Estimation Risk in Morningstar Fund Ratings

H.D. VINOD AND MATTHEW R. MOREY

H.D. VINOD
is a professor of economics at Fordham University in New York City.
vinod@fordham.edu

MATTHEW R. MOREY
is an associate professor of finance at Pace University in New York City.
mmorey@pace.edu

With over 12,000 mutual funds now available, many investors are attracted to mutual fund ratings to help them identify the best funds. The most popular of different mutual fund rating companies and systems is undoubtedly Morningstar's five-star rating system. Morningstar rates mutual funds on scale of one to five stars, where one star is the lowest rating and five the highest.

The star rating system has become part of the vocabulary in mutual funds (see Morey [2000]). There is now strong empirical evidence that investment flows into and out of mutual funds are closely related to the Morningstar star rating.

Del Guercio and Tkac [2001] find that an initial five-star rating for a fund resulted in an inflow of funds that was 53% above the normal expected flow. Damato [1996] reports that 97% of the money flowing into no-load equity mutual funds between January and August 1995 was invested into funds rated at four or five stars, while funds with under three stars actually suffered a net outflow during the same period.

The common use of Morningstar ratings in mutual fund advertising suggests that mutual fund companies believe investors care about Morningstar ratings. In some cases, the only mention of performance in a mutual fund advertisement is the Morningstar rating.

Whatever its widespread use, Morningstar makes no assertion that its star rating system can predict future mutual fund performance. Rather, it regards the star ratings as "achievement scores" that investors should use to help focus their search for the best mutual fund. While this advice is certainty sensible, the fact is that many people do use the star ratings as signals of future performance.[1] One has only to open a newspaper to that see many mutual funds advertise their Morningstar ratings in the belief that they will attract new investors.

We investigate *estimation risk* in the popular Morningstar rating system. While estimation risk is used in performance metrics such as Jensen's alpha or value at risk, it has received very little or no attention in mutual fund ratings, which ironically are the performance metrics that many individual investors actually use. We would like to meet this need by illustrating the way the Morningstar rating system suffers from estimation risk.

Specifically, we show that since Morningstar rates funds regardless of age differences, the estimates upon which younger fund ratings are based have significantly higher estimation risk than the estimates upon which the ratings of older funds are based. As a result, investors can be somewhat less confident that the ratings of younger funds are truly what they are estimated to be. We illustrate the point in an investigation of 1,281 international equity mutual funds.

MORNINGSTAR METHODOLOGY

Our description of the basics of the Morningstar rating system is a synthesis of what can be found in Morningstar manuals, as described in Blume [1998], Sharpe [1998], Blake and Morey [2000], and Morey [2000, 2001].

Morningstar has a number of different rating systems. We focus on what we call the overall Morningstar star rating. It is the most well known of the Morningstar ratings systems—the rating that Morningstar has popularized and marketed in its data since the company started providing ratings.

To calculate the overall Morningstar star rating, Morningstar classifies funds into one of four categories: Domestic Equity, International Equity, Municipal Bond, and Taxable Bond. After categorization, the overall star ratings are based on risk-adjusted returns.

To calculate the risk-adjusted return, Morningstar calculates an expense- and load-adjusted return for each fund first by adjusting the returns for expenses such as 12b-1 fees, management fees, and other costs automatically taken out of the fund, and then by adjusting for front-end and deferred loads.[2] 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 90-day U.S. Treasury bill rate:[3]

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

Morningstar divides through by one of these two variables to prevent distortions caused by low or negative average excess returns in the denominator of Equation (1). This might occur in a protracted down market.[4]

Morningstar then calculates a *Morningstar Risk* measure. The calculation is different from traditional risk measures, such as beta and standard deviation. Traditional statistical measures treat greater-than- and less-than-expected returns as added volatility. Morningstar believes that the greatest fear for most investors is losing money, which it defines as underperforming the risk-free rate of return any investor can earn from the 90-day Treasury bill. Hence, its risk measure focuses only on downside risk.

To calculate Morningstar Risk, the system plots the monthly returns in relation to T-bill returns. The amounts by which the fund trails the T-bill return each month are summed and then divided by the total number of months in the time horizon. 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 compared to the average fund in its category.

Exhibit 1 provides a hypothetical example for a time horizon of 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 would be subtracted from the three-year Morningstar return; for the five-year score, the five-year Morningstar risk would be subtracted from the five-year Morningstar return; and so on. The resulting time horizon score is then compared to the time horizon score of other funds in that fund category: the three-year score of the fund compared to all the other funds' three-year scores, and so on.

Then, for each of the three time horizons, star ratings are allocated in an ad hoc manner. If the fund's score lands in the top 10%, it receives a time-specific rating of five stars; if it falls in the next 22.5%, it receives a time-specific star rating of four stars. In the middle 35%, it receives a time-specific rating of three stars; in the next 22.5% a time-specific rating of two stars, and in the bottom 10% a time-specific rating of one star.

Note that if a fund does not have five years or ten years of return history, the time-specific ratings for those time horizons are not calculated. Since all included 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 uses a weighting system that depends upon the age of the fund. For funds with ten years or more of returns (seasoned funds), Morningstar weights the three-year star rating by 20%, the five-year star rating by 30% and the ten-year rating by 50%. For funds with five to under ten years of return data (middle-aged funds), Morningstar weights the three-year star rating by 40% and the five-year star rating by 60%. For funds with under five but at least three years of return data (young funds), Morningstar weights the three-year star rating by 100%.

Morningstar then takes this average number and rounds it up if it has a decimal value of 0.5 or above. For

Exhibit 1
Understanding Morningstar Risk

Month	Fund Return (%)	T-Bill Return	Underperformance
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**

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

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

example, a seasoned fund with 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; the decimal value is 0.5. If the fund had received a 3.4, it would get a three-star overall rating.

It is this second step in the overall star rating calculation that accounts for our motivation. In Morningstar's drive to account for all types of funds, both young and old, and to include as much information as possible, its rating system requires younger funds to be rated on less information than older funds. As there is less information used in rating younger funds, there will likely be more estimation risk in their estimates.

DATA

To illustrate our case, we use data from the January 2001 Morningstar Principia Data Disk, which at the time of writing was the latest available disk. With this disk we collect data on all the funds in Morningstar's International Equity fund category that have received a Morningstar overall star rating. Hence this sample represents an entire category of funds.[5]

Exhibit 2 presents summary data on the sample. Of the 1,281 total, there are 508 young funds, 619 middle-aged funds, and 154 seasoned funds. There are fewer seasoned funds because international equity funds tend to be relatively new funds.

Exhibit 2 illustrates the Morningstar rating methodology we have described. It shows that there are 1,281 funds that receive the three-year time-specific rating (all the funds in the sample, as each fund must have three years of historical returns). Of these funds, approximately 10% receive five-star ratings for the three-year time-specific star ratings; approximately 22.5% receive four-star ratings for the three-year time-specific ratings; approximately 35% receive three-star ratings for the three-year time-specific ratings; approximately 22.5% receive two-star ratings for the three-year time-specific ratings; and approximately 10% receive one-star ratings. The same breakdown takes place for the five-year and ten-year time-specific ratings. The overall ratings are then calculated using the weights and rounding procedure as described.

We then collect from the Morningstar CD-ROM the excess non-load-adjusted monthly returns for each fund for the last ten years (January 1991 through December 2000) or for the history of the fund if the fund has under ten years of historical data.[6] As we have noted, Morningstar uses load-adjusted returns to calculate its star rating, but provides only the non-load-adjusted returns on the CD-ROM disks.[7]

A METHODOLOGY FOR CALCULATING ESTIMATION RISK

Calculating the estimation risk is certainly not straightforward. Moreover, the use of discrete stars for the ratings limits the range from which we determine confidence intervals. As a result, we use a methodology that does not directly determine the estimation risk in the star ratings themselves, but rather calculates the estimation risk in the estimates that Morningstar uses to calculate the star ratings.

While this is somewhat different from the actual estimation risk of the star ratings, we are able to calculate the estimation risk in the *estimates* that are then used to calculate the ratings. In this way, we provide an indirect method of determining the rating estimation risk.

First, using the non-load-adjusted monthly returns, we calculate our own Morningstar return number for each fund for each of the three time horizons. We calculate only the three-year horizon for young funds and

EXHIBIT 2
Characteristics of International Equity Fund Sample

Ratings

	3-Year Rating	5-Year Rating	10-Year Rating	Overall Rating
# of 5-star funds	128	77	15	129
# of 4-star funds	288	174	35	306
# of 3-star funds	448	270	53	454
# of 2-star funds	288	174	35	282
# of 1-star funds	129	78	16	110
Total Number of Funds that receive the rating	1281	773	154	1281

Age

Age of Funds	
Number of Young Funds (at least 3 and under 5 years of data)	508
Number of Middle-Aged Funds (at least 5 years and under 10 years of data)	619
Number of Seasoned Funds (at least 10 years of data)	154
Total Number of Funds	1281

Source: January 2001 disk.

only the three-year and five-year horizons for middle-aged funds.

Our Morningstar return numbers vary slightly from what Morningstar provides in two important ways. First, our data are based on excess *non-load-adjusted* returns, while Morningstar's are based on excess *load-adjusted* returns. Second, for each time period we use the average of monthly returns, while Morningstar uses the compound return.

For example, to compute a fund's three-year Morningstar return, we divide its three-year average monthly excess return by the higher of the three-year average monthly excess return for the category of funds *or* the three-year average monthly return for the 90-day T-bill. Morningstar would calculate the three-year return by dividing the three-year excess compounded return of the fund by the higher of the average three-year excess compounded return for the category of funds or the three-year 90-day T-bill compounded return. We use the average monthly return rather than the compounded return because that gives us a distribution of monthly returns to work with. The compounded return is just a product of returns.[8]

While this difference is notable, we show in an appendix that the two methods are close to each other, so our method does not change our overall conclusions about the estimation risk differences between different ages of funds. Thus we define modified Morningstar return as $\overline{y}$.[9]

The Morningstar risk is then computed from the same non-load-adjusted data as the Morningstar return. As with the Morningstar return, we compute the Morningstar risk for each time horizon for which there are data. We define the Morningstar risk as $\overline{z}$.[10]

The Morningstar time-specific star rating is then based on criterion $C = \overline{y} - \overline{z}$. Remember again that Morningstar simply rank-orders all funds according to its estimates of C, and then allocates the time-specific star ratings in fixed percentages as described.

Our next task is to find the confidence interval for $C = \overline{y} - \overline{z}$. While the technical statistical issues regarding the sampling distributions of $\overline{y}$ and $\overline{z}$ are discussed in the appendix, we can explain here the basic method of calculating the confidence interval. First of all, for our modified Morningstar return $\overline{y}$, we use the central limit theorem (CLT) of elementary statistics. Roughly speaking, the CLT states that the sample mean from any distribution is normally distributed, especially if the sample size n is sufficiently large ($n \geq 30$). It is clear that $\overline{y}$ is normally distributed, as we have at least 36 observations for the three-year period. Although we have simplified matters by considering an average of simple returns rather than compound ones, the appendix provides an example to

EXHIBIT 3

Weighted 95% Confidence Interval Widths Organized by Age—Dummy Variable Analysis

Age of Fund	Number of Funds	Average Weighted Confidence Interval
Young	508	11.36
Middle-Aged	619	9.28
Seasoned	154	7.06

EXHIBIT 4

Confidence Interval Widths Using Dummy Variable Regressions

Equation (A): $CI_i = \gamma_0 + \gamma_1 DMA_i + \gamma_2 DOLD_i + u_i$ (Comparison Group is Young Funds)

Equation (B): $CI_i = \gamma_0 + \gamma_1 DMA_i + \gamma_2 DYOUNG_i + u_i$ (Comparison Group is Seasoned Funds)

where:

CI_i = 95% weighted confidence interval width of i-th fund;
DMA_i = dummy variable for middle-aged funds (middle-aged funds receive 1, and other funds receive 0);
$DOLD_i$ = dummy variable for seasoned funds (seasoned funds receive 1, and other funds receive 0); and
$DYOUNG_i$ = dummy variable for young funds (young funds receive 1, and other funds receive 0).

Equation Examined	γ_0 (constant)	γ_1 (Middle-Aged)	γ_2 (Seasoned)	γ_2 (Young)
A	11.36***	-2.07***	-4.30***	NA
B	7.06***	2.23***	NA	4.30***

*** *Indicates the difference between dummy group and reference group is significantly different at the 5% level.*

show that both the average return and standard error by the two methods (simple or compound) are generally close to each other.

Now consider the sampling distribution of Morningstar risk, $\bar{z}$. It is an average monthly underperformance statistic that receives a value only when the fund underperforms. This is akin to a truncated distribution. It has been known since the 1940s that the sampling distribution of sums of truncated normals involves several integrals and is generally not normal. The appendix explains through a simulation that one can still approximate it by a normal distribution for our sample sizes and refine this in future work.

Thus, C is a difference of two possibly correlated approximately normal variables. Since the Morningstar return $\bar{y}$ may well be correlated with the Morningstar risk, $\bar{z}$, we consider a multivariate normal (MVN) random variable, instead of the usual univariate random variable. Note that the correlation makes the off-diagonal term of the variance-covariance matrix non-zero.

The simplest way to handle the MVN is with the use of vectors and matrices. Accordingly, let us define two 2×1 column vectors, $x = \{\bar{y}, \bar{z}\}$ and $a = \{1, -1\}$. We can thus verify that the Morningstar criterion $C = \bar{y} - \bar{z} = a'x$, where the prime denotes a transpose of the vector. As a result of the prime, a' becomes a 1×2 row vector.

From the properties of the normal distribution, if x is approximately MVN with corresponding true values of $\xi = \{\bar{\eta}, \bar{\vartheta}\}$, as the 2×1 vector of true unknown means, and V as the 2×2 variance-covariance matrix. Now, we write $x \sim MVN(\xi, V)$ as a good approximation.

From the properties of the MVN we can write the Morningstar criterion:

$$C = \bar{y} - \bar{z} = a'x \sim N(a'\xi, a'Va) \quad (2)$$

where dimensions of both $a'\xi$, and $a'Va$ are 1×1, which are scalars (not vectors or matrices), and the MVN becomes a simple univariate normal N(.,.). In practice, sample data are used to compute the *estimate* of V (the covariance matrix of the MVN) denoted by $\hat{V}$. This means that we need to use the Student t distribution instead of the normal distribution.

Since the variance of C is the quadratic form $a'\hat{V}a$, the standard error is its square root: SE = $[a'\hat{V}a]^{0.5}$. Now, we use the Student t distribution to compute the $100(1-\alpha)\%$ confidence interval by using the value from t-tables. Usually $\alpha = 0.05$, and we consider a two-sided interval leaving $\alpha/2$ on each side. Denote by df = $n - 1$ the degrees of freedom, and denote the t-table values as $t_{\alpha/2,\mathrm{df}}$.

Now, the desired confidence interval is given as:

$$\text{Confidence Interval of } C(\tau) = \{C - t_{\alpha/2,\mathrm{df}}\,\mathrm{SE},\ C + t_{\alpha/2,\mathrm{df}}\,\mathrm{SE}\} \qquad (3)$$

where we have introduced (τ), with $\tau = 3$, 5, and 10, to represent the three-year, five-year, and 10-year time horizons.

So far we have constructed a confidence interval on C, the measure that Morningstar uses to calculate the time-specific star ratings. Simply because of the larger sample sizes ($n = 120$) used in calculating $C(10)$, the confidence intervals of the estimate of $C(10)$ will generally be much narrower than the confidence intervals of the estimates of $C(3)$ with $n = 36$ and $C(5)$ with $n = 60$.

The next question is how to approximate the estimation risk in the overall rating. To do this, we calculate "a weighted confidence interval," where the weights are based on Morningstar's weightings. For seasoned funds, we take a weighted average of the width of confidence interval of $C(3)$, the confidence interval of $C(5)$, and the confidence interval of $C(10)$, where the weights are 0.2, 0.3, and 0.5 for the three τ values. For middle-aged funds, we take the weighted average of the width of confidence interval of $C(3)$ and the confidence interval of $C(5)$ using weights 0.4 and 0.6 for the two τ values. For young funds, we take the confidence interval width for $C(3)$.

Again we need to be cautious about the interpretation of our approach. This weighted confidence interval width is again not a confidence interval width on the overall rating itself. Rather it is a weighted average measure of the estimation risk *in the numbers* that are used to calculate the overall rating. It provides only an approximation of the extent of estimation risk in the actual ratings.

EXAMPLE

To illustrate our estimation risk procedure, we show calculation of the weighted confidence interval of one fund, 59 Wall Street Pacific Basin Equity, the first seasoned fund (alphabetically) in the sample. In January 2001, this seasoned fund received a three-year time-specific star rating of four stars, a five-year time-specific star rating of two stars, and a ten-year time-specific star rating of three stars. Given the 20%, 30%, and 50% weightings on the time-specific ratings, this fund received an overall star rating of three stars: 4 stars(0.20) + 2 stars(0.30) + 3 stars(0.50) = 3.1, or three stars.

Using the non-load-adjusted monthly returns, we then calculate our own values of $C(3)$, $C(5)$, and $C(10)$. Again, it is these values that Morningstar uses to calculate the time-specific star rating. We find the weighted width from three 95% confidence intervals as follows: The confidence interval width for $C(3)$ is equal to 12.03, of $C(5)$ equal to 8.44, and of $C(10)$ equal to 5.77. To form the weighted confidence interval width, we then apply the weights to the confidence interval width of $C(3)$, $C(5)$, and $C(10)$. This produces a value of 0.2(12.03) + 0.3(8.44) + 0.5(5.77) = 7.82. Hence the weighted width of a 95% confidence interval for this fund was 7.82.

RESULTS—RELATIONSHIP BETWEEN FUND AGE AND ESTIMATION RISK

Exhibits 3 and 4 show the relationship between fund age and the weighted 95% confidence interval width using the sample of 1,281 international funds. Exhibit 3 uses simple averages, and Exhibit 4 uses a dummy variable regression where the dummies represent different ages of funds and the reference groups are young funds (Equation (A) in Exhibit 4) or seasoned funds (Equation (B) of Exhibit 4).

The coefficients in Exhibit 4 show the mean difference in the weighted confidence intervals between two age groups of funds. For example, the negative and significant coefficient for γ_1 in Equation (A) signifies that that the average weighted confidence interval width of middle-aged funds is significantly lower than that of young funds.

The results in Exhibits 3 and 4 show that as age increases, the average width of the weighted confidence interval falls dramatically. Exhibit 3 shows that the average width of the confidence interval falls drastically for an older fund, while Exhibit 4 illustrates that there is a significant reduction in the weighted confidence interval width as we move from younger to older funds.

As a result of a methodology that rates younger and older funds in one rating system, and in its the drive to use as much data as possible, Morningstar causes the estimates from which the ratings on older funds are derived to have significantly less estimation risk than the estimates from which the ratings of young funds are derived.

The end result is that investors can be somewhat less confident that the ratings of young funds are truly as they are given. To understand this better, consider a young fund and a seasoned fund in our sample of 1,281 funds. The young fund's overall ratings are equal to its three-year time-specific star rating. The three-year time-specific star rating is then based on the fund's estimate of $C(3)$ in comparison to the other 1,280 estimates of $C(3)$. Since all these estimates of $C(3)$ are based on the same amount of data, i.e., 36 months of return data, the estimation risk is probably somewhat similar across all funds.

Now consider the seasoned fund. Its overall rating is based on the weighted combination of the three-year, five-year, and ten-year time-specific star ratings. While a portion of the overall rating (20%) is based on the three-year time-specific rating, 80% of the overall rating is based on the five-year and ten-year time-specific ratings. Simply because these time-specific ratings are based on estimates of $C(5)$ and $C(10)$, which require more data, the estimates will have less estimation risk.

That means that if a young fund receives an overall rating of, say, four stars based on its value of $C(3)$, it may be the case that the confidence interval on $C(3)$ is so wide that the true value of $C(3)$ could range from a value that would give the fund five stars to a value that would give the fund only two stars. We can indeed be more confident that a seasoned fund that receives four stars for the overall rating truly merits four stars (by the methodology), even after considering estimation risk. This again is because the values on which the five-year and ten-year time-specific ratings are based will have much less estimation risk.

CONCLUSIONS

Since the pioneering work of Kolmogorov, the Russian mathematician of the late nineteenth century, statisticians regard any time series as a random realization of an underlying true stochastic process. Accordingly, observed values of sample characteristics (e.g., average returns, volatility of returns, risk measures) are random estimates of corresponding true values and are subject to estimation risk. Note that estimation risk should not be confused with risk arising from volatility of asset returns or data errors. After all, the estimated volatility or any other risk measure itself is subject to estimation risk.

Yet little attention has been given to estimation risk in mutual fund ratings. To remedy this oversight, we investigate estimation risk in the popular Morningstar mutual fund rating system. We show that the distinct estimation risks in the ratings associated with a fund are easily calculated. Our results then indicate that because it uses a system that rates funds regardless of age differences, Morningstar in fact creates uncertainty; investors should be somewhat less confident that the ratings of young funds are truly what they are estimated to be.

Our findings have a number of implications. First, the results imply that investors should heed Morningstar's own advice that the ratings are not signals of future performance. The fact that there is significantly more estimation risk in the estimates used to form younger funds' ratings than in the estimates used to form seasoned funds' ratings implies that funds are not treated equally in terms of calculation of ratings.

Such results combined with those of Morey [2001] that funds may be rated differently not because of performance per se, but rather because of the age of the fund, and those of Blake and Morey [2000] that the overall star rating does not have much ability in predicting winning funds, provide investors with ample evidence that they need to look beyond the ratings when deciding on the funds to invest in.[11]

Second, our results suggest that mutual fund rating services may want to use common horizons to evaluate funds. Systems that weight time horizons depending upon the age of the fund, such as those employed in the Morningstar overall star ratings, can lead to systematic differences that can cause a ratings methodology to vary across different funds.

To its credit, Morningstar itself has tried to deal with this situation in at least two ways. First, it has put more emphasis on the time-specific ratings that compare all funds with the same time history. That is, these time-specific ratings data are now provided, while before only the overall Morningstar rating data were provided. Second, Morningstar has developed a "category rating" that uses a somewhat different methodology from the overall rating, but most important bases all ratings on three-year historical returns.

Finally, if mutual fund rating services want to use different ages of funds in the same rating system, they may want to consider estimation risk as a category on which to rate funds. Indeed, one easy method would be to rank funds on the lower limit of the confidence interval. Using Morningstar's methodology, one could construct ratings that are based the lower limit of the weighted 95% confidence interval. This would improve or complement the current Morningstar ratings.

APPENDIX

Morningstar Risk and Return: Sampling Distributions and Possible Extensions

The appendix describes the asymptotic sampling distributions of Morningstar return and an approximation of the exact sampling distribution of Morningstar risk from a statistical theory viewpoint. For a clearer exposition, the derivations ignore load adjustments or scale factors that cause notational clutter. All computations use the GAUSS package.

Morningstar Return

If the percent return in month i is denoted by 100 times R_i and there are n months in the data set, Morningstar return is not a simple arithmetic average of monthly returns: $\bar{y} \approx (R_1 + R_2 + \ldots + R_n)/n$. Rather, it is based on the product of one plus monthly returns: $B = (1 + R_1)(1 + R_2) \ldots (1 + R_n)$. The average compound return is simply the return R_{avcm} that solves the non-linear equation: $(1 + R_{avcm})^n = B$.

Now consider the natural logarithm (ln) of both sides to yield:

$$n\, ln(1 + R_{avcm}) = ln(1 + R_1) + ln(1 + R_2) + \ldots + ln(1 + R_n) = W_1 + W_2 + \ldots + W_n$$

which defines the notation for W_i, and denote the average by ϖ. Next, divide both sides by n to yield:

$$R_{avcm} = \exp[(W_1 + W_2 + \ldots + W_n)/n] - 1 = \exp(\varpi) - 1$$

This shows that the average compound return is a differentiable function $g(\varpi)$ of the sample mean ϖ of the W. Since ϖ is asymptotically normal by the central limit theorem (CLT), we use the Slutsky theorem to show that $R_{avcm} = g(\varpi)$ is also asymptotically normal.

The important point is that the simple asymptotics of $\bar{y}$ in the text remain valid, in a slightly modified form (see Mittelhammer [1996, Ch. 5]). Thus the standard error can be readily estimated.

As a check, we randomly select one mutual fund (No. 60). It has 120 data points, and its average return by the method described in the text is: $\bar{y} = 0.0064747500$. The standard error, $SE(\bar{y}) = 0.0039428310$. Using the formula for average compound returns, $R_{avcm} = 0.0055427074$, and $SE(R_{avcm}) = 0.0039837729$ for this fund. Note that if we round to three places, $\bar{y} = R_{avcm} = 0.006$ and $SE(\bar{y}) = SE(R_{avcm}) = 0.004$ hold.

Although all funds cannot have similarly close results, the differences are likely to be small and unlikely to be systematic. Using R_{avcm} instead of $\bar{y}$ in the text to represent Morningstar return will make all computations burdensome. We do not expect any differences to reverse our conclusions.

Morningstar Risk

The probability distribution of returns $f(R)$ is generally not known. Without loss of generality and for simplicity, let us initially assume that the $f(R)$, after some standardization, is unit normal, N(0, 1).

Now focusing on only the money-losing months, as Morningstar does, is akin to considering a truncated $f(R)$, i.e., deleting all money-making positive realizations of N(0, 1). Francis [1946, pp. 223-232] proves that the sampling distribution of the *sum* of truncated $f(R)$ will still have n times the mean and n times the variance of the truncated $f(R)$. Accordingly, with analytically known first two moments, one can use Cornish-Fisher normalizing transformation to achieve closeness to normality.

We use a simulation study to assess the normal approximation with the number of months as sample sizes: $n = 36, 37, \ldots, 120$. For each n, we simulate 1,000 truncated sample means to assess the properties of the true sampling distribution for each n. Francis [1946] proves that the sampling distribution is analytically known for $n = 1, 2$, and that it involves numerical integrations for larger n. The object of simulation is to check whether a normal approximation is reasonable for the n values relevant for us.

We use the Bera-Jarque two degrees of freedom (BJ2) χ^2 test of normality, even though we recognize that skewness and kurtosis parameters are difficult to estimate reliably. If we set the Type I error $\alpha = 0.01, 0.05$ the χ^2 table yields the critical values 9.2103, 5.9915.

We simulate all 85 sample sizes in the [36, 120] interval, to yield 85 BJ2 χ^2 statistics. The average of BJ2 values over 85 realizations is 7.9251587 (median = 6.1165398), both not exceeding the 1% critical value 9.2103. Since the mean is larger than the median, the distribution of BJ2 is skewed to the left, suggesting that many BJ2 values are low, implying that the sampling distribution of Morningstar risk is approximately normal.

Unfortunately, at the usual 5% level it is not normal. A larger n does not yield a monotonically smaller BJ2. That is, we do not always have greater closeness to normality as n increases. This may be because of 1) sampling variation or 2) unreliable estimates of higher moments in the BJ2 statistic, or 3) that the true (not asymptotic) sampling distribution may exhibit distinct behavior for different n values.

If we use Cornish-Fisher normalizing transformation, we find that a similar average of 85 BJ2 values is 5.1900123 (median = 3.9167191). Since these are less than the critical value for $\alpha = 0.05$, we recommend using Cornish-Fisher methods to improve the approximation. For more general $f(R)$, a bootstrap uses the empirical distribution functions, and a double bootstrap provides further flexibility (see Vinod [1995, pp. 287-302]). Such refinements may be unappealing in this present context, and are left for future work.

We interpret the simulation to suggest that the refinements are arguably unnecessary for the intuitively straightforward point we make: that young funds have less data and therefore higher estimation risk than older funds.

ENDNOTES

[1]Khorana and Nelling [1998] and Blake and Morey [2000] have both investigated the out-of-sample performance of the Morningstar ratings.

[2]Blume [1998, pp. 4-5] provides an excellent description of the Morningstar accounting for loads. Assume L is the load adjustment. If there is no load of any type, L is equal to 1. If there is a load, L is less than one; e.g., a 4% 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 lengthens. Blake and Morey [2000] also explain the load adjustment in detail.

[3]Note that the excess return is a compound return. Hence the three-year return is the three-year compound return rather than the three-year average monthly return.

[4]"Principia Manual" [1998].

[5]We chose this category because it has significantly fewer funds than Domestic Equity: 1,281 funds compared to 4,164.

[6]The 90-day T-bill rate is used to calculate the excess returns.

[7]The returns provided do account for management, administrative, and 12b-1 fees and other expenses automatically taken out of fund assets. That is, the monthly returns are not adjusted for loads but are adjusted for expenses.

[8]Note that Morningstar's description of its return calculation method is not completely clear on this point. In fact it does not use the term, compound returns, in its definition. We thank an anonymous referee for noting this point.

[9]We use $\bar{y}$ since the Morningstar return is essentially a mean divided by a constant.

[10]We use $\bar{z}$ since the Morningstar risk is essentially a mean divided by a constant.

[11]Specifically, Blake and Morey [2000] examine the ratings on annual Morningstar disks from 1992 through 1997, and then examine the out-of-sample performance of the funds. Using many different samples, and different methodologies, they find that highly rated funds (funds receiving four and five stars) do no better than median-rated funds (three-star funds) in terms of out-of-sample performance. They also find that low-rated funds (one- and two-star funds) have worse out-of-sample performance than higher-rated funds. Khorana and Nelling [1998] examine this same issue.

REFERENCES

Blake, Christopher, and Matthew R. Morey. "Morningstar Ratings and Mutual Fund Performance." *Journal of Financial and Quantitative Analysis*, September 2000, pp. 451-481.

Blume, Marshall. "An Anatomy of Morningstar Ratings." *Financial Analysts Journal*, March/April 1998, pp. 19-27.

Damato, Karen. "Morningstar Edges Toward One-Year Ratings." *The Wall Street Journal,* April 5, 1996, p. C1.

Del Guercio, Diane, and Paula Tkac. "Star Power: The Effect of Morningstar Ratings on Mutual Fund Flows." Federal Reserve Bank of Atlanta Working Paper No. 2001-15, August 2001.

Francis, V.J. "On the Distribution of the Sum of n Sample Values Drawn from a Truncated Normal Population." *Journal of the Royal Statistical Society*, Ser. B (2) (1946), pp. 223-232.

Khorana, Ajay, and Edward Nelling. "The Determinants and Predictive Ability of Mutual Fund Ratings." *The Journal of Investing*, November 1998, pp. 61-66.

Mittelhammer, R.C. *Mathematical Statistics for Economics and Business*. New York: Springer Verlag, 1996.

Morey, Matthew R. "Morningstar Mutual Fund Star Ratings: What Investors Should Know." Research Dialogue, TIAA-CREF Institute, 2000.

——. "Mutual Fund Age and Morningstar Ratings." *Financial Analysts Journal*, Vol. 58, No. 2 (March/April 2002), pp. 56-63.

Principia, Principia-Plus, and Principia-Pro Manuals, various editions. Chicago, IL: Morningstar Publications, 1991-2000.

Sharpe, William F. "Morningstar Performance Measures." *Financial Analysts Journal*, July/August 1998, pp. 21-33.

Vinod, H.D. "Double Bootstrap for Shrinkage Estimators." *Journal of Econometrics,* 68(2) (1995), pp. 287-302.