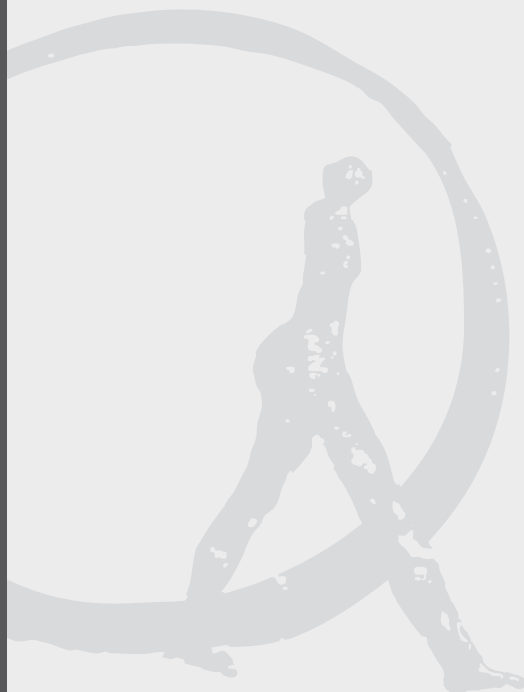


Rating the ratings

A Critical Analysis of Fund Rating Systems



Noël Amenc

Professor of Finance and Director of the EDHEC Risk
& Asset Management Research Centre

Véronique Le Sourd

Research Engineer for the EDHEC Risk
& Asset Management Research Centre

Abstract

Fund ratings play a significant role in fund selection for investors: the funds being made available are more and more numerous and increasingly diversified in terms of assets and style management. Ratings are in fact a simple tool with which to evaluate and compare the respective qualities of funds with regard to their risk and return. Moreover, numerous studies performed in the United States have shown that investors are widely influenced by fund ratings in making their choice, whether it is for an initial subscription or with a view to readjusting their investment in the event of rating upgrades or downgrades. Meanwhile, the various ratings do not offer the same properties in terms of performance evaluation and risks taken into account. This article provides a comparative analysis of a selection of leading rating systems that are very familiar to investors, including Standard & Poor's star rating, the Morningstar rating, and the Lipper Leader rating, and also addresses the required properties for an optimal rating system. This critical analysis was carried out in 2004 to allow for the development of a new European rating system, the Europerformance-EDHEC style ratings, which are not evaluated here.

EDHEC is one of the top five business schools in France. Its reputation is built on the high quality of its faculty (104 professors and researchers from France and abroad) and the privileged relationship with professionals that the school has been developing 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. Its 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.

Introduction

Fund ratings are a widely used tool for fund promoters and fund subscribers. They serve to evaluate fund performance on a risk and return basis in an easily understandable way, and allow the performance of different funds to be compared. In this context, the quality and the robustness of the ratings is a critical subject for both investment management firms and investors. Though the predictive capability of fund ratings has not been proved, numerous studies performed on US mutual funds have concluded that fund subscribers are widely influenced by fund ratings in making their choice. Moreover, if the rating of a fund is downgraded, one observes a transfer of money towards other funds.

We carried out a comparative analysis¹ of three well known rating methodologies: Standard and Poor's star rating, the Morningstar rating and the Lipper Leader rating, whose results are published extensively on the promoters' web sites. From this analysis it appears that these ratings do not deal adequately with three aspects of fund evaluation:

- the first concerns risk, both in terms of measuring the risks that were really taken by the manager and the necessity of taking extreme risks into account;
- the second concerns the fact that ratings are used by investors to make investment decisions and not only to reward past performance. Unfortunately, despite its usefulness, performance persistence measurement is not yet a major concern for rating agencies;
- the third relates to confidence in the results. Most of the rating systems use a relative ranking of fund performance within each category of funds, which makes the ratings attributed totally dependent on the definition of the categories.

After introducing the various quantitative methods, we provide a detailed comparison of the various methodologies, underlining their respective strengths and weaknesses. The aim of this study is to supply precise information on each rating system, so that they may be used with all due caution.

Overview of the various quantitative fund ratings²

Standard & Poor's Star Rating³

Standard & Poor's Star Rating is based on the evaluation of the fund information ratio relative to the mean performance of funds belonging to the same category. Funds are divided into about 160 asset categories, which relate to the country, sector, asset type, etc. A minimum of 3 years of historical data is required for the fund to be rated, as well as a minimum of 6 funds belonging to the same category during the evaluation period.

Star attribution is based on the observation that the higher the information ratio, the greater the fund's ability to consistently outperform its peers during the evaluation period, and so the more stars it should be given. The stars are attributed as detailed in the table below:

Fund percentile	Rating
Top 10%	5*
10% to 30%	4*
30% to 50%	3*
50% to 75%	2*
Bottom 25%	1*

1 - See Amenc and Le Sourd (2005) for a more detailed analysis of the ratings.

2 - More details concerning the methodologies are provided in the appendix.

3 - <http://www.funds-sp.com>

The stars are recalculated every month. Standard & Poor's stress that their star rating is not intended to be predictive and that past performance does not give information on future performance.

This methodology appears simple to understand. However, the existence of numerous categories may make it difficult to compare the performance of funds between categories. The aim of Standard & Poor's is to reward funds that behave in the same way as the funds belonging to the same category, while exhibiting outperformance, so only the relative risk of the fund with regard to the average of the category is taken into account and no indication of the fund's absolute risk is given. The information ratio used to determine the rating is highly dependent on the benchmark chosen to calculate this ratio: different benchmarks yield different information ratios. The use of an average of the funds belonging to a specific category as a benchmark does not ensure that the risk of all the funds in the category will be perfectly represented.

Morningstar Rating⁴

The Morningstar Rating was introduced in 1985 in the United States. In 2002, Morningstar revised its methodology, replacing broad fund categories with narrow groups, in order to reduce the influence of style management in fund ratings. These categories are intended to define groups of funds whose members are similar enough in their risk factor exposures for the return comparisons between them to be useful. For example, the broad domestic equity category has been divided into large-cap, mid-cap and small-cap equity categories, which prevents large-cap funds from receiving a disproportionate percentage of high star ratings in the equity class when large-cap stocks dominate small-cap stocks. This improvement allows manager skill, rather than market circumstances, to be integrated better in its relative rating. Funds are grouped by the type of investments that are predominant in their portfolios.

For star attribution, each fund is placed in the specific Morningstar category indicated in the most recent monthly record. The rating is based on a risk-adjusted performance measure. For each category, Morningstar calculates a three-year star rating for all member funds that have at least 36 continuous months of total return data, up to and including the evaluation month, by computing the load-adjusted MRAR(2)⁵ of total returns for the 36 months ending in the evaluation month. All the funds in a peer group are then ranked in descending order of their MRAR(2) and a number of stars is attributed to each fund according to its position in the distribution of MRAR(2) values, as detailed in the table below:

Fund percentile	Rating
Top 10%	5*
10% to 32.5%	4*
32.5% to 67.5%	3*
67.5% to 90%	2*
Bottom 10%	1*

The rankings are updated each month. Five-year and ten-year ratings are also produced in the United States, if the data is available, together with an overall star rating based on the weighted average of the number of stars assigned to it in the three-year, five-year and ten-year rating periods.

A problem of age bias in risk estimation was described in Blume (1998), Morey (2002) and Vinod and Morey (2002) before the new methodology was introduced. According to Adkisson and Fraser (2003b, 2004), despite the improvements to the methodology, a potential age bias still exists in the ratings. This problem is linked to the weighting of the different periods, which is quite arbitrary and puts relatively little weight on the most recent performance. This can lead to a high rating for a mutual fund with relatively poor recent performance, but strong longer-term returns.

4 - <http://www.morningstar.com> and <http://datalab.morningstar.com>
 5 - See the appendix for the definition of this indicator.

Lipper Leader Rating⁶

The Lipper Leader rating system uses some criteria related to investors' concerns to deliver a description of a fund's success in meeting their goals. Consistency of Return and Preservation of Capital are the two major criteria. A further two have been added more recently: one to measure Total Return and the other to measure Expense Ratios. In addition, a Tax Efficiency criterion is used in the United States only. According to Lipper, the aim of this rating system is to provide investors with a scorecard to help them find funds which meet their priorities. Funds are compared with similar funds, and only those that truly stand out are awarded Lipper Leader status.

Scores for Total Return, Consistent Return, Expense and Tax Efficiency are computed for all Lipper classifications with five or more distinct portfolios and span both equity and fixed income funds (e.g., large-cap core, general U.S. Treasury, etc.). Scores for Preservation are computed from three broad asset classes – equity funds, mixed-equity funds and bond funds (e.g., equity includes U.S. diversified, sector, and world equity funds). Each fund is ranked against its peers based on each of the metrics (Total Return, Consistent Return, Preservation, Expense and Tax Efficiency) and scores are attributed for each metric in each peer group as detailed in the table below:

Fund percentile	Scores
Top 20%	Lipper Leaders
20% to 40%	2
40% to 60%	3
60% to 80%	4
Bottom 20%	5

The scores are subject to change every month and are calculated for the following periods: three-year, five-year, ten-year and overall. The overall calculation is based on an equal-weighted average of percentile ranks for each measure over three-, five-, and ten-year periods (if applicable), which gives greater weight to more recent results, as the latest three years are included in each of the three periods.

The strength of the Lipper scores is their use in conjunction with one another. They can be used together to identify funds that meet the particular characteristics of the investor. On the other hand, all the different scores are computed independently and there is no single summary score for each fund. This may render it difficult for an investor to reach an opinion about a specific fund, as Lipper do not supply a global rating.

Critical Analysis

In this section, we consider each aspect of the various ratings and analyse them with regard to the expected requirements of users.

Fund Categories

The first point to consider is how the fund categories are derived and how funds are split into the various categories. This is an important point as the results of the ranking may be dependent on this split. The promoters of ratings group funds into peer groups to calculate the return and risk of the funds relative to these peer groups. For rating systems that use *a priori* defined categories, the most important criticism concerns the difficulty in defining homogenous peer groups, together with the dilemma of choosing between wide or narrow categories. This problem is shared by the three rating systems, Standard & Poor's, Lipper and Morningstar.

Broad Categories versus Narrow Categories

- Broad categories do not take a specialised manager's style into account. They include funds practicing very different investment management, with no comparable performance. In that way, a "domestic stocks" or "international stocks" category does not recognise allocation by style. Morningstar's previous rating system used broad categories. As a result, instead of rewarding the manager's skill, the attributed stars rewarded the funds whose investment style was favoured by the market during the evaluation period.

- The intention of narrow groups is to take style management into account better, but in that case it is difficult to identify funds that truly behave in a similar way in order to group them together. Moreover, narrow categories do not capture the characteristics of strategies based on diversification between multiple styles or asset classes and used by numerous funds. However, some categories had to be defined broadly to incorporate enough funds for a meaningful comparison group. This implies that there are larger style differences among funds in some categories. In its new methodology, Morningstar uses narrow groups.

Split of funds between categories: the issues of fund classification

- Knowing the composition of the fund

In order to define precise categories corresponding to a specific style, it is appropriate to have information concerning the fund composition that is reliable and representative of the fund's investment policy over the period. When one knows the portfolio holdings, the classification between the groups may be based on the study of the characteristics of each asset the fund is made up of. This is the method used by Morningstar. The style of a fund is then determined as the weighted average of the style of the various assets. This is performed each month and the fund is assigned to a category based on its three-year "normal" style. Meanwhile, in order to describe the fund's style according to the assets it is made up of, the portfolio composition should be known not only at the beginning of the period, but throughout the analysis period. Most of the time, this information is not delivered by managers. For instance, a study carried out in 1995 on the data provided by the two main disseminators of information on mutual funds showed that only 1/5 of the portfolios contained in "Morningstar Ondisc" had updated less than three months previously and only half of the funds in the "Value Line Fund Analyser" detailed their composition.

- Correctly identifying the fund style: value versus growth

The difficulty of *a priori* classification of securities according to their characteristics has been highlighted by numerous studies⁷. Although it is relatively easy to obtain a consensus on the segmentation of classes, sectors or countries, style analysis, on the other hand, relies on more subjective classification criteria. Commonly used attributes such as the book/price or earnings/price ratios are unstable and depend as much on market conditions as on company-specific qualities. Moreover, the characteristics of a large number of securities do not allow satisfactory discriminatory analysis to be carried out. The growth or value classification of a security, for example, is neither stable nor objective. In a Bienstock and Sorensen study (1992), it appears that only 20% of the stocks, from a sample of 3,000, can, to a significant degree, be classified as value or growth. On this subject, Haslem and Schegara (2001) showed that the Morningstar style classification for funds invested in large caps – large growth, large blend and large value – was not consistent with a "clustering" type statistical approach, which allows the funds to be grouped together according to their behaviour, without any *a priori* choice. To take this problem into account in its new system, Morningstar measures the value-growth dimension using ten factors – five value factors and five growth factors. Still, no academic research has been published on the subject to show that returns can be better explained when the value-growth dimension is measured using value and growth factors instead of value factors alone.

- Can we trust the manager's self-declared style?

The difficulty in knowing the reality of the funds' investments can lead the person responsible for measuring the funds to rely on the manager's stated objectives and investment styles. However, the category declared is not always the true category of the fund, as was shown by diBartolomeo and Witkowski (1997), who performed a style analysis on 748 funds and found that 40% of them were in a category other than the one declared. Another difficulty is related to the fact that the style of a fund may not be stable over time, so the category in which the fund is classified may differ from its current style category. Kim, Shukla and Tomas (2000) sought to determine whether mutual funds actually adhered to their stated objective or whether their style drifted. To test for style drift, the authors evaluated funds based on many well-established investment criteria, such as P/E and B/M ratios, market capitalization, dividend yield, and so on. They then compared the fund's actual style to its Morningstar classification. They used Morningstar's database for the period 1993-1996 and studied 1,043 funds.

7 - Cf. notably Lucas and Riepe (1996).

Their results are the following. Only 46% of funds had investment attributes that were consistent with the fund's stated objectives; 54% of funds were misclassified. Over one third of funds were severely misrepresented. Over the very short period covered by the study, 57% of the funds that survived changed their investment style at some point. Only 27% of funds held their investment attributes constant throughout the period. 33% of funds were found to be less risky than their stated objectives. 31% of funds were found to be more risky than their stated objectives. Jin and Yang (2004) have carried out an analysis of mutual fund objective classification in China using statistical methods of distance analysis and discriminant analysis. They found that 50% of mutual funds were not consistent with their objective groups.

In order to illustrate the difficulty of categorising funds for rating agencies using *a priori* defined categories, we investigated the style exposure of funds distributed in France from various Morningstar Equity categories, using Sharpe's (1988, 1992) returns-based style analysis. This analysis was performed at the beginning of 2005 using three years of weekly returns. As a general conclusion, we observe that large cap categories are quite heterogeneous, while mid and small cap categories, made up of fewer funds, appear more homogeneous. The most surprising results have been obtained with the Morningstar Equity Euroland Large Cap category, which appears to contain a very large number of funds with a very pronounced exposure to the small cap style. *De facto*, the Morningstar categories do not allow the funds' style choices to be taken into account, even though these choices condition the essential part of fund performance. This weakness in the method leads to a comparison between funds whose risks are not comparable.

Fund Evaluation Criteria

Risk-adjusted measure and risk reference

For Standard & Poor's, Morningstar and Lipper ratings, the risk-adjusted measure, which is the core of fund ratings, refers to the average performance of a category, without the rating agency being completely sure that the category or the index really describes the fund's strategic allocation, and thus the risks to which the fund has been exposed. In the case of Lipper and S&P, this difficulty is exacerbated by the fact that the information ratio they use in their calculation is extremely sensitive to the choice of index.

These rating methodologies *a priori* derive a mean exposure to risks, based on an arbitrarily defined category classification that often leads to erroneous analysis of fund performance. Thus, a fund declared as large cap that in fact exhibits a significant exposure to the small cap style may outperform the large cap category during some periods, though this gap is not related to the manager's added value with regard to stock picking or market timing. Empirical studies have shown that risk-adjusted performance measures relative to an index or a category defined *a priori* generally lead to rewarding atypical managers whose allocation is not in line with that of their category.

Measure of extreme risks

With the noticeable exception of the "Lipper Leader for Preservation", which is based on a risk of loss measure, the rating systems do not explicitly take into account or do not try to evaluate the fund's extreme loss risks. Whether it is the indicator chosen (information ratio, volatility) or how risk aversion is modelled, risk-adjusted measures consider that the investor is averse to the investment's average risk, but do not have any specific point of view with regard to extreme risks.

Performance Persistence

Influence of fund rating on investor's choice

Most of the rating systems stress that their ratings are not intended to predict future performance, but serve to evaluate past results in terms of risk and return. This is the case for Standard & Poor's and Morningstar. So investors will be able to check *a posteriori* whether they made good choices or not, and fund managers will be able to communicate on the ability they have demonstrated in the previous periods, but this does not say anything on what will happen during the periods to come. Meanwhile, despite the warnings given by rating agencies, most investors are greatly influenced by ratings and tend to use them as a predictor of future performance. Investors avoid funds with the lowest ratings and choose funds with the highest ratings, in the expectation that the best

funds will continue to outperform. For the same reasons, they tend to sell funds that have been downgraded and to keep funds that maintain the highest ratings.

For example, according to a study reported in the Boston Globe and the Wall Street Journal in 1995, about 97% of new investments that year were in mutual funds that had previously been awarded four or five stars. In addition, Del Guercio and Tkac (2001) found that an initial Morningstar 5-star rating results, on average, in six months of abnormal flows (53% above the normal expected flow). They also found significant abnormal flow in the case of rating changes, with positive flow for rating upgrades and negative flow for rating downgrades. Adkisson and Fraser (2003a) also present significant evidence that investors withheld funds from mutual funds that lost stars, but did not proportionately reward funds that gained stars.

The importance of measuring fund performance persistence

Indeed, investors appear to be concerned with the persistence of ratings, i.e. the capacity of a fund to reproduce its performance, as well as the initial rating of the funds in which they invest money. In fact, knowing the rankings of funds is of little interest if it is not possible to use the results for future investment choices. The capacity of rating systems to detect funds exhibiting performance persistence is controversial. Using Morningstar ratings before the methodology revision, Khorana and Nelling (1998) found that fund ratings exhibit persistence over a thirty-month period: highly rated funds tend to continue to be successful in the future, and poorly rated funds continue to underperform relative to their peers. Blake and Morey (2000) found somewhat different results. They found that low ratings indicate relatively poor future performance, while there is little evidence that the highest-rated funds outperform the medium-rated funds. Moreover, the ratings are not a better predictor of performance than the fund's past average monthly returns. Finally, Morey (2003) showed that three years after a fund receives its initial Morningstar 5-star rating, fund performance falls off severely.

Ratings that really take performance persistence into account are thus intended to meet investors' expectations. Unfortunately, in spite of its usefulness, and in spite of the numerous studies dedicated to performance persistence, the persistence measure is not of major concern to most rating agencies. Only Lipper, which includes the Lipper Leader for Consistent Return in its five rating scores, takes this aspect into account in its fund evaluation. It uses the Hurst exponent⁸ as a criterion. This indicator measures the capacity to produce outperformance smoothly, without excessive volatility. Lipper establishes the rating relative to a peer group. Funds are ranked from the best to the worst according to the persistence criteria and the highest 20% receive the best rating. Lipper separates the funds into three groups according to whether their H exponent is greater than 0.55, between 0.45 and 0.55, or less than 0.45. As this value of 0.55 is more demanding than the threshold of 0.5, which ensures that there is persistence in performance, this ensures that the performance of the funds in the first group is strongly persistent. Unfortunately, attributing the best mark with regard to persistence to the 20% best H value rated funds does not ensure that all these funds exhibit a H exponent higher than 0.5. Nothing prevents the last funds in this first quintile of funds from having a Hurst exponent lower than 0.5. If they exist, these funds should be identified, so that they are not ranked as a Lipper Leader for Persistence.

Measuring the fund's true alpha

In order to correctly identify fund performance persistence, it is appropriate to evaluate the true fund performance. However, in the rating methodologies described above, the fund performance is considered as a whole, without making a distinction between risk premiums, i.e. the reward for the various risk factors (the betas) to which the fund is exposed, and the specific performance due to manager skill (the alpha). In that case, it is difficult to try to identify any kind of persistence in the performance. Assuming that the performance as a whole can persist would be equivalent to supposing that style and asset class returns are auto-correlated and persistent, which does not conform with the weak form of market efficiency.

Alternatively, in separating alphas from beta risk premiums, it is possible to isolate the success of management decisions from the returns due to the economic situation. We recall that the return of a portfolio or a fund may be written in the following way:

$$R_{it} - r_f = \alpha_i + \beta_{i1}(F_{1t} - r_f) + \beta_{i2}(F_{2t} - r_f) + \dots + \beta_{iK}(F_{Kt} - r_f) + e_{it}$$

8 - The definition of Hurst exponent is provided in the appendix.

where α_i is the abnormal return of the fund due to stock picking and/or market timing; r_f is the risk free rate; β_{ik} is the sensitivity of the fund to factor k; F_{kt} is the return of factor k during period t and the products $\beta_{ik} F_{kt}$ evaluate the return due to the risks taken.

Alpha is in fact the only reliable performance measure for active management. While the persistence of style and asset class returns over a long period has not been proved, there is a higher probability that the manager's ability may persist over time and it is possible to seek to evaluate this persistence. Indeed, the fact that an investment style performs well or badly should not be confused with the manager's skill in picking the right stocks within the style he or she has chosen. A manager's skill in practising a well-defined style should thus be evaluated in comparison with a benchmark that is adapted to that style, so that his or her ability to achieve returns for the different risk factors may be identified.

Ratings

The rating systems considered reward funds by using a star system, attributing one to five stars. Lipper use scores instead of stars but the principle is similar as there are five different scores. In addition, Lipper attribute several ratings, one for each of the criteria used to evaluate funds. Investors are free to consider the indicator or the combination of indicators that best suit their needs. The drawback is that there is no general indicator to sum up the others in order to obtain general advice on the funds.

Influence of percentile distribution

All three rating systems use a pre-defined percentile distribution to attribute their stars. For a given rating system and depending on the period, the quality of funds that receive five stars may differ. It is also difficult to compare the same number of stars between the various rating systems. Depending on the percentile distribution chosen, the total percentage of funds that receive a medium to high rating may differ greatly and the same number of stars does not therefore correspond to the same relative standing of a fund, as can be seen from the table below.

S&P	Morningstar	Lipper	
Number of funds with 3 stars or more	50%	67.5%	60%
Position of a 3-star fund	Within the top 31% to 50%	Middle third of ranked funds	Between 40% and 60% deciles
Position of 1-star fund	Bottom quartile	Bottom decile	Bottom 20%

Compared with Standard & Poor's, the Morningstar rating tends to lead one to believe that the number of good funds is higher, so the choice of the percentile distribution influences investors' perceptions of the results.

Influence of fund category definition

These stars are attributed within each category. This approach makes the ratings attributed totally dependent on the definition of the categories. However, the increasing numbers of investment products and managers prevent the rating agency from being sure that all the funds in a category are alike. There may be style bias. Ratings of funds that do not have the same risk profile are thus compiled, and consequently the information given to investors is not always reliable. Therefore, a large number of the stars attributed by Morningstar in reference to a ranking of fund performance relative to a pre-defined category are erroneously attributed, since the rated funds have an allocation that differs from the initial allocation, and have not therefore been exposed to the same risks during the whole period. Whether the choice concerns an index or a category, this kind of approximation leads more to rating whether the style to which the fund has been exposed has been favoured by the market than really evaluating the manager's skill, so these ratings do not have any predictive power, unless one thinks that style returns tend to repeat or trusts in the random walk of markets. Moreover, as there is a ranking by category, the more categories there are, the more funds receive a high star rating. Even the weakest sector is guaranteed its share of five-star funds. The size of the categories is likely to influence the rating of the funds. As a result, some funds may obtain a different number of stars depending on whether they are integrated into a relatively narrow category or a larger one.

Transparency and Comprehensibility of the Ratings for Investors and Managers

As a major source of information for investors and management firms, fund ratings have to be clear and easily understandable. It is not only a matter of describing the calculations and the rating criteria chosen, but also allowing investors to recalculate and verify the ratings. This requirement for transparency is essential as it ensures that the information disseminated can be trusted. Moreover, as a communication tool, the ratings must be easily understandable and rely on calculations that are significant for professionals and individuals. This is a guarantee of possible arbitrage between the quality of various methodologies and ratings, which can help rating agencies to improve their methodology and correct their possible mistakes.

Conclusion

It appears that various points should be improved in the existing ratings, so that they are not only an appraisal of past results, but also a tool serving future investment choices. First, greater attention has to be given to risk evaluation, and specifically to a fund's extreme loss potential. Second, ratings appear too dependent on the definition of fund categories. Finally, persistence in performance is not clearly identified.

The table below sums up our evaluations of the various criteria contributing to the ratings.

Name	Standard & Poor's Stars	Morningstar Stars	Lipper Leader
Measure of Risk Actually Taken	Poor	Fair	Poor
Measure of Extreme Risks	No	No	Fair
Measure of Performance Persistence	No	No	Good
Robustness and Confidence	Poor	Poor	Poor
Transparency and Comprehensibility	Very Good	Very Good	Good

In view of these results, we propose several improvements:

- In terms of risk evaluation, it seems to us that the introduction of an evaluation of extreme risks into the rating system, with measures such as Value-at-Risk, is essential and should be considered by rating providers, in view of the fact that funds are increasingly turning to derivative instruments and may be invested in assets that present non-normal distributions, which makes the use of volatility as the only risk measure illusory.
- We also propose that ratings should be based on alpha measurement, which does not depend on the prevailing situation in the financial markets, but only on the manager's skill, the persistence of which can be measured, by considering the capacity of managers to repeat alpha. For this purpose, we suggest the use of Sharpe's returns-based style analysis both to identify the benchmark that best suits the fund style and to evaluate the fund alpha. Using style analysis enables portfolio performance to be evaluated using a relevant style peer group. This approach allows the style of the portfolio to be identified according to its behaviour, whether or not it is made up of the asset classes that describe its style. Ratings in which each fund is compared to its own customised benchmark are more reliable than ones that compare funds to a benchmark shared by all funds in the same category. Moreover, this avoids the problem of fund misclassification in predefined categories.
- Finally, it seems to us that fund ratings should be free of any specific choice of categories. We think that categories should only serve as a common way to display the results and should not influence fund rankings in any way. Moreover, funds should satisfy defined rules to receive a certain number of stars. More specifically, the percentile number of funds that will receive the maximum number of stars should not be pre-determined, so that the quality of five-star funds remains equal all the time. This is required to ensure robustness and confidence in the results.

Appendix

Standard & Poor's Star Rating

Ratings are based on a fund's monthly performance relative to the sector average over each of the preceding 36 months.

The return of each fund in the category is computed for each of the 36 months in the period. The return of fund i during month t is denoted by R_{it} . The mean average return of all the funds in the category is also calculated for each month. Assuming there are N funds in category S , the monthly mean average return R_{St} is given by:

$$R_{St} = \frac{1}{N} \sum_{i=1}^N R_{it}$$

The relative monthly return of each fund is computed as the difference between the fund's monthly return and the monthly mean average return of the funds in the category, i.e. $R_{it} - R_{St}$. The calculation of the average of the fund's latest 36 relative monthly returns enables its average outperformance or underperformance compared to the other funds in the same sector to be measured. The higher the average, the more the fund has outperformed its peers over the 36 months.

The volatility of the fund's relative return is then computed as:

$$\sigma(R_i - R_S) = \sqrt{\frac{1}{36} \sum_{t=1}^{36} ((R_{it} - R_{St}) - (\bar{R}_i - \bar{R}_S))^2}$$

$$\text{where } \bar{R}_i = \frac{1}{36} \sum_{t=1}^{36} R_{it} \text{ and } \bar{R}_S = \frac{1}{36} \sum_{t=1}^{36} R_{St}. \text{ We note that } \bar{R}_S = \frac{1}{N} \sum_{i=1}^N \bar{R}_i.$$

Volatility measures how much a series of values deviates around its average. The calculation of the volatility of the fund's 36 relative monthly performances allows for the measurement of how consistently the fund has outperformed or underperformed its peers. The higher the volatility, the less consistent the fund's performance relative to its peers. It can be seen as a tracking error indicator between the fund and the average of the category.

Finally, Standard & Poor's Relative Risk Adjusted Ratio is calculated by dividing the fund's average relative performance by the volatility of its relative performance. This is in fact an information ratio. The information ratio is equivalent to a Sharpe ratio where the riskless asset has been replaced by the return of a benchmark. Here, the benchmark is the average return of the sector. The information ratio (IR) can be formally written as follows:

$$IR = \frac{\sum_{t=1}^{36} (R_{it} - R_{St})}{\sigma(R_i - R_S)}$$

This ratio measures the ability of the fund to outperform its peers on a consistent basis. The funds are ranked on the basis of this ratio. Funds that tend to do well under this method are those that behave similarly to their sector (they rise and fall together) but steadily pull ahead of their peers over time.

For this calculation, each month is given an equal weight. Management fees, custodian charges, and all other running costs are taken into account in the computation of fund returns.

Morningstar Rating

Performance measurement

Morningstar calculates a fund's total return for a given month as follows:

$$R = \left\{ \frac{P_e}{P_b} \prod_{i=1}^n \left(1 + \frac{D_i}{P_i} \right) \right\} - 1$$

where:

- P_e is the NAV per share at the end of the month;
- P_b is the NAV per share at the beginning of the month;
- D_i is the distribution per share at time i ;
- P_i is the NAV reinvested per share at time i ;
- n is the number of distributions during the month.

Distributions include dividends and distributed capital gains, where applicable. All calculations are based on pre-tax return data.

If there were no loads or redemption fees, the cumulative value of one dollar over a period of T months would be:

$$V_u = \prod_{t=1}^T (1 + R_t)$$

If there are loads or redemption fees, the cumulative value adjusted for loads and redemption fees is given by:

$$V = (1 - F)(1 - R)V_u - D(1 - F) \frac{\min(P_0, P_T)}{P_0}$$

where:

- F is the front load;
- D is the deferred load;
- R is the redemption fee;
- P_0 is the NAV per share at the beginning of the period;
- P_T is the NAV per share at the end of the period.

Morningstar Risk-Adjusted Return

The Morningstar Risk-Adjusted Return (MRAR) measure has the following characteristics. It assumes no particular distribution of excess returns and risk is penalised in all cases.

MRAR is motivated by expected utility theory, according to which an investor ranks alternative portfolios using the mathematical expectation of a function (called the utility function) of the ending value of each portfolio. According to financial theory, each individual has his own utility function. Nevertheless, in order to be able to establish a fund ranking, Morningstar assumes that all investors use the same utility function. The function chosen is the iso-elastic (or power) function and is defined by:

$$u(W) = \begin{cases} -\frac{W^{-\gamma}}{\gamma} & \gamma > -1, \gamma \neq 0 \\ \ln(W) & \gamma = 0 \end{cases}$$

where W is the ending value of the portfolio being considered and γ is a parameter that describes the degree of risk aversion. The choice of this particular function is based on the desirable properties required to describe investors' attitudes towards risk. To be meaningful, the utility function must satisfy two conditions. First, it must

always be positively sloped; i.e. $u'(W) > 0$. That is, more expected wealth is always preferred to less expected wealth. Second, it must imply that investors are risk averse. This means that the investor prefers a riskless portfolio with a known end-of-period value to a risky portfolio that is expected, but not certain, to have the same end-of-period value. This is described by the fact that the utility of the expected wealth must be greater than the expected utility of the wealth:

$$u(E[W]) > E[u(W)]$$

From probability theory, it follows that this can be true only if the utility function is concave, i.e. $u''(W) < 0$. The degree of risk aversion of investors can be measured by the coefficient of relative risk aversion (RRA), defined by:

$$RRA(W) = -\frac{Wu''(W)}{u'(W)}$$

In the specific case of the utility function defined above, the relative risk aversion is equal to the constant $\gamma + 1$. Constant relative risk aversion utility functions are particularly suitable, as they imply that the risk aversion of the individual does not depend on the amount of his wealth.

Assuming W_0 is the value of the portfolio at the beginning of the investing period and R the return of the portfolio during the period, the terminal wealth of the investor will be $W = W_0(1 + R)$. Replacing W with this expression in the utility function will lead to:

$$u(W) = \begin{cases} -\frac{W_0^{-\gamma}}{\gamma} (1 + R)^{-\gamma} & \gamma > -1, \gamma \neq 0 \\ \ln(W_0(1 + R)) & \gamma = 0 \end{cases}$$

Instead of holding a risky portfolio, the investor could buy a risk-free asset. Let R_f be the return on the risk-free asset. To compare risky portfolios to the risk-free asset, we assume that the investor initially has all his wealth invested in the risk-free asset and that the beginning-of-period wealth is such that the end-of-period wealth will be \$1. Hence W_0 has the following value:

$$W_0 = \frac{1}{1 + R_f}$$

and

$$u(W_0(1 + R)) = \begin{cases} -\frac{1}{\gamma} \left(\frac{1 + R}{1 + R_f} \right)^{-\gamma} & \gamma > -1, \gamma \neq 0 \\ \ln\left(\frac{1 + R}{1 + R_f} \right) & \gamma = 0 \end{cases}$$

Defining the geometric excess return as: $r_G = \frac{1 + R}{1 + R_f} - 1$

we have:

$$u(1 + r_G) = \begin{cases} -\frac{(1 + r_G)^{-\gamma}}{\gamma} & \gamma > -1, \gamma \neq 0 \\ \ln(1 + r_G) & \gamma = 0 \end{cases}$$

The certainty equivalent geometric excess return of a risky investment is the guaranteed geometric excess return that the investor would accept as a substitute for the uncertain geometric excess return of that investment. Letting $r_G^{CE}(\gamma)$ denote the certainty equivalent geometric excess return for a given value of γ , this means that:

$$u(1 + r_G^{CE}(\gamma)) = E[u(1 + r_G)]$$

hence:

$$(1 + r_G^{CE}(\gamma))^{-\gamma} = E[(1 + r_G)^{-\gamma}]$$

and:

$$r_G^{CE} = \begin{cases} (E[(1 + r_G)^{-\gamma}])^{-\frac{1}{\gamma}} - 1 & \gamma > -1, \gamma \neq 0 \\ e^{E[\ln(1 + r_G)]} - 1 & \gamma = 0 \end{cases}$$

MRAR(γ) is defined as the annualised value of r_G^{CE} using the time series average of $(1 + r_G)^{-\gamma}$ as an estimate of $E[(1 + r_G)^{-\gamma}]$. With $\gamma \neq 0$, we have:

$$MRAR(\gamma) = \left[\frac{1}{T} \sum_{t=1}^T (1 + r_{Gt})^{-\gamma} \right]^{\frac{12}{\gamma}} - 1$$

where:

$$r_{Gt} = \frac{1 + R_t}{1 + R_{ft}} - 1 \quad \text{is the geometric excess return in month } t;$$

R_{ft} is the return on the risk-free asset in month t .

When $\gamma = 0$, MRAR is the annualised geometric mean of r_G :

$$MRAR(0) = \left[\prod_{t=1}^T (1 + r_{Gt}) \right]^{\frac{1}{T}} - 1$$

A rating system based solely on performance would rank funds on their geometric mean return, or equivalently, MRAR(0). A rating system that provides a heavier penalty for risk requires that $\gamma > 0$. Morningstar considers that $\gamma = 2$ results in fund rankings that are consistent with the risk tolerances of typical retail investors. Hence, Morningstar uses a γ equal to 2 in the calculation of its star ratings. It also calculates a risk indicator as MRAR(0) – MRAR(2), i.e. the difference between the mean annual geometric return and the certainty equivalent annual geometric excess return.

To calculate MRAR when there are loads and redemption fees, total monthly returns must be adjusted. Let a be the adjustment factor:

$$a = \left(\frac{V}{V_u} \right)^{\frac{1}{T}}$$

where:

V is the cumulative value adjusted for loads and redemption fees;

V_u is the cumulative value not adjusted for loads and redemption fees.

The adjusted total return for month t AR_t is given by:

$$AR_t = a(1 + R_t) - 1$$

where R_t is the total return for month t . Loads and fees can then be incorporated into the calculation of MRAR by replacing R_t with AR_t .

Lipper Leader

Lipper Scores for Total Return

Lipper Scores for Total Return reflect funds' historical total return performances relative to peers. A Lipper Leader for Total Return may be the best fit for investors who want the best return, without looking at risk. This measure alone may not be suitable for investors who want to avoid downside risk. For more risk-averse investors, Total Return scores can be combined with Preservation and/or Consistent Return scores to make a risk-return trade-off decision.

Total Return is defined as the return after (net of) expenses and includes reinvested dividends.

Lipper Scores for Consistent Return

Lipper Scores for Consistent Return measure which funds have delivered consistently superior returns compared with a group of similar funds. A Lipper Leader for Consistent Return may be the best fit for investors who value a fund's year-to-year consistency relative to other funds in a particular peer group. The measure is based on two indicators: the Hurst-Holder (H) exponent and the Effective Return.

The H exponent⁹ is a measure of risk or the jaggedness of the price series line. Funds with high H exponents tend to be less volatile than funds with low H exponents. H, then, identifies funds that are consistent in terms of volatility or risk.

Lipper identify three types of volatility behaviour: funds with high H values (greater than or equal to 0.55) tend to be less volatile than funds with a medium H exponent (H between 0.55 and 0.45), while funds that have low H exponents (H less than 0.45) tend to be the most volatile.

The Effective Return is a risk-adjusted return measure which is similar to the information ratio¹⁰. This ratio is based on the excess return of the fund, calculated by subtracting the relevant Lipper Global sector index return from the fund's return.

Lipper use both H and Effective Return to establish the ranking. If a fund has a high exponent and a positive Effective Return, this is an indication that the fund is a good one to purchase. If a fund has a high H exponent and a negative Effective Return, then the fund is probably exhibiting a negative P&L. If a fund has a H exponent in the low to medium range and a positive Effective Return, since H is the intrinsic measure of volatility for a fund, investors would expect the fund to have a higher Effective Return to compensate for the greater risk they will be taking compared with a fund with a higher H value. The extent to which the return needs to be higher will depend on investors' risk aversion. The ranking methodology is as follows:

1. All the funds in the classification are sorted in descending order based on their H exponent.
2. The funds with H values greater than or equal to 0.55 are separated out, as are those that are between 0.55 and 0.45 and those below 0.45.
3. In each group, the funds are now re-ranked in descending order based on their Effective Return.
4. The high (greater than or equal to 0.55) H funds with negative Effective Return are removed from the list of other high H funds and placed after the last fund in the less than 0.45 category.
5. The groups are now regrouped in their original order (high, medium, and low H), with the high H funds that have a negative Effective Return bringing up the rear.

Lipper Scores for Preservation

Lipper Scores for Preservation reflect funds' historical loss avoidance relative to other funds within the same asset class.

The utility of popular tools such as the Sharpe ratio may not be high for investors whose investment decisions are a function of absolute loss avoidance. The Preservation measure aims to help investors at various levels of risk averseness by providing a distinctive one-parameter measure of downside risk.

⁹ - More details on the H exponent are provided below.

¹⁰ - More details on the calculation of an information ratio are given above in the section describing Standard & Poor's Star Rating.

The Preservation model is defined as the sum of negative monthly returns over three-, five-, and ten-year performance periods, or:

$$\sum_{t=1}^n \text{Min}(0, \tau_t)$$

where $n = 36, 60$ or 120 months and τ_t is the return in month t .

For ease of use, the Preservation model assumes that the investor is more concerned with negative performance than below-target returns. In fact, a 1996 Investment Company Institute survey found that 51% of mutual fund investors think of risk as the chance of losing money.

Unlike the other Lipper measures, the Preservation measure is calculated from three broad asset classes (equity funds, mixed equity funds and bond funds) instead of from the investment classification level.

According to Lipper, choosing a Lipper Leader for Preservation may help to minimise downside risk compared to other fund choices in the same asset class.

Lipper Scores for Expense

Lipper Scores for Expense reflect funds' expense minimization relative to peers with similar load structures. A Lipper Leader for Expense may be the best fit for investors who want to minimize their total cost and can be used in conjunction with Total Return or Consistent Return scores to identify funds with above-average performance and lower-than-average cost.

One of the reasons given for looking at expenses (in conjunction with other measures such as total return, risk-adjusted return, etc.) is that higher costs could lead to lower net returns and hence poorer performance, unless they are more than offset by higher gross performance.

In an unpublished study available on his web site¹¹, William Sharpe found that the average Sharpe ratio for funds with the lowest expense ratios was over 75% greater than that of funds with the highest expense ratios. Although Professor Sharpe's study was not exhaustive (it did not include funds that were in existence at the start of the test period but "died" before the end of the test period), it is fair to say that at least directionally funds with higher expenses add more (and possibly far more) to expenses than they add to performance.

Lipper Scores for Expense differentiate funds that have minimized their expenses compared to competing funds. These costs are subtracted directly from a fund's return, so if two funds have equal returns before expenses, the lower cost fund will deliver higher net returns to an investor.

To score an individual fund, Lipper first look at the classification the fund is in and then at the load structure (funds within a classification are grouped into one of three load classifications –no-load/front-end load, back-end load/level-load, and institutional load). The fund is then ranked against its peers (similar classification, similar load structure) so a fund that has the lowest expense levels within a given classification and time period is designated as a Lipper Leader for Expense.

Lipper Scores for Tax Efficiency

Lipper Scores for Tax Efficiency reflect funds' historical success in postponing taxable distributions relative to peers. A Lipper Leader for Tax Efficiency may be the best fit for tax-conscious investors who hold investments that are not in a defined benefit or retirement plan account.

The impact of taxes on mutual fund returns has received increasing attention over the past few years, so much so that the Securities and Exchange Commission (SEC) and the Association for Investment Management and Research have issued papers outlining the preferred ways of computing tax efficiency.

Lipper have chosen one of these preferred methodologies – Relative Wealth – as the measure to compute Tax Efficiency. Lipper calculate Relative Wealth over three-, five-, and ten-year periods. Relative Wealth is defined as:

$$\text{Relative Wealth} = \left[\frac{1 + R_{at}}{1 + R_{bt}} - 1 \right] * 100$$

where:

R_{at} = pre-liquidation standard after-tax return

R_{bt} = SEC return (before-tax return)

Relative Wealth is a way of looking at both before- and after-tax returns in an intuitive and powerful manner. The actual numerical result of the Relative Wealth calculation explains how much value, as a percentage of change, is lost because of taxes as compared to pre-tax value (net of all expenses and loads). As a side note, Relative Wealth can be computed on a pre-liquidation or a post-liquidation basis. From an investor's standpoint, pre-liquidation relative wealth is a computation of the percentage lost to taxes prior to liquidating or cashing out the fund. Post-liquidation calculates the percentage lost to taxes including the gains (or losses) realized at liquidation. Lipper use pre-liquidation Relative Wealth as the methodology for computing Lipper Scores for Tax Efficiency, since they believe that most investors are long-term investors by nature and are not interested primarily in cashing out their fund holdings at a particular time.

Management fees, custodian charges, and all other running costs are taken into account for the calculation of all criteria.

Hurst exponent

The Hurst exponent relates to various areas of mathematics (e.g., fractals, autocorrelation, and wavelets, to name a few). The intuitively fractal nature of financial data such as the weekly return series has led many people to apply the mathematics of fractals and chaos in analysing these time series.

If funds' excess returns with regard to their benchmarks follow a random walk, this means that there is no correlation between the excess returns observed at time t and time $t + 1$. This leads one to believe that a portfolio manager is not exhibiting persistence in his/her returns. A high return is followed by a return which may be low or high with the same probability of $1/2$. Conversely, if there is some persistence in the series of returns, one may expect to find a positive correlation between the returns at time t and the returns at time $t + 1$. The process is no longer a random walk. As it happens, the Hurst exponent is specifically designed as a measure of deviation from the random walk, which corresponds to the case where the exponent is equal to $1/2$. If the Hurst exponent is higher than $1/2$, there is persistence and positive correlation between the series of returns. If the Hurst exponent is lower than $1/2$, there is no persistence and the correlation between the series of returns is negative.

The Hurst exponent therefore provides a measure of whether a time series has a long memory character or not. For each 3-year rolling window of data, the series of returns is recalculated by subtracting the mean return of the sample:

$$Z(t) = R_t - m$$

where $m = \frac{1}{T} \sum_{t=1}^T R_t$

Defining:

$$Y(t) = \sum_{s=0}^t Z(s)$$

$$Y_1 = \max_{0 < t < T} Y(t)$$

$$Y_2 = \min_{0 < t < T} Y(t)$$

exponent H is obtained through the following formula:

$$H = \frac{1}{\ln T} \ln \left(\frac{(Y_1 - Y_2)}{\sigma} \right)$$

where $\sigma = \sqrt{\frac{1}{T-1} \sum_{t=1}^T (R_t - m)^2}$.

As demonstrated above, estimating the Hurst exponent for a series of returns provides a measure of whether this data is a pure random walk or has underlying trends. As a result, three types of funds can be identified:

- If $H > 0.50$, funds have persistent performance;
- If $H = 0.50$, funds have a random process;
- If $H < 0.50$, funds have no performance persistence (the returns are negatively correlated).

References

- Adkisson, J.A. and Don R. Fraser, "Realigning the Stars", Working Paper, 2003a.
- Adkisson, J.A. and Don R. Fraser, "Reading the Stars: Age Bias in Morningstar Ratings", *Financial Analysts Journal*, vol. 5, n°3, September/October 2003b, pp. 24-27.
- Adkisson, J.A. and Don R. Fraser, "Is there still an Age Bias in the Morningstar Ratings?", Working Paper, January 2004.
- Amenc, Noël and Véronique Le Sourd, "Rating the Ratings", EDHEC Risk and Asset Management Research Centre, 2005.
- Amenc, Noël, Lionel Martellini and Daphné Sfeir, "An Integrated Framework for Style Analysis and Performance Measurement", *Journal of Performance Measurement*, vol. 7, n°4, summer 2003, pp. 35-41.
- Bienstock, Sergio and Eric Sorensen, "Segregating Growth from Value: It's not Always Either/Or", Salomon Brothers, Quantitative Equities Strategy, July 1992.
- Blake, Christopher R. and Matthew R. Morey, "Morningstar Ratings and Mutual Fund Performance", *Journal of Financial and Quantitative Analysis*, vol. 35, n°3, September 2000, pp. 451-483.
- Blume, Marshall E., "An Anatomy of Morningstar Ratings", *Financial Analysts Journal*, vol. 54, n°2, March/April 1998, pp. 19-27.
- Chiang, Kevin C.H, Kirill Kozhevnikov and Craig H. Wisen, "The Ranking Properties of the Morningstar Risk-Adjusted Rating", Working Paper, 2003.
- Del Guercio, Diane and Paula Tkac, "Star Power: The Effect of Morningstar Ratings on Mutual Fund Flows", Working Paper n° 2001-15, Federal Reserve Bank of Atlanta, March 2001.
- diBartolomeo, Dan and Erik Witkowski, "Mutual Fund Misclassification: Evidence Based on Style Analysis", *Financial Analysts Journal*, vol. 53, n°5, September/October 1997, pp. 32-43.
- Haslem, John A. and Carl A. Scheraga., "Morningstar's Classification of Large-Cap Mutual Funds", *Journal of Investing*, vol. 10, n°1, spring 2001, pp.79-84.
- Jin, Xue-jun and Xiao-lan Yang, "Empirical Study on Mutual Fund Objective Classification", *Journal of Zhejiang University SCIENCE*, vol.5, n°5, 2004.
- Khorana, Ajay and Edward Nelling, "The Determinants and Predictive Ability of Mutual Fund Ratings", *Journal of Investing*, vol. 7, n°3, 1998, pp. 61-66.
- Kim, Moon, Ravi Shukla and Michael Tomas, "Mutual Fund Objective Misclassification", *Journal of Economics and Business*, vol. 52, n°4, July/August 2000, pp. 309-323.
- Lucas, Lori and Mark W. Riepe, "The Role of Returns-Based Style Analysis: Understanding, Implementing and Interpreting the Technique", Ibbotson Associates, May 1996.
- Morey, Matthew R., "Mutual Fund Age and Morningstar Ratings", *Financial Analysts Journal*, vol. 58, n°2, March/April 2002, pp. 56-63.

- Morey, Matthew R., "The Kiss of Death: A 5-Star Morningstar Mutual Fund Rating?", Working Paper, Pace University, 2003.
- Morningstar, "The New Morningstar Rating Methodology", Morningstar Research Report, April 2002.
- Reichenstein, William, "Morningstar's New Star-Rating System: Advances and Innovation", *Journal of Financial Planning*, vol. 17, n°3, March 2004, pp. 40-47.
- Sharpe, William F., "Determining a Fund's Effective Asset Mix", *Investment Management Review*, vol.2, December 1988, pp. 59-69.
- Sharpe, William F., "Asset Allocation: Management Style and Performance Measurement", *Journal of Portfolio Management*, vol. 18, winter 1992, pp. 7-19.
- Stutzer, Michael, "Mutual Fund Ratings: What is the Risk in Risk-Adjusted Fund Returns?", Working Paper, 2003.
- Vinod, H.D. and Matthew R. Morey, "Estimation Risk in Morningstar Fund Ratings", *Journal of Investing*, vol. 11, n°4, winter 2002, pp. 67-75.

