the three conditions in l). These weights correspond to the marginal utilities of a quadratic utility estor. They do not satisfy the constraint that the weights be nonnegative ause quadratic utility functions exhibit satiation for sufficiently high wealth els, making the marginal utilities (the weights) negative for benchmark returns t are very high. What this means is that successful timers that have very high as when market returns are high may be penalized rather than rewarded by sen's alpha for this astute behavior. This gives an additional interpretation for Jensen's alpha can provide misleading inferences for successful timers.

here are reasons to believe that the positive period weighting measure may iore robust to different behavioral assumptions than the other measures that lop timing adjustments. First, it is very flexible. Any set of weights satisfying conditions (11) above will reward rather than penalize timing ability. This is that we are not restricted to the somewhat implausible binomial or linear

beta adjustment behavior as being behavior that is properly adjusted for by the measure. If we select a plausible utility function to represent the behavior of the portfolio, then the measure is tailor made for that form of beta adjustment. Moreover, Grinblatt & Titman [1989b] have shown that irrespective of the monotone concave utility function used to compute the weights, the measure will reward timing behavior provided that (i) the portfolio's beta is monotonically related to the forecast of the benchmark return and (ii) returns are multivariate normal.

Grinblatt & Titman [1994] computed a positive period weighting measure using the marginal utilities from a power utility function. They found that the positive period weighting measure was highly correlated with Jensen's alpha when the same benchmark is used (e.g. the correlation coefficient with the equal-weighted index was 0.99). Since the two measures are asymptotically identical in the absence of market timing, the high correlation between these two measures suggests that there is at best a negligible amount of market timing prevalent in their sample. They also found a statistically significant relation between the difference between the positive period weighting measure and the quadratic term slope coefficient in the Treynor–Mazuy regression. This implies that the positive period weighting measure differentiated itself from Jensen's alpha when market timing did appear to exist.

Cumby & Glen [1990] also used the positive period weighting measure to study a small sample of international mutual funds. They found no evidence of superior performance within the fifteen funds that they looked at. Consistent with Grinblatt & Titman [1994], they found that the positive period weighting measure and Jensen's alpha provided virtually identical inferences.

3. Computing performance when portfolio weights are observable

3.1. Stationarity and measures based on portfolio holdings

When portfolio weights are observable, performance measures can be developed that substantially reduce or eliminate the problems relating to timing and benchmark choice. Although benchmark portfolios can be used with these measures to increase power and to test for robustness, they are not required. However, the requirement that asset return distributions be stationary is probably more important for the measures that utilize portfolio weights than for those that rely on a benchmark portfolio. For this reason, the addition of a benchmark portfolio may enhance the robustness of these measures with respect to certain nonstationarities.

The assumption that return distributions are stationary implies that the portfolio holdings of an investor with no special abilities or information cannot be correlated with future asset returns. However, since an informed investor can predict when certain assets will tend to have higher than average and lower than average returns, asset returns are nonstationary from his perspective. His percentage holdings for a particular investment, x_j, will tend to be large in periods when

the investment's subsequent return, r_j, is large and vice versa implying that the covariance between the percentage holdings and the subsequent returns will be positive.

3.2. The event study measure and the Grinblatt–Titman measure

Grinblatt & Titman [1989b] have shown that the total contribution of timing and selectivity to an increased return is identical to the summed covariances between the portfolio weights and the returns of the securities in the portfolio:

$$\text{Cov} = \sum_{j=1}^{N} \left(E[\tilde{x}_j \tilde{r}_j] - E[\tilde{x}_j]E[\tilde{r}_j]\right). \tag{12}$$

This sum equals the expected return of the investor's portfolio, given his information, less what the portfolio's expected return would be if his portfolio weights and asset returns were uncorrelated. Rewriting the expression in (12) in two equivalent ways leads to two alternative performance measures that have been implemented in the literature:

$$\text{Cov} = \sum_{j=1}^{N} E\left[\tilde{x}_j\left(\tilde{r}_j - E[\tilde{r}_j]\right)\right] \tag{13a}$$

$$\text{Cov} = \sum_{j=1}^{N} E\left[(x_j - E[x_j])r_j\right] \tag{13b}$$

Expression (13a) is the foundation for the event study measure used by Copeland & Mayers [1982] to evaluate Value Line rankings. Implementing the event study measure requires a proxy for the expected return of each investment included in the evaluated portfolio. Expression (13b) is the foundation for the measure used by Grinblatt & Titman [1993] to evaluate the performance of mutual funds. Intuitively, this measure compares the return of a managed portfolio in each month with a 'passive' portfolio formed in an earlier time period. Implementing this measure requires an estimate of the expected portfolio weight for each of the investments in the evaluated portfolio.

In practice, Grinblatt & Titman [1993] use an investment's portfolio weight in an earlier period as the proxy for the expected portfolio weight of an investment.[17] Similarly, a reasonable proxy for the expected return in a given period is the investment's actual return in a later period. (Future average returns were the benchmarks implemented by Copeland & Mayers in their Value Line study.)

[17] The critical assumption is that the proxy for the expected portfolio holding be independent of the security return. This assumption is not necessarily valid if future holdings are used as a proxy for the expected holding. For example, in this case, an investor that selects past winners for his portfolio will induce a positive correlation between 'expected holdings' and returns, which will in turn downwardly bias the measure. Similarly, if past returns are used as a benchmark for expected returns, the event study measure will be downwardly biased.

Hence, to simplify a comparison of the two measures we will assume that the period $t+k$ return for each asset is used as a proxy for its period t expected return in the event study measure and that its period $t-k$ portfolio holding is used as a proxy for its expected holdings for the Grinblatt & Titman [1993] measure. The event study measure and the Grinblatt–Titman measure can thus be expressed as follows:

$$\text{The event study measure} = \sum\sum \frac{x_{jt}(r_{jt} - r_{j,t+k})}{T} \tag{14a}$$

$$\text{The Grinblatt–Titman measure} = \sum\sum \frac{r_{jt}(x_{jt} - x_{j,t-k})}{T} \tag{14b}$$

For a constant universe of risky assets, the two measures are asymptotically identical. (In finite samples they differ at the k first and k last time series entries.) However, there are several advantages to the measure described in equation (14b). One advantage is statistical. This measure is the average dollar return (i.e. end-of-period value per unit of investment) of a zero-cost, zero-systematic risk portfolio. Under the null hypothesis that asset returns are serially uncorrelated, the returns of this zero-cost portfolio are serially uncorrelated, which makes computation of test statistics for the significance of the average return a trivial exercise. In contrast, the event study measure uses future returns to calculate excess returns. As a result, serial correlation is induced in the time series of excess returns, which makes tests of statistical significance more difficult.

Another weakness of the event study measure is its sensitivity to the future survival of assets currently in the evaluated portfolio. If a particular asset fails to exist shortly after it is included in an evaluated portfolio, the investor's holding of that asset cannot be used to assess the portfolio's performance. This creates a bias in large samples as well as small samples. The problem is especially critical for evaluating portfolio managers who hold near bankrupt stocks or stocks in takeover plays. Grinblatt and Titman's measure, (14b), which for each time period applies current and past portfolio weights to returns in the coming period, cannot have survivorship bias by construction.

Both measures are sensitive to the stationarity of the returns of the individual assets in the evaluated portfolio. A portfolio will spuriously generate positive performance using either measure if it systematically selects assets that temporarily have high risk (and thus high expected return). An example would be a portfolio that buys either near bankrupt stocks or takeover plays. With the Grinblatt–Titman measure, however, it is possible to test whether or not this is a problem by regressing the time-series of return differences on the returns of various market indexes. The intercepts from these regressions are also performance measures. These intercepts will be as robust to various nonstationarities as Jensen's alpha.

3.3. Empirical work using performance measures that require portfolio weights

In practice, portfolio performance evaluators restrict their attention to portfolio returns and ignore information about the portfolio holdings of the evaluated

funds. This is unfortunate since the timing related and benchmark related problems of the performance measures that do not require the observation of portfolio holdings can be substantially reduced with measures that employ portfolio holdings.

Although data on portfolio holdings are available to professional portfolio evaluators, the data is relatively expensive for academics to obtain. For this reason there are a limited number of academic articles that empirically examine measures that require portfolio holdings. We are aware of only three. The first was an article by Copeland & Mayers [1982] that examined the performance of portfolios formed on the basis of Value Line rankings. This article uses two measures similar to that in equation (14a). The first measure, as suggested by equation (14a), uses the difference in the raw returns of a stock between two time periods. The second looks at the difference over time between the Jensen's alphas of the stocks. There is also a follow up article by Chen, Copeland & Mayers [1987] that uses the same data set but employs APT-based alphas, rather than CAPM alphas for this subtraction. The third article, by Grinblatt & Titman [1993], examined the quarterly holdings of a sample of mutual funds.

Copeland and Mayers' sample included the rankings for each stock covered by Value Line at 26 week intervals for the 1965 to 1978 period. These rankings range from 1 to 5 with stocks ranking 1 considered the best choices and those ranked 5 considered the worst choices. In the past, those stocks ranked 1 have realized higher returns on average than those stocks ranked 5. In the Copeland and Mayers sample period, the stocks ranked 1 yielded average yearly returns of 17.7% while those ranked 5 yielded average yearly returns of 3.6% per year for portfolios formed on a six month basis. The average betas with respect to a market proxy for both groups were close to one, indicating that a Jensen's alpha would reveal an excess return of about 14% per year from a strategy of buying stocks ranked 1 and shorting stocks ranked 5.

The event study measure reveals somewhat reduced performance measures for this strategy. For example, Copeland and Mayers comparison of the Jensen measure of each stock in the 6 month holding period to its Jensen measure in the following 6 month period revealed a yearly risk-adjusted return of 0.7% for the rank 1 stocks and −6.1% for the rank 5 stocks. The measure that compares the raw returns in the holding period to the raw returns in the benchmark period reveals a risk-adjusted return of 4.3% per year for the rank 1 stocks and −4.7% per year for the rank 5 stocks.

Grinblatt & Titman [1993] used the measure described in equation (14b) to examine the performance of the quarterly mutual fund holdings considered in their earlier [1989a] paper. They considered two measures. The first, based on quarterly changes in portfolio holdings ($k = 1$), calculates the mean return of a zero cost portfolio that includes a long position in the current quarter's portfolio holdings and is short in the previous quarter's holdings. The second, based on yearly changes in portfolio holdings ($k = 4$), is the same as the first except that the short position in the zero cost portfolio is the previous year's, rather than the previous quarter's, holdings. If mutual fund managers have superior

information that gets revealed to the market within one quarter, the quarterly measure provides the most power. However, if the information is incorporated into market prices more slowly, the quarterly measure may be biased downwards, (due to the correlation between the past holdings and current returns), making the yearly measure the preferred alternative.

Grinblatt and Titman found that the yearly change measure revealed statistically significant abnormal performance, (2% per year on average). However, the quarterly change measure revealed insignificant performance on average. These findings were consistent with the superior information used by the funds being revealed beyond one quarter following the initial purchase decision. With the yearly change measure, the magnitude of the average abnormal performance for the funds categorized by investment objectives were very similar to Grinblatt & Titman's [1989a] earlier results that used the Jensen measure with P8 benchmark. Specifically, the aggressive growth and growth funds revealed significant abnormal performance with both measures. The income funds, which did not exhibit significant superior performance in the earlier study, showed small (1.19% per year), but statistically significant, performance with the yearly change measure. The performance of funds with other investment objectives did not achieve significant abnormal performance in either study.

Grinblatt & Titman [1993] also used their measure to analyze whether or not differential performance existed within the various investment categories. They did this with the joint intercept tests and the persistence tests described earlier. The joint intercept tests revealed differential performance for the Aggressive Growth, Growth, Growth–Income and the Venture Capital/Special Situation funds. They also found evidence of persistence for the entire sample of funds; the second subperiod abnormal returns of funds that did well in the first subperiod exceeded the abnormal returns of the funds that did poorly in the first subperiod by a statistically significant 2.6% per year. These differences in excess returns were also found for subsamples of funds grouped by investment objective, however, the differences were statistically significant only for the Growth–Income funds.

4. Conclusions and directions for future research

In this chapter, we have described various methods for evaluating the performance of a managed portfolio. These methods assess whether the investment strategy of the managed portfolio achieves a higher return than a passive strategy with the equivalent risk. The major difficulty in implementing these techniques, and the source of most of the debate and criticism about them, revolves around the identification of the relevant passive portfolio.

When portfolio holdings are unobservable, it is necessary to make use of an asset pricing theory to derive a benchmark portfolio(s) that in combination with a risk-free security determines this passive portfolio. The passive portfolio estimated with the asset pricing methodology should maximize expected returns

given its level of risk. The empirical evidence suggests that the equally-weighted index and the value-weighted index are not mean–variance efficient and thus would not be appropriate benchmark portfolios. Multiple portfolio benchmarks should be more reliable than single portfolio benchmarks and our own research suggests that forming factor portfolios based on the characteristics of stocks (like dividend yields and size) may be more reliable than using factor analysis to form benchmarks. However, the formation of benchmark portfolios is likely to remain controversial. It is always difficult to verify that your factor portfolios adequately capture all relevant factors. Perhaps future research that combines the various approaches will develop factor portfolios that are more widely accepted.

Although portfolio weights may not always be available to academic researchers, they are available to practitioners who are in the business of evaluating performance. In the past, this information may have been ignored because of the costs of handling such large data bases. However, computing power is now very cheap and the necessary software is now readily available so we would expect portfolio holdings data to be utilized more in the future. When portfolio holdings are used, a unique passive portfolio, based on the evaluated portfolio's past holdings, can be constructed for each point in time and each fund. We think that these individualized benchmarks are more reliable than the single benchmarks used in the more traditional approach.

Other issues also tend to favor the portfolio holdings methodology. If the evaluated investor has the ability to successfully time the benchmark portfolio, evaluation techniques based on the asset pricing methodology may generate biased estimates of risk. Although a number of asset pricing based measures have been developed that account for timing, each require that returns be normally distributed. The basic problem is that it is difficult to distinguish a managed portfolio with returns that are positively co-skewed with the benchmark portfolio from one that really can time the market. The portfolio holdings methodology cannot be subject to this timing-related bias because it does not require a benchmark portfolio.

Given the problems associated with evaluating timing performance it is not surprising that at this point there is no convincing evidence of mutual funds systematically timing the market. However, there is evidence that some mutual funds consistently achieve abnormal returns by systematically picking stocks that subsequently do well. How should we interpret this evidence?

One view is that there are certain skilled investors who are very good at uncovering and interpreting fundamental information. This view would suggest that the market is only weakly efficient; smart investors are earning what looks like a lot of money, but they have to be talented and they have to earn it. The second view is that the abnormal performance was generated by technical trading rules that exploit what we will call time-series anomalies. The performance measures examined in this chapter are designed to eliminate the possibility that abnormal performance is generated by exploiting what we call cross-sectional anomalies, like the size effect. However, since the techniques essentially compare managed returns to equivalent passive returns, abnormal performance can be generated by

active trading if return distributions are not stationary, i.e., stocks with high past returns also have high future returns.[18]

For a variety reasons, it is important to understand the extent to which abnormal performance is generated as a result of technical trading rules rather than fundamental analysis. First, if the abnormal performance is generated by very simple technical rules rather than fundamental analysis, we would be less willing to attribute the performance to skill rather than luck. In addition, if the nonstationarities observed in past studies are due to market inefficiencies, they are likely to disappear over time. For these reasons, we would probably be less willing to hire a portfolio manager based on his past performance if we thought the performance was generated from exploiting simple technical rules. We would think that performance generated by careful fundamental analysis would be more likely to persist over time.

Grinblatt, Titman & Wermers [1995] examines the extent to which mutual funds generated abnormal performance by exploiting momemtum strategies.[19] The paper shows that mutual funds on average do have a tendency to buy stocks that did well in the past, and that this tendency is greatest among the Aggressive Growth funds, which was the category that showed the best performance. Moreover, the study shows that mutual funds that did not show this tendency to buy past winners did not realize significant abnormal performance. This evidence suggests that at least part of the abnormal performance of these funds comes from their tendency to buy past winners.

Research on portfolio performance evaluation has clearly progressed over the past 10 years, benefiting tremendously from the recent advances in the asset pricing literature. We expect similar strides to be made in the next 10 years. This area of research should benefit from the availability of much better mutual fund data sets that are both broader, in terms of the number of funds included, and longer, in terms of the length of the time-series. This literature should also benefit from our increased understanding of both the cross-sectional and time-series properties of stock returns which should enable us to develop new performance measures that account for both. With improved data and improved measures researchers should be able to achieve a very good understanding of what determines superior portfolio performance.

References

Admati, A., and S. Ross (1985). Measuring investment performance in a rational expectations equilibrium model. *J. Bus.* 58, 1–26.

Admati, A., S. Bhattacharya, P. Pfleiderer and S. Ross (1986). On timing and selectivity. *J. Finance* 41, 715–30.

[18] The P8 benchmark mitigates this somewhat by using a low past returns portfolio for estimating risk.

[19] Jegadeesh & Titman [1993] shows that strategies that buy stocks that performed well over the past 3, 6, 9 or 12 months continue to outperform the market over the subsequent 12 months.

Beebower, G., and G. Bergstrom (1977). A performance analysis of pension and profit-sharing portfolios: 1966–1975. *Financ. Anal. J.* (May/June), 31–38.
Brown, S., and W. Geotzmann (1995). Performance persistence, *J. Finance* 50, 679–98.
Brown, S., W. Goetzmann, R. Ibbotson and S. Ross (1992). Survivorship bias in performance studies. *Rev. Financ. Studies* 5, 553–580.
Chang, E., and W. Lewellen (1984). Market timing and mutual fund investment performance. *J. Bus.* 57, 57–72.
Chen, N.-f., T. Copeland and D. Mayers (1987). A comparison of single and multifactor portfolio performance methodologies. *J. Financ. Quant. Anal.* 22, 401–17.
Connor, G., and R. Korajzcyk (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *J. Financ. Econ.* 15, 374–94.
Connor, G., and R. Korajzcyk (1991). The attributes, behavior and performance of U.S. mutual funds. *Rev. Quant. Finance Account.* 1, 5–26.
Connor, G., and R. Korajzcyk (1995). The arbitrage pricing theory and multifactor models of asset returns, in: R. Jarrow, V. Maksimovic and W.T. Ziemba (eds.), *Finance*, Handbooks in Operations Research and Management Science, Vol. 9, North-Holland, Amsterdam, pp. 87–144 (this volume).
Copeland, T., and D. Mayers (1982). The Value Line enigma (1965–1978): A case study of performance evaluation issues. *J. Financ. Econ.* 10, 289–321.
Cornell, B. (1979). Asymmetric information and portfolio performance measurement. *J. Financ. Econ.* 7, 381–390.
Cumby, R., and J. Glen (1990). Evaluating the performance of international mutual funds. *J. Finance* 45, 497–521.
De Bondt, W.F.M., and R.H. Thaler (1995). Financial decision-making in markets and firms: A behavioral perspective, in: R. Jarrow, V. Maksimovic and W.T. Ziemba (eds.), *Finance*, Handbooks in Operations Research and Management Science, Vol. 9, North Holland, Amsterdam, pp. 385–410 (this volume).
Dybvig, P., and S. Ross (1985a). Differential information and performance measurement using a security market line. *J. Finance* 40, 383–399.
Dybvig, P., and S. Ross (1985b). The analytics of performance measurement using a security market line. *J. Finance* 40, 401–16
Elton, E., M. Gruber, S. Das and M. Hlavka (1993). Efficiency with costly information: A reinterpretation of evidence from managed portfolios. *Rev. Financ. Studies* 6(1), 1–22.
Ferson, W.E. (1995). Theory and empirical testing of asset pricing models, in: R. Jarrow, V. Maksimovic and W.T. Ziemba (eds.), *Finance*, Handbooks in Operations Research and Management Science, North Holland, Vol. 9, Amsterdam, pp. 145–200 (this volume).
Grant, D. (1977). Portfolio performace and the 'cost' of timing decisions. *J. Finance* 32, 837–46.
Grinblatt, M., and S. Titman (1987). The relation between mean-variance efficiency and arbitrage pricing. *J. Bus.* 60, 97–112.
Grinblatt, M., and S. Titman (1988). The evaluation of mutual fund performance: An analysis of monthly returns, working paper, University of California.
Grinblatt, M., and S. Titman (1989a). Mutual fund performance: An analysis of quarterly portfolio holdings. *J. Bus.* 62, 393–416.
Grinblatt, M., and S. Titman (1989b). Portfolio performance evaluation: Old issues and new insights. *Rev. Financ. Studies* 2(3), 393–421.
Grinblatt, M., and S. Titman (1992). The persistence of mutual fund performance. *J. Finance* 47, 1977–1984.
Grinblatt, M., and S. Titman (1993). Performance measurement without benchmarks: An examination of mutual fund returns. *J. Bus.* 66, 47–68.
Grinblatt, M., and S. Titman (1994). A study of monthly mutual fund returns and performance evaluation techniques. *J. Financ. Quant. Anal.* 29, in press.
Grinblatt, M., S. Titman and R. Wermers (1995). Momentum investment strategies, portfolio performance and herding: A study of mutual fund behavior, *Am. Econ. Rev.*, forthcoming.

Hawawini, G., and D.B. Keim (1995). On the predictability of common stock returns: World-wide evidence, in: R. Jarrow, V. Maksimovic and W.T. Ziemba (eds.), *Finance*, Handbooks in Operations Research and Management Science, Vol. 9, North Holland, Amsterdam, (this volume).
Hendricks, D., J. Patel and R. Zeckhauser (1993). Hot hands in mutual funds: The persistence of performance, 1974–88. *J. Finance* 48, 93–130.
Henriksson, R. (1984). Market timing and mutual fund performance. *J. Bus* 57, 73–96.
Henriksson, R., and R. Merton (1981). On market timing and investment performance II: Statistical procedures for evaluating forecasting skills. *J. Bus.* 54, 513–33.
Ippolito, R. (1989). Efficiency with costly information: A study of mutual fund performance, 1965–84. *Q. J. Econ.* 104, 1–23.
Jagannathan, R., and R.A. Korajczyk (1986). Assessing the market timing performance of managed portfolios. *J. Bus.* 59, 217–235.
Jegadeesh, N., and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *J. Finance* 48, 65–91.
Jensen, M. (1968). The performance of mutual funds in the period 1945–1964. *J. Finance* 23, 389–416.
Jensen, M. (1969). Risk, the pricing of capital assets, and the evaluation of investment portfolios. *J. Bus.* 42, 167–247.
Jensen, M. (1972). Optimal utilization of market forecasts and the evaluation of investment performance, in: G.P. Szego and K. Shell (eds.), *Mathematical Methods in Investment and Finance*. Elsevier, Amsterdam.
Kon, S. (1983). The market-timing performance of mutual fund managers. *J. Bus.* 56, 323–48.
Kon, S., and F. Jen (1979). The investment performance of mutual funds: An empirical investigation of timing, selectivity, and market efficiency. *J. Bus.* 52, 263–89.
Kraus, A., and R. Litzenberger (1976). Skewness preference and the valuation of risk assets. *J. Finance* 311, 1085–1100.
Lee, C.-f., and S. Rahman (1990). Market timing, selectivity, and mutual fund performance: An empirical investigation. *J. Bus.* 63, 261–78.
Lehmann, B., and D. Modest (1987). Mutual fund performance evaluation: A comparison of benchmarks and benchmark comparisons. *J. Finance* 42, 233–265.
Lehmann, B., and D. Modest (1989). The empirical foundations of the arbitrage pricing theory. *J. Financ. Econ.* 21, 213–54.
Mains, N. (1977). Risk, the pricing of capital assets, and the evaluation of investment portfolios: Comment. *J. Bus.* 50, 371–84.
Malkiel, B. (1995). Returns from investing in equity mutual funds 1971–1991, *J. Finance* 50, 549–72.
Mayers, D., and E. Rice (1979). Measuring portfolio performance and the empirical content of pricing models. *J. Financ. Econ.* 7, 3–28.
McDonald, J. (1974). Objectives and performance of mutual funds, 1960–1969. *J. Financ. Quant. Anal.* 9, 311–33.
Merton, R. (1981). On market timing and investment performance I: An equilibrium theory of value for market forecasts. *J. Bus.* 54, 363–406.
Roll, R. (1978). Ambiguity when performance is measured by the securities market line. *J. Finance* 33, 1051–69.
Sharpe, W. (1966). Mutual fund performance. *J. Bus.* 39, 119–38.
Sharpe, W. (1992). Asset allocation: Management style and performance measurement. *J. Portfolio Manage.* 18, 7–19.
Treynor, J. (1965). How to rate management of investment funds. *Harvard Bus. Rev.* 43, 63–75.
Treynor, J., and F. Black (1973). How to use security analysis to improve portfolio selection. *J. Bus.* 46, 66–86.
Treynor, J., and K. Mazuy (1966). Can mutual funds outguess the market? *Harvard Bus. Rev.* 44, 131–36.
Verrecchia, R. (1980). The Mayers–Rice conjecture. A counterexample. *J. Financ. Econ.* 8, 87–100.