

RATING THE RATERS: AN INVESTIGATION OF MUTUAL FUND RATING SERVICES

By Matthew R. Morey, Ph.D.

This paper examines two aspects of ratings that have heretofore received little attention in the academic literature. Using methodology robust to survivorship bias and loads, Dr. Morey reports on whether or not mutual fund rating services are able to predict winning funds and whether there is evidence of their ability to predict losing funds.

Abstract

This paper examines two aspects of ratings that have heretofore received little attention in the academic literature. First, the paper documents the methodology of three well-known mutual fund rating/ranking services: Morningstar, Value Line, and Lipper Analytical. Second, the paper examines which of two rating systems, Morningstar and Value Line, is a better predictor of future performance. Using methodology robust to survivorship bias and loads, we find that the ratings of both systems show little ability to predict winning funds. However, there is some evidence, particularly using the Value Line system, of ability to predict losing funds.

1. Introduction

With thousands of mutual funds to choose from, investors and financial consultants are understandably drawn to mutual fund ratings that help them differentiate good funds from bad. Indeed, the widespread use of the ratings is confirmed by the fact that almost every major financial publication now has regular coverage of mutual fund ratings. However, what is not as apparent to many is how influential the ratings are on the behavior of the mutual fund industry. For example, due to their popularity, the ratings

have attained a status akin to a "Good Housekeeping seal of approval."¹ The result is that many investors restrict their financial advisors to investing only in funds that have high ratings. Funds that have lower ratings cannot even be considered. In addition, many people now believe investment flows in and out of mutual funds are closely related to the mutual fund ratings. For example, a recent study by the Financial Research Corporation of Boston and reported in *The Wall Street Journal* found that in 1999, funds with five or four Morningstar stars received inflows of \$223.6 billion while funds with three or fewer stars had outflows of \$132 billion.² Moreover, the heavy use of ratings in mutual fund advertising suggests that mutual fund companies believe that investors care about ratings. Indeed, in some cases, the only mention of return performance in the mutual fund advertisement is the rating.

Given the importance of mutual fund ratings, this paper examines two aspects of ratings that have heretofore received little attention in the academic literature. First, the paper documents the methodology of three well-known mutual fund rating services: Morningstar, Value Line, and Lipper Analytical. Such information is important to financial consultants and investors because it provides information on where the ratings originate and how changes in the ratings may be more

a result of the methodology than actual performance. Second, the paper attempts to answer a question that many financial consultants have probably asked: which of the rating systems is the best predictor of future performance? This is an important question because, despite ratings services stating that their ratings should *not* be used as signals of future performance, the simple fact is that investors and financial consultants *do* use the ratings to choose the funds in which to invest. Hence, in this paper we conduct a straightforward examination of the out-of-sample performance of two of these mutual fund rating services (Morningstar and Value Line).

The paper is organized as follows. Section 2 presents some background research and describes how this paper fits into this literature. Section 3 describes the methodology of the ratings systems used by Morningstar, Value Line, and Lipper Analytical. Section 4 explains the data for the out-of-sample analysis. Section 5 provides the methodology. Section 6 presents our results. Section 7 concludes the paper.

2. Background and Significance

While there has been a great deal of research conducted on the performance of mutual funds, relatively little work has been conducted on mutual fund ratings. Moreover, the previous work is almost exclusively centered on the Morningstar ratings rather than on the Value Line or Lipper approaches. The literature on Morningstar ratings can be broken down into three groups: predicting performance, methodology, and fund flows as related to ratings. The following provides a very brief review of this research.

In terms of predicting performance, the most relevant paper is Blake and Morey (2000). This paper examines the out-of-sample performance of the Morningstar ratings using a wide range of in-sample and out-of-sample performance measures that are robust to survivorship bias. The basic conclusions of the paper are twofold. First, in out-of-sample tests, high-star-rated funds (five and four stars) were not able to outperform median-rated funds. That is, out of sample, a three-star fund performed just as a five- or four-star rated fund did. Second, there was some

evidence that low Morningstar ratings implied weak future performance.

Another area of research on the Morningstar ratings has been on the documentation/evaluation of the Morningstar ratings methodology. Blume (1998), Sharpe (1998), Warshawsky, DiCarlantonio and Mullan (2000), and Morey (2002a) all provide extensive documentation of the Morningstar methodology and, in doing so, relate certain idiosyncrasies in the Morningstar methodology. For example, Morey (2002a) finds that the Morningstar system has a built-in bias toward giving older funds slightly higher ratings than younger funds.

One last major area of research regarding the Morningstar ratings is the effect of Morningstar ratings on fund flows. A recent paper by Del Guercio and Tkac (2002) finds that even after controlling for performance, changes in a fund's Morningstar ratings have a significant effect on the fund's flows. More specifically, they find that following a ratings decline (increase), there is a statistically significant outflow (inflow) of funds.

While the current paper does not examine fund flows, it is related to the previous work on predicting performance with Morningstar ratings and documenting the Morningstar methodology. Indeed, it uses much of the same methodology for examining the out-of-sample performance of the Morningstar ratings system as does Blake and Morey (2000) except that it uses later data and a longer out-of-sample evaluation period. The way this paper differs from the others in the literature is that it attempts to examine the Morningstar rating system within the context of other rating systems. For example, in describing the methodology, this paper explains the Morningstar ratings methodology side by side with a description of the Value Line and Lipper Analytical methods. As such, it presents a fuller view of how the mutual fund rating industry operates and how the Morningstar methodology differs from other methodologies. The same can be said of the out-of-sample performance analysis. This paper examines not only the Morningstar rating system but also the Value Line system and a simple alternative rating system. In this way, the Morningstar system can be examined in relation to real alternatives.

3. The Methodology of the Rating Systems: Morningstar, Value Line, and Lipper Analytical

In this section we describe the methodology of the overall ratings systems used by the three rating services. It should be noted that these ratings services do have other ratings systems; however, this paper examines only the most popular rating system of each service. For Morningstar this is its overall rating, for Value Line it is its overall rating, and for Lipper Analytical it is its mutual fund rankings. (Lipper Analytical provides only a ranking system; it does not formally rate funds.)

3.1 The Morningstar Overall Rating

To calculate the overall Morningstar star rating, Morningstar currently classifies funds into one of four categories: domestic equity, foreign equity, municipal bond, and taxable bond. The domestic equity is the broadest category as it includes diversified U.S. equity funds as well as hybrid and specialty funds.³ Note that before November 1996, Morningstar did not separate foreign equity funds into their own category and instead simply grouped them into the domestic equity category. Also, before November 1996, hybrid funds were given their own category and/or rated differently from other funds. Indeed, when the November 1996 change took place, a number of funds went from being three-star funds in October 1996 to five-star funds in January 1997 simply because their comparison groups changed.⁴

After categorization, the overall star ratings are (and always have been) based on risk-adjusted returns. More specifically, for funds with ten or more years of return history, the ratings are based on an aggregation of the three-year, five-year, and ten-year risk-adjusted returns. For funds with five to fewer than ten years of return data, the ratings are based on an aggregation of the three-year and five-year risk-adjusted returns. Finally, for funds with three to fewer than five years of return data, the ratings are based only on three-year risk-adjusted returns. Note that Morningstar does not give overall star ratings to funds with fewer than three years of historical returns.

The risk-adjusted return is calculated in the following manner. First, Morningstar calculates an expense- and load-adjusted return for each fund by adjusting the returns for such expenses as 12b-1 fees, management fees, and other costs auto-

matically taken out of the fund, and then by adjusting for front-end and deferred loads.⁵ Next, it calculates a *Morningstar Return* by dividing the expense- and load-adjusted excess return by the higher of two variables: the excess average return of the fund category, e.g., domestic equity, or the average ninety-day U.S. Treasury bill rate:

$$\frac{(\text{Expense and Load-Adjusted Return on the Fund} - \text{T-Bill})}{\text{Higher of } (\text{Average Category Return} - \text{T-Bill} \text{ or } \text{T-Bill})} \quad (1)$$

Morningstar divides through by one of these two variables to prevent distortions caused by having low or negative average excess returns in the denominator of the equation (1). Such a situation might occur in a protracted down market.⁶

Morningstar then calculates a *Morningstar Risk* measure. This measure is calculated differently from traditional risk measures, such as beta and standard deviation, which both see greater-than and less-than-expected returns as added volatility. Morningstar believes that the greatest fear of most investors is losing money, which they define as underperforming the risk-free rate of return an investor can earn from the ninety-day Treasury bill. Hence, their risk measure focuses only on downside risk. To calculate the Morningstar risk, the company plots the monthly returns in relation to T-bill returns. It then adds up the amounts by which the fund trails the T-bill return each month and then divides that total by the time horizon's total number of months. This number, the average monthly underperformance statistic, is then compared with those of other funds in the same broad investment category to assign the risk scores. The resultant Morningstar risk score expresses how risky the fund is relative to the average fund in its category.⁷ To more clearly illustrate the Morningstar risk calculation, table 1 provides a hypothetical example where we define the time horizon as one year.

To calculate a fund's *overall* star rating, Morningstar then follows two more steps. In the first step, Morningstar calculates a fund's star rating for each of three different time horizons: three years, five years, and ten years. We call these the *time-specific* star ratings. For each separate time horizon, a "score" is calculated by subtracting the fund's Morningstar risk from the Morningstar return. Hence, for the three-year score, the three-year Morningstar risk is subtracted from the three-year Morningstar return; for the five-year score, the five-year Morningstar risk is subtracted from the

five-year Morningstar return; and for the ten-year score, the ten-year Morningstar risk is subtracted from the ten-year Morningstar return. The resulting time-horizon score is then compared to the time-horizon score of other funds in that fund category. Hence, the three-year score of the fund will be compared to all the other funds' three-year scores, the five-year score will be compared to all the other funds' five-year scores, and the ten-year score will be compared to all the other funds' ten-year scores. Then, for each of the three time horizons, star ratings are allocated in the following manner. If the fund's score lands in the top 10 percent, it receives a time-specific rating of five stars; if the fund falls into the next 22.5 percent, it receives a time-specific star rating of four stars; if it falls in the middle 35 percent, it receives a time-specific rating of three stars; if it lies in the next 22.5 percent, the fund receives a time-specific rating of two stars; and if it is in the bottom 10 percent, it receives a time-specific rating of one star. Note that if a fund does not have five years and/or ten years of return history, then the time-specific ratings for those time horizons are not calculated. Since all funds must have three years worth of returns, all funds have at least a three-year time-specific star rating.

For the second step in the overall star rating calculation, Morningstar then uses a weighting system that depends upon the *age of the fund*. For funds with ten years or more of returns, Morningstar weights the three-year star rating by 20 percent, the five-year star rating by 30 percent, and the ten-year rating by 50 percent. For funds with five to fewer than ten years of return data, Morningstar weights the three-year star rating by 40 percent and the five-year star rating by 60 percent. For funds with fewer than five but at least three years of return data, Morningstar weights the three-year star rating by 100 percent. Morningstar then takes this average number and rounds it up if it has a decimal value of 0.5 or above and rounds it down if it has a decimal value of below 0.5. Hence, for example, a fund that had a four-star rating for the three-year time horizon, a four-star rating for the five-year time horizon, and a three-star rating for the ten-year time horizon, would receive a 3.5 [4 stars(0.2) + 4 stars(0.3) + 3 stars(0.5) = 3.5]. Morningstar would give this fund a four-star overall rating since the decimal value was .5 or higher. If the fund had instead received a 3.4, the fund would have received a three-star overall rating.

Table 1: Understanding Morningstar Risk

<u>Month</u>	<u>Fund Return(%)</u>	<u>T-Bill Return</u>	<u>Underperformance</u>
1	2.0	0.5	NA
2	-1.5	0.5	2.0
3	3.2	0.5	NA
4	1.2	0.4	NA
5	-4.0	0.6	4.6
6	2.1	0.5	NA
7	0.7	0.5	NA
8	2.3	0.5	NA
9	-1.7	0.5	2.2
10	2.4	0.4	NA
11	1.2	0.6	NA
12	-3.1	0.5	3.6
Total Underperformance			13.2

$$\frac{\text{Total Underperformance}}{\text{Total Number of Months}} = \frac{13.2}{12} = 1.10 \text{ is the average monthly underperformance}$$

$$\frac{\text{Average Monthly Underperformance}}{\text{Average Monthly Underperformance of Investment Category}} = 1\text{-year Morningstar risk}$$

3.2 Changes in the Morningstar Ratings

In April of 2002, Morningstar announced that it was going to change its overall rating system as of June 30, 2002. The changes were twofold. First, and most important, instead of rating funds within four broad areas—domestic equity, international equity, municipal bonds, and taxable bonds—the new system measures funds only against those that have similar investment styles. That is, instead of the four broad investment categories, there now will be forty-eight different categories into which a fund can be grouped. For example, all Latin American funds will be grouped into one category, and the ratings will be derived from how a Latin American fund performs relative to other Latin American funds. Second, there will be a slight change in the methodology of the Morningstar risk measures to more accurately measure downside risk. (At the time of writing this article, it was not clear what this change would be.) Apart from these changes, the rest of the methodology of the Morningstar ratings system will remain.⁸

The change from four to forty-eight categories has significant implications for the ratings system. Before this upcoming change, funds could have received high ratings simply because the style of the fund was in vogue. For example, in 1999 almost all growth funds were highly rated funds because growth funds generally outperformed other domestic equity funds during the previous few years. Indeed, one growth fund might have actually performed quite poorly relative to other growth funds, but because the fund in question was being compared to other styles of funds, the overall Morningstar rating of the fund remained high. Conversely, a high-performing value fund would have received a relatively low rating simply because growth funds were included in the comparison group. Now, because of the upcoming change in the methodology, these situations will not exist.

While this change in the ratings methodology should improve on the comparison group issue, it is not without problems of its own. Indeed, it now becomes critical that Morningstar define a fund's style correctly.⁹ If it does not, then a fund could be placed into an inappropriate style classification and easily receive unwarranted high or low ratings. Furthermore, because the ratings are so important

to the funds, the changes in the ratings methodology may make it more advantageous for funds to practice style drift or window dressing so as to receive the best possible ratings.

3.3 The Value Line Overall Rating

Although Value Line is a much older company than Morningstar, it started formally rating mutual funds only in 1993, a full eight years after Morningstar began. The product that contains the Value Line ratings is the *Value Line Mutual Funds Survey*, which is published each month.

In the survey, Value Line rates funds using an "overall Value Line rank." Instead of using the five-star to one-star method of Morningstar, Value Line instead uses a 1 to 5 rating system where 1 is the best fund and 5 is the worst. To calculate the Value Line overall rating it first classifies funds into three groups: equity/partial equity, municipal bond, and taxable bond. To do this, it uses the funds' prospectuses and its own analysis of the fund style. Then using non-expense-, non-load-adjusted returns (remember that Morningstar uses expense- and load-adjusted returns), three performance measures are calculated for each fund. These three performance measures are as follows:

1. A five-year growth persistence number. This is a measure of how consistently a fund has outperformed all other funds in the broad category. To best understand how this number is calculated, consider the following example. First, assume the broad category is equity/partial equity. To calculate the five-year growth persistence number, Value Line first divides the month-end net asset value of each equity fund by the average net asset value of all equity/partial funds. This provides them with relative net asset values. Then using each month as a starting point, Value Line counts the number of subsequent months that the fund's relative net asset value rose in each of the past sixty months. For example, consider a fund whose net asset value rose in each of the sixty months. The count would be sixty for the current month, fifty-nine for last month, fifty-eight for two months ago, and so on down to one for the month that took place fifty-nine months ago. Conversely, a fund

that had sixty straight months of declining relative net assets would receive a zero for this month, a zero for last month, a zero for two months ago, etc. Indeed, all the numbers would be zero for this fund.

The sixty monthly counts are then added up for each fund in the sample to form the growth persistence number. For a fund that always had increasing relative net assets, the number would be 1,830; for a fund that always had declining relative net assets, the number would be zero. The funds are then ranked from top to bottom, with the best being the fund(s) with the highest growth persistence number.

2. A one-year growth persistence number. This is the same measure as listed above, but it uses only the past twelve months.
3. A three-year Sharpe ratio based on the fund's past three years of actual returns. The monthly returns are *not* load-adjusted (unlike in Morningstar), nor are they excess returns. Hence, for the ratio, Value Line simply takes the mean monthly return divided by the standard deviation of the thirty-six monthly returns.

Then for each of the three measures, Value Line rank-orders the funds from top to bottom. It then calculates a percentile rank for the fund for each of the three measures. Hence, each fund receives three percentile ranks (e.g., a fund could have the following percentile ranks: 97th percentile in terms of the five-year growth persistence measure, the 81st percentile in terms of the one-year growth persistence measure, and the 83rd percentile for the three-year Sharpe ratio). Value Line then takes an equally weighted average of the three percentile ranks to come up with an overall percentile rank. Hence, in our example the fund has a $(97+81+83)/3 = 87$ th overall percentile rank.

If a fund does not have five years of return history at the time it is rated but has at least three years of data, then the calculation is based only on the one-year growth persistence and three-year Sharpe ratio percentile ranks. The two percentile ranks are then equally weighted to find the overall percentile rank. Similar to Morningstar, funds with fewer than three years of return history are not rated by Value Line.

With these overall percentile ranks, Value Line then assigns the overall rating. Funds that are in the top 10 percent receive an overall rating of 1; funds in the next 20 percent receive a 2; the next 40 percent receive a 3; the next 20 percent receive a 4; and the last 10 percent receive a 5.

Finally, three other items should be mentioned about the methodology of the Value Line Overall Rating system. First, the methodology of the Value Line ratings system has not changed since it was started in 1993. Second, in the data products, there is little explanation of the methodology used to calculate the ratings. Indeed, only through talking to executives at Value Line was I able to ascertain this information. This is quite different from Morningstar, which provides a very clear explanation of their methodology. Third, while many high-rated funds in the Morningstar system are also high rated by Value Line, the Value Line ratings system does indeed give significantly different ratings for many funds relative to what Morningstar indicates. For example, there are tens of funds that receive the highest rating in one rating system and yet receive the median rating (3) in the other system. Case in point, the Invesco Emerging Growth Fund—this fund was rated a five-star fund by Morningstar in January 1995, yet at the same time was given a rating of only 3 by Value Line.

3.4 The Lipper Analytical Ranking

The Lipper Analytical approach to evaluating mutual funds is quite different from that used by Morningstar and Value Line. Instead of rating mutual funds, Lipper classifies funds into a style and then simply rank-orders the funds in each style by their one-year, five-year, ten-year, and fifteen-year returns. No rating system itself is used.

Despite the lack of ratings, the real value added in the Lipper approach is in its style classification methodology. As mentioned above, the current Morningstar and Value Line rating systems do not emphasize style differences in their ratings methodologies because they rate all funds in a broad category. On the other hand, Lipper currently goes to great lengths to classify funds according to style. Specifically, what it does for domestic equity funds is the following.

First, in an attempt to add stability to the style classifications, Lipper uses *weighted* portfolio holdings as a base for the style classification. For funds with at least three years of return history, it uses

portfolio holdings weightings of 60 percent for the current period portfolio (heretofore called P); 30 percent for the prior fiscal year-end portfolio (heretofore called P-1); and 10 percent for the preceding prior fiscal year-end portfolio (heretofore called P-2). Hence, both P-1 and P-2 must be the portfolio holdings on the actual fiscal year-ends, while P is the fiscal year-end or half fiscal year-end portfolio holdings. For example, if a fund's fiscal year-end is June 30, and the current time is August 2002, then P would be the June 30, 2002 holdings, P-1 would be the June 30, 2001 holdings, and P-2 would be the June 30, 2000 holdings. However, by February 2003, P would be the December 31, 2002 holdings while P-1 and P-2 would be the same as used in the August calculation.

For funds that are fewer than three years old, Lipper uses a slightly different weighting system. For funds with enough history to have P and P-1, it uses a 60 percent weighting for P and a 40 percent weighting for P-1. For new funds, it uses a 100 percent weighting on P. Hence, unlike Morningstar and Value Line, Lipper incorporates new funds into the rankings.

Lipper then uses these weighted portfolio holdings to calculate the percentage of each fund's holdings that fall into each of Lipper's three market capitalization ranges: large-cap, medium-cap, and small-cap.¹⁰ To be classified as a large-cap fund, at least 75 percent of the weighted holdings of the funds must be in large-cap companies. For a mid-cap classification, 75 percent of the weighted holdings must be in mid-cap companies; and for a small-cap classification, 75 percent of the weighted holdings must be in the small-cap companies. Funds with less than 75 percent of their weighted holdings in any of the three market capitalization ranges are considered multi-cap funds.¹¹

Following the market capitalization classification, Lipper examines three primary characteristics of each fund: the price-to-earnings ratio, the price-to-book ratio, and a three-year sales-per-share growth value.¹² Lipper then compares the values of these three characteristics to values found from a market index. For large-cap funds, the index is the Standard and Poor's 500; for medium-cap funds, it is the Standard and Poor's MidCap 400; for small-cap funds, it is the Standard and Poor's SmallCap 600; and for multi-cap funds, it is the Standard and Poor's Super Composite 1500 Index. For each of the three primary characteristics, Lipper then calculates a "Z-score" that is the difference between the fund's

and the index's value divided by the standard deviation of the index. For example, if the price-to-earnings ratio of the fund was 12, the price-to-earnings ratio of the index was 6, and the standard deviation of the price-to-earnings ratio for the market index was 3, then the Z-score would be 2.

Upon calculation of the three Z-scores (one for each of the primary characteristics), Lipper then takes an equally weighted average of the three to arrive at an "L-score" for the period.

The style classification process then continues on. For each of the three periods—P, P-1, and P-2—Lipper calculates this L-score. It then uses the weighting system described before (60 percent for P, 30 percent for P-1, and 10 percent for P-2) to calculate "final weighted L-score." The final weighted L-scores are then fitted onto a normal distribution curve with funds with weighted final L-scores in the top 42 percent receiving growth classifications, funds in the bottom 42 percent receiving value classifications, and funds in the middle 16 percent receiving core classifications.¹³

Finally, after the market capitalization and style classifications, Lipper comes up with twelve styles of funds based on the methodology here (three style and four market capitalization classifications) plus another three styles for specialized equity funds. It is then within these styles that Lipper rank-orders the return performance of funds.

4. Data for the Out-of-Sample Analysis

There are two distinct data sets that we examine in the paper: (1) a data set composed of diversified domestic equity funds listed on the *Morningstar* data disk as of January 1, 1995, and (2) a data set composed of diversified domestic equity funds listed on the *Value Line* data disk from January 1, 1995. In this section we first describe in detail the creation of the *Morningstar* data set, followed by a description of the *Value Line* data set.¹⁴

4.1 The Morningstar Data Set

4.1.a Fund Selection Criteria

To select funds, we use the January 1995 *Morningstar On-Disk*. This disk provides data for all funds that were available to U.S. investors as of December 31, 1994. Our rationale in using the 1995 disk rather than a later disk is

that it enables us to examine the six-year out-of-sample performance of these funds (1995–2000) while at the same time using the most current data (this study was started in 2001). Furthermore, using earlier disks would have entailed several other problems. First, the Value Line ratings began only in 1993, so only a year or two of extra historical data could be added. Second, the process of following funds in the out-of-sample period is extremely onerous, as we have to trace the funds through name changes, mergers, and liquidations. Hence, starting in 1995 brought some relief from this process. Nevertheless, in spite of these difficulties, the six-year out-of-sample period used in this study represents the longest such sample used in the mutual fund rating literature.¹⁵

From this disk we then select all open funds with at least three years of return history that are within each of the following five Morningstar styles: aggressive growth, equity-income, growth, growth and income, and small company.¹⁶ This produced a sample of 770 funds.¹⁷ Our rationale for selecting only open funds is that we wanted all of the selected funds to be actually available to investors as of December 31, 1994. We use the three-year history criterion because it ensures that each fund will contain enough in-sample data to calculate some of our performance metrics (see Section 4.3 for more on this issue) and because of the fact that Morningstar does not rate funds with fewer than three years of returns. Finally, we use the styles of funds described above, as these styles are popular for domestic investors.

We then narrow the sample by eliminating replicate funds. Replicate funds are funds that are exactly the same as other funds in our sample. These replicates appear due to the different mutual fund share classes. In each such case, we exclude the fund that had the highest load charges given a six-year holding period. Hence, if one of the funds was a deferred-load fund whose load declined as the holding period increased and the other was a front-load fund, we always excluded the front-load fund, as its load was the highest given a six-year holding period. Our rationale here is that the investor would choose that share class as they were planning to hold the fund for six years. There were thirty-two of these replicate funds that were excluded from the sample, for a total final sample of 738 funds.

4.1.b Types of Data Used in the Out-of-Sample Analysis

With this sample of 738 funds, we then acquire the following data for each fund:

1. *The in-sample monthly return history from 1992–1994.* This again is available from the January 1995 Morningstar disk. These returns data account for management, administrative, and 12b-1 fees and other expenses automatically taken out of fund assets; however, they do not account for loads. Note again that all funds in our sample possess these historical returns, as we required that all funds have three years of return history at the time of selection.
2. *The out-of-sample monthly return history from 1995–2000.* This information is taken from later Morningstar data disks (quarterly data disks ranging from 1995 to 2000). As with the monthly returns from 1992–1994, these returns account for management, administrative, and 12b-1 fees and other costs, but do not account for loads.
3. *The Overall Morningstar Rating for the funds as of December 31, 1994.*
4. *The front and/or deferred loads of the fund (as of December 31, 1994) and the structure of how the deferred load was to be reduced over the next six years.* This information is taken from the January 1995 Morningstar disk.
5. *The style of the fund as of December 31, 1994.* This information is taken from the January 1995 Morningstar disk.

4.1.c Problem Funds

As described earlier, we select funds at the time the funds were listed by Morningstar. To examine the out-of-sample performance, we obtain the out-of-sample monthly returns of these funds. For a majority of the funds, obtaining the out-of-sample returns is simply a matter of following the funds' future performance. However, a minority of funds has experienced a name change, a merger, a combination of both, or has been liquidated during the out-of-sample period. Identifying

the out-of-sample returns for such funds is a complicated process, which we describe in this section.

To identify name changes, we use the Morningstar data disks. We then simply use the renamed funds' returns as the out-of-sample returns.

For the merger funds, we use the quarterly Morningstar disks from 1995–2000 to ascertain the month of each fund's merger. If these two sources did not provide the necessary information, we called the individual mutual fund companies to ascertain the merger information. Once the merger month was identified, we collected the out-of-sample returns by the following procedure. First, until the fund merged, we simply use the out-of-sample returns of the fund in question. After the fund has merged into its partner fund, we assume the investor randomly reinvests into one of the other surviving funds of the same style where the style is defined as of the January 1995 Morningstar disk. Hence, the out-of-sample returns from the merger month onward are the equally weighted average returns of all the other surviving funds in our sample with the same style. For example, the returns from the merger month onward of a growth fund would be the equally weighted average returns of all the other surviving growth funds.¹⁸

For the liquidated funds, we first identify when the funds were liquidated. Again, this information was obtained from the Morningstar disks. As with the merger funds, from the month of liquidation onward, we assume the investor randomly reinvests into those surviving funds in our sample with the same investment style.

4.1.d Returns Data and Load Adjustments

The out-of-sample returns and the in-sample returns consist of monthly returns from the Morningstar data disk. However, while these returns data are adjusted to account for management, administrative, and 12b-1 fees and other costs automatically taken out of fund assets, they are *not* adjusted for loads. That is, although the Morningstar ratings are based on load-adjusted returns, the monthly returns that Morningstar provides on its data disks do not account for loads.

As stated in section 3.1, Morningstar's overall ratings are based on *load-adjusted* returns. Hence, it is important for us to adjust the out-of-sample returns for loads when investigating whether the ratings are able to predict future performance.

Moreover, for investors, load-adjusted performance is what they are typically concerned with.

To adjust the returns for loads, we use an approach similar to that in Rea and Reid (1998), in Blake and Morey (2000), and Morey (2002b). This approach is described in Appendix A.

4.2 The Value Line Data Set

The Value Line data set is created in a fashion very similar to that of the Morningstar data set. This was purposely done so as to be able to compare the two ratings systems' relative out-of-sample performances.

Like the Morningstar data set, we choose all the funds from the Value Line data as of January 1, 1995. Similarly, we choose funds with three years of return history at the time they were rated and that have styles similar to those in the Morningstar sample. This means we included five Value Line styles of funds (aggressive growth, income, growth, growth-income, and small company). Again, these were the styles as defined by Value Line and not Morningstar. Finally, we exclude any multiple-class shares funds. This left a total of 653 funds.¹⁹

For the actual data we use the Morningstar data disks to obtain all the data for the Value Line sample except for the style and the Value Line overall rating of the funds, which are taken from the Value Line mutual fund survey.

For the problem funds we used an approach similar to that used in creating the Morningstar data set. Before a fund merged or liquidated, we simply use the out-of-sample returns of the fund. After the fund had merged into its partner fund, we assume the investor randomly reinvested into one of the other surviving funds of the same Value Line style where the style was defined as of January 1995. Hence, the out-of-sample returns from the merger onward are equally weighted averages of the returns of all the other surviving funds in our sample with the same Value Line style.

Finally, we use exactly the same load-adjustment process in the Value Line data set that we used in the Morningstar data set.

5. Methodology of Out-of-Sample Analysis

To measure out-of-sample performance we use four performance metrics: the mean monthly excess returns, the Sharpe ratio, a modified version of Jensen's alpha, and a

four-index alpha. For each performance metric we examine both *non-load-adjusted* and *load-adjusted* versions. We now explain, in detail, the four out-of-sample performance metrics.

5.1 Excess Mean Monthly Returns

The *non-load-adjusted* excess monthly returns for the *i*th mutual fund during the out-of-sample period are signified by $R_{it} - R_{ft}$, where R_{ft} is the thirty-day T-bill rate. The *non-load-adjusted mean* monthly excess return for the *i*th mutual fund during the out-of-sample period is $\overline{R_i - R_f}$.

The *load-adjusted* excess monthly returns for the *i*th mutual fund during the out-of-sample period are signified by $R_{it}^{LA} - R_{ft}$, where $R_{it}^{LA} = R_{it} - f^m$. The *load-adjusted mean* monthly excess returns are simply equal to $\overline{R_i^{LA} - R_f}$.

5.2 The Sharpe Ratio

The *non-load-adjusted* Sharpe ratio is

$$\text{Sharpe}_i = \frac{\overline{R_i - R_f}}{\sigma_i} \quad (2)$$

where

σ_i = the standard deviation of $R_{it} - R_{ft}$.

The *load-adjusted* Sharpe ratio for fund *i* is

$$\text{Sharpe}_i = \frac{\overline{R_i^{LA} - R_f}}{\sigma_i^{LA}} \quad (2a)$$

where

σ_i^{LA} = the standard deviation of $R_{it}^{LA} - R_{ft}$.

5.3 Modified Jensen and Four-Index Alphas

Two additional performance metrics used are the Jensen single-index and four-index alphas. The single-index model is a very known measure of performance, and the four-index alpha provides a *style-adjusted* measure of performance.

To calculate these alphas, the following time-series regression model is used:

$$R_{it} - R_{ft} = \alpha_i + \sum_{k=1}^K \beta_{ik} I_{kt} + \varepsilon_{it} \quad (3)$$

where

$R_{it} - R_{ft}$ = the excess total return (net of the thirty-day T-bill return) for fund *i* in in-sample month *t*

α_i = the alpha for fund *i*, used as a performance predictor

β_{ik} = the sensitivity of fund *i*'s excess return to index *k*

I_{kt} = the return for index *k* in in-sample month *t*

ε_{it} = the random error for fund *i* in in-sample month *t*

For Jensen alphas, $K = 1$ and I_{1t} = the excess total return of the S&P 500 in month *t*. For the four-index alphas, $K = 4$, I_{1t} = the excess total return of the S&P 500 in month *t*, I_{2t} = the excess total return of the Lehman Aggregate Bond Index in month *t*, I_{3t} = the difference in return between a small-cap and large-cap stock portfolio based on Prudential Bache indexes in month *t*, and I_{4t} = the difference in return between a growth and value stock portfolio based on Prudential Bache indexes in month *t*. We utilize this four-index model because, as shown in Elton, Gruber, and Blake (1996), this model provides for better risk adjustment for mutual funds than does the single-index model.²⁰

The non-load-adjusted modified Jensen and four-index alphas are calculated using a methodology similar to that of Elton, Gruber, and Blake (1996). Specifically, we utilize a time series period of monthly non-load-adjusted returns going back three years from the selection date and forward to the end of the out-of-sample evaluation period to obtain an estimate of the intercept from either the single-index or four-index model regression (equation 3). As mentioned in section 2, to be included in the sample each fund had to have three years of in-sample returns.

To obtain the alphas, we add the average monthly residual during the evaluation period to the intercept. For example, to obtain the modified Jensen alpha, we run the one-index model on monthly returns starting in January 1992 and ending in December 2000 (nine years) to obtain an estimate of the intercept. We then add the average of the fund's residuals during the six years after the selection date (1995–2000) to the estimated intercept to obtain the fund's modified Jensen alpha.

To obtain alphas for funds that merged or liquidated during the evaluation period, we proceed as follows. First, we run two regressions: (1) a regression using the fund's returns starting in January 1992 and ending in the month prior to the fund's disappearance and (2) a regression run over the

entire regression period (1992–2000) using the returns of an equally weighted portfolio formed each month from the existing funds of the same style where again the style was defined as of January 1995.²¹ We then form a weighted average of (1) the fund's estimated intercept plus the fund's average residual during the time it survived in the evaluation period and (2) the estimated intercept plus the average residual during the remaining time in the evaluation period of the equally weighted portfolio, where the fund's weight is the fraction of the evaluation period it survived and the equally weighted portfolio's weight is the remaining fraction. This provides a performance measure for an investor who buys a remaining fund in the sample at random if the original fund merges or liquidates.²²

For the load-adjusted modified Jensen and four-index alphas, we use the same methodology described above except that we use the excess load-adjusted returns, $R_{it}^{LA} - R_{ft}$, for the out-of-sample returns. That is, we use the excess non-load-adjusted returns for the in-sample data (1992–1994) and the excess load-adjusted returns for the out-of-sample period (1995–2000). Our rationale for not using load-adjusted returns during the in-sample period is that we assume the investor has not yet bought the fund, and hence, a load should not be subtracted from the returns. Moreover, the loads may be quite different during the in-sample period than during the out-of-sample period. As a result, it would be difficult to know what load to apply and for how long to apply it.

5.4 Alternative Predictors

In our study we compare the predictive ability of the Morningstar and Value Line ratings with those of two alternative predictors, an alternative Morningstar rating and an alternative Value Line rating. The rationale for using these alternative predictors is to see whether ratings based on a simpler methodology predict future performance as well as the actual Morningstar and Value Line ratings.

Both alternatives are calculated in the following manner. First, for all 738 funds in the Morningstar sample and all 653 funds in the Value Line sample, we calculate the 1992–1994 non-load-adjusted Sharpe ratio in a manner similar to that in equation 3. Then, for each of the two samples, we rank-order the funds from top to bottom according to the in-sample Sharpe ratio. Hence, we have two lists of funds, each ranked by the in-

sample Sharpe ratio. Then, to ensure that we use the same distribution of the ratings found in the original samples, we allocate the ratings based on the same distribution of ratings found in the original Morningstar and Value Line samples. For example, in the Morningstar sample there were 54 five-star funds. Hence, for the alternative Morningstar rating system we gave the five-star ratings to the funds with the 54 highest in-sample Sharpe ratios. In the Value Line sample there were 80 1-rated funds, so for the alternative Value Line rating system we gave ratings of 1 to the funds with the 80 highest in-sample Sharpe ratios.

5.5 Dummy Variable Regressions

The method we use to actually examine the out-of-sample predictive performance is a cross-sectional dummy variable regression analysis. This approach allows us to examine the differences in performance predictability among the rated funds.

For the dummy variable regression analysis, we estimate the following equation for each of the four predictors (Morningstar, Value Line, alternative Morningstar and alternative Value Line).

$$S_i = \gamma_0 + \gamma_1 D4_i + \gamma_2 D3_i + \gamma_3 D2_i + \gamma_4 D1_i + u_i \quad (4)$$

where

S_i = out-of-sample performance metric for fund i , i.e., non-load-adjusted and load-adjusted Sharpe ratios, non-load- and load-adjusted Jensen alphas, non-load- and load-adjusted four-index alphas, and non-load- and load-adjusted excess mean monthly returns.

$D4$ = 1 if a four-star fund (for the Morningstar and alternative Morningstar samples) or a 2-rated fund (for the Value Line and alternative Value Line samples), 0 if not.

$D3$ = 1 if a three-star fund (for the Morningstar and alternative Morningstar samples) or a 3-rated fund (for the Value Line and alternative Value Line samples), 0 if not.

$D2$ = 1 if a two-star fund (for the Morningstar and alternative Morningstar samples) or a 4-rated fund (for the Value Line and alternative Value Line samples), 0 if not.

$D1 = 1$ if a one-star fund (for the Morningstar and alternative Morningstar samples) or a 5-rated fund (for the Value Line and alternative Value Line samples), 0 if not.

$i = 1$ through N , where N is the total number of funds in the sample.

In equation 4, the five-star fund group for the Morningstar sample and the 1-rated fund group for the Value Line sample are the respective reference groups for the dummy variable regressions. Hence, when using the load-adjusted Sharpe ratio as the out-of-sample performance measure, the coefficient γ_0 represents the expected load-adjusted Sharpe ratio when all the dummy variables are equal to 0, and the coefficients γ_1 through γ_4 represent the differences between the dummy variables and the reference group. For example, a negative γ_1 implies that the group of four-star funds (or 2-rated funds for the Value Line sample) performs worse than the group of five-star funds (1-rated funds for the Value Line sample); a positive γ_1 implies that the group of four-star funds (2-rated funds for the Value Line sample) outperforms

the five-star fund group (1-rated funds for the Value Line sample). The t -statistics on the coefficients provide a test of the significance of the difference between an individual dummy group and the reference group.

We use the five-star funds (Morningstar and alternative Morningstar) or 1-rated funds (Value Line or alternative Value Line) as a reference group because they provide a ceiling from which we can compare the performance of the lower-rated group funds. If the ratings accurately predict out-of-sample performance, we should see increasingly negative (and significant) coefficients as we move from γ_1 to γ_4 .

6. Results

6.1 Summary Statistics of the Samples

Table 2 presents the summary statistics of the Morningstar and Value Line samples. It shows the number of funds in each style, the number of young, middle-aged, and seasoned funds, the number of problem funds, and the number of funds rated in each rating grade. The two samples are relatively similar except in the

Table 2: Summary Statistics on the Data Sets
All ages, ratings, and styles are as of January 1, 1995.

	The Morningstar Data Set		The Value Line Data Set
Total Number of Funds	738	Total Number of Funds	653
By Style:			
Aggressive Growth	44	Aggressive Growth	62
Equity-Income	50	Income	55
Growth	332	Growth	277
Growth-Income	210	Growth-Income	181
Small Company	102	Small Company	78
By Age:			
Young (3 to less than 5 years)	151	Young (3 to less than 5 years)	116
Middle-Aged (5 to less than 10 years)	263	Middle-Aged (5 to less than 10 years)	231
Seasoned (10 years or more)	324	Seasoned (10 years or more)	306
By Problem Funds:			
Liquidated Funds	26	Liquidated Funds	21
Merger Funds	95	Merger Funds	88
By Ratings:			
5-Star Funds	54 (7.31% of sample)	1-Rated Funds:	80 (12.25% of sample)
4-Star Funds	192 (26.02% of sample)	2-Rated Funds:	145 (22.20% of sample)
3-Star Funds	308 (41.73% of sample)	3-Rated Funds:	258 (39.51% of sample)
2-Star Funds	157 (21.27% of sample)	4-Rated Funds:	114 (17.46% of sample)
1-Star Funds	27 (3.67% of sample)	5-Rated Funds:	55 (8.42% of sample)
Average Rating:	3.121	Average Rating:	2.877
Standard Deviation of Ratings:	0.948	Standard Deviation of Ratings:	1.100

Table 3: Results of Dummy Variable Regressions on Morningstar and Value Line Mutual Fund Ratings (Using Non-Load-Adjusted Performance Metrics)

The ratings are the Morningstar and Value Line ratings on January 1, 1995 for all open diversified domestic equity funds with three years of historical returns. This consisted of 738 funds for the Morningstar sample and 653 funds for the Value Line sample. For the Morningstar system, 5 is the highest and 1 is the lowest rating; for the Value Line system, 1 is the highest rating and 5 is the lowest. The out-of-sample period is 1995–2000. The T-statistics are in parentheses.

Rating System Evaluated	Out-of-Sample Performance Measure	γ_0 (constant)	γ_1 (4-star funds)	γ_2 (3-star funds)	γ_3 (2-star funds)	γ_4 (1-star funds)	R ²	F-stat
Morningstar	Non-Load-Adjusted Sharpe Ratio	0.194*** (19.697)	0.024** (2.152)	0.035*** (3.303)	0.017 (1.524)	-0.011 (0.670)	0.028	5.29
Morningstar	Non-Load-Adjusted Jensen Alpha	-0.147*** (3.620)	0.051 (1.121)	0.052 (1.186)	-0.022 (0.479)	-0.031 (0.437)	0.012	2.31
Morningstar	Non-Load-Adjusted 4-Index Alpha	0.083 (1.92)	-0.015 (0.030)	-0.066 (1.420)	-0.079 (1.61)	-0.007 (0.092)	0.009	1.62

Rating System Evaluated	Out-of-Sample Performance Measure	γ_0 (constant)	γ_1 (2-rated funds)	γ_2 (3-rated funds)	γ_3 (4-rated funds)	γ_4 (5-rated funds)	R ²	F-stat
Value Line	Non-Load-Adjusted Sharpe Ratio	0.230*** (28.257)	0.003 (0.314)	-0.017 (1.848)	-0.027** (2.582)	-0.028** (2.203)	0.026	4.354
Value Line	Non-Load-Adjusted Jensen Alpha	-0.067** (2.007)	-0.008 (0.195)	-0.054 (1.396)	-0.116*** (2.649)	-0.133*** (2.522)	0.022	3.654
Value Line	Non-Load-Adjusted 4-Index Alpha	0.019 (0.526)	-0.001 (0.029)	0.019 (0.458)	0.015 (0.315)	-0.033 (0.599)	0.002	0.359

*** Indicates significance at the 1 percent level.
 **Indicates significance at the 5 percent level.

Table 4: Results of Dummy Variable Regressions on Morningstar and Value Line Mutual Fund Ratings (Using Load-Adjusted Performance Metrics)

The ratings are the Morningstar and Value Line ratings on January 1, 1995 for all open diversified domestic equity funds with three years of historical returns. This consisted of 738 funds for the Morningstar sample and 653 funds for the Value Line sample. For the Morningstar system, 5 is the highest and 1 is the lowest rating; for the Value Line system, 1 is the highest rating and 5 is the lowest. The T-statistics are in parentheses.

Rating System Evaluated	Out-of-Sample Performance Measure	γ_0 (constant)	γ_1 (4-star funds)	γ_2 (3-star funds)	γ_3 (2-star funds)	γ_4 (1-star funds)	R ²	F-stat
Morningstar	Load-Adjusted Sharpe Ratio	0.187*** (19.156)	0.025** (2.247)	0.034*** (3.229)	0.016 (1.377)	-0.014 (0.827)	0.029	5.493
Morningstar	Load-Adjusted Jensen Alpha	-0.178*** (4.350)	0.057 (1.233)	0.049 (1.123)	-0.030 (0.635)	-0.044 (0.628)	0.015	2.834
Morningstar	Load-Adjusted 4-Index Alpha	0.052 (1.20)	-0.009 (0.188)	-0.068 (1.461)	-0.087 (1.747)	-0.020 (0.273)	0.011	1.993

Rating System Evaluated	Out-of-Sample Performance Measure	γ_0 (constant)	γ_1 (2-rated funds)	γ_2 (3-rated funds)	γ_3 (4-rated funds)	γ_4 (5-rated funds)	R ²	F-stat
Value Line	Load-Adjusted Sharpe Ratio	0.223*** (27.448)	0.001 (0.111)	-0.018 (1.919)	-0.027*** (2.623)	-0.029** (2.264)	0.024	4.095
Value Line	Load-Adjusted Jensen Alpha	-0.094*** (2.797)	-0.020 (0.475)	-0.063 (1.620)	-0.127*** (2.875)	-0.136*** (2.571)	0.022	3.676
Value Line	Load-Adjusted 4-Index Alpha	-0.008 (0.236)	-0.013 (0.293)	0.009 (0.238)	0.004 (0.082)	-0.037 (0.656)	0.002	0.303

*** Indicates significance at the 1 percent level.
 **Indicates significance at the 5 percent level.

Table 5: Results of Dummy Variable Regressions on Alternative Ratings (Using Non-Load-Adjusted Performance Metrics)

The alternative Morningstar and alternative Value Line ratings are used in the below regressions. All ratings are as of January 1, 1995. The funds included are all open diversified domestic equity funds with three years of historical returns. This consisted of 738 funds for the Morningstar sample and 653 funds for the Value Line sample. For the alternative Morningstar system, 5 is the highest and 1 is the lowest rating; for the alternative Value Line system, 1 is the highest rating and 5 is the lowest. The T-statistics are in parentheses.

Rating System Evaluated	Out-of-Sample Performance Measure	γ_0 (constant)	γ_1 (4-star funds)	γ_2 (3-star funds)	γ_3 (2-star funds)	γ_4 (1-star funds)	R ²	F-stat
Alternative Morningstar	Non-Load-Adjusted Sharpe Ratio	0.210*** (21.218)	0.001 (0.092)	0.019 (1.839)	0.010 (0.085)	-0.037** (2.157)	0.027	5.159
Alternative Morningstar	Non-Load-Adjusted Jensen Alpha	-0.091** (2.243)	-0.024 (0.518)	-0.006 (0.145)	-0.053 (1.121)	-0.180*** (2.566)	0.014	2.563
Alternative Morningstar	Non-Load-Adjusted 4-Index Alpha	0.037 (0.872)	-0.009 (0.193)	0.016 (0.347)	-0.005 (0.094)	-0.181** (2.444)	0.013	2.472

Rating System Evaluated	Out-of-Sample Performance Measure	γ_0 (constant)	γ_1 (2-rated funds)	γ_2 (3-rated funds)	γ_3 (4-rated funds)	γ_4 (5-rated funds)	R ²	F-stat
Alternative Value Line	Non-Load-Adjusted Sharpe Ratio	0.201*** (24.490)	0.012 (1.261)	0.025*** (2.716)	0.019 (1.767)	0.002 (0.127)	0.016	2.664
Alternative Value Line	Non-Load-Adjusted Jensen Alpha	-0.138*** (4.110)	0.029 (0.709)	0.041 (1.062)	-0.003 (0.076)	-0.064 (1.215)	0.010	1.664
Alternative Value Line	Non-Load-Adjusted 4-Index Alpha	-0.005 (0.154)	0.023 (0.519)	0.057 (1.435)	0.046 (0.999)	-0.061 (1.103)	0.011	1.919

*** Indicates significance at the 1 percent level.
**Indicates significance at the 5 percent level.

Table 6: Results of Dummy Variable Regressions on Morningstar and Value Line Fund Ratings (Using Load-Adjusted Performance Metrics)

The alternative Morningstar and alternative Value Line ratings are used in the below regressions. All ratings are as of January 1, 1995. The funds included are all open diversified domestic equity funds with three years of historical returns. This consisted of 738 funds for the Morningstar sample and 653 funds for the Value Line sample. For the alternative Morningstar system, 5 is the highest and 1 is the lowest rating; for the alternative Value Line system, 1 is the highest rating and 5 is the lowest. The T-statistics are in parentheses.

Rating System Evaluated	Out-of-Sample Performance Measure	γ_0 (constant)	γ_1 (4-star funds)	γ_2 (3-star funds)	γ_3 (2-star funds)	γ_4 (1-star funds)	R ²	F-stat
Alternative Morningstar	Load-Adjusted Sharpe Ratio	0.202*** (20.728)	-0.001 (0.094)	0.019 (1.840)	0.009 (0.818)	-0.039** (2.329)	0.031	5.852
Alternative Morningstar	Load-Adjusted Jensen Alpha	-0.118*** (2.875)	-0.034 (0.737)	-0.010 (0.227)	-0.059 (1.252)	-0.191*** (2.697)	0.015	2.781
Alternative Morningstar	Load-Adjusted 4-Index Alpha	0.011 (0.252)	-0.019 (0.402)	0.012 (0.265)	-0.011 (0.225)	-0.191*** (2.564)	0.014	2.657

Rating System Evaluated	Out-of-Sample Performance Measure	γ_0 (constant)	γ_1 (2-rated funds)	γ_2 (3-rated funds)	γ_3 (4-rated funds)	γ_4 (5-rated funds)	R ²	F-stat
Alternative Value Line	Load-Adjusted Sharpe Ratio	0.192*** (23.577)	0.012 (1.160)	0.027*** (2.881)	0.019 (1.877)	0.001 (0.072)	0.02	3.169
Alternative Value Line	Load-Adjusted Jensen Alpha	-0.174*** (5.127)	0.025 (0.597)	0.045 (1.155)	-0.004 (0.084)	-0.071 (1.331)	0.12	1.916
Alternative Value Line	Load-Adjusted 4-Index Alpha	-0.041 (1.147)	0.018 (0.414)	0.062 (1.521)	0.046 (0.984)	-0.067 (1.212)	0.014	2.236

*** Indicates significance at the 1 percent level.
**Indicates significance at the 5 percent level.

number of funds rated very high and very low. In the Morningstar sample, only 11 percent of funds are rated as five-star or one-star, while in the Value Line sample the percentage of funds rated 1 or 5 is about 21 percent.²³

6.2 Risk-Adjusted Results

The risk-adjusted results are reported in tables 3 through 8. Table 3 shows the results of our dummy variable analysis on the Morningstar and Value Line systems using non-load-adjusted performance metrics, while table 4 shows the results using load-adjusted performance metrics. Tables 5 and 6 then show the results using the alternative ratings, with table 5 using non-load-adjusted performance metrics and table 6 using load-adjusted performance metrics. Finally, tables 7 and 8 provide summaries of the out-of-sample analysis on the risk-adjusted measures. Specifically, table 7 shows the number of times that the predictors produce significant coefficients with the correct and wrong signs, and table 8 presents the number of cases for which the signs on the coefficients are negative or positive without regard to statistical significance.

The results show several interesting findings. First, while the Morningstar results for the four-index alpha do show the correct negative sign on tables 3 and 4, for the most part the results for Morningstar and Value Line systems do not show much statistical ability to predict winning funds. For all the performance metrics (non-load-adjusted and load-adjusted), both rating systems show little difference between out-of-sample performance of highest-rated funds (five-star for Morningstar or 1-rated for Value Line) compared with the median-rated funds (three-star or 3-rated). Indeed, as showcased in table 7, there is not a single case where the γ_2 coefficient is significant and negative, indicating that top-rated funds do not perform significantly better out of sample than median-rated funds. In fact, the only significant coefficient for γ_2 is in the Morningstar non-load-adjusted Sharpe ratio case (table 3); however, in this case the coefficient is positive, indicating that median-rated funds actually perform significantly *better* than the top-rated funds.

Table 8 provides even more evidence of the inability of top-rated funds to perform better than the median-rated funds as it shows that in six of the twelve total cases (two for Morningstar and

four for Value Line), the γ_2 coefficient is positive. In other words, in half the cases, the coefficient sign actually shows that median-rated funds are performing *better* than top-rated funds. By showing that it is difficult to predict winning funds, these results are consistent with those of Blake and Morey (2000), who find that high-rated funds do not perform better out-of-sample than median-rated funds.

Second, in terms of predicting poor-performing funds, our results indicate that the Value Line system has some ability to predict these funds whereas the Morningstar method shows less ability. These results can be best illustrated by examining tables 7 and 8. In table 7, the far right two columns show the number of cases (six for each rating system) in which the coefficients γ_3 and γ_4 are significant and negative. Again, a negative and significant coefficient implies that low-rated funds performed significantly worse than top-rated funds. For the Morningstar system, we find that there are no cases where the coefficients are negative and statistically significant. On the other hand, for the Value Line system we find four (out of six total) with negative and significant coefficients (note that these coefficients were found for the non-load-adjusted and load-adjusted Sharpe and Jensen measures). Table 8 shows that in terms of the signs alone, the Value Line system also seems to work slightly better in predicting the low-performing funds, as in all six cases the coefficients for the lowest-rated funds are lower than those of median-rated funds. The Morningstar results show only four of the six cases having coefficients that meet this condition.

Third, the Value Line system is better at not completely mispredicting future performance. Table 7 (columns 1 and 2) shows that the examination of the Value Line system produces eight (out of a possible twenty-four) coefficients that are negative and significant and zero coefficients that are positive and significant. Conversely, the Morningstar method produces zero negative and significant coefficients and four positive and significant coefficients. Table 8 provides similar evidence as eighteen of the twenty-four coefficients are negative for the Value Line results while only ten of the twenty-four are negative for the Morningstar results.

Fourth, in terms of the specific performance metrics, our results show that the Value Line system does a better job of predicting future

Table 7: Summary of the Ability of the Ratings to Forecast Out-of-Sample Performance: Significance Levels

There are twenty-four total γ_1 , γ_2 , γ_3 , and γ_4 coefficients (four ratings systems: Morningstar, Value Line, alternative Morningstar, alternative Value Line) and six out-of-sample performance measures for each rating system (non-load-adjusted and load-adjusted Sharpe indexes, non-load- and load-adjusted Jensen alphas, and non-load- and load-adjusted four-index alphas).

Rating System Evaluated	Number of times (out of 24) that the rating system produces a significantly* negative coefficient for γ_1 , γ_2 , γ_3 or γ_4	Number of times (out of 24) that the rating system produces a significantly* positive coefficient for γ_1 , γ_2 , γ_3 or γ_4	Number of cases (out of 6) in which the coefficient γ_1 (4-star or 2-rated) is significant* and negative	Number of cases (out of 6) in which the coefficient γ_2 (3-star or 3-rated) is significant* and negative	Number of cases (out of 6) in which the coefficient γ_3 (2-star or 4-rated) is significant* and negative	Number of cases (out of 6) in which the coefficient, γ_4 (1-star or 5-rated), is significant* and negative
Morningstar	0	4	0	0	0	0
Value Line	8	0	0	0	4	4
Alternative Morningstar	6	0	0	0	0	6
Alternative Value Line	0	1	0	0	0	0

* At the 5 percent level of significance or higher.

Table 8: Summary of the Ability of the Ratings to Forecast Out-of-Sample Performance: Coefficient Signs

There are twenty-four total γ_1 , γ_2 , γ_3 , and γ_4 coefficients (four ratings systems: Morningstar, Value Line, alternative Morningstar, alternative Value Line) and six out-of-sample performance measures for each rating system (non-load-adjusted and load-adjusted Sharpe indexes, non-load- and load-adjusted Jensen alphas, and non-load- and load-adjusted four-index alphas).

Predictor	Number of cases (out of 24) in which the coefficient signs for γ_1 , γ_2 , γ_3 and γ_4 are negative	Number of cases (out of 24) in which the coefficient signs for γ_1 , γ_2 , γ_3 , and γ_4 are positive	Number of cases (out of 6) in which the coefficient signs are such that: $\gamma_0 > \gamma_2$ (highest rated versus median rated)	Number of cases (out of 6) in which the coefficient signs are such that: $\gamma_1 > \gamma_3$ (second highest versus second lowest rated)	Number of cases (out of 6) in which the coefficient signs are such that: $\gamma_2 > \gamma_4$ (median rated versus lowest rated)
Morningstar	14	10	2	6	4
Value Line	18	6	4	4	6
Alternative Morningstar	16	8	2	2	6
Alternative Value Line	6	18	0	2	6

performance than the Morningstar system when using the Sharpe ratio and the Jensen alpha, whereas the Morningstar system does a better job when using the four-index alpha. For example, for the Sharpe and Jensen performance measures, the results for the Value Line system show negative coefficients that generally grow larger as we move from γ_1 to γ_4 , indicating that lower-rated funds have worse out-of-sample performance. These results again hold up whether examining non-load-adjusted or load-adjusted returns. The same generally holds true for the Morningstar system when using four-index alphas as the performance metric.

Fifth, in terms of the alternative predictors, we find that while the alternative Morningstar predictor shows little ability to predict winning funds, it actually works slightly better at predicting poor future performance than the actual Morningstar system. In all of the six cases, the Morningstar alternative produces negative and significant coefficients for γ_4 , whereas in the Morningstar system there are no cases of negative and significant coefficients. The alternative Value Line system, on the other hand, is not able to predict better than the original Value Line system. Indeed, in eighteen of the twenty-four coefficients, the alternative Value Line rating system actually produces coefficients with positive signs.

In sum, the risk-adjusted results suggest that none of the four ratings systems evaluated (Morningstar, Value Line, alternative Morningstar, and alternative Value Line) is able to significantly predict winning funds. We also find that, when looking at aggregate results, the Value Line system slightly outperforms the Morningstar system. This is especially the case when examining the system's ability to predict losing funds. In terms of the alternative predictors, the alternative to the Morningstar predictor predicts better than the original while the alternative Value Line system is inferior to the Value Line system.

7. Conclusions

In this paper we examine two issues. First, we document the mutual fund ratings/rankings methodology of the Morningstar, Value Line, and Lipper Analytical systems. Second, we investigate the out-of-sample predictive ability of the Morningstar and Value Line ratings. Our findings are as follows:

1. In terms of the ratings methodologies, we find that the Morningstar system emphasizes expense-, load-, and risk-adjusted returns where risk is defined as downside risk. On the other hand, the Value Line system emphasizes the persistence of fund performance, i.e., the ability of a fund to consistently outperform other funds in terms of simple (non-expense-, non-load-, non-risk-adjusted) returns. There is also a difference in the time horizons that the two systems examine: Morningstar uses a system that emphasizes a fund's long-term performance (up to ten years) while Value Line uses a shorter period of time (up to five years). While both Morningstar and Value Line rate mutual funds on a scale of 1 to 5, the Lipper Analytical system, conversely, separates funds into many styles that are determined by Lipper itself, and then ranks funds in each of the defined styles from top to bottom by their total return. Lipper does not provide any formal ratings for funds.
2. In terms of the predictive ability of the Morningstar and Value Line ratings, we use an approach that is robust to survivorship bias and load-adjusted returns. Furthermore, we use four different measures of out-of-sample performance and also examine alternative ratings systems that are based on simpler methodology than the Morningstar and Value Line systems. We find three results in this analysis:
 - A. Neither of the ratings systems, nor the alternative ratings systems, is able to successfully predict winning funds. Specifically, we find that the difference in out-of-sample performance between top-rated and median-rated funds is never significant with the correct signs and sometimes actually has the incorrect signs.
 - B. There is some ability using risk-adjusted measures to predict losing funds, as the lowest-rated funds do have lower levels of out-of-sample performance than other funds. This result is consistent with Carhart (1997) and Blake and Morey (2000), who find that poor-performing funds consistently perform worse in the future.
 - C. There is some weak evidence that the Value Line system actually predicts future performance better than the Morningstar system. However, this result holds only for the poor-performing funds and only for the Sharpe

index and Jensen alpha performance metrics. The Morningstar system is able to better predict future performance using the more complicated four-index alpha performance metric.

These findings, when combined with research by others (e.g., Carhart 1997 and Blake and Morey 2000), provide strong evidence that investors should be very cautious about interpreting the ratings as signals of future winning performance. Instead, they should use the ratings as measures of past performance and remember that ratings are of dubious help in predicting future winning funds.

This study was funded by a grant from the Investment Management Consultants Association (IMCA). The author wishes to thank Edward Baker and Bonny Brill of IMCA and Reuben Gregg Brewer and Samuel Eisenstadt of Value Line for helpful comments.

Appendix A: Methodology for Calculation of Load-Adjusted Returns

For front loads, we consider an investor who buys and holds the load shares for our holding period of seventy-two months (six years). As with most front loads, we assume that the investor buying the fund pays a load in a lump sum at the time the fund is purchased. To spread the front load across the period that the shares are held, we use Rea and Reid's assumption that the investor borrows the amount necessary to pay the load up front and then repays the loan as an annuity in equal, monthly installments during the holding period. Hence, the monthly load adjustment reflects the amount that was borrowed and the interest on the loan.

Mathematically, our front-load adjustment process is the following:

$$f^m = \frac{f}{\sum_{j=1}^h (1+r)^j}$$

where

r = the monthly interest rate (the monthly periodic interest rate of five-year Treasury yields in January 1995)

f = the front load (expressed as a percentage)

h = the number of months the fund is held

f^m = the monthly front-load adjustment

Hence, the front-load-adjusted (for front loads) returns are

$$R_{it}^{LA} = R_{it} - f^m$$

where

R_{it} = the non-load-adjusted monthly return of fund i in month t , where t goes from 1 to 72

R_{it}^{LA} = the monthly front-load-adjusted return of fund i in month t

As an example of the above adjustment, consider a six-year investment in Fidelity's Magellan fund starting December 31, 1994. As of December 31, 1994, this fund had a front load of 3 percent, and the six-year Treasury yield was 7.76 percent, giving a monthly periodic rate of 0.6467 percent.²⁴ Therefore, for the five-year holding (out-of-sample) period, $f = 3\%$, $r = 0.006467$, and $h = 72$, giving $f^m = 0.05225\%$. We then subtract 0.05225% from each of the Magellan fund's seventy-two monthly returns from 1995 to 2000 to obtain the front-load-adjusted returns.

In terms of deferred loads, we did not have to make any adjustments, as all the funds with deferred loads had loads that went to zero if the fund was held for six years. If we had chosen a shorter time horizon to evaluate the performance of the funds, it would have been necessary to adjust the returns for deferred loads.

Endnotes

1. See David Franekki, "Fund Ratings and Recent Results Diverge," *Wall Street Journal*, 3 May 2000, C27.

2. *Ibid.*

3. Hybrid funds are funds classified by Morningstar as those that do not fit cleanly into either the equity or fixed-income categorization. They are usually balanced funds that invest in stocks and bonds or high-yield bond funds that are technically fixed-income funds but are much more speculative in nature. Specialty funds are those that seek capital appreciation by concentrating in a single industry or sector, such as natural resources, utilities, real estate, or precious metals. The *Morningstar Principia Manual* has a more detailed description of these definitions.

4. For example, the Ivy International A Fund was rated as a three-star fund in October 1996, yet in January 1997, after the broad asset class reorganization, the fund received a five-star rating. Another international equity fund, the T. Rowe Price International Stock Fund, was rated a two-star fund in October 1996 and yet in January 1997 received a four-star rating.

5. Blume (1998), p. 4–5, provides an excellent description of how Morningstar accounts for loads in the Morningstar returns. The load-adjustment process is the following. Assume L is the load adjustment. If there is no load of any type, then L is equal to 1. If there is a load, L is less than 1, i.e., a 4 percent front-end load would make L equal to 0.96. The load-adjusted return is then the (return of the fund)* L . Note that the front-end load is always assumed to be the maximum possible load. The deferred load adjustment is reduced as the holding period is increased.

6. See *Morningstar Principia Manual* (1998).

7. *Ibid.*

8. See Floyd Norris, "Morningstar to Grade on a Curve," *New York Times*, 23 April 2002, C 11 and Mark Hulbert, "A Realignment of the Stars in the Mutual Fund Sky," *New York Times*, 19 May 2002, Sec. 3, p. 7.

9. As of this paper writing (June 2002), it was not clear what methodology Morningstar was going to use to define the styles of funds.

10. The three ranges are updated each month to adjust for market conditions. For example, as of December 31, 2000, in the medium-cap range were companies with market capitalizations between \$10.89 billion and \$2.31 billion.

11. Note that all funds that have between 73 percent and 74.99 percent of their holdings in a specific market capitalization range have their weighted portfolio holdings recalculated. Instead of using the 60 percent, 30 percent, and 10 percent weightings described above, they instead use a simple average of P, P-1, and P-2. If the simple average puts the amount of the holdings at 75 percent or above in a specific market capitalization range, the fund is reclassified.

12. The three-year sales-per-share growth value is a weighted average of the three-year sales-per-share growth values for each individual security held by the fund.

13. As with the market capitalization classifications, Lipper also calculates border tests for the style classification. See Lipper (2001) for more.

14. Note that we do not examine the out-of-sample performance of the Lipper rankings. The reason is that the style classification methodology of Lipper described in Section 3.4 is completely different from the methodology that Lipper used in 1995, the time when we begin the out-of-sample analysis. At that time, Lipper used a style classification methodology that was based largely on the language of the funds' prospectuses. In fact, the Lipper styles for domestic equity funds in 1995 were defined as completely different styles from those defined in 3.4. In sum, the fact that the past methodology is so different from that used today makes a study of the out-of-sample performance of the previous methodology inconsequential for investors interested in using the current Lipper methodology.

15. For example, Blake and Morey (2000) use five years of out-of-sample data. Note also that a research note by a Morningstar employee, Laura Lалlos, investigated the ten-year out-of-sample performance of Morningstar funds and found that funds that had been rated as five-star funds ten years ago had outperformed other funds. However, because the study provides little description of the methodology used (including whether any adjustment was made for survivorship bias), it is difficult to ascertain its validity.

16. These Morningstar styles are based on the wording of the investment prospectuses of the individual funds. Morningstar technically uses the term "investment objectives" instead of "styles" on the January 1995 data disk.

17. The onerous data collection process of tracing the funds through name changes, mergers, and liquidations forced us to include only these diversified domestic equity funds and to not look at other funds, such as bond funds. An inclusion of all funds in the Morningstar database with three years of returns in January 1995 would have resulted in over 2000 funds and would have been well beyond the scope of this study. Moreover, the use of diversified domestic equity funds is consistent with that of Blume (1998) and Blake and Morey (2000).

18. Of course, an alternative approach would be to use the "follow-the-money" approach introduced in Elton, Gruber, and Blake (1996), where a merged fund's returns are spliced to its "merge partner" fund's returns to form a complete time series. We did not use this method because we require a complete in-sample time series of returns, i.e., 1992–1994, for the "merge partner" fund, and in some cases the partner fund did not exist long enough to obtain such a series.

19. The reason for the smaller number of funds in the Value Line system is the differences in how the two systems define the style of a fund. In 1995, Morningstar typically used the prospectus of the fund whereas Value Line used the prospectus and its own in-house analysis.

20. In addition to the relatively standard use of the S&P, bond, and size indexes, the model includes an index to account for the performance of growth versus value stocks. This index is added due to the large number of mutual funds that state either growth or value as an objective and because the growth and value categorizations have been shown by Fama and French (1993, 1995) to be highly related to stock returns. See Elton, Gruber, and Blake (1996), p. 137 for more on this issue. We use the Prudential Bache indexes as opposed to other indexes, as the Prudential Bache indexes have been used in a number of well-known mutual fund performance studies including Elton, Gruber, and Blake (1995, 1996).

21. The Morningstar data set uses the Morningstar styles while the Value Line data set uses the Value Line styles.

22. As is well known, betas can be time-varying, and hence, a sufficiently long sample is needed to account for such variation. This is the reason for the combination of in-sample and out-of-sample data for the construction of the alphas. We choose the three-year in-sample rule, as every fund had at least three years of in-sample data. Of course, some of the funds in our sample had many more than three years of in-sample data; however, this would have meant that the number of observations used to construct the alphas would have been significantly different for some funds. As a compromise we used the three-year in-sample rule for all funds.

23. It should also be noted that both the samples have an average rating above the median rating (3.121 for Morningstar and 2.877 for Value Line). Again the reason for this upward bias in the ratings is that, in both the Morningstar and Value Line cases, we are not sampling the entire category of funds. Rather, we are only selecting diversified domestic equity funds, i.e., aggressive growth, equity-income (income for Value Line), growth, growth-income, and small company. For example, in 1995 the Morningstar equity category included balanced, hybrid and international funds. Hence, in both the Morningstar and Value Line systems, these diversified domestic equity funds received higher-than-average ratings than the other funds in the category.

24. These data are from the Federal Reserve Database on the World Wide Web at <http://www.econmagic.com>. This method is, of course, somewhat sensitive to the interest rate at the time of purchase. Lower rates mean lower front-load costs.

References

- Blake, Christopher R., and Matthew R. Morey. "Morningstar Ratings and Mutual Fund Performance." *Journal of Financial and Quantitative Analysis* 35, no. 3 (September 2000): 451–481.
- Blume, Marshall. "An Anatomy of Morningstar Ratings." *Financial Analysts Journal* (March/April 1998): 19–27.
- Carhart, Mark. "On Persistence in Mutual Fund Performance." *Journal of Finance* 52 (1997): 1–33.
- Del Guercio, Diane, and Paula A. Tkac. "Star Power: The Effect of Morningstar Ratings on Mutual Fund Flows." Working paper, Federal Reserve Bank of Atlanta, 2002.
- Fama, Eugene, and Kenneth French. "Common Risk Factors in the Returns of Stocks and Bonds." *Journal of Financial Economics* 33 (1993): 3–56.
- . "Size and Book-to-Market Factors in Earnings and Returns." *Journal of Finance* 50 (1995): 131–156.
- Elton, Edwin, J., Martin J. Gruber, and Christopher R. Blake. "Fundamental Variables, APT, and Bond Fund Performance." *Journal of Finance* 50 (1995): 257–276.

- . "The Persistence of Risk-Adjusted Mutual Fund Performance." *Journal of Business* 69, no. 2 (1996): 133–157.
- Lipper U.S. Diversified Equity Fund Classification Source Document, 2001. Obtained from the World Wide Web at <http://www.lipperweb.com>.
- Morey, Matthew, R. "Mutual Fund Age and Morningstar Ratings." *Financial Analysts Journal* (March/April 2002a): 56–63.
- . "Should You Carry the Load: A Comprehensive Analysis of the Out-of-Sample Performance of Load and No-Load Mutual Funds." *Journal of Banking and Finance* (forthcoming 2002b).
- Morningstar On-Disk and Principia Manuals. Chicago: Morningstar Publications, 1992–2001.
- Rea, John D., and Brian K. Reid. "Trends in the Ownership Cost of Equity Mutual Funds." *Investment Company Institute Perspective* 4, no. 3 (1998): 1–15.
- Sharpe, William. "Morningstar Performance Measures." *Financial Analysts Journal* (July/August 1998): 21–33.
- Warshawsky, Mark, Mary DiCarlantonio, and Lisa Mullan. "The Persistence of Morningstar Ratings." *Journal of Financial Planning* 13, no. 9 (September 2000): 110–121.