

How Do Performance Measures Perform?

March 2007



Georges Hübner

Affiliate Professor of Finance, EDHEC Business School.

Abstract

The relevance of the information ratio and the alpha, two leading performance measures for multi-index models, depends on the type of portfolio held by investors. We compare these measures with the generalized treynor ratio (*GTR*) on the quality of the rankings they produce. A precise measure yields similar rankings with alternative benchmarks. A stable measure leaves unchanged rankings with different model specifications. The outcome indicates the types of skills emphasized by portfolio managers. The *GTR* provides superior results with our sample of mutual funds, suggesting that managerial skills relate to the ability to generate alpha while controlling for systematic risk.

I thank Pascal François, Eric Jacquier, Denis Larocque, Nicolas Papageorgiou and Peter Schotman for helpful comments. The author acknowledges financial support from Deloitte Luxemburg and a research grant of the Belgian Funds for Scientific Research (FNRS). Part of this research was done while I was visiting HEC Montréal. All errors remaining are mine.

EDHEC is one of the top five business schools in France. Its reputation is built on the high quality of its faculty (110 professors and researchers from France and abroad) and the privileged relationship with professionals that the school has cultivated since its establishment in 1906. EDHEC Business School has decided to draw on its extensive knowledge of the professional environment and has therefore focused its research on themes that satisfy the needs of professionals.

EDHEC pursues an active research policy in the field of finance. The EDHEC Risk and Asset Management Research Centre carries out numerous research programmes in the areas of asset allocation and risk management in both the traditional and alternative investment universes.

How do performance measures perform ?

Among the performance measures for managed portfolios with directional strategies developed in the framework of the capital asset pricing Model proposed by Treynor (1961), Sharpe (1964) and Lintner (1965), three of them directly relate to the beta of the portfolio through the security market line (SML). Jensen's (1968) alpha is defined as the portfolio excess return earned in addition to the required average return, while the Treynor (1965) ratio and the information ratio (IR) are defined as the alpha divided by the portfolio beta and by the standard deviation of the portfolio residual returns. There exists no other widely used alternative measure that sticks to the SML. Indeed, most recent performance measures developed along with the increasing popularity of hedge funds, such as the Sortino ratio (Sortino and Price [1994]), the M2 (Modigliani and Modigliani [1997]) and the Omega (Keating and Shadwick [2002]) focus on a measure of total risk, in the continuation of the Sharpe (1966) ratio applied to the capital market line.

In the context of the extension of the CAPM to linear multi-factor asset pricing models, performance measurement has not been very innovative. Empirical research has only made use of a single measure that uses systematic risk as an ingredient, namely the portfolio abnormal return computed in the same way as Jensen's (1968) alpha. Being the intercept of a linear regression, the alpha is flexible enough to be used in most asset pricing specifications. As a matter of fact, Kothari and Warner (2001) stick to this performance measure in their empirical comparison of multi-index asset pricing models. Thereby, they implicitly assume that the alpha represents the leading performance metric based on systematic risk for multi-index models.

Meanwhile, in their quest for positive abnormal returns accompanying active portfolio management strategies, many finance practitioners have gradually adopted the information ratio (IR) as a standard performance measure (see, Grinold and Kahn [1992]; Goodwin [1998]). Underlying this choice, the IR of a portfolio is claimed to provide *ex ante* a proxy for its return potential, and thus represents *ex post* a consistent tool for assessing the manager's ability to realize abnormal returns.

To analyze the empirical quality of these various measures, *i.e.* how do performance measures perform, it is important to understand what do these performance measures try to measure. If an investor chooses a managed portfolio which is her exclusive investment vehicle, only total risk should matter, and the Sharpe ratio or any measure based on total portfolio risk is relevant. If the investor holds several funds, the performance evaluation of a particular portfolio depends on the type of complementary portfolio held by the investor, *i.e.* the rest of her investment vehicles. If she holds a passive portfolio like an index fund, the performance of the managed portfolio is measured by the information ratio. On the other hand, if the portfolio under study is just one among many other actively managed portfolios, then the alpha of the portfolio rightly measures its contribution to the overall performance. In this context, as emphasized by Bodie, Kane and Marcus [2005], "*an even better solution, however, is to use Treynor's measure*".

For portfolio managers themselves, the motivations for striving to generate high information ratios, high alpha or high Treynor ratios are likely to be different. Consider a style-based managed portfolio. This specialization precludes the use of a performance measure based on total risk, such as the Sharpe ratio, because no rational investor would use this portfolio as her sole diversification vehicle. The absolute performance potential of this fund, measured by the alpha, is directly related to the freedom given to the manager to take specific risks. Therefore, the quality of the manager should be assessed through the information ratio. On the other hand, if the manager considers that her clientele focuses on the fund's absolute performance—as is commonly argued in the hedge fund industry—she cares about her alpha. Provided that fees are related to the fund's excess return over a benchmark, the manager might even be induced to enhance the level of the alpha by magnifying systematic risk exposures.

To identify the dimensions along which portfolio managers differentiate themselves, this article examines the empirical performance of performance measures for multi-index models through the quality of the ranking scheme they produce. We use a sample of directional mutual funds with different styles, so that the *IR*, the alpha and the generalized Treynor ratio proposed by Hübner (2005) can adequately measure performance. The quality of a performance measure is assessed through two dimensions: its precision in reproducing the true ranking and the stability of the rankings it produces under alternative asset pricing models. These two dimensions provide complementary information about which performance measure is most likely best to describe managerial skills in active portfolio management.

On the practical side, there is nonetheless a serious difficulty raised with the assessment of the quality of performance measures. Statistical inference with measures based on ratios, such as the Treynor performance measure, is rather ticklish. If the denominator tends to zero, the Treynor ratio goes to infinity. In particular, this measure provides unstable and imprecise performance measures for non-directional portfolios, such as market neutral hedge funds. When the denominator is negative, the economic interpretation of the Treynor ratio is even corrupt as it would assign positive performance to portfolios with negative abnormal returns. Even for directional portfolio strategies, the risk of measurement error does not bound the denominator away from zero and the expectation of the ratio is therefore infinite. This undesirable behavior only disappears asymptotically, but this property is hardly usable for most practical applications.

We tackle this empirical issue in three ways. The dataset used in the study, with only directional managed portfolios, guarantees a suitable context for the application of the generalizations of Jensen's [1968] alpha, of the information ratio and of the Treynor ratio. Furthermore, to ensure fairness of the comparisons between performance measures, we control for the risk exposures and measurement errors of the passive index portfolios used in the study. Finally, the empirical methods allow us to avoid most of the drawbacks generated by the use of ratios, by looking at ordinal relationships (portfolio rankings) and using non-parametric concordance measures that are relatively insensitive to outliers.

Performance measurement in multi-index models

We consider performance measures that apply to the *ex post* version of the security market line (SML). The empirical derivation of the SML corresponds to the market model $r_{it} = \alpha_i + \beta_i r_{mt} + \varepsilon_{it}$ where $r_i = R_i - R_f$ denotes the excess return on security i . The *ex post* (realized) equation is:

$$\bar{r}_i = \hat{\alpha}_i + \hat{\beta}_i \bar{r}_m \quad (1)$$

where $\hat{\beta}_i = \frac{\widehat{\text{cov}}(r_i, r_m)}{\widehat{\sigma}^2(r_m)}$ and $\hat{\alpha}_i = \bar{r}_i - \hat{\beta}_i \bar{r}_m$ are the estimators of β_i and α_i , respectively.

Jensen's alpha is measured by $\hat{\alpha}_i$ in equation (1). As originally defined, the Treynor ratio (*TR*) is the ratio of Jensen's alpha over the stock beta: $TR_i = \frac{\hat{\alpha}_i}{\hat{\beta}_i}$. Finally, the Information Ratio (*IR*) is a measure of Jensen's alpha per unit of portfolio specific risk, measured as the standard deviation of the market model residuals: $IR_i = \frac{\hat{\alpha}_i}{\hat{\sigma}(\varepsilon_i)}$.

For many portfolios, the use of a single index in a market model is not sufficient to keep track of the systematic sources of portfolio returns in excess of the risk free rate. The families of linear multi-index unconditional asset pricing models are numerous and heterogenous. However, in spite of their differences, these approaches share a common affine model specification that can be summarized by this *ex-post* multidimensional equation:

$$\bar{r}_i = \hat{\alpha}_i + \sum_{j=1}^K \hat{\beta}_{ij} \bar{r}_j = \hat{\alpha}_i + \hat{\mathbf{B}}_i \bar{\mathbf{R}} \quad (2)$$

where $j = 1, \dots, K$ denotes the number of distinct risk factors, the line vector $\hat{\mathbf{B}}_i = (\hat{\beta}_{i1}, \dots, \hat{\beta}_{iK})$ and the column vector $\bar{\mathbf{R}} = (\bar{r}_1, \dots, \bar{r}_K)^\top$ represent risk loadings and average returns for the factors, respectively.

Under this specification, the alpha remains a scalar, and the standard deviation of the regression residuals underlying equation (2) is also a positive number. Therefore, the multiindex counterparts of Jensen's alpha and the information ratio are similar to the performance measures applied to the single index model.

The generalization of the Treynor ratio is also provided by a simple ratio (Hübner [2005]).

It is given by:

$$G\hat{T}R_i = \hat{\alpha}_i \frac{\hat{\mathbf{B}}_m \bar{\mathbf{R}}}{\hat{\mathbf{B}}_i \bar{\mathbf{R}}} \quad (3)$$

where m denotes the benchmark portfolio against which portfolio i is compared.

This measure bears the same intuitive interpretation as the original Treynor ratio; namely, that it provides the amount of abnormal return of portfolio i per unit of systematic risk. The use of the required return on the benchmark portfolio, $\hat{\mathbf{B}}_m \bar{\mathbf{R}}$, ensures that the systematic risk exposure is normalized with respect to a comparable passive investment strategy.

Data and empirical methods

Unlike Kothari and Warner [2001] who use the alpha as instruments to assess the ability of asset pricing specifications to detect the presence or the absence of abnormal performance, our objective is exactly the converse. The performance measures are under study, and the pricing models are used as instruments.

The way active portfolios are managed influences the empirical quality of a performance measure. If managers consider that investors use their fund as a complement to passive portfolios, they must care about their information ratio. Otherwise, they should primarily consider the systematic risk exposures of their portfolio. If the level of alpha is truly independent of the exposure to systematic risks, normalizing the alpha by an estimate of systematic risk exposure would introduce measurement error and the alpha alone would dominate any other measure. But if absolute excess returns are proportional to systematic risks, then the generalized Treynor ratio (GTR) properly measures performance.

Thus, the ability to identify differential skills among portfolio managers depends on the choice of the performance measure. The adoption of the measure that best corresponds to the common behavior of portfolio managers will yield the most reliable ranking. Consider the case where the return generating process of their portfolios is known. Under this process, the IR , the alpha or the GTR – depending on how managers generate their performance—should produce a ranking scheme that perfectly corresponds to their skills. Unlike the rankings obtained using less accurate measures, the best scheme should be rather insensitive to random perturbations in returns. On the other hand, consider the alternative case where the return generating process is unknown. The performance measure that adequately accounts for the managers' skills still has to deliver the most reliable ranking as it is less sensitive to measurement errors. We translate these qualitative features into some workable criteria.

We assess the quality of a performance measure along two dimensions: (i) the ability of a criterion to rank funds in a way that is as close as possible to the theoretical ranking induced by the correct asset pricing model; (ii) the stability of these rankings when confronted with alternative asset pricing models, whether over-specified or under-specified. The first dimension

represents the precision with which a performance measure yields the replication of a perfect ranking of portfolio managers. The second dimension corresponds to a complementary matter, namely, the reliability of the ranking obtained with a performance measure when uncertainty prevails *vis-à-vis* the correct asset pricing specification. The outcome of this analysis yields information about the stability of the rankings based on a given performance measure.

Style-based mutual funds

We select a sample of US directional mutual funds data over an 11-year period, with end-of-month prices recorded from December 1993 to December 2004. The Morningstar and Lipper (subsidiary of Reuters) Large caps/Midcaps/Small caps and Growth/Blend/Value mutual funds strategies have been mixed so as to yield nine style-based categories. To be considered for inclusion in the sample, a mutual fund had to fit in the corresponding Morningstar and Lipper classifications together on January 2004. For each style category, we selected those funds that had achieved the largest five-year return prior to the selection date with the requirements that they could display a complete set of consecutive returns over the preceding 10-year period and that only the best performing fund per promoter would appear in each category. This yields a sample of 72 funds with 132 returns per fund. Mutual funds data have been retrieved from Thomson Financial Datastream.

The index portfolios corresponding to the nine style categories are the corresponding Standard and Poor's and Barra indexes: Barra-S&P 500 – Growth, S&P 500 and Barra-S&P 500–Value for the Large/Growth, Large/Blend, and Large/Value categories; S&P Midcap 400 –Growth, S&P Midcap 400 and S&P Midcap 400 – Value for the Midcap/Growth, Midcap/Blend and Midcap/Value categories; and S&P Small caps 600, S&P Small caps 600 – Growth and S&P Small caps 600–Value for the Small/Blend, Small/Growth and Small/Value categories. In addition to these nine style indexes, we construct a composite index by taking the equally-weighted average monthly returns. Although we presumably ignore whether mutual fund performance should be ideally measured with the alpha, the *GTR* or the *IR*, we use alternative index characterizations to assess the precision of a performance measure. The underlying motivation for this choice is the claim that if a performance measure truly accounts for the realized financial performance of a given portfolio, the ranking of funds produced by this performance measure should be insensitive to the choice of the benchmark, which is by definition a passive portfolio with no performance related to managerial skills.

We use the three-month T-Bill rate, the excess return of the weighted US market index, and the SMB, HML and UMD factors obtained from Kenneth French's website. We also compute an additional factor by taking the difference between the monthly returns of the equally weighted portfolio of stocks with the top 30% highest dividend yield and the monthly returns of a zero dividend payout equally weighted portfolio. The choice of this additional candidate factor, called "HDMZD" (High Dividend Minus Zero Dividend) is motivated by the hypothesized positive relationship set forth by Litzenberger and Ramaswamy (1979), theoretically explained by tax differential arguments, but challenged by Christie and Huang (1994) and Goyal and Welch (2003) on empirical grounds. Due to its controversial and at best weak contribution to explaining equity returns, this source of risk is an ideal candidate for playing the role of a "supernumerary" factor in our test of the stability of performance measures. The alternative use of a three-factor, four-factor or five-factor model to measure mutual fund performance will enable us to assess the stability of the candidate performance measures, in a spirit similar to the analysis carried out in the previous section.

Exhibit 1: Descriptive statistics for the empirical sample

This exhibit reports the average descriptive statistics for the sample of funds for each Morningstar - Lipper category (8 funds and 132 monthly observations), and for the corresponding style benchmark index, for the 1994–2004 period. Returns are in percentage and are computed in excess of the three-month T-Bill rate. The last line reports the average statistics over the 72 funds and for the composite index constructed by averaging returns of the nine style benchmark indexes.

Category	\bar{r}_i	S.D.	Skew.	Kurt.	Benchmark	\bar{r}_m	S.D.	Skew.	Kurt.
Large/Growth	0.17	5.79	-0.44	3.73	S&P500 Growth	0.53	4.83	-0.56	3.00
Large/Blend	0.36	4.71	-0.34	3.78	S&P500	0.50	4.40	-0.60	3.44
Large/Value	0.38	4.38	-0.49	4.01	S&P500 Value	0.45	4.43	-0.63	4.08
Mid/Growth	0.48	6.73	-0.09	5.27	S&P400 Mid G.	0.86	6.01	-0.16	4.36
Mid/Blend	0.30	5.15	-0.61	5.03	S&P400 Mid	0.87	4.91	-0.59	4.32
Mid/Value	0.18	4.95	-0.71	5.13	S&P400 Mid V.	0.90	4.55	-0.47	4.63
Small/Growth	0.63	7.51	-0.33	3.77	S&P600 Small G.	0.61	5.98	-0.29	4.22
Small/Blend	0.32	5.35	-0.50	4.69	S&P600 Small	0.77	5.28	-0.67	4.29
Small/Value	0.43	4.92	-0.65	4.79	S&P600 Small V.	0.86	4.96	-0.91	5.11
All	0.36	5.50	-0.46	4.47	Composite	0.70	4.57	-0.81	4.31

Exhibit 1 shows that all categories have displayed on average positive excess returns over the sample period, but the index portfolio returns exceeds all averages across strategies except the Small/Growth one, without any systematically greater exposure to variance, skewness or kurtosis risks. Jarque-Bera statistics indicate that the hypothesis of normally distributed excess returns can be rejected for 53 mutual funds (73.6%) at the 5% confidence level, while it is rejected for all benchmark indexes.

To compute the performance measures for each fund, we confront funds returns to the Fama and French [1993] three-factor model, the Carhart [1997] four-factor model and the five-factor model with the dividend yield factor added to the original Market, SMB, HML and UMD factors. The average exposures of the mutual funds to these various model specifications are summarized in exhibit 2.

Exhibit 2: Time-series regressions on the factor specifications

This exhibit reports the average coefficient estimates for the 1994–2004 period for the set of funds in each category using the three-factor Fama-French (1993) specification, the four-factor Carhart (1997) specification and the five-factor specification with the “High Dividend Minus Zero Dividend” factor. The last two lines report, the mean value of the factor during the sample period and the incremental adjusted R-squared obtained by adding the factor in the corresponding column. The last two columns report the number of funds in the sample for which the alpha of the corresponding regression is significantly positive and negative, respectively, at the 10% significance level. Detailed results for the individual funds, including t-stats, are available upon request.

	$\hat{\alpha}$	Mkt	SMB	HML	UMD	HDMZD	R_{adj}^2	$\hat{\alpha} > 0$	$\hat{\alpha} < 0$
L/G	-0.13	0.85	-0.22	-0.42			65.41	0	0
	-0.13	0.85	-0.22	-0.42	0.00		65.80	0	0
	-0.06	0.72	-0.21	-0.26	-0.07	-0.20	69.57	0	0
L/B	-0.11	0.83	-0.14	-0.13			71.74	0	0
	-0.06	0.81	-0.15	-0.13	-0.05		74.12	0	0
	-0.08	0.85	-0.16	-0.17	-0.03	0.05	75.00	0	0
L/V	-0.27	0.82	0.09	0.24			59.98	0	0
	-0.08	0.78	0.04	0.25	-0.22		67.21	0	0
	-0.16	0.92	0.02	0.08	-0.15	0.21	72.63	0	1
M/G	-0.11	1.08	0.11	-0.06			56.97	0	0
	-0.29	1.13	0.16	-0.07	0.22		60.81	0	0
	-0.12	0.82	0.20	0.31	0.07	-0.46	71.54	0	0
M/B	-0.36	0.91	0.20	0.24			56.55	0	2
	-0.27	0.88	0.18	0.25	-0.11		58.34	0	1
	-0.26	0.87	0.18	0.27	-0.12	-0.02	58.91	0	1

M/V	-0.56	0.83	0.28	0.44			45.37	0	3
	-0.37	0.78	0.22	0.45	-0.21		51.00	0	1
	-0.41	0.84	0.22	0.37	-0.18	0.09	51.69	0	2
	$\hat{\alpha}$	Mkt	SMB	HML	UMD	HDMZD	R_{adj}^2	$\hat{\alpha} > 0$	$\hat{\alpha} < 0$
S/G	-0.11	1.31	0.37	0.07			65.89	0	0
	-0.28	1.35	0.42	0.06	0.21		68.50	0	0
	-0.09	1.00	0.46	0.49	0.04	-0.52	79.50	0	0
S/B	-0.42	0.91	0.44	0.45			51.81	0	2
	-0.38	0.90	0.43	0.45	-0.05		53.38	0	0
	-0.31	0.78	0.44	0.60	-0.11	-0.18	59.25	0	0
S/V	-0.34	0.89	0.57	0.57			62.08	0	1
	-0.29	0.87	0.56	0.58	-0.06		62.69	0	1
	-0.24	0.80	0.56	0.67	-0.10	-0.11	64.04	0	2
\bar{r}_j		0.58%	0.17%	0.53%	0.90%	0.25%			
ΔR_{adj}^2		+54.77	+2.07	+3.16	+2.74	+5.68			

Under the three- and four-factor specifications, the average coefficients for size and book-to-market factors are perfectly in line with the assumed characteristics of the categories. For the SMB factor, mean coefficients for the large funds categories, midcap funds and small-cap funds are ranked in decreasing order. Within each of these families, average HML coefficients for the value funds, blend funds and growth funds are themselves ranked in decreasing order. This monotonic ordering is broken when the HDMZD factor is introduced for the midcap funds. This behavior is due to the very high correlation of 59.6% between the factors HML and HDMZD, introducing multicollinearity problems.

Nevertheless, every factor is shown to add significant explanatory power to the returns generating process specification. Under the one-factor model, the average significance level of the individual regressions exceeds 50%; it reaches 60% nearly with the three-factor model and nearly 70% with the five-factor model. The introduction of the HDMZD factor improves the average adjusted R-squared for all nine categories, with increments ranging from 0.57% for the Mid/Blend category to as much as 11% for the Small/Growth funds. Thus, no proposed asset pricing specification displays obvious signals of over-specification.

Style-adjusted performance

Usually, portfolio performance is measured either by reference to an asset pricing model or by comparison with a proper benchmark. Here, we compute each performance measure through the use of the linear asset pricing models presented above. However, we study the performance that is purely attributable to each portfolio manager across different styles. But each of these style indexes has had different levels of required returns, abnormal performance, and volatility of regression residuals during the test period. By merely comparing the performance measures defined above, we would mix the pure skill effect that we want to capture with the performance of the passive style index under the corresponding asset specification. Therefore, we control for style-performance effects by defining the following style-adjusted performance measures:

$$\hat{\alpha}_{i,adj} = \hat{\alpha}_i - \hat{\alpha}_{m_i} \quad (4a)$$

$$GT\hat{R}_{i,adj} = \hat{\alpha}_i \frac{\hat{\mathbf{B}}_{m_i} \bar{\mathbf{R}}}{\hat{\mathbf{B}}_i \bar{\mathbf{R}}} - \hat{\alpha}_{m_i} \quad (4b)$$

$$I\hat{R}_{i,adj} = \hat{\alpha}_i \frac{\hat{\sigma}(\varepsilon_{m_i})}{\hat{\sigma}(\varepsilon_i)} - \hat{\alpha}_{m_i} \quad (4c)$$

where the return of the style index portfolio m_j associated to fund i follows the market model

$$r_{m_i t} = \hat{\alpha}_{m_i} + \hat{\mathbf{B}}_{m_i} \mathbf{R}_t + \varepsilon_{m_i t}$$

Equation (4a) allows us to control for the index alpha, $\hat{\alpha}_{m_i}$ to isolate the level of absolute performance specifically related to management skills. Equations (4b) and (4c) further account for the risk exposures of the funds in relation with that of their index portfolio. Hence, these last two measures provide the risk-adjusted performance of portfolios that accounts for the risk-return characteristics of their corresponding style index.

To provide a first check of the stability of the performance measures, we compute the difference between the style-based performance of the funds under the four-factor and the three-factor models. Similarly, examining the differences between performances obtained in the four-factor model with style index portfolios versus the composite index portfolio provides insights about the precision of the performance measures. The results of these comparisons are summarized in Exhibit 3.

Exhibit 3: Statistics of performance measures across benchmarks and models

This exhibit reports the mean and standard deviation of the alpha, the generalized Treynor ratio and the information ratio for the sample of mutual funds during the period 1994-2004. In panel A, the performance measure of each individual fund is computed against its corresponding style benchmark index. In panel B, the performance measures are computed using the Carhart (1997) four-factor asset pricing model. All numbers are in monthly percentages.

Panel A: four-factor versus three-factor model						
Measure	Mean			Standard deviation		
	4-factor	3-factor	Difference	4-factor	3-factor	Difference
$\hat{\alpha}$	-0.162	-0.200	0.038	0.263	0.285	0.106
<i>GTR</i>	-0.250	-0.263	0.013	0.370	0.379	0.104
<i>IR</i>	-0.056	-0.083	0.027	0.182	0.183	0.084

Panel B: Style benchmarks versus Composite benchmark						
Measure	Mean			Standard deviation		
	Style	Composite	Difference	Style	Composite	Difference
$\hat{\alpha}$	-0.162	-0.162	0.000	0.263	0.225	0.159
<i>GTR</i>	-0.250	-0.236	-0.014	0.370	0.409	0.192
<i>IR</i>	-0.056	0.010	-0.066	0.182	0.068	0.149

Panel A of exhibit 3 shows that using a three-factor or a four-factor specification for the returns generating process of each fund induces a very low average difference when the *GTR* measures the mutual fund performance. The average difference recorded for a switch from the three-factor to the four-factor model is 19% (0.038/0.200) of the mean fund performance for the alpha, and 32.5% (0.027/0.083) for the *IR*. Moreover, a closer look at the standard deviation reveals that, although the cross-sectional variation of the *GTRs* is approximately 33% to 40% higher than the one of the alphas, the variability of the differences in *GTRs* and in alphas from one pricing model to another is almost identical. A quite similar finding obtains for panel B, where the use of a composite benchmark induces a greater cross-sectional variability in *GTRs*, but the standard deviation of the differences arising from changes in benchmarks is proportionally lower for the *GTR* than for the alpha. The results achieved by the *IR* in both panels are rather poor. This indicates that the *GTR* spans a greater range than the alpha, but that it displays relatively more precision against different benchmarks and more stability when different asset pricing models are used. These conjectures are further investigated below.

Methods for statistical inference

The empirical analysis is performed by pairwise comparing the series of performance measures yielded under various specifications.

Two alternative types of hypotheses are investigated: the hypothesis of imprecision or instability—henceforth generically termed “lack of association”—postulates the absence of relationship between alternative classifications. On the contrary, the hypothesis of precision or stability—henceforth generically termed “perfect association”—posits that a given measure will produce the same ranking under alternative but equivalent specifications. In reality, in this particular context, not only the perfect or lack of association is important, but also its intensity is informative about the quality of each performance measure. Therefore, when confronted with an intensity-based measure of association, we will perform statistical inference on the basis of confidence intervals around the observed value.

For each measure of association ρ , we identify the theoretical values $\bar{\rho} = 1$ and $\underline{\rho} = 0$ corresponding to perfect concordance and lack of concordance, respectively. The observed value of ρ , denoted $\hat{\rho}$, must satisfy $\hat{\rho} \leq \bar{\rho}$ but has no constraint with respect to $\underline{\rho}$. Consider further different threshold levels θ_i , $i = 1, \dots, k$ ranked by their intensity of association: $\underline{\rho} < \theta_1 < \theta_2 < \dots < \theta_k < \bar{\rho}$. We can construct a one-sided confidence interval below $\hat{\rho}$ with a confidence level of ψ and denote it L_ψ : if $\Pr[\rho \leq L_\psi] = \psi$, we cannot reject the hypothesis that ρ displays at least the intensity of association corresponding to θ_i if $\theta_i \geq L_\psi$. Naturally, the hypotheses of perfect association and lack of association will not be rejected if $\underline{\rho} \geq L_\psi$ and $\bar{\rho} \leq U_\psi$ respectively, where U_ψ is the one-sided upper interval.

For the parametric approach, we first use the Pearson product-moment correlation coefficient. However, since the variables under study are continuous, the “concordance correlation coefficient” ρ_C proposed by Lin [1989, 2000] is more appropriate than Pearson's correlation. The estimator value in finite sample is given by:

$$\hat{\rho}_C = \frac{2\hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 + (\hat{\mu}_X - \hat{\mu}_Y)^2} \quad (5)$$

where X and Y are the random variables, $\hat{\mu}_X$, $\hat{\mu}_Y$ are the sample means, $\hat{\sigma}_X^2$, $\hat{\sigma}_Y^2$ are the sample variances and $\hat{\sigma}_{XY}$ is the sample covariance. The lower bound of the confidence interval is provided by Lin [2000].

For the non-parametric approach, we can consider the rankings produced by the performance measures or a dichotomous approach based on contingency. To test for association with rankings, we rely on the Spearman rank correlation coefficient ρ_S .

The second test is based on contingency table statistics. We rank funds by their performance under each specification and identify the median; funds with higher performance are the “Winners” (W) and the rest are “Losers” (L). Therefore, the funds are partitioned in four categories: the number of funds that are winners or losers under both classifications are equal to WW and LL, respectively, and the number of winners under one and losers under the other ranking are denoted by WL and LW.

To assess the association level of classifications using $K \times K$ contingency tables, Cohen (1960) introduces the kappa statistic.¹ This measure is suitable to assess the concordance of judgments made by different raters. In our case, these raters are the performance measures themselves. The kappa aims at measuring the proportion of matching pairs in the table that is not merely due to chance. This statistic cannot be used for statistical inference except for the 2×2 contingency table. In this case, the estimate of Cohen's kappa, denoted $\hat{\kappa}$, is given by:

1 - Contingency table statistics such as the cross-product ratio or the chi-square for independence (see Brown and Goetzmann [1995], Carpenter and Lynch [1999]) would be useless for the current issue as they are only well-specified under the null hypothesis of lack of association. This null hypothesis is not under study here.

$$\hat{\kappa} = 2 \frac{WW + LL}{N} - 1. \quad (6)$$

and the lower bound for the one-sided confidence interval is given by Fleiss [1981].

We define in turn four thresholds $\theta_1 = 0.2$, $\theta_2 = 0.4$, $\theta_3 = 0.6$, $\theta_4 = 0.8$, on the basis of a refinement of the standard scaling proposed by Landis and Koch [1977]. Let $i = P, C, S, \kappa$.

They propose the following interpretations: if $\rho_i \in [0, 0.2]$, $]0.2, 0.4]$, $]0.4, 0.6]$, $]0.6, 0.8]$, $]0.8, 1]$ the association is "slight", "fair", "moderate", "substantial" and "almost perfect", respectively. We split these intervals further in lengths of 0.1 and define 10 zones, zone 1 corresponding to "slight–lower range" and zone 10 corresponding to "almost perfect–upper range". This procedure ensures that each performance measure will be classified in a hypothesis-free scheme.

Empirical comparison of performance measures

Precision

To examine the precision of the performance measures, we are confronted with the difficulty that it is impossible to determine with certainty on what risk dimension mutual fund managers tend to distinguish themselves, *i.e.* what is the "true" measure of performance. Yet, if this measure could be identified, the rankings produced using alternative benchmarks would reveal invariant under this particular performance measure, while the use of imprecise metric would only produce random classifications. Different benchmarks would thus yield unrelated classifications. Therefore, we rely on comparisons between rankings of funds obtained for a given asset pricing specification but with alternative benchmarks to assess the precision of each performance measure. Exhibit 4 shows the corresponding tests results.

Exhibit 4: Empirical precision of performance measures

This exhibit reports association measures between classifications using style-based benchmark portfolios and the composite reference index using the three-factor Fama-French [1993] specification, the four-factor Carhart [1997] specification and the five-factor specification with the "High Dividend Minus Zero Dividend" factor with monthly returns during the 1994–2004 period. For each measure, the observed value $\hat{\rho}$ and lower bound with a one-sided confidence at 5% level L_ψ are reported. For each measure and association measure, the reported zone corresponds to the lowest value i such that $L_\psi \leq 0.1i$, $i = 1, \dots, 10$. Panel A presents the parametric product-moment (Pearson) correlation and the concordance (Lin [1989, 2000]) coefficients. Panel B presents the nonparametric rank (Spearman) correlation and Cohen's [1960] kappa coefficients.

Panel A: Parametric association measures

<i>Model</i>	<i>Measure</i>	Pearson correlation			Lin's concordance		
		$\hat{\rho}_P$	$L_{\psi,P}$	zone	$\hat{\rho}_C$	$L_{\psi,C}$	zone
three-factor	$\hat{\alpha}$	0.899	0.795	8	0.884	0.826	9
	<i>GTR</i>	0.890	0.782	8	0.873	0.810	9
	<i>IR</i>	0.695	0.524	6	0.493	0.365	4
four-factor	$\hat{\alpha}$	0.798	0.655	7	0.777	0.673	7
	<i>GTR</i>	0.883	0.772	8	0.866	0.800	9
	<i>IR</i>	0.622	0.436	5	0.403	0.276	3
five-factor	$\hat{\alpha}$	0.882	0.770	8	0.859	0.792	8
	<i>GTR</i>	0.974	0.920	10	0.833	0.805	9
	<i>IR</i>	0.673	0.497	5	0.464	0.345	4

Panel B: Nonparametric association measures							
<i>Model</i>	<i>Measure</i>	Spearman rank			Cohen's kappa		
		$\hat{\rho}_S$	$L_{\psi,S}$	zone	$\hat{\rho}_\kappa$	$L_{\psi,\kappa}$	zone
three-factor	$\hat{\alpha}$	0.894	0.786	8	0.722	0.560	6
	<i>GTR</i>	0.909	0.809	9	0.833	0.704	8
	<i>IR</i>	0.664	0.485	4	0.389	0.173	2
four-factor	$\hat{\alpha}$	0.799	0.655	7	0.667	0.492	5
	<i>GTR</i>	0.920	0.828	9	0.889	0.782	8
	<i>IR</i>	0.610	0.423	5	0.500	0.297	3
five-factor	$\hat{\alpha}$	0.873	0.756	8	0.778	0.631	7
	<i>GTR</i>	0.931	0.844	9	0.833	0.704	8
	<i>IR</i>	0.655	0.474	5	0.444	0.234	3

In general, the lower bounds obtained with non-parametric statistics reported in panel B, especially when using Cohen's kappa, are more conservative than those corresponding to parametric statistics. A noticeable exception, however, is the set of values of the Spearman rank correlation obtained by the generalized Treynor ratio for the three-factor and the four-factor models. This is mainly due to the presence of a fund with a very low required return, and thus a very high value of the *GTR*. Since this outlier value is very sensitive and has a great impact on parametric association measures, the use of rankings precisely aims at correcting the weight of extreme values in the association coefficient.

The results of exhibit 4 indicate that the *GTR* achieves a very high precision level. Regardless of the measure or the asset pricing model considered, the lower bound of the confidence interval always belongs to zones 8 to 10, corresponding to the upper range of "substantial" and the ranges of "almost perfect" associations. The observed values themselves lie in zones 9 to 10, always supporting an "almost perfect" relationship. In particular, considering a symmetric upper bound for the confidence intervals would not lead to rejecting the perfect association hypothesis for the Spearman as well as the Pearson (except for the four-factor model) coefficients.

In contrast with the homogenous behavior of the *GTR*, the different measures of association estimated with alphas are not stable. The poorest association intensities are observed with the four-factor model, where the observed Pearson, Spearman and concordance coefficients are very close to 0.80, while Cohen's kappa falls to 0.667, supporting only the "moderate" association hypothesis. As a matter of fact, no measure is able to support the hypothesis of perfect association ($\rho_i = 1$) when performance is measured with the alpha. Observed association measures for the alpha are systematically below the corresponding ones for the *GTR*, except for Lin's concordance coefficient measured with the three-factor and the five-factor models. The difference deteriorates when one compares lower bounds of confidence intervals, as the estimators of association coefficients always display a larger variance when alpha is used than when the *GTR* is the performance measure considered for the rankings.

The information ratio displays by far the poorest levels of precision. The Pearson and Spearman coefficients culminate at 0.695 and 0.664 respectively, and no metric supports a hypothesis greater than a "moderate" association between rankings produced with this performance measure, except for the Pearson coefficient with the three-factor model.

Stability

The empirical analysis of stability of performance measures does not suffer from the same obstacle as the one of precision. Considering a relevant reference portfolio—the style-based benchmark indexes for this study—we need to check the behavior of the rankings produced with each performance measure when switching from one asset pricing specification to another. Exhibit 5 reports the results obtained by comparing the four-factor model with the more parsimonious three-factor model and with the more sophisticated five-factor model.

Exhibit 5: Empirical stability of performance measures

This exhibit reports association measures between classifications using style-based benchmark portfolios and the composite reference index using the three-factor Fama-French [1993] specification, the four-factor Carhart [1997] specification and the five-factor specification with the "High Dividend Minus Zero Dividend" factor with monthly returns during the 1994–2004 period. For each measure, the observed value $\hat{\rho}$ and lower bound with a one-sided confidence at 5% level L_{ψ} are reported. For each measure and association measure, the reported zone corresponds to the lowest value i such that $L_{\psi} \leq 0.1i$, $i = 1, \dots, 10$. Panel A presents the parametric product-moment (Pearson) correlation and the concordance (Lin [1989, 2000]) coefficients. Panel B presents the nonparametric rank (Spearman) correlation and Cohen's [1960] kappa coefficients.

Panel A: Parametric association measures							
Model	Measure	Pearson correlation			Lin's concordance		
		$\hat{\rho}_P$	$L_{\psi,P}$	zone	$\hat{\rho}_C$	$L_{\psi,C}$	zone
3-f vs. 4-f	$\hat{\alpha}$	0.927	0.839	9	0.912	0.865	9
	<i>GTR</i>	0.962	0.897	9	0.948	0.923	10
	<i>IR</i>	0.894	0.788	8	0.882	0.819	9
4-f vs. 5-f	$\hat{\alpha}$	0.974	0.920	10	0.960	0.939	10
	<i>GTR</i>	0.675	0.500	6	0.464	0.349	4
	<i>IR</i>	0.968	0.908	10	0.954	0.928	10

Panel B: Nonparametric association measures							
Model	Measure	Spearman rank			Cohen's kappa		
		$\hat{\rho}_S$	$L_{\psi,S}$	zone	$\hat{\rho}_\kappa$	$L_{\psi,\kappa}$	zone
3-f vs. 4-f	$\hat{\alpha}$	0.924	0.833	9	0.833	0.704	8
	<i>GTR</i>	0.974	0.919	10	0.889	0.782	8
	<i>IR</i>	0.869	0.750	8	0.611	0.426	5
4-f vs. 5-f	$\hat{\alpha}$	0.971	0.913	10	0.944	0.867	9
	<i>GTR</i>	0.982	0.937	10	0.889	0.782	8
	<i>IR</i>	0.970	0.912	10	0.833	0.704	8

Panel B of exhibit 5 displays the same picture, although less pronounced, as exhibit 4. With non-parametric association measures, the stability of the generalized Treynor ratio is confirmed, whereas the information ratio displays the lowest association values. The differences between performance measures are not very large though, especially when one shifts from the four-factor to the five-factor model. When asset pricing models are more under-specified, the relative outperformance of the *GTR* picks up. The lower bound for the confidence level of the Spearman rank coefficient is to the order of 10% greater with the *GTR* than with the alpha.

Interpretation

Overall, the generalized Treynor ratio appears to obtain both results for the precision and stability criteria. Panel A of exhibit 5 introduces a very important caveat in this respect: the passage from

the four-factor to the five-factor model rather harms the stability of the generalized Treynor ratio, whereas the other measures maintain very high stability levels. The explanation must be found in the impact of the fund's required return in its measure of performance. With the introduction of the fifth factor, two of the funds experience a significant drop in its required return—these funds have a very negative loading on the dividend factor. Consequently, their risk exposures make them look like overall non-directional funds. Hence, their *GTR* tend to rocket up in absolute value and to drive down parametric coefficients of association. This behavior indicates that the use of non-parametric statistical methods is much more advisable to assess the reliability of highly non-linear performance measures such as the *GTR* in the presence of non-directional fund returns and significant measurement error.

Comparing the *GTR* with the alpha, the results in favor of the former measure indicate that the correction of the leverage bias inherent to the use of alpha does indeed matter for performance evaluation. Despite controlling for the sources of risks used in its derivation, the alpha itself is not risk-adjusted. As it is not invariant to a change in portfolio leverage, it could lead to false classifications of actively managed portfolios, magnifying the abnormal performance of managers proportionally to the leverage of their portfolios.

The heavy reliance on variance as a measure of risk underlying the *IR* is probably responsible for its poor empirical results. Richardson and Smith (1993) and Zhou (1993) show that the distribution of regression residuals significantly departs from normality. Measurement issues lead to strong overestimation of the residual variance, in turn biasing downward the information ratio (Ledoit and Wolf [2004]). However, measurement issues alone cannot fully explain the poor values obtained by the *IR* on non-parametric association measures.

This study shows that overall, portfolio managers distinguish themselves more clearly with the generalized Treynor ratio than with the other two measures. This result suggests that the managers of style-based portfolios consider their fund as a complement to other active portfolios held by investors. Furthermore, unlike the way alphas are measured, management skills are more accurately assessed when portfolios' required returns are controlled for.

Conclusion

The empirical use of the alpha as performance measure adjusted for systematic risk has never really been questioned for multi-index models. This paper has attempted to provide some empirical evidence in the field of the assessment of performance measures. With a sample of directional mutual fund returns, our results provide strong support for the Generalized Treynor Ratio. Due to its nature of a slope measure however, rankings produced by the *GTR* are slightly more reliable than cardinal values as this ratio can produce extremely sensitive outputs for portfolios with a very small level of required return. These results indicate that when it is applicable, *i.e.* when the required return on managed portfolios is expected to be positive, the generalized Treynor ratio displays superior ranking abilities over its competitors, the alpha and the information ratio, provided that comparisons are done with proper instruments. This advantage overcomes the drawback of the sensitivity of the *GTR* to measurement error of the *ex-post* factor.

More fundamentally, these results raise the question of the risk dimensions that should be accounted for in performance measurement with multi-index models. The phenomenon of "leverage bias", originally identified by Modigliani and Pogue [1974], diminishes the ability of Jensen's alpha to rank leveraged portfolios on the basis of their performance. Our study provides some support for the suspicion of the relevance of this phenomenon for the selected sample.

Important additional work remains to be done with respect to the assessment of the quality of performance measures. The proper measurement of performance for non-directional funds

has not been addressed in this paper, and cannot be accounted for by the generalized Treynor ratio. The alpha and the information ratio could be relevant then, but they should be compared with measures designed on total risk as well. This research topic is part of our ongoing research agenda.

References

- Bodie, Z., A. Kane, and A. J. Marcus. *Investments* (6th edition). New York: McGraw Hill Irwin, 2005.
- Brown, S. J. and W. N. Goetzmann. "Performance persistence." *Journal of Finance*, vol. 50, no. 2 (1995), pp. 679-698.
- Carhart, M. M. "On persistence in mutual funds performance." *Journal of Finance*, vol. 52, no. 1 (1997), pp. 57-82.
- Carpenter, J. N., and A. W. Lynch. "Survivorship bias and attrition effects in measures of performance persistence." *Journal of Financial Economics*, vol. 54, no. 3 (1999), pp. 337-374.
- Christie, W. G., and R. D. Huang. "The changing functional relation between stock returns and dividend yields." *Journal of Empirical Finance*, vol. 2, no. 1 (1994), pp. 161-191.
- Cohen, J. "A coefficient of agreement for nominal scales." *Educational and Psychological Measurement*, vol. 20 (1960), pp. 37-46.
- Fama, E. F. and K. French. "Common risk factors in the returns of stocks and bonds." *Journal of Financial Economics*, vol. 33, no. 1 (1993), pp. 3-56
- Fleiss, J. L. *Statistical methods for rates and proportions*. New York: Wiley and Sons, 1981.
- Goodwin, T. H. "The information ratio." *Financial Analysts Journal*, vol. 54, no. 4 (July/August 1998), pp. 34-43.
- Goyal, A. and I. Welch. "Predicting the equity premium with dividend ratios." *Management Science*, vol. 49, no. 5 (2003), pp. 639-654.
- Grinold, R. C. and R. N. Kahn. "Information analysis." *Journal of Portfolio Management*, vol. 18, no. 3 (Spring 1992), pp. 14-21.
- Hübner, G. "The Generalized Treynor Ratio." *Review of Finance*, vol. 9, no.3 (2005), pp. 415-435.
- Jensen, M. J.. "The performance of mutual funds in the period 1945-1964." *Journal of Finance*, vol. 23, no. 2 (1968), pp. 389-416.
- Keating, C. and W. F. Shadwick. "Omega: A universal performance measure." *Journal of Performance Measurement*, vol. 6, no. 3 (Spring 2002), pp. 59-84.
- Kothari, S.P. and J. B. Warner. "Evaluating mutual fund performance." *Journal of Finance*, vol. 56, no. 5 (2001), pp. 1985-2010.
- Landis, J. R. and G. G. Koch. "The measurement of observer agreement for categorical data." *Biometrics*, vol. 33 (1977), pp. 159-174.
- Ledoit, O. and M. Wolf. "Honey, I shrunk the sample covariance matrix." *Journal of Portfolio Management*, Vol. 30, no. 4 (Summer 2004), pp. 110-119.
- Lin, Li-K. "A concordance correlation coefficient to evaluate reproducibility." *Biometrics*, vol. 45 (1989), pp. 255-268.
- Lin, Li-K. "A note on the concordance correlation coefficient." *Biometrics*, vol. 56 (2000), pp. 324-325.

- Lintner, J. "The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets." *Review of Economics and Statistics*, vol. 47, no. 1 (1965), pp. 13-37.
- Litzenberger, R. H. and K. Ramaswamy. "The effect of personal taxes and dividends on capital asset prices: theory and empirical evidence." *Journal of Financial Economics*, vol. 7, no. 2 (1979), pp. 163-195.
- Modigliani, F. and L. Modigliani. "Risk-adjusted performance." *Journal of Portfolio Management*, vol. 23, no. 2 (Winter 1997), pp. 45-54.
- Modigliani, F. and G. A. Pogue. "An introduction to risk and return – II." *Financial Analysts Journal*, vol. 30, no. 3 (May/June 1974), pp. 69-86.
- Richardson, M. and T. Smith. "A test for multivariate normality in stock returns." *Journal of Business*, vol. 66, no. 2 (1993), pp. 295-321.
- Sharpe, W. F. "Capital asset prices: A theory for market equilibrium under conditions of risk." *Journal of Finance*, vol. 19, no. 3 (1964), pp. 425-442.
- —. "Mutual fund performance." *Journal of Business*, vol. 39, no. 1 (1966), pp. 119-138.
- Sortino, F. A., and L. N. Price. "Performance measurement in a downside risk framework." *Journal of Investing*, vol. 3, no. 3 (Fall 1994), pp. 59-65.
- Treynor, J. L. "Toward a theory of market value of risky assets." Mimeo (1961), subsequently published in Korajczyk, R. A. , *Asset Pricing and Portfolios Performance: Models, Strategy and Performance Metrics*. London: Risk Books, 1999.
- Treynor, J. L. "How to rate management investment funds." *Harvard Business Review*, vol. 43, no. 1 (Jan/Feb 1965), pp. 63-75.
- Zhou, G. "Asset-pricing tests under alternative distributions." *Journal of Finance*, vol. 48, no. 5 (1993), pp. 1927-1942.