How Do Performance Measures Perform?

Examining precision and stability.

Georges Hübner

GEORGES HÜBNER
is the Deloitte professor of financial management in the HEC Management School of the University of Liège in Belgium, an associate professor of finance at Maastricht University, and an academic expert in the Luxembourg School of Finance.
g.hubner@ulg.ac.be

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 relate directly to the beta of the portfolio through the security market line (SML). Jensen's [1968] alpha is defined as the portfolio excess return earned beyond the required average return, while the Treynor ratio [1965] 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, respectively.

There is no other widely used alternative measure that builds on the SML. More recent performance measures developed to deal with the increasing popularity of hedge funds focus on a measure of total risk, in a similar fashion with the Sharpe [1966] ratio applied to the capital market line. These include the Sortino ratio (Sortino and Price [1994]), the M^2 (Modigliani and Modigliani [1997]), and the omega (Keating and Shadwick [2002]).

In extending the CAPM to linear multifactor asset pricing models, performance measurement has not been very innovative. Empirical research has made use of only one single measure that uses systematic risk as a component, namely, portfolio abnormal return computed in the same way as Jensen's [1968] alpha. As the intercept of a linear regression, the alpha is flexible enough to be used in most asset pricing specifications. Kothari and Warner [2001] choose this performance measure in their empirical comparison of multi-index asset pricing models. 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, e.g., Grinold and Kahn [1992] and Goodwin [1998]). 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 performance measures, it is important to understand what they try to measure. If an investor chooses a managed portfolio as the 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 also on the type of complementary portfolio held by the investor, i.e., the other investment vehicles. If the investor holds a passive portfolio like an index fund, the performance of the managed portfolio is measured by the information ratio. If the portfolio is just one among many other actively managed portfolios, the alpha of the portfolio rightly measures its contribution to the overall performance. In this context, as Bodie, Kane, and Marcus [2005] note, "an even better solution . . . is to use Treynor's measure."

For portfolio managers themselves, the motivations for striving to generate high information ratios, high alphas, 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 the 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 using 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, I examine the empirical performance of performance measures for multi-index models through the quality of the ranking scheme they produce. I use a sample of directional mutual funds with different styles, so that the *IR*, the alpha, and the generalized Treynor ratio that I propose in Hübner [2005] can adequately measure performance. The quality of a performance measure is assessed on 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 to best describe managerial skills in active portfolio management.

On the practical side, there is nonetheless a serious difficulty with regard to 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. It 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 disappears only asymptotically, but this property is hardly usable for most practical applications.

I tackle this empirical issue in three ways. My dataset, including only directional managed portfolios, guarantees a suitable context for application of the generalizations of Jensen's [1968] alpha, the information ratio, and the Treynor ratio. Furthermore, to ensure fairness of the comparisons of performance measures, I control for the risk exposures and measurement errors of the passive index portfolios. Finally, the empirical methods allow me 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

I 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

$$\hat{\beta}_i = \frac{\widehat{\text{cov}}(r_i, r_m)}{\hat{\sigma}^2(r_m)}$$

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:

$$TR_i = \frac{\hat{\alpha}_i}{\hat{\beta}_i}$$

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 heterogeneous. Despite their differences, however, these approaches share a common affine model specification that can be summarized by this ex post multidimensional equation:

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

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

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 multi-index 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:

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

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

This measure has 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, $\hat{\mathbf{B}}_m \overline{\mathbf{R}}$, 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 an instrument to assess the ability of asset pricing specifications to detect the presence or the absence of abnormal performance, my objective is exactly the converse. The performance measures are the issue 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 yields the most reliable ranking. Suppose the return-generating process of the portfolios is known. In this case, either 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.

Or, consider that 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. I translate these qualitative features into some workable criteria.

I assess the quality of a performance measure on two dimensions: 1) its ability to rank funds in a way that is as close as possible to the theoretical ranking induced by the correct asset pricing model, and 2) the stability of these rankings when alternative asset pricing models are used, whether overspecified or underspecified. The first dimension represents how closely a performance measure yields the replication of a perfect ranking of portfolio managers. The second dimension is complementary, 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

I select a sample of U.S. directional mutual funds data over an 11-year period, recording end-of-month prices from December 1993 to December 2004. The Morningstar and Lipper (subsidiary of Reuters) Large cap/Midcap/Small cap and Growth/Blend/Value mutual fund strategies are mixed so as to yield nine style-based categories. To be considered for inclusion in the sample, a mutual fund had to fit in both Morningstar and Lipper classifications on January 2004. For each style category, I select the funds that had achieved the highest five-year return prior to the selection date with the requirements that there had to be a complete set of consecutive returns over the preceding ten-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 fund data come from Thomson Financial Datastream.

The index portfolios corresponding to the nine style categories are the corresponding Standard & 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. Beyond these nine style indexes, I construct a composite index by taking the equally weighted average monthly returns.

Although I presumably ignore whether mutual fund performance should be ideally measured with the alpha, the *GTR*, or the *IR*, I use alternative index characterizations to assess the precision of a performance measure. The underlying motivation 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.

I use the three-month T-bill rate, the excess return of the weighted U.S. market index, and the SMB, HML, and UMD factors obtained from Kenneth French's website. I 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.

Because of 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 my 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 me to assess the stability of the candidate performance measures, in a spirit similar to the analysis carried out in the previous section.

Exhibit 1 shows that all categories have displayed on average positive excess returns over the sample period, but the index portfolio returns exceed all averages across strategies except the Small/Growth portfolio 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 compare the performance measures for each fund, I confront fund 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

Exhibit 1
Descriptive Statistics for the Empirical Sample—1994-2004

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

factors. The average exposures of the mutual funds to these various model specifications are summarized in Exhibit 2.

Under the three- and four-factor specifications, the average coefficients for size and book-to-market factors are consistent with the assumed characteristics of the categories. For the SMB factor, mean coefficients for the large fund categories, the midcap funds, and the small-cap funds are ranked in declining order. Within each of these families, average HML coefficients for the value funds, blend funds, and growth funds are themselves ranked in declining 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 return-generating process specification. Under the one-factor model, the average significance level of the individual regressions exceeds 50%; it reaches almost 60% with the three-factor model and almost 70% with the five-factor model.

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 overspecification.

Style-Adjusted Performance

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

Therefore, I control for style-performance effects by defining the style-adjusted performance measures:

$$\hat{\alpha}_{i,adj} = \hat{\alpha}_i - \hat{\alpha}_{m_i} \tag{4-A}$$

$$G\hat{T}R_{i,adj} = \hat{\alpha}_i \frac{\hat{\mathbf{B}}_{m_i}\overline{\mathbf{R}}}{\hat{\mathbf{B}}_i\overline{\mathbf{R}}} - \hat{\alpha}_{m_i} \tag{4-B}$$

$$I\hat{R}_{i,adj} = \hat{\alpha}_i \frac{\hat{\sigma}(\varepsilon_{m_i})}{\hat{\sigma}(\varepsilon_i)} - \hat{\alpha}_{m}. \tag{4-C}$$

where the return of the style index portfolio m_i associated with 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 (4-A) allows me to control for the index alpha, $\hat{\alpha}_{m_i}$ to isolate the level of absolute performance specifically related to management skills. Equations (4-B) and (4-C) further account for the risk exposures of the funds in relation to one of their index portfolios. 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, I 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 performance obtained with the four-factor model and 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. Panel A shows that using a three-factor or a four-factor specification for the return-generating process

EXHIBIT 2
Time Series Regressions on the Factor Specifications

	$\hat{\alpha}$	Mkt	SMB	HML	UMD	HDMZD	R^2_{adj}	$\hat{\alpha} > 0$	$\hat{\alpha} < 0$
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
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^2_{adj}		+54.77	+2.07	+3.16	+2.74	+5.68			

The last two lines report, respectively, 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 or negative, at the 10% significance level. Detailed results for individual funds, including t-stats, are available upon request.

of each fund induces a very low average difference when the generalized Treynor ratio measures 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 *GTR* is approximately 33% to 40% higher than the variation of the alphas, the variability of the differences in *GTR* and in alphas from

EXHIBIT 3
Statistics of Performance Measures Across Benchmarks and Models

Panel A: Four-factor versus three-factor model

	Mean			Standard deviation		
Measure	4-factor	3-factor	Difference	4-factor	3-factor	Difference
$\hat{\alpha}$	-0.162	-0.200	0.038	0.263	0.285	0.106
GTR	-0.250	-0.263	0.013	0.370	0.379	0.104
IR	-0.056	-0.083	0.027	0.182	0.183	0.084

Panel B: Style benchmarks versus Composite benchmark

	Mean			Standard deviation		
Measure	Style	Composite	Difference	Style	Composite	Difference
$\hat{\alpha}$	-0.162	-0.162	0.000	0.263	0.225	0.159
GTR	-0.250	-0.236	-0.014	0.370	0.409	0.192
IR	-0.056	0.010	-0.066	0.182	0.068	0.149

All numbers are in monthly percentages.

one pricing model to another is almost identical. A similar finding obtains for Panel B, where the use of a composite benchmark induces a greater cross-sectional variability in *GTR*, 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 with regard to different benchmarks and more stability when different asset pricing models are used.

Methods for Statistical Inference

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

Two different types of hypotheses are investigated. The hypothesis of imprecision or instability—henceforth generically termed *lack of association*—postulates the absence of a relation between alternative classifications. 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, more than perfect association or lack of association is important; its intensity is also informative as to the quality of each performance measure. Therefore, for an intensity-based measure of association, I will perform statistical inference on the basis of confidence intervals around the observed value.

For each measure of association ρ, I identify the theoretical values $\overline{\rho} = 1$ and $\underline{\rho} = 0$ corresponding to perfect concordance and lack of concordance. Note that the observed value of ρ, denoted $\widehat{\rho}$, must satisfy $\widehat{\rho} \le \overline{\rho}$, but is unconstrained with respect to $\underline{\rho}$. Consider further different threshold levels θ_i, $i = 1,\ldots,k$ ranked by their intensity of association: $\underline{\rho} < \theta_1 < \theta_2 < \ldots < \theta_k < \overline{\rho}$. I can construct a one-sided confidence interval below $\widehat{\rho}$ with a confidence level of ψ and denote it L_ψ. If $\Pr[\rho \le L_\psi] = \psi$, I cannot reject the hypothesis that ρ displays at least the intensity of association corresponding to θ_i if $\theta_i \ge L_\psi$. Naturally, the hypotheses of perfect association and lack of association will not be rejected if $\underline{\rho} \ge L_\psi$ and $\overline{\rho} \le U_\psi$, respectively, where U_ψ is the one-sided upper interval.

For the parametric approach, I first use the Pearson product-moment correlation coefficient. As the variables under study are continuous, however, the *concordance correlation coefficient* ρc proposed by Lin [1989, 2000] is more appropriate than Pearson's correlation. The estimator value in the finite sample is given by:

$$\widehat{\rho}c = \frac{2\widehat{\sigma}_{XY}}{\widehat{\sigma}_X^2 + \widehat{\sigma}_Y^2 + (\widehat{\mu}_X - \widehat{\mu}_Y)^2} \tag{5}$$

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

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

The second test is based on contingency table statistics. I rank funds by their performance under each specification and identify the median; funds with better performance are the *winners* (W), and the rest are *losers* (L). Thus there are four categories of funds: the funds that are winners or losers under both classifications are WW and *LL*, and the winners under one and losers under the other ranking are *WL* and *LW*.

To assess the association level of classifications using $K \times K$ contingency tables, Cohen [1960] introduces the kappa statistic.* This measure is useful for assessing the agreement of judgments made by different raters. In our case, these raters are the performance measures themselves. The kappa aims to measure the proportion of matching pairs in the table that is not due merely 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 $\widehat{\kappa}$, is given by:

$$\widehat{\kappa} = 2\frac{WW + LL}{N} - 1 \tag{6}$$

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

I define in turn four thresholds: $\upsilon_1 = 0.2$, $\upsilon_2 = 0.4$, $\upsilon_3 = 0.6$, and $\upsilon_4 = 0.8$, on the basis of a refinement of the standard scaling proposed by Landis and Koch [1977]. Let $i = P, C, S, k$. Landis and Koch propose this interpretation: 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, in that order. The intervals are further split into lengths of 0.1 to create ten zones. Zone 1 corresponds to slight—lower range, and zone 10 corresponds 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

I evaluate both the precision and the stability of performance measures.

Precision

To examine the precision of the performance measures, one is confronted with the difficulty that it is impossible to determine with certainty on what risk dimension mutual fund managers tend to distinguish themselves; that is, what the "true" measure of performance is. Yet, if this measure could be identified, the rankings produced using alternative benchmarks would remain invariant under this particular performance measure, while the use of an imprecise metric would produce only random classifications. Different benchmarks would thus yield unrelated classifications. Therefore, I rely on comparisons of 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 test results. In general, the lower bounds obtained with the non-parametric statistics reported in Panel B, especially when using Cohen's kappa, are more

Exhibit 4
Empirical Precision of Performance Measures

Panel A: Parametric association measures

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

Panel B: Non-parametric association measures

		Spearman rank			Cohen's kappa		
Model	*Measure*	$\hat{\rho}_S$	$L_{\psi,S}$	zone	$\hat{\rho}_\kappa$	$L_{\psi,\kappa}$	zone
Three-factor	$\hat{\alpha}$	0.894	0.786	8	0.722	0.560	6
	GTR	0.909	0.809	9	0.833	0.704	8
	IR	0.664	0.485	4	0.389	0.173	2
Four-factor	$\hat{\alpha}$	0.799	0.655	7	0.667	0.492	5
	GTR	0.920	0.828	9	0.889	0.782	8
	IR	0.610	0.423	5	0.500	0.297	3
Five-factor	$\hat{\alpha}$	0.873	0.756	8	0.778	0.631	7
	GTR	0.931	0.844	9	0.833	0.704	8
	IR	0.655	0.474	5	0.444	0.234	3

For each measure, the observed value $\hat{\rho}$ and lower bound with a one-sided confidence at the 5% level L_ψ are reported. For each measure and association measure, the reported zone corresponds to the lowest value i such that $L_\psi \leq 0.1i$, $i = 1, \ldots, 10$.

conservative than the ones obtained with parametric statistics. A noticeable exception is the set of values of the Spearman rank correlation obtained by the generalized Treynor ratio for the three-factor and 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 aims to correct for 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 is in zones 8 to 10, or the upper range of substantial and almost perfect associations. The observed values themselves lie in zones 9 to 10, always supporting an almost perfect relationship. Considering a symmetric upper bound for the confidence intervals would not lead to rejecting the perfect association hypothesis for the Spearman or Pearson coefficients (except for the four-factor model).

Contrary to the homogeneous behavior of the *GTR*, the different measures of association estimated with alphas are not stable. The poorest association intensities are observed for 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. In fact, no measure is able to support the hypothesis of perfect association ($\rho_i = 1$) when performance is measured using alpha.

Observed association measures for alpha are systematically below the associations for the *GTR*, except for Lin's concordance coefficient measured with the three-factor and the five-factor models. The difference worsens when one compares lower bounds of confidence intervals, as the estimators of association coefficients always display greater 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 highest Pearson and Spearman coefficients are 0.695 and 0.664, respectively, and no metric supports a hypothesis better than a moderate association between rankings produced with this performance measure, except the Pearson coefficient with the three-factor model.

Stability

The empirical analysis of stability of performance measures does not suffer from the same problems as analysis of precision. I use the style-based benchmark indexes 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. Panel B displays the same picture, although less pronounced, as Exhibit 4. With non-parametric association measures, the stability of the generalized Treynor ratio is confirmed, while the information ratio displays the lowest association values. The differences between performance

Exhibit 5
Empirical Stability of Performance Measures

Panel A: Parametric association measures

		Pearson correlation			Lin's concordance		
Model	*Measure*	$\hat{\rho}_P$	$L_{\psi,P}$	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
	GTR	0.962	0.897	9	0.948	0.923	10
	IR	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
	GTR	0.675	0.500	6	0.464	0.349	4
	IR	0.968	0.908	10	0.954	0.928	10

Panel B: Non-parametric association measures

		Spearman rank			Cohen's kappa		
Model	*Measure*	$\hat{\rho}_S$	$L_{\psi,S}$	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
	GTR	0.974	0.919	10	0.889	0.782	8
	IR	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
	GTR	0.982	0.937	10	0.889	0.782	8
	IR	0.970	0.912	10	0.833	0.704	8

For each measure, the observed value $\hat{\rho}$ and lower bound with a one-sided confidence at the 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, ..., 10.

measures are not very great, though, especially when one shifts from the four-factor to the five-factor model. When asset pricing models are more underspecified, the relative outperformance of the *GTR* picks up. The lower bound for the confidence level of the Spearman rank coefficient is on the order of 10% higher with the *GTR* than with the alpha.

Interpretation

Overall, the generalized Treynor ratio appears to produce better results in terms of both precision and stability. Panel A of Exhibit 5 reveals a very important caveat in this respect. The change from the four-factor to the five-factor model harms the stability of the generalized Treynor ratio, but the other measures maintain very high stability levels.

The explanation may be found in the impact of the fund's required return on its measure of performance. With introduction of the fifth factor, two of the funds experience a significant drop in 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 *GTRs* tend to rocket up in absolute value and to drive down parametric coefficients of association.

This behavior indicates that non-parametric statistical methods are much better for assessing 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 correction of the leverage bias inherent in 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, amplifying the abnormal performance of managers proportionally with the leverage of their portfolios.

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 a significantly non-normal distribution of regression residuals. Measurement issues lead to strong overestimation of the residual variance, in turn biasing the information ratio downward (Ledoit and Wolf [2004]). Yet measurement issues alone cannot fully explain the poor values obtained by the *IR* on non-parametric association measures.

I show overall that portfolio managers distinguish themselves more clearly according to the generalized Treynor ratio than the other two measures. This result suggests that the managers of style-based portfolios consider their funds complements to other active portfolios held by investors. Furthermore, contrary to the way alphas are measured, controlling for portfolios' required returns allows more accurate assessment of management skills.

CONCLUSION

The empirical use of the alpha as a performance measure adjusted for systematic risk has never really been questioned for multi-index models. I have attempted to provide some empirical evidence in the assessment of performance measures.

With a sample of directional mutual fund returns, the results provide strong support for the generalized Treynor ratio. Rankings produced by the *GTR* are slightly more reliable than cardinal values, as the ratio can produce extremely sensitive outputs for portfolios with a low level of required return. When it is applicable, i.e., when the required return on managed portfolios is expected to be positive, the generalized Treynor ratio has better ranking abilities than its competitors, the alpha and the information ratio, provided that comparisons are made 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], hampers the ability of Jensen's alpha to rank leveraged portfolios on the basis of their performance. My study provides some support for suspicion of the relevance of this phenomenon for the selected sample.

Important additional work remains with respect to assessment of the quality of performance measures. The proper measurement of performance for non-directional funds is not addressed here, 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.

ENDNOTES

The author thanks Pascal François, Eric Jacquier, Denis Larocque, Nicolas Papageorgiou, and Peter Schotman for helpful comments. He acknowledges financial support from Deloitte Luxembourg and a research grant of the Belgian Funds for Scientific Research (FNRS). Some of this research was carried out while visiting HEC Montréal.

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

REFERENCES

Bodie, Z., A. Kane, and A. J. Marcus. *Investments,* 6th ed. 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 Fund 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: John Wiley & 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." *The 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." *The 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.

————. "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." *The 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." *The Journal of Investing*, vol. 3, no. 3 (Fall 1994), pp. 59-65.

Treynor, J. L. "How to Rate Management Investment Funds." *Harvard Business Review*, vol. 43, no. 1 (January/February 1965), pp. 63-75.

————. "Toward a Theory of Market Value of Risky Assets." Working paper, 1961. Subsequently published in R. A. Korajczyk. *Asset Pricing and Portfolio Performance: Models, Strategy and Performance Metrics*. London: Risk Books, 1999.

Zhou, G. "Asset-Pricing Tests Under Alternative Distributions." *Journal of Finance*, vol. 48, no. 5 (1993), pp. 1927-1942.