Working Paper Series, N. 20, December 2009

On the role of risk in the Morningstar rating for mutual funds

Francesco Lisi
Department of Statistical Sciences
University of Padua
Italy

Massimilano Caporin
Department of Economics and Management "Marco Fanno",
University of Padova
Italy

Abstract: In the mutual funds industry the rating process is very important, and Morningstar is surely the most influential international rating agency . In this work we consider the problem of evaluating if the risk component is adequately accounted for in the Morningstar rating. To face this problem we compare the ratings produced giving different weights to the risk component. The focus of the analysis is on testing the hypothesis that two similar rating procedures with different risk parameters (or, in statistical terms, two raters) are equivalent. To that end, first the notion of $\beta-$equivalence is introduced and then a Monte Carlo test for the hypothesis of $\beta-$equivalence is described. Finally, to answer the question on the role of risk in the Morningstar rating, we analyze 1763 monthly return time series of US mutual funds. Results show that the current Morningstar classification, based on a risk-adjusted measure, only marginally accounts for risk and that if we want that risk really matters, the risk parameter should be increased.

Keywords: Risk, Morningstar rating, Rating agreement, $\beta-$equivalence, mutual funds

This version: september 2009

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Francesco Lisi
tel: +39 049 827 4182
francesco.lisi@unipd.it

On the role of risk in the Morningstar rating for mutual funds

Francesco Lisi
Department of Statistical Sciences
University of Padua
Italy

Massimilano Caporin
Department of Economics and Management "Marco Fanno",
University of Padova
Italy

Abstract: In the mutual funds industry the rating process is very important, and Morningstar is surely the most influential international rating agency . In this work we consider the problem of evaluating if the risk component is adequately accounted for in the Morningstar rating. To face this problem we compare the ratings produced giving different weights to the risk component. The focus of the analysis is on testing the hypothesis that two similar rating procedures with different risk parameters (or, in statistical terms, two raters) are equivalent. To that end, first the notion of $\beta-$equivalence is introduced and then a Monte Carlo test for the hypothesis of $\beta-$equivalence is described. Finally, to answer the question on the role of risk in the Morningstar rating, we analyze 1763 monthly return time series of US mutual funds. Results show that the current Morningstar classification, based on a risk-adjusted measure, only marginally accounts for risk and that if we want that risk really matters, the risk parameter should be increased.

Keywords: Risk, Morningstar rating, Rating agreement, $\beta-$equivalence, mutual funds

1 Introduction

A rating is a score given to some subjects by a rater, which can be either a person, i.e. a judge or an expert, or a tool, such as a diagnostic test, a performance measure etc. The rating, which in some ways is similar to a classification system, is a matter of great interest in finance, where ratings are assigned to countries, to credits, to bonds, to managed portfolios, etc. (Krink *et al.* 2007; Krishnan and Lawrence, 2007; Jewell and Livingston, 2002; Blake and Morey, 1999). In the mutual funds industry, then, the rating is particularly important because the score given to funds by rating agencies affects and leads the investment decisions of both private and institutional financial agents (Del Guercio and Tkac, 2008; Knuutila *et al.*, 2006).
The number of agencies providing funds evaluations is not large. Among them, stand

out Morningstar, Standard & Poor, Lipper and Fitch[1]. Another rating system for mutual funds has been recently proposed by Bechmann and Rangvid (2007). Each produces a rating differing for characteristics and for methodologies used. Among these, the Morningstar rating system is surely the most widespread and the most influential, so much so that a "Morningstar effect" on fund flows, which has been widely documented in the financial literature (Del Guercio and Tkac, 2008; Knuutila *et al.*, 2006).

Morningstar classifies funds in 5 categories, giving them from 1 to 5 stars according to a specific performance measure called Morningstar Risk-Adjusted Return (MRAR). Such a measure considers risk-adjusted and load-adjusted returns: this means that, in principle, the final evaluation of a fund is affected by the level of risk and costs, beside of the profitability component.

To be an efficient and operative tool, the rating must be continuously updated and, fo this reason, Morningstar updates its ratings at a monthly frequency. Del Guercio and Tkac (2008) report positive abnormal flows following rating upgrades, and negative abnormal flows following rating downgrades, ranging from 13 to 30 percent of normal flows. In particular, an upgrade from four to five stars would result in an increasing of fund subscriptions of 25 per cent above normal. A much smaller impact has been found for a downgrade to four from five stars. Adkisson and Fraser (2006) present significant evidence that investors witheld funds from mutual funds that lost stars, but did not proportionately reward funds that gained stars.

These results explain and motivate the interest for a deep analysis of the Morningstar rating system. Since, the risk component has been often underestimated by raters - not only in the mutual funds field - one can legitimately and usefully wonder how much risk really weights in the Morningstar's final evaluation, and whether it is adequately accounted for. Answering this question is the main objective of this work.

The MRAR measure derives from an utility function that accounts for risk through a parameter, γ, representing the investor risk aversion. To assess the relevance of the risk component, in this paper we compare the ratings obtained with different values of γ, that is using different intensities of risk aversion and, thus, assigning different weight to the risk component in the whole evaluation.

Although other works were concerned about the role of risk in the Morningstar rating (Amenc and Le Sourde (2007), Vinod and Morey (2002), among others), to our best knowledge, this kind of investigation is new.

In statistical terms the problem we face is a rater agreement one, where the two raters are given by a same performance measure with different values of a parameter (γ). We are interested in testing the hypothesis that two raters are, in some sense, equivalent. Given the way the Morningstar rating is implemented, the hypothesis of identical raters is too strong because this implies a perfect agreement and, thus, is only relatively interesting. So, we introduce the weaker condition of β-equivalence: we say that two raters are β-equivalent if the probability that both of them classify a fund in the same category is β and the probability that their ratings differ for just one class is $1 - \beta$. Building on this definition, we introduce a suitable

[1] See, respectively, www.morningstar.com, www.funds-sp.com, www.lipperweb.com and www.fitchratings.com

measure of β-equivalence, which is a modification of the weighted Cohen's Kappa statistic (Cohen, 1960; Cohen, 1968), called κ^*. Finally, we outline a Monte Carlo procedure to obtain the distribution of κ^* under the hypothesis of β-equivalence. This distribution allows us to do a formal test of β-equivalence for a given value of β and to find a suitable upper confidence bound for β.
Using this methodology we find that the ratings obtained with the setting of Morningstar are very similar to those obtained by assuming that the investor is risk-indifferent and that the similarity decreases for higher values of the risk aversion parameter. This suggests that the Morningstar rating system is mainly influenced by profitability, and only marginally by riskiness.
The paper proceeds as follows. Section 2 contains a summary of the Morningstar methodology; Section 3 introduces the problem of the raters agreement, the notion of $\beta-$equivalence as well as the statistic κ^*, useful for measuring the rater agreement. In Section 4 a Monte Carlo procedure for testing the $\beta-$equivalence is provided. Section 5 answers the question about the role of risk in the Morningstar rating by analyzing a dataset of US mutual funds. Section 6 concludes the work and suggests some additional lines of research.

2 The Morningstar rating

The Morningstar rating methodology is based on two key characteristics: the consideration of peer groups, that are categories of fund styles defined by Morningstar, and the use of risk-adjusted and load-adjusted returns. Peer groups are used to classify mutual funds in coherent categories with respect to reference financial markets (US, Europe...), investment styles (Large, Medium, Small, Value, Blend, Growth...), and exposure to risk factors. With this approach, funds within the same groups can be considered as perfect substitutes and this provides the need for a rating system to rank them. Instead, the comparison across groups is not considered nor is possible with the Morningstar rating.
Morningstar ranks funds inside each category using a specific performance measure: the Morningstar Risk-Adjusted Return (MRAR). Morningstar motivates MRAR using the expected utility theory and assuming that an investor ranks alternative portfolios using the mathematical expectation of a power utility function, based on the terminal value of a given investment.
In deriving MRAR, Morningstar uses some additional elements which affect the computation of mutual funds returns. First, all returns are adjusted for the impact of sales loads. Second, Morningstar recognizes that the investor always has the choice to buy a risk-free asset instead of holding a risky portfolio. Therefore, Morningstar measures a fund's excess returns over and above the return on the risk-free asset (RF) taking into account investment costs that are charged to agents.
The definition of the Morningstar Risk-Adjusted Return is the following:

$$MRAR(\gamma) = \begin{cases} \left[\frac{1}{n} \sum_{t=1}^{n} (1 + ER_t)^{-\gamma} \right]^{-12/\gamma} - 1 & \gamma \neq 0 \\ \left[\prod_{t=1}^{n} (1 + ER_t) \right]^{12/n} - 1 & \gamma = 0 \end{cases} \tag{1}$$

where $ER_t = [(1 + LR_t)/(1 + RF_t)] - 1$ is the monthly geometric excess return and RF_t and LR_t are the monthly return of a risk-free asset and the load-adjusted monthly return for the fund, respectively. Expression (1) is an annualized value. The load-adjusting permits to consider front loads, deferred loads, or redemption fees applied during the month-end under consideration.
Finally, the parameter γ defines the degree of risk aversion. When $\gamma < -1$ the investor is risk-lover rather than risk-averse.
For $\gamma = -1$, the degree of risk aversion is zero, meaning that the investor is indifferent between a risk-free choice and a risky choice as long as the arithmetic average expected return is the same.
For $\gamma = 0$, the investor is indifferent between a risk-free choice and a risky choice as long as the geometric average expected return is the same.
When $\gamma > 0$, the investor is risk-averse and demands a risk premium for choosing a risky portfolio.
"A rating system based solely on performance would rank funds on their geometric mean return, or equivalently, on MRAR(0)" (Morningstar, 2007, pag.12). Evaluation systems that provide a heavier penalty for risk require that $\gamma > 0$. Morningstar's fund analysts concluded that $\gamma = 2$ results in fund rankings that are consistent with the risk tolerances of typical retail investors. Hence, Morningstar uses $\gamma = 2$ in the calculation of its star ratings.
By converting all return series to their riskless equivalents, Morningstar can compare one fund to another on a risk-adjusted basis. This equalizes the playing field for funds in the same category that have different exposures to risk factors.
Once the funds are ranked inside their category, they are scored from one to five "stars" according to their position in the category. The score follows the bell-curve listed in Table 1.
Morningstar calculates ratings for the three-, five-, and 10-year periods, and then the Overall Morningstar Rating is based on a weighted average of the available time-period ratings. Investments must have at least 36 continuous months of total returns in order to receive a rating. Additional details can be recovered in Morningstar (2007).

Table 1: The Morningstar score.

Score	1	2	3	4	5
Percent	bottom 10%	next 22.5%	next 35.5%	next 22.5%	top 10%

3 The rater agreement and the $\beta-$equivalence

In this section we describe the statistical framework of our analyses. Let us consider two raters that evaluate n subjects on a common ordinal scale composed by m categories. In our context, the two raters are given by $MRAR(\gamma_1)$ and $MRAR(\gamma_2)$. By means of $MRAR(\gamma_1)$ and $MRAR(\gamma_2)$ we can obtain two different ratings; then, we can consider a cross-classification table like Table 2, where f_{ij} denote the number

of subjects (funds) classified in the $i-$th category by the first rater and rater in theS $j-$th category by the second rater.

Table 2: Example of cross-classification table.

		$MRAR(\gamma_2)$		Total
$MRAR(\gamma_1)$	1	...	m	
1	f_{11}	...	f_{1m}	$f_{1.}$
$\vdots$	$\vdots$	f_{ij}	$\vdots$	$\vdots$
m	f_{m1}	...	f_{mm}	$f_{m.}$
Total	$f_{.1}$	...	$f_{.m}$	n

Let $d_k = [MRAR(\gamma_1) = i \cap MRAR(\gamma_2) = j \cap \mid i - j \mid = k]$, $(k = 0, 1, ..., m-1; i, j = 1, ..., m)$ be the difference between ratings, that is, the variable describing the circumstance where the two raters give an evaluation that differs for k categories. The variable d_k assumes absolute frequencies $f_k = \sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij} \cdot I(\mid i - j \mid = k)$, where $I(\cdot)$ denotes the indicator function, and relative frequencies $p_k = f_k/n$. We have that $n = \sum_{i,j=1}^{m} f_{ij} = \sum_{k=0}^{m-1} f_k$.
We define two raters as $\beta-$*equivalent* when the distribution of d_k is:

$$P(d_k) \equiv \pi_k = \begin{cases} \beta & k = 0 \\ 1 - \beta & k = 1 \\ 0 & k > 1 \end{cases} \qquad k = 0, ..., m - 1. \tag{2}$$

Thus, for $\beta-$equivalent raters, the probability that they give to a fund the same rating is β, the probability that ratings differ for just a category is (1-β) and the probability they disagree for more than one category is zero. Thus, for $\beta-$equivalent raters, the probability that they give to a fund the same rating is β, the probability that ratings differ for just a category is (1-β) and the probability they disagree for more than one category is zero. Distribution (2) is suitable for evaluation scales with a relatively small number of categories, for instance $m = 4$ or $m = 5$ which are widespread cases. In the mutual fund context, Morningstar (Morningstar, 2007) as well as Lipper (Lipper, 2007) and Standard & Poor, give an evaluation based on 5 ordinal categories.
In the mutual fund context, Morningstar (Morningstar, 2007) as well as Lipper (Lipper, 2007) and Standard & Poor (Standard & Poor, 2009), give an evaluation based on 5 ordinal categories.
When the number of categories is substantially higher, and if reasonable, the notion of $\beta-$equivalence can be generalized allowing d_k to follow some specific multinomial distribution $\text{Mult}(\pi_0, ..., \pi_{m-1})$, with $\pi_0 = \beta$ and $\sum_k \pi_k = 1$. In this work, however, we will not pursue this case but we focus on definition (2) of $\beta-$equivalence.
To study the level of agreement connected to the $\beta-$equivalence between raters $MRAR(\gamma_1)$ and $MRAR(\gamma_2)$, we propose the following variant of the weighted Cohen's κ (Cohen, 1968) statistic:

$$\kappa^*(\beta) = \frac{\sum_{k=0}^{m-1} w_k \pi_k - \sum_{k=0}^{m-1} w_k p_k}{\sum_{k=0}^{m-1} w_k \pi_k} \tag{3}$$

where w_k is a weighting scheme such that $w_0 = 1$, $0 \leq w_i \leq 1$ for $i > 0$ and $w_i \geq w_j$ if $j > i$. The weighting system is important in order to modulate the severity of disagreement. Indeed, in some contexts, it would not be reasonable to consider only a full agreement or a full disagreement.
The statistic κ^* measures the 'distance' between the observed weighted relative frequencies p_k and the expected weighted frequencies π_k and, thus, the 'distance' from the $\beta-$equivalence.
When considering the statistic κ^*, the following cases occur:

- $\kappa^* = 0$ if $\sum_{k=0}^{m-1} w_k \pi_k = \sum_{k=0}^{m-1} w_k p_k$, when the observed frequencies are exactly those expected under the hypothesis of $\beta-$equivalence;
- κ^* reaches its maximum value (κ^*_{max}), when there is the maximum deviation from the $\beta-$equivalence. In particular, if $\sum_{k=0}^{m-1} w_k p_k = 0$ then $\kappa^* = 1$. In the case of definition (1), this may occur if the rater evaluations always differ for two or more categories;
- $0 < \kappa^* < \kappa^*_{max}$ for intermediate cases, when frequencies do not support a complete accord nor a complete disagreement with respect to the $\beta-$equivalence;
- $\kappa^* < 0$ if $\sum_{k=0}^{m-1} w_k \pi_k < \sum_{k=0}^{m-1} w_k p_k$, when the observed agreement is higher than that expected by the definition of $\beta-$equivalence for a given β.

Of course, also the classical κ statistic can be used to measure dependence. However, it differs from κ^* in that κ is based on the expected frequencies under independence, whereas κ^* is built using the expected frequencies under the hypothesis of $\beta-$equivalence and, thus, it seems more appropriate. But the main difference between κ and κ^* is their interpretation: while, for example, $\kappa^* = 0.90$ has a precise and clear interpretation according to distribution (2), $\kappa = 0.90$ means only a generic high level agreement.
The value assumed by the statistic κ^* depends also on the weighting system w_k. In turn, this may depend both on the number of categories, m, and on the features of the specific problem that is being studied.
The rater agreement literature contains several proposals of weighting schemes (see Vanbelle and Albert, 2009 and references therein). Although they share the nice feature of having a statistical interpretation, for some applications they do not seem appropriate since they decrease too slowly with k. Instead, when m is small, also small values of k may indicate important level of disagreement. Thus, we propose the following function for the weights:

$$w_k = exp\ (-a\ \frac{k^b}{m}) \qquad (k = 0, 1, ..., m-1) \tag{4}$$

where $a, b > 0$ are suitable parameters which control how fast weights decrease. This kind of function always gives $w_0 = 1$ and depends both on k and m.
The weighting function and its calibrated parameters could be chosen with respect to the problem under study and the purposes of the work. In our case, we specified a weighting function allowing us to give a relevant weight to the mis-rating by one category and weights close to zero to differences larger than one category. This is

obtained by setting $a = 3.5$ and $b = 3$. Indeed, this gives $w_0 = 1$ and makes function (4) decreasing very quickly with k, depending also on m. In particular, for $m = 5$, it leads to $w_k = (1, 0.497, 0.003, \sim 0, \sim 0)$. Different parameters could be chosen if the number of categories is higher or when it is accaptable to consider disagreement for more than one class as a partial agreement.

4 A Monte Carlo test of $\beta-$equivalence

We want now to test - at a significance level α - the null hypothesis that two raters R_1 and R_2, are β_0-equivalent, for a given β_0 and for a given weighting system. In particular, we want to consider the following hypothesis system:

$$\begin{cases} H_0 : R_1 \ \ and \ \ R_2 \ are \ \beta_0 - equivalent, \\ H_1 : R_1 \ \ and \ \ R_2 \ are \ not \ \beta_0 - equivalent, \end{cases} \tag{5}$$

where the raters can be not β_0-equivalent because they are, for example, β-equivalent with $\beta < \beta_0$ or because they are not β-equivalent at all and d_k follows some multinomial distribution.
Since the standard likelihood ratio test for $H_0 : d_k \sim Mult(\beta, 1-\beta, 0, ..., 0)$ against $H_1 : d_k \sim Mult(\pi_0, \pi_1, ..., \pi_{m-1})$, is not feasible due to the zero expected frequencies under the null hypothesis, we test system (5) through the statistic κ^*.
In order to obtain the distribution of κ^* under the hypothesis of β_0-equivalence, we suggest the following Monte Carlo procedure:

1. let $n = \sum_k f_k$ be the number of subjects to be evaluated and κ^*_{obs} the value of the statistic κ^* computed for the observed data;

2. fix a degree of equivalence β_0 and draw a Monte Carlo sample of size n from the multinomial distribution π_k with $\beta = \beta_0$;

3. denoted by p^{mc}_k the relative frequency for d_k on the Monte Carlo sample, compute the statistic:

$$\kappa^*(\beta) = \frac{\sum_{k=0}^{m-1} w_k \pi_k - \sum_{k=0}^{m-1} w_k p^{mc}_k}{\sum_{k=0}^{m-1} w_k \pi_k}$$

4. repeat steps (b) and (c) N times (with N large) in order to obtain N realizations κ^*_i, $i = 1, ..., N$. The N values κ^*_i represent a sample of size N from the distribution of κ^* under the null of β_0-equivalence;

5. let us denote by

$$pval = \frac{\sum_{i=1}^{N} I(\kappa^*_i > \kappa^*_{obs})}{N}$$

the p-value of the test. If $pval > \alpha$ then, H_0 is accepted at the significance level α .

The acceptance region of this test allows us to identify a (1-α) upper confidence bound for β, that is the highest value of β_0, called β_u, such that H_0 is accepted at level α.
Note that, since we are looking for the highest value of β_0 for which H_0 is accepted, κ^* will be always positive and this is why the test is one-sided.
The above procedure is nonparametric and does not require any distributional assumption on the scores. Since the test is not based on asymptotic considerations, but is calibrated on the sample size n, it is expected to work better for small samples.
One could argue that considering the upper confidence bound for the traditional kappa calculated following Fleiss and Cicchetti (1978) will give similar results. For the application considered in this work it tends to give very high values of κ, sometimes larger than 1. However, they cannot be directly compared with the value of β_u, because they refer to κ rather than to β and, thus, they have different interpretations.

4.1 Validation of the procedure

To assess the performance of the test just described, we conducted a series of simulation trials. They have three purposes. The first was to analyze how the test behaves for different sample sizes and for different levels of β-equivalence. In particular, we are interested in studying the behavior of the test when the true distribution of π_k cannot be fully described by distribution (2). The other two goals were to study the effective level and power of the test and the coverage of the confidence upper bound. We always assumed that two raters were assessing n subjects with two methods that classify them into $m = 5$ mutually exclusive categories. The variable d_k described the difference between ratings and called π_k the true and unknown distribution of d_k. For π_k, two groups of settings were considered: in the first group, data were generated for different values of β according to Definition 2, while in the second one no value of β fully satisfied the Definition of $\beta-$equivalence. The specific values of π_k (k=0,...,4) for each setting (S) are given in column two of Tables 3, 5 and 4. In all the simulations, the distribution of κ^* has been obtained by drawing $N = 10000$ Monte Carlo realizations were generated. Also, to set the weights we used function (4) with $a = 3.5$ and $b = 3$.
As a first step, to better understand how the procedure works, for each setting we generated 500 data sets of length $n = 100, 500$ and 1000 and, for each data set, we found the upper confidence bound, β_u, at the 95% level. Columns three to eight of Table 3 give the mean ($\bar{\beta}_u$) and the standard error of β_u over the 500 simulations. For the first group of settings (S_1 to S_4), results show that for n increasing, the mean of β_u, $\bar{\beta}_u$, tends to get closer to the true value of β, with decreasing standard error. As expected, when the underlying data generator satisfies the definition of $\beta-$equivalence, $\bar{\beta}_u$ is always greater than the true value of β, that is, of the true proportion of full agreement given by π_0. On the contrary, when the data generator does not satisfy the definition of $\beta-$equivalence (2), the procedure tries to force the $\beta-$equivalence, leading to values of β_u that are - on average - smaller than π_0. The bigger the 'distance' from the $\beta-$equivalence, the smaller the value of $\bar{\beta}_u$. As an

example, the setting S_5 is not very far from $\beta-$equivalence. As a result the $\bar{\beta}_u$ is only a little smaller that π_0. Instead, settings S_8 and - even more - S_9 represent distributions very different from (2). This entails very small values of $\bar{\beta}_u$ and in the case of S_9 a large number of cases for which $\beta_u = 0$. In these cases we conclude that the two raters cannot be considered $\beta-$equivalent for any beta. The effective

Table 3: Simulation results for different sample sizes and settings of π.

Setting	$(\pi_0, \pi_1, \pi_2, \pi_3, \pi_4)$	n=100		n=500		n=1000	
		$\bar{\beta}_u$	se	$\bar{\beta}_u$	se	$\bar{\beta}_u$	se
S_1	(0.8,0.2,0,0,0)	0.854	0.035	0.826	0.017	0.815	0.011
S_2	(0.6,0.4,0,0,0)	0.667	0.046	0.632	0.021	0.621	0.015
S_3	(0.4,0.6,0,0,0)	0.474	0.051	0.432	0.022	0.421	0.0161
S_4	(0.2,0.8,0,0,0)	0.261	0.044	0.225	0.019	0.219	0.013
S_5	(0.8,0.1,0.1,0,0)	0.768	0.054	0.729	0.027	0.721	0.020
S_6	(0.7,0.1,0.1,0.1,0)	0.574	0.079	0.535	0.035	0.520	0.025
S_7	(0.6,0.2,0.15,0.05,0)	0.473	0.084	0.434	0.035	0.421	0.025
S_8	(0.4,0.3,0.2,0.1,0)	0.136	0.097	0.119	0.041	0.41	0.028
S_9	(0.3,0.3,0.2,0.1,0.1)	0.006	0.028	0.000	0.000	0.000	0.000

Notes: se=standard error of β_u.

level and the power of the test are analyzed referring to a similar simulation framework. Again, two groups of data generators were involved. The first one, defined by settings S_1, S_2 and S_3, generates data under the null hypothesis allowing the study of the effective level. The second group, defined by settings S_5, S_6 and S_7, produces data that, with different intensities, are not fully consistent with the definition of $\beta-$equivalence and allows us to study the power of the test. In this case, for each setting, 2000 data sets were generated and the hypothesis H_0: A and B are β_0-equivalent, with $\beta_0 = \pi_0$, is tested. We considered sample sizes $n = 100$, $n = 500$ and $n = 1000$ and significance levels $\alpha = 0.01$, $\alpha = 0.025$ and $\alpha = 0.05$. Table 4 lists the effective levels and powers, rounded to the third decimal figure. Results show that nominal levels are basically respected for all settings, sample sizes and levels; moreover, the test has good power against alternatives close to the null and for relatively small sample sizes as, for example, for setting S_4 and $n = 100$. For more distant alternatives and/or for larger sample sizes, the power is always very high. Several other settings were considered in the Monte Carlo simulations but results were not reported since they basically confirm those listed in this paper. Finally, for studying the effective coverage, $(1-\alpha)_{obs}$, of the upper confidence bound with respect to the nominal one, $(1-\alpha)$, a third set of simulations was performed. In this case, only settings S_1, S_2 and S_3 were considered. For each of these, we generated 2000 data sets of size n and, for each data set, we computed upper confidence bounds at the nominal level $(1-\alpha)$, $\beta_u^{(i)}$ (i=1,...,2000) and the effective coverage, defined as:

$$(1-\alpha)_{obs} = \frac{\sum_{i=1}^{2000} I(\beta_u^{(i)} - \pi_0)}{2000} \tag{6}$$

Table 4: Effective levels and powers for different nominal levels (α), sample sizes (n) and settings (S) of π.

Setting	$(\pi_0, \pi_1, \pi_2, \pi_3, \pi_4)$	n	$\alpha = 0.01$	$\alpha = 0.025$	$\alpha = 0.05$
S_1	(0.8,0.2,0,0,0)	n=100	0.011	0.030	0.053
		n=500	0.009	0.023	0.049
		n=1000	0.009	0.023	0.049
S_2	(0.6,0.4,0,0,0)	n=100	0.011	0.029	0.062
		n=500	0.013	0.030	0.059
		n=1000	0.009	0.023	0.054
S_3	(0.4,0.6,0,0,0)	n=100	0.016	0.031	0.061
		n=500	0.012	0.028	0.056
		n=1000	0.010	0.025	0.053
S_5	(0.8,0.1,0.1,0,0)	n=100	0.472	0.597	0.683
		n=500	0.982	0.991	0.995
		n=1000	1	1	1
S_6	(0.7,0.1,0.1,0.1,0)	n=100	0.857	0.905	0.941
		n=500	1	1	1
		n=1000	1	1	1
S_7	(0.6,0.2,0.15,0.05,0)	n=100	0.869	0.915	0.946
		n=500	1	1	1
		n=1000	1	1	1

Notes: All the computation are based on 2000 replications.

where $I(u) = 0$ for $u \geq 0$ and 0 otherwise. Table 5, lists the results for $n = 100$ and $n = 500$ and for levels $(1 - \alpha) = 0.99$, $(1 - \alpha) = 0.975$ and $(1 - \alpha) = 0.95$. It shows that effective and nominal coverages are quite close, confirming the correctness of the procedure.

5 Does really Morningstar account for risk?

The notion of $\beta-$equivalence between raters and the procedure previously described provide a tool allowing to analyse the role of risk in the Morningstar rating and to determine to what extent risk is relevant in the current practice.
To that end, we analyze the degree of $\beta-$equivalence between couples of ratings obtained giving different relevance to risk in the final rating. The idea is to evaluate how much different is the final result of ratings that weight differently the risk component of a fund. In particular, we wish to compare the rating bases on the setting currently used by Morningstar and the rating obtained ignoring the risk.
To answer the original question about the role of risk in Morningstar rating for mutual funds, we apply the methodology of Section 4 to 1763 monthly return time series of US mutual funds for the period January 2003 - December 2007.
Our 'Morningstar rating' slightly differs from the original one in so far as loads are not considered. This implies that (1) was applied considering simple returns instead

Table 5: Effective coverages, $(1-\alpha)_{obs}$, against the nominal coverage $(1-\alpha)$, for different settings, sample sizes and levels.

Setting	$(\pi_0, \pi_1, \pi_2, \pi_3, \pi_4)$	n	$1-\alpha = 0.99$	$1-\alpha = 0.975$	$1-\alpha = 0.95$
S_1	(0.8,0.2,0,0,0)	n=100	0.990	0.970	0.954
		n=500	0.985	0.971	0.941
S_2	(0.6,0.4,0,0,0)	n=100	0.987	0.977	0.941
		n=500	0.985	0.974	0.947
S_3	(0.4,0.6,0,0,0)	n=100	0.986	0.972	0.940
		n=500	0.987	0.968	0.937

Notes: All the computation are based on 2000 replications.

of LR_t, the load-adjusted returns. We are forced to this because the data referring to fund loads were not available to us. However, this does not affect our analysis, because our interest is centered in comparing different MRARs with respect to risk, rather than to study the performance of the rating itself. Thus, it is possible to work conditionally to a given level of costs. To further simplify the analysis, we did not consider the entire set of Morningstar categories, but only classes implicit in the Morningstar style box, which classify funds with respect to market capitalization (Large, Medium, Small) and investment style (Value, Blend, Growth). Crossing the capitalization and the investment style leads to nine classes, that will be denoted by LV(245), LB(343), LG(412), MV(60), MB(197), MG(119), SV(80), SB(170) and SG(137), where the numbers in brackets are the number of fuds in each class.
Results of our analyses are listed in Tables 7 - 12 where, for each category, we reported: the number of funds; the 95% upper confidence bound; β_u, for which the null hypothesis is accepted, at a 5% level; the value of the statistic κ^*; p_0 and p_1, the observed relative frequencies for d_0 and d_1.
For each of the nine categories the rating was performed by using the Morningstar risk-adjusted return as a function of the parameter γ, which represents the risk component in the final evaluation. In our analysis, we compared the ratings resulting from MRAR(γ), with $\gamma = -1, 0, 2, 5, 10$, considering them different raters. Since Morningstar is a five-class rating we set $m = 5$. When testing for the $\beta-$equivalence, in the computation of κ^* we used function (4), with $a = 3.5$ and $b = 3$ to set weights. This gives: $w_0 = 1$, $w_1 = 0.497$, $w_2 = 0.003$, $w_3 = 1 \times 10^{-8}$ and $w_5 = 1 \times 10^{-19}$. To practical purposes, this is equivalent to define as a full agreement evaluations that coincide, "an half agreement" evaluations that differ of just one category and a full disagreement evaluations differing for more than one category. Finally, in the computation of the test p-values, $N = 10000$ Monte Carlo replications were considered.
The first step in our analyses consists in comparing the ratings produced by $\gamma = -1$ and $\gamma = 0$ because both parameters imply a condition of indifference to risk. Thus, we expect that the corresponding ratings are very similar. Indeed, the column β_u in Table 7 shows that the degree of $\beta-$equivalence between this two raters is very large and ranges from 0.94 to 0.98. Note also that, when $\gamma = -1$ and $\gamma = 0$ are involved, we have always $p_0 + p_1 = 1$, meaning that there are not funds for which the two ratings differ for more than one class. On the whole the difference between

these two raters, in terms of final rating, is negligible.
Bearing this comparison in mind, in subsequent analysis we focus on comparing the rating produced by MRAR(0) with those achieved with different values of γ.
The parameter $\gamma = 2$ is that used in the current practice by Morningstar (Morningstar 2007) because it is believed that this adequately represents the risk aversion of the typical investor. However, as Table 9 shows, the level of agreement - in terms of $\beta-$equivalence - between the rating obtained in a framework of indifference to risk ($\gamma = 0$) and that produced by MRAR(2) is quite elevated and not so different from the case in which to parameters not accounting for risk were involved. This is true, in particular, for the Large and Medium classes, whereas for the Small class, differences seem to be a little higher. Moreover, the percentage of cases for which the rating differs for at most one "star" is almost always equal to 100%.
On the whole, these results lead us to believe that in the Morningstar rating the risk component plays only a marginal role and that - perhaps - a greater weight of riskiness in the rating procedure would be suitable in order to avoid to underestimate the risk.
These conclusions are supported by the results obtained comparing ratings produced by $\gamma = 0$ with those derived by considering $\gamma = 5$ and $\gamma = 10$. Indeed, increasing the value of γ, that is increasing the weight of risk, the level of $\beta-$equivalence sensibly decreases, with β_u ranging from 0.69 to 0.85 in the case of $\gamma = 5$ (Table 10) and from 0.405 to 0.685 in the case of $\gamma = 10$ (Table 12). Also the number of cases where the raters disagree for more than one class increases . This suggests that the raters are becoming really different.
This further results point out that if we want that risk really matters in the Morningstar rating, the value of γ should be increased.
Note that, even though β_u is generally higher than the observed relative frequency of d_0 (p_0), this is not always true. For example, for MB in Table 12 the $p_0 = 50.4$, but $\beta_u = 0.45$. This is because, what we actually do is testing the whole distribution (2) and not only π_0.
With respect to the macro-categories of funds Large, Medium and Small, we found that the agreement is always higher for the Large category, followed by Medium and Small.
Finally, to study the impact on the $\beta-$equivalence of scoring funds following Table 1, we repeat the analyses scoring funds according the equally-spaced scheme of Table 6. The results are not reported here, apart from those related to the comparison $\gamma = 0$ and $\gamma = -1$ (see Table 8). On the whole they point out that considering an equally-spaced scheme slightly decreases the level of the $\beta-$equivalence, but it does not change the conclusion reached by analyzing the Morningstar's approach.

Table 6: The equally-spaced score.

Score	1	2	3	4	5
Percent	bottom 20%	next 20%	next 20%	next 20%	top 20%

Table 7: MRAR(0) vs MRAR(-1): Morningstar categories.

Category	n	β_u	κ^*_{oss}	p_0	p_1
LV	245	0.975	0.0075	95.9	4.1
LG	343	0.975	0.0051	96.5	3.5
LB	412	0.98	0.0046	97.1	2.9
MV	60	0.965	0.0161	93.3	6.7
MG	197	0.94	0.0109	91.9	8.1
MB	119	0.945	0.0149	91.6	8.4
SV	80	0.94	0.0206	90	10
SV	170	0.94	0.0215	90.3	9.7
SB	137	0.965	0.0119	94.2	5.8

Notes: β_u=95% upper confidence bound, for which the null hypothesis is accepted, at the 5% of significance level; p_0 and p_1=relative frequencies of d_0 and d_1.

Table 8: MRAR(0) vs MRAR(-1), 5 equally spaced categories.

Category	n	β_u	κ^*_{oss}	p_0	p_1
LV	245	0.945	0.0095	92.7	7.3
LG	343	0.95	0.0073	93.6	6.4
LB	412	0.96	0.0068	94.7	5.3
MV	60	0.965	0.0161	93.3	6.7
MG	197	0.935	0.0136	90.9	9.1
MB	119	0.945	0.0149	91.6	8.4
SV	80	0.92	0.0234	87.5	12.5
SG	170	0.91	0.0176	88.2	11.2
SB	137	0.975	0.0095	95.6	4.4

Notes: β_u=95% upper confidence bound, for which the null hypothesis is accepted, at the 5% of significance level; p_0 and p_1=relative frequencies of d_0 and d_1.

6 Conclusions

This paper focuses on testing the level of agreement between two raters when ordinal scales of rating are involved. The test we propose is based on the notion of $\beta-$equivalence between raters that is useful in defining the level of agreement between two rankings. First, a suitable statistic for measuring $\beta-$equivalence has been defined, then a Monte Carlo procedure for testing the $\beta-$equivalence and finding an upper confidence bound for β has been outlined.
The usefulness of this approach has been shown in the context of mutual fund rating, in particular referring to the Morningstar rating. The application of the test led us to conclude that risk plays only a marginal role in the final Morningstar rating and that it is probably underestimated. We think that our results are important because the literature suggests that individual investors, as well as many financial advisors, believe the Morningstar rating and base their investment decisions "following stars". In periods when the concepts of risk and volatility appear to be less and less abstract

Table 9: MRAR(0) vs MRAR(2), Morningstar classes.

Category	n	β_u	κ^*_{oss}	p_0	p_1
LV	245	0.95	0.012	92.7	7.3
LG	343	0.925	0.0095	90.7	9.3
LB	412	0.945	0.0092	92.7	7.3
MV	60	0.94	0.0206	90	10
MG	197	0.925	0.0138	89.8	10.2
MB	119	0.875	0.023	83.2	16.8
SV	80	0.875	0.0267	82.5	17.5
SG	170	0.85	0.0207	81.2	18.8
SB	137	0.88	0.0216	84.7	14.6

Notes: β_u=95% upper confidence bound, for which the null hypothesis is accepted, at the 5% of significance level; p_0 and p_1=relative frequencies of d_0 and d_1.

Table 10: MRAR(0) vs MRAR(5), Morningstar categories.

Category	n	β_u	κ^*_{oss}	p_0	p_1
LV	245	0.825	0.0204	79.6	19.6
LG	343	0.775	0.0178	75.5	23.3
LB	412	0.85	0.0134	83.5	15.5
MV	60	0.83	0.0528	73.3	26.7
MG	197	0.78	0.0247	76.1	21.3
MB	119	0.695	0.0382	66.4	30.3
SV	80	0.63	0.0491	57.5	40
SG	170	0.63	0.0328	60.6	36.5
SB	137	0.74	0.031	70.1	28.5

Notes: β_u=95% upper confidence bound, for which the null hypothesis is accepted, at the 5% of significance level; p_0 and p_1=relative frequencies of d_0 and d_1.

it is crucial that investors are conscious of their choices and of the level of risk they take.

Note that, even though we showed an application in the financial field, scenarios where raters give categorical ratings to subjects commonly occur in several other fields and thus, the outlined procedure may be useful in a wide range of applications.

References

[1] Adkisson, J.A. and Fraser, D.R (2003) Realigning the Stars: The Reaction of Investors and Fund Managers to Changes in the Morningstar Rating Methodology for Mutual Funds. *Working Paper, Mays Business School, Texas A&M University.*

[2] Adkisson, J.A. and Fraser, D.R (2003) Realigning the Stars: The Reaction of Investors and Fund Managers to Changes in the Morningstar Rating Method-

Table 11: MRAR(0) vs MRAR(10), Morningstar classes.

Category	n	β_u	κ^*_{oss}	p_0	p_1
LV	245	0.68	0.0258	65.7	32.2
LG	343	0.46	0.0295	48.7	44.3
LB	412	0.685	0.019	68.2	28.9
MV	60	0.53	0.0632	50	43.3
MG	197	0.515	0.0384	52.8	40.1
MB	119	0.45	0.0496	50.4	37
SV	80	0.405	0.0569	47.5	37.5
SG	170	0.41	0.0405	45.9	43.5
SB	137	0.55	0.0346	54.7	40.1

Notes: β_u=95% upper confidence bound, for which the null hypothesis is accepted, at the 5% of significance level; p_0 and p_1=relative frequencies of d_0 and d_1.

Table 12: MRAR(2) vs MRAR(10), Morningstar categories.

Category	n	β_u	κ^*_{oss}	p_0	p_1
LV	245	0.75	0.0251	72.2	26.1
LG	343	0.535	0.0257	55.1	39.4
LB	412	0.74	0.018	73.1	24.8
MV	60	0.62	0.0535	56.7	40
MG	197	0.595	0.0357	58.4	37.1
MB	119	0.58	0.0427	56.3	38.7
SV	80	0.585	0.0536	55	40
SG	170	0.565	0.0378	55.3	40
SB	137	0.65	0.0356	62	35

Notes: β_u=95% upper confidence bound, for which the null hypothesis is accepted, at the 5% of significance level; p_0 and p_1=relative frequencies of d_0 and d_1.

ology for Mutual Funds. *Working Paper, Mays Business School, Texas A&M University.*

[3] Agresti, A. (2002) *Categorical data analysis.* Wiley, New York.

[4] Amenc, N. and Le Sourd, V. (2007) Rating the ratings. A critical analysis of fund rating systems. *EDHEC Working paper.*

[5] Banerjee, M., Capozzoli, M., McSweeney, L., and Sinha, D. (1999) Beyond kappa: a review of interrater agreement measures. *The canadian journal of statistics*, **27**, 3-23.

[6] Blake, C.R. and Morey, M.R (1999) Morningstar ratings and mutual fund performance. *Journal of Financial and Quantitative Analysis*, **35**, 451-483.

[7] Bechmann and Rangvid (2007) Rating mutual funds: Construction and information content of an investor-cost based rating of Danish mutual funds. *Journal of Empirical Financial*, **14**, 662-693.

[8] Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37-46.

[9] Cohen, J. (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **20**, 213-220.

[10] Del Guercio, D. and Tkac, P.A. (2008) Star Power: The Effect of Morningstar Ratings on Mutual Fund Flow. *Journal of Financial and Quantitative Analysis*, **43**, 907-936.

[11] Fleiss, J.L., Cicchetti (1978) Inference about weighted kappa in the non-null case. *Applied psychology measurement*, **2**, 113-117.

[12] Goodman, L.A. and Kruskal, W.H. (1954) Measures of association for cross classifications. *Journal of American Statistical Association*, **49**, 732-764.

[13] Jewell, J. and Livingston, M. (2002) A comparison of bond ratings from Moody's S&P and Fitch IBCA. *Financial markets, institutions and instruments*, **8**, 1-45.

[14] Knuutila, M., Puttonen, V. and Smythe, T. (2006) The effect of distribution channels on mutual fund flows. *Journal of Financial Services Marketing*, **12**, 88-96 .

[15] Krink, T., Paterlini S. and Resti A. (2007) Using differential evolution to improve the accuracy of bank rating systems. *Computational Statistics & Data Analysis*, **52**, 68-87.

[16] Krishnan, D. and Lawrence, E.R. (2007) Examining Split Bond Ratings: Effect of Rating Scale. *Quarterly Journal of Finance and Accounting*, Spring 2007.

[17] Lipper (2007) The Lipper Leader Rating System. `http://www.lipperweb.com`.

[18] Morningstar (2007) The Morningstar Rating Methodology. Morningstar methodology paper.

[19] Standard & Poor's (2009) Mutual Fund Ranking Methodology. `http://www2.standardandpoors.com/ pf/pdf/equity/MFMethodology.pdf`.

[20] Vanbelle, S. and Albert A. (2009) A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology*, **6**, 157-163.

[21] Vinod, H.D., and Morey, M.R. (2002) Estimation risk in Morningstar Fund Ratings. *Journal of Investing*, **11**, 67-75.

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it
Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Department of Statistical Sciences
University of Padua
Italy