Performance Evaluation

Kevin C. Kaufhold
2006:12; Latest Update Jan. 2012

Framework for Evaluation [1]

The general goals for portfolio managers include the ability to derive above average returns for a given risk class, the ability to diversify to eliminate unsystematic risk, appropriate asset allocation, and portfolio composition.

In evaluating performance, several factors other than return will affect performance, and should be taken into account. These include:

- Differential risk levels
- Different time periods
- Comparison to the relevant benchmark
- Objectives and constraints of the investor and / or manager.

Return Calculations

Generally, the investor is normally concerned with total change in wealth. The proper measure for this is total return, which captures both the capital and income components of return. In accord with this general belief, the CFA Institute has developed comprehensive standards governing performance evaluation. The standards use total return with time weighting on a post-tax, post expense basis.

Dollar weighted rates of return equates all cash flows with the beginning value of the portfolio. This is the traditional measure of performance, and constitutes the Internal rate of return (IRR). This measure should not be used to compare to benchmarks or to other portfolios, since it measures an investor's return inclusive of all inflows and outflows of CF (which the manager does not have any control over) and thus can distort the return assessment. This measure is appropriate when only the perspective of the portfolio manager is considered.

A time-weighted return measures the actual rate of return earned by an investment manager, and thus is more appropriate to use as a comparison to a benchmark or to other portfolios. The AIMR uses a time-weighted rate of return, with geometric linking of periods, since time weighted returns can generate a return regardless of size of funds. CF's are excluded from the calculations, and only the return on the portfolio is actually measured. This measure is more appropriate when comparing the manager's performance.

[1] This paper has been drawn from several standard portfolio textbooks, including: Jones (2002); Reilly and Brown (2000); Bodie, Kane and Marcus (2004); and Gruber, Elton, et al (2002).

Fund performance is often broken down by levels of risk. Using equally weighted return averages for funds of low, medium, and high risk, a mean return and variance can be generated on each of the three classes of funds. Then, an individual fund can be compared with funds of the same risk class (Elton, Gruber, at 621).

Portfolio Evaluation Measurements

Several portfolio evaluation measurements have been developed over the years. Portfolio evaluation prior to the early 1960's was generally done on the basis of return, with little thought given to non-systematic risk. With the advent of several of the following measures and rations, the emphasis has changed towards that of risk-adjusted returns. The following is a summary of the various performance related measures that are used with professional level portfolios.

Peer Group comparisons. The portfolio performance will be compared with peers of the same style and length of time. There is no express adjustment for risk.

The Treynor Measure. Jack Treynor developed the first composite measure of portfolio performance that included risk. A characteristic line is developed that compares and defines the risk - return relationship. The slope of the line is the beta coefficient. Deviations from the line indicate unique returns for the portfolio versus the market. Treynor used the Capital Asset Pricing Theory to show that risk averse investors would always prefer characteristic lines with larger slopes because the investors would move to higher indifference lines. $T = (R_i - RFRR) / \beta_i$, where R = the average return rate for the portfolio. RDR = the average return rate for the RRFR; and B = the port's relative volatility or beta. The risk premium then is $R_i - RFRR_i$. The Treynor measure actually looks at the risk premium per unit of volatility risk with respect to the market risk. All investors would want to maximize this value. The measure assumes diversification of assets. The characteristic line intercept would be the T-bill rate, and the slope of the SML would be the beta coefficient.

T2 is a variant of Treynor, where the beta of the portfolio is matched to the market by mixing in T-bills, similar to the M2 method. Then, the returns of the adjusted port and the market can be made directly.

The Sharpe Ratio.[2] The most widely known and used measurement is the Sharpe measure, which is: $S = (r_p - r_f) / \sigma_p$. Sharpe used standard deviation instead of beta, as Treynor did. Otherwise the formula is the same as Treynor. The Sharpe ratio seeks to measure the total risk by including standard deviation, or the spread between rates of return, rather than the systematic risk of beta. It measures the risk premium per unit of total risk. Sharpe therefore would compare to the CML instead of Treynor's SML. The

[2] William F. Sharpe, "Mutual Fund Performance," *Journal of Business* 39, no. 1, part 2, January 1966:119-138.

Sharpe Ratio is also known as the reward to risk ratio, or more descriptively, the excess return to variability measure.

A variant of the Sharpe ratio is the M2 measure, developed in part by Modigliani and his granddaughter. This measure provides for a better interpretation of the standard deviation result of Sharpe. This method adjusts the portfolio return to match the volatility of the portfolio with the market index by introducing T-bills into the portfolio until the port deviation equals the market. Then, returns of the port are compared to that of the market to see if the port has out-performed. Basically, we would move down the CML from the optimal point until the standard deviation of the port and the market were the same. Then, the returns of the port and the market could be directly compared.

Elton, Gruber, at 630, then takes this process further with a Differential return measure. A regression line is initially established for funds or portfolios with varying risk and returns. Each fund's return is compared with the return of the regression line for the same level of risk. Funds are then ranked by their differential returns to the regression line.

The Jensen Measure. This is based on the CAPM, too.[3] This measure is concerned with the time series of expected returns. Jensen defines abnormal returns as alpha, or α p. Then, $ri - rf = \alpha + \beta (rm - rf)$, which is generally in the form of $y = a + bx$. The intercept is then mathematically derived as: $\alpha_p = (Rp - RF) - (B_p (R_m - RF))$. Alpha becomes the difference between the excess return on portfolio p and the risk premium on the portfolio, given the portfolio's level of risk.

Superior managers are able to consistently select undervalued securities and thus earn higher risk premiums than those implied by the CAPM model. A positive intercept in the equation is caused by positive differences from the model, while a negative intercept is caused by negative differences from the model. Even where alpha is positive or negative, however, it may still not be statistically significant. The regression line of Jensen will sketch out the characteristic line, and levels of confidence can then be obtained.

The degree of intercept off of zero, expressed as $\acute{\alpha}_j$ (or lambda of security j), is the measure of superior or inferior performance. It measures how much return is attributable to the manager ability to derive above average returns adjusted for risk. The positive or excess returns of the port compared to a benchmark are considered to be alphas. The Jensen measure then can be applied to the market model, with the Y intercept as alpha. If alpha is positive, then outperformance has occurred on a risk adjusted basis. Sharpe suggests to then measure the statistical significance of alpha, since outperformance may be statistically insignificant.

The Information Ratio (IR). This was recently developed by Theodore H. Goodwin. This is also called the Appraisal Ratio. It measures the average return in excess of a benchmark. $IRj = Erj / \sigma_{Er}$. Then, this is $= (Rj - Rb) / \sigma_{Er}$, where Rj is the return of the

[3] (Michael C. Jensen, "The Performance of Mutual Finds in the Period 1945 – 64", *Journal of Finance* 23, no. 2 (May 1968), 389-416).

port, Rb is the return of the benchmark, and ER is the excess return above the benchmark. This equation is a measure of outperformance / unsystematic risk. The deviation of the excess return is a measure of the tracking error. IR is a benefit to cost ratio assessing the quality of information deflated by the unsystematic risk. Goodwin felt that the Sharpe ratio may be thought of as a special case of the IR where the RFRR is the benchmark. If excess returns are estimated with historical data using Jensen's alpha, then IR becomes = outperformance / standard deviation error of regression.

R^2 (R square). When used in the context of comparison to a market benchmark index, the coefficient of determination, cv, can be stated as R^2, which measures the extent of co-movement between the portfolio and a market-level index. As the R square value approaches 1.00, the more complete the diversification, but the more similar the portfolio is to the benchmark index. If the R^2 is over 90%, the portfolio or mutual fund under review may be considered a "closet" index fund, but with the extra costs of active management. Diversification can be measured by the, and this is measured by R^2. This indicates the % of variance of the portfolio returns explained by market returns. As R^2 approaches 1.00, returns are completely explained by the market.

Style analysis, initially thought up by Sharpe, can also be used for management evaluation.[4] This measure is important in monitoring and preventing style drift of a manager.

Another measure of style is the Morningstar Risk-Adjusted Rating (RAR). The rating produces a star ranking, from 1 to 5. The ratings will compare funds within peer groups, providing more stars and a higher rating for funds in the upper percentiles of the peer group. The ratings use procedures similar to the Sharpe ratio. A high rating is greatly prized among the active funds.

The Appraisal Ratio stems from the Treynor-Black Model, [5] which uses means-variance to optimize active management decisions. The ratio uses the Sharpe ratio: $S(P) = [S^2(M) + \alpha^2_A / \sigma^2 (e_A)]^{1/2}$. The critical value is the comparison between alpha to nonsystematic risk, $\alpha^2_A / \sigma^2 (e_A)$. The position that is takes by the active portfolio depends upon the ratio of the active portfolio's excess returns, α^2_A, to the diversifiable risk, $\sigma^2 (e_A)$. The higher the alpha in relation to the diversifiable risk, the greater should be the percentage of the active portfolio in a combined fund composed of both active and passive investments.

Components of Investment Performance. Fama suggests that overall performance is the excess return beyond a RFRR, with that being: Er + RFRR = port risk + selectivity. Fama looks at component parts of a CML, as being investor's risk, manager's risk, selectivity, diversification, all resulting in what he calls "net selectivity". That is then

[4] William F. Sharpe, "Asset Allocation: Management Style and Performance Evaluation," *Journal of Portfolio Management*, Winter 1992: 7-19.

[5] The model is described in Bodie, Kane, and Marcus, at 709-712; For the original article, see: Jack L. Treynor and Fischer Black, "How to Use Security Analysis to Improve Portfolio Selection," *Journal of Business* 46 (January 1973),

compared to the required rate of return for the risk assumed by the manager. If the overall performance > the required return for the risk, the port has experienced a positive return for the selectivity. You can also do this for lack of diversification too, to arrive at the positive or negative return for the net selectivity. So, it is best to consider composite performance figures collectively. There is a close correlation between the various performance measures.

Performance Attribution Analysis. Attempts have been made to identify the factors that are a source of the port's overall performance. Allocation effects measure the manager's decision to overweight or underweight a sector or segment. The selection effect measures the manager's ability to form superior returns relative to the way in which the market is defined. So, you can then compare whether active management is costing a port or helping the port. This is a value added concept in relation to the benchmark. Two studies showed that the initial strategic AA choice rather than any of its active management decisions was the primary determinant of portfolio performance.

Measuring Market Timing Skills. Tactical AA shifts between stocks, bonds, and cash. Attribute Analysis is ill suited to TAA since tactical attempts to index the asset classes, so there will be no selection effect. It will all be allocation effect. Managers may attempt to improve performance by changing the allocation percentage in anticipation of a market change. Timing can also occur by adjusting the average beta in anticipation of market changes. Equity allocations and/or beta are increased (decreased) in expected equity market advances (declines). One study evaluates TAA by looking back to returns of the best performance asset class. It showed .30 (out of 1.00) timing ability by the managers of TAA funds. Another study (Treynor and Mazuy) found that only one fund out of 37 exhibited any significant timing ability. The equation used was: $Rit - Rft = ai + bi\ (Rm - Rft) + ci(\ Rm - Rft)^2 + ei.$ Ai, bi, and ci are constants, and the equation tested for fit in a quadratic curve.

Bond Performance. A bond market line can be divided into return for: 1) investment policy effect; 2) rate anticipation effect; 3) analysis effect; 4) trading effect. One study showed that .35% was for investment policy; .40% for interest rates; .40% for analysis; and 1.10% for trading effect. This does not consider however the differences in the risk of default. Another study showed that known maturity yield effects occurred, coupled with unknown effects for quality /sector considerations, interest rates, and a residual effect for superior bond selection capabilities. Still another study shows that $R = I + C,$ where $I = E + U$ and $C = M + S + B,$ where E is the expected treasury yield, U = the unexpected treasury yield, M = the maturity management, S= spread and quality management, and B = selectivity. Also one can measure the performance of bonds. One study showed no relationship between past price performance and future performance of bond pricing (ala EMH for equities).

Multi-Index Performance Measures. Many funds hold assets that go beyond a single index benchmark comparison. For example, a S&P 500 benchmark may not entirely appropriate for use with a primarily large-cap fund that also holds a significant portion of assets in small caps. One multi-index approach is that of Morningstar, which categorizes

funds by size of portfolio assets and fund style. A multi-benchmark could calculate alpha by (from Elton, Gruber, at 646): $R_{it} - R_{ft} = a_i + b_{iL} (R_{Lt} - R_{ft}) + b_{is} (R_{st} - R_{ft}) + + b_{iB} (R_B - R_{ft}) + eit.$ This can be viewed as a generalization of the Jensen measure.

APT models. Multi-index measures combined with APT models can also be developed for performance measurement. See, 2006:11 for more details on the APT.

A Comparison of Performance Measurements

The various risk-adjusted measures can be compared. Sharpe uses standard deviation, while both Treynor and Jensen are based on the CAPM, with beta being a part of the equation. There can be situations where there will be disagreement between the measures. This occurs where beta is so low, or so high, to overcome standard deviation differences. If diversification possibilities are limited, or where only one asset is being reviewed (such as with a private business), then Sharpe is more appropriate to use. Where an entire portfolio of assets is being analyzed, then Treynor and Jensen is better, with their references to the market beta.

With a completely diversified portfolio, both the Sharpe and Treynor measures would be identical because the total variance of the port would only be systematic in nature. The difference between Treynor and Sharpe is thus the difference in diversification levels of the portfolio. Both measures could be used to ascertain diversification levels of the portfolio. If Sharpe and Treynor differ in value, then the portfolio is not completely diversified. Sharpe is a measure of how well a fund is diversified, since a portfolio will contain only systematic risk levels when it is as diversified as the entire market. Sharpe should be used when measuring an entire portfolio, because it measures total return to total risk, including any unsystematic risk in the portfolio. Treynor should be used where a portfolio contains only a small portion of an investor's total assets, since it measures the beta of the particular assets in the portfolio.

Sharpe, Treynor, and Jensen can also be compared. Sharpe measures return and diversification. Treynor measures only the systematic risk of the portfolio. They can both be used to rank portfolio's return on a risk-adjusted basis, but cannot tell us the percentage that a fund outperformed. Jensen's alpha does not rank a fund's performance, but can be made to be identical to Treynor, with some modification, since both are based on the CAPM and beta. If a portfolio is completely diversified, all three measures will be identical. Without complete diversification, Treynor and Jensen measures undiversifed portfolios and will rank them much higher than Sharpe. Shaper and Treynor both use average returns over a measurement period. Jensen uses period-by-period returns.

Difficulties in the Measurements. A common problem with all performance measures is that they assume constant or uniform portfolio risk measures over time. But, the standard deviation or betas of the assets themselves may change over time, thus shifting the risk-adjusted performance measurements. On a related concern, funds that deliberately change their risk exposure (i.e. market timing strategies) will cause a fund's evaluation in time

period 1 to be skewed or non-relevant for time period 2, since the evaluation in two different time periods will have two different base-risk levels for the fund.

Also note that Jensen uses a different RFRR for each time period versus Treynor and Sharpe using an average RFRR for the total of all variables. Jensen may therefore have somewhat different conclusions due to the use of different inputs and data.

There is also a potential bias of the measures. One study showed a significant inverse relationship of low risk portfolios and out-performance versus high risk and under-performance. Another study showed a positive relationship, however.

Almost all measures use the CAPM and the assumptions of the efficient frontier. There is a problem with a proxy for risky assets being a benchmark. This is referred to as the benchmark error. There are also global benchmark problems, since there are various SML's for the various international indexes. There is a possibility of multiple market indexes, which merge numerous world markets together. Also, there is a search for customized benchmarks reflecting the style of alternative managers.

Elton, Gruber, at 634-635 also notes that empirical Beta lines have lower slopes than do theoretical CAPM driven beta lines, with the y intercept occurring above that of the RFRR. The empirical line thus intersects the theoretical line at some point, since it starts out at a higher Ri but then has a lower slope to the theoretical line. Using the empirical line would generally result in more negative differential return evaluations of funds, through most of the return versus beta dimensional space.

Other problems also exist. Most people do not borrow or lend at the riskless rate, as assumed with most of the performance measures. The regions of dominance would be affected by a change in riskless rate assumptions.

Also, the performance period may be too short to truly measure superior performance. Luck can result by too short of a time period to measure performance. At least 10 years is needed to measure performance according to CFA standards.

When analyzing load mutual funds, load funds overstate performance abilities compared with no-load funds. This is due to the load fee taken from the investor at the initial purchase.

Performance evaluation also suffers from survivorship bias, since funds in existence throughout the test period are the only funds studied.

<u>Results of Performance studies.</u> In numerous studies, actively managed mutual funds generally have been found to have poorer performance than a simple strategy of random selection of passive portfolios and the riskless asset. No-load mutual funds performed better than load funds, but still have a negative alpha. Active funds with low turnover fare better than high turnover funds, but all active funds of varying turnover will have negative alpha. Active funds with low expense ratios do better than high-expense funds,

but even the low expense funds possess negative alpha. The asset size of the fund was found to be inconclusive of performance. Bond funds have a similar pattern of performance problems. (Elton, Gruber, at 651-655). The general recommendation is therefore to invest in funds that are no-load, having low expenses ratios and low turnover. This generally implies passive index funds, although some value-oriented funds may also be considered to be low expense, low turnover in nature.

References

Zvi Bodie, Alex Kane, and Alan J. Marcus, *Essentials of Investments,* McGraw-Hill Irwin, 2004

Edwin J. Elton, Martin J. Gruber, Stephen J. Brown, and William N. Goetzmann, *Modern Portfolio Theory and Investment Analysis,* John Wiley & Sons, 6th ed., 2003.

Eugene F. Fama, "Components of Investment Performance," *Journal of Finance* 27, no. 3, June 1972: 551-567.

Theodore H. Goodwin, "The Information Ratio," *Finance Analysis Journal* 54, no. 4, July–Aug 1998: 34-43.

Michael C. Jensen, "The Performance of Mutual Finds in the Period 1945 – 64," *Journal of Finance* 23, no. 2 (May 1968): 389-416.

Charles P. Jones, *Investments: Analysis and Management,* Wiley & Sons, 9th ed. 2004.

Franco Modigliani and Leah Modigliani, "Risk-Adjusted Performance," *Journal of Portfolio Management,* Winter 1997: 45-54.

Frank K. Reilly, and Keith C. Brown, *Investment Analysis and Portfolio Management,* The Dryden Press, 6th Ed., 2000.

William F. Sharpe, "Mutual Fund Performance," *Journal of Business* 39, no. 1, part 2, January 1966:119-138.

William F. Sharpe, "Asset Allocation: Management Style and Performance Evaluation," *Journal of Portfolio Management,* Winter 1992: 7-19.

Jack L. Treynor, "How to Rate Management of Investment Funds," *Harvard Business Review* 43, no. 1, Jan-Feb 1965: 63-75.

Jack L. Treynor and Fischer Black, "How to Use Security Analysis to Improve Portfolio Selection," *Journal of Business* 46 (January 1973).