# The Ratings Game: What is the Risk in Risk-Adjusted Fund Returns?

Michael Stutzer [1]

Professor of Finance and Director, Burridge Center for Securities Analysis and Valuation

University of Colorado, 419 UCB, Boulder, CO 80309

michael.stutzer@colorado.edu ph. 303 492 4348

ABSTRACT

Several firms use historical returns to rate or rank the performance of mutual funds, for the benefit of individual investors desiring to invest in one or more of them. Because the past will not exactly repeat itself, such ratings reflect different notions of "risk" faced by investors. This paper describes and analyzes the performance measures employed by some advisory firms. Approximations are derived to facilitate understanding and comparison of these measures. Simple modifications are proposed to transform these measures into a fund performance measure that is directly related to the fund's probability of outperforming a user-selected benchmark's cumulative return. Perhaps the most important problem plaguing all these measures is the difficulty of accurately estimating expected returns from historical averages, even when a high number of years is used. A proposal to use filtering to alleviate this difficulty is explained and discussed.

# 1 Introduction

Investors are stuck at a crossroads. They could follow the direction preferred by many financial economists by investing in a diversified value-weighted portfolio (either through an index fund or exchange-traded fund), or they could follow others in selecting an actively-managed fund from among the thousands that are readily available. The largest index fund (Vanguard 500) had an average annual total return of $-1.65\%$ over the five years ending in October 2002, while the largest actively-managed equity fund (Fidelity Magellan) had an average annual total return $-1.84\%$ over the same period. Let us pray that investors have the wisdom to choose wisely![2]

In an attempt to help investors choose wisely, some advisory firms (e.g. Morningstar and Lipper) publish relative performance ratings of funds. Perhaps you have seen those full page ads in the New York Times or Wall Street Journal, where a fund management firm trumpets one or more of its funds that have obtained the highest (i.e. five star) ratings from Morningstar. If you haven't paid attention to those ratings, plenty of individual investors have: a recent analysis by Del Guercio and Tkac [6] concluded that "Overall, our results indicate that Morningstar ratings have unique power to affect asset flow."

The raw data used for these and other ratings are historical fund returns. But rating firms crunch these numbers differently when producing their respective ratings. Moreover, the number crunching is complex, and its rationale obscure to most anyone other than some financial economists specializing in performance evaluation. Further complicating matters is the belief of many financial economists that the notion and weighting of "risk" be applied to the invested wealth and possibly even human capital of the investor, rather than to just the alternatives under consideration. More specifically, returns would need to be calculated on the investor's total personal wealth including

that from the fund analyzed. The fund's benefits and costs would be evaluated in terms of its effects on total investor wealth. Hence in general, one would need to know–more realistically, make unrealistic assumptions – about the pre-existing composition of investor portfolios. In the absence of these assumptions, ratings firms adopt other criteria which attempt compare the rated funds to each other and to observable benchmark portfolios, e.g. the S&P 500 or its more easily investable counterparts (mutual index funds or exchange-traded index funds). Without more investor-specific knowledge, the best that can be hoped for is that fund ratings will partially contribute to solving investors' problems.

This paper has a similarly modest goal. It is intended to help financial professionals better understand some of the performance rating systems devised to rank a fund's performance *relative to other funds, including observable benchmark portfolios.*. Section 2 provides a canonical comparison of a fund with a benchmark index, that is used throughout the paper. We analyze the (holding period-dependent) probabilities that the cumulative return from investing in one will exceed that of the other. Section 3 provides an overview of some alternative performance ratings systems in use, develops approximations of them that facilitate our understanding of them, and develops simple modifications required to turn them into measures consistent with the outperformance probability desiderata developed in section 2. Section 4 discusses the practical problems facing all performance ratings systems based on historical fund returns. Typical implementations do not overcome these problems, but another possibility is discussed that could. Section 5 concludes.

## 2 A Typical Comparision

Figure 1 contrasts the hypothetical results of investing $1.00 in 1980 in a broad-based index to investing it in a particular managed fund or fund strategy (hereafter dubbed "the fund"), through 2001.[3] We see that the two were basically in a dead heat until 1995, after which the fund generally excelled. On this basis, perhaps the fund should be ranked higher than the index. But Figure 2, which graphs their raw monthly returns, clearly depicts the higher volatility of the fund's monthly returns. The high volatility resulted in particularly bad month during 1998 that virtually wiped out its advantage to that point. Due to this high volatility, that might happen again, possibly dragging the fund's cumulative return back below the index.

Just how much does this higher volatility put the fund investor at-risk of underperforming the index? To investigate this, let us employ a popular statistical tool known as *bootstrapping*. While we know that the exact sequence of past monthly returns will not repeat itself, we think of the return numbers themselves as indicative of what could happen in the future. That is, we randomly sample months (with replacement) from the (258) months for which we have historical returns, and string the fund's returns during those months together, representing a hypothetical future for the fund's returns. The same months are used to construct a hypothetical future for the index returns, and the two hypothetical futures contrasted to see if the fund still outperforms the index over the number of months sampled.[4] Constructing 10,000 hypothetical 120 month future periods in this way, one finds that the fund failed to beat the index in only 35% of those hypothetical futures, i.e. there is a 65% chance (i.e. almost 2:1 odds) that the fund will beat the benchmark after a 10 year holding period.

But it is also important to examine what happens over shorter investment horizons, which

are perhaps more relevant for investors who are a bit jumpy and/or nearing retirement. Using the same method to simulate 10,000 hypothetical 5 year future holding periods, it turns out that the fund still underperforms the index 38% of the time. So there is still a 62% chance (almost 2:1 odds) that the fund will outperform the index over a 5 year holding period. But note that the 38% underperformance probability over 5 year future scenarios is slightly higher than the 35% underperformance probability over 10 year scenarios. This pattern continues to hold when analyzing even shorter holding periods of 3 years and 1 year. In the latter case, an investor who puts funds on a really short 1 year leash will face an an even higher, 43% chance that the fund will underperform the index over the next year. Of course, this still leaves a 57% chance (about 4:3 odds) of beating the index over that year. In summary, the fund's underperformance probabilities decrease (i.e. improve) as the holding period grows longer – a type of *time diversification* – but non-negligable underperformance probabilities persist over surprisingly long holding periods. This is illustrated by the very slow decrease in bar heights at the left of Figure 3.[5] This typifies funds that have historically outperformed less volatile, more diversified benchmarks (e.g. highly diversified index funds).

Now let us conduct similar analyses of both the fund and index' probabilities of underperforming a cumulative investment in one-month T-Bills, a benchmark that has generally lower and certainly less volatile returns than both the fund and the index. Figure 4 shows that the index has underperformance probabilities that decay more rapidly than the fund's do as the holding period lengthens. Even after a 10 year holding period, there is still a better than 20% chance that the fund will underperform a cumulative investment in one-month T-Bills, due to the slow decay of its underperformance probabilities. After seeing Figure 4, someone who is primarily concerned with beating a one-month T-Bill benchmark would probably rank the index over the fund, because

its lower probabilities of underperforming (the T-Bill benchmark) imply higher *out*performance probabilities.

Valid generalizations from considering Figures 3 and 4 are summarized in the following proposition:

**Proposition 1**: *A fund that achieved a higher historical cumulative return than a benchmark may have significant probabilities of underperforming that benchmark in the future, even though its underperformance probabilities may steadily decay toward zero as the holding period lengthens. The problem is exacerbated in funds that have high volatility relative to their average historical return. Those primarily interested in beating the benchmark will prefer funds whose underperformance probabilities decay most rapidly toward zero as the holding period increases.*

## 2.1 Absolute or Relative Performance?

While the bootstrap analysis contrasts the possible future *relative* performance of the fund and index, it can (with somewhat more reservations) be used to examine the possible future performance of the fund itself. This is a more problematic exercise, because of the possibly atypical high growth of both fund and index during the 1980-2001 period exhibited in Figure 1. A separate bootstrap analysis of the fund will not reflect this. But if the fairly tight historical statistical connection ( e.g. relatively high correlation coefficient of 87%) between the fund and index returns illustrated in Figure 2 continues to hold, the *relative* future cumulative return analysis conducted above is less problematic. [6]

With that caveat, the underperformance probabilities for the possible 10 year future cumulative fund returns is depicted in Figure 5. The figure shows that there is still a better than even chance

that 1 dollar invested in the fund will grow to 5 dollars after just 10 years, reflecting the bootstrap's projection of the fund's generally high average returns over 1980-2001. Readers may find these projections to be implausible, yet they are not just due to the bootstrap methodology. They are implicit in any performance analysis of a fund in isolation, rather than its performance *relative* to a fairly closely related benchmark, when that analysis is based solely on historical returns.

# 3 Risk-Adjusted Performance Measures

Some performance measures used by advisory firms are now described. When needed, an approximation of each performance measure is constructed to show how it adjusts for the standard textbook notion of "volatility risk". In accord with Proposition 1, each performance measure is simply modified to provide a performance measure that is directly related to the underperformance probability's rate of decay to zero, i.e. that is adjusted for the underperformance risk defined in section 2.

## 3.1 The Sharpe and Information Ratios

Perhaps the best known performance measure is the *Sharpe Ratio* [18]. It is used by some advisory firms, e.g. in the quantitative components of both Standard and Poors "SelectFunds" and Charles Schwab's "Select List" rating systems.[7] It is also widely used outside of fund advisory firms, e.g. in the Hulbert Financial Digest's performance ratings of newsletters' investment strategies. Its definition depends on unobservable quantities: it is the ratio of the mathematical expectation of a fund's returns in excess of a "riskfree" rate, divided by the standard deviation of this excess return.

But these and other determinants of the future distribution of relative performance are never known with certainty. If they could be known with certainty, they could be used to perform a

different bootstrap analysis that would produce the probabilities of possible future performance levels with higher accuracy than that produced from raw historical returns in section 2. Instead, advisory firms typically use the past return histories to estimate the required unobservable numbers. The Sharpe Ratio (dubbed "SR") performance measure is most commonly estimated by calculating the historical average monthly fund return in excess of a one-month T-Bill's return (this is the same as the difference between the average fund and T-Bill returns), and dividing by the standard deviation of these historical excess returns.[8] The fund's average monthly return over the 1980-2001 period was 1.99% per month, well above the index' 1.21% per month. The standard deviation of the managed fund's returns over the same period was 10.83% per month, well in excess of the index' 4.52% per month. Subtracting the one-month T-Bill returns from both fund and index returns produces lower *excess* returns, used in the comparison below:

$$SR^{fund} = \frac{\frac{1}{T}\sum_{t=1}^{T} R_t^{fund} - \frac{1}{T}\sum_{t=1}^{T} R_t^{bill}}{StdDev^{fund-bill}} \quad (1)$$

$$= \frac{1.99\% - 0.54\%}{10.83\%} = 13.4\%$$

$$SR^{index} = \frac{\frac{1}{T}\sum_{t=1}^{T} R_t^{index} - R_t^{bill}}{StdDev^{index-bill}} \quad (2)$$

$$= \frac{1.21\% - 0.54\%}{4.52\%} = 14.9\%$$

where $T$ denotes the number of historical monthly returns in the 1980-2001 period used in section 2.[9]

So according to the estimated Sharpe Ratio comparison (1) - (2), the fund is ranked *lower* than the index, despite Figure 3, which indicates that the more volatile fund has a better than even chance of outperforming the broad stock index (i.e. a less than half chance of underperforming it) over all horizons examined, with the risk of underperforming it declining as the holding period

lengthened. Both Figure 3 and (1)- (2) were dependent solely on the historical data between 1980-2001, so the disconnection between the outperformance probability ranking (i.e fund beats index) and the Sharpe Ratio ranking (index beats fund) must be attributable to something else. The Sharpe Ratio is adjusting for some sort of risk, but evidently not the risk of underperforming the index that is analyzed in section 2.

One of the reasons for the conflicting rankings is that the "benchmark" used in the Sharpe Ratio is a one-month T-Bill, rather than the broad stock index. Substituting the stock index return $R^{index}$ for $R^{bill}$ in (1) results in a different measure, commonly referred to as an *Information Ratio* [9] (dubbed "IR" below) with an index benchmark. The Sharpe Ratio is the key performance index in the textbook mean-variance portfolio theory (in the presence of a riskless asset, as we will discuss later in section 4), while the Information Ratio is the key performance index in the tracking error variance (TEV) theory of Roll [17]. In this case, use of the Information Ratio yields:

$$IR = \frac{\frac{1}{T}\sum_{t=1}^{T} R_t^{fund} - \frac{1}{T}\sum_{t=1}^{T} R_t^{index}}{StdDev^{fund-index}} = \frac{1.99\% - 1.21\%}{7.24\%} = 10.8\% > 0 \tag{3}$$

The positive Information Ratio (3) implies that the fund should be ranked above the index, consistent with section 2's outperformance probability analysis. The agreement of the two is not guaranteed, but it isn't completely accidental. A simple modification of the Information Ratio, which replaces the net returns denoted $R$ in (3) with continuously compounded returns $\log 1 + R$, provides an estimated *Log-Modified Information Ratio* (dubbed "LIR"). When positive, the Appendix shows that the LIR usually implies that the fund's underperformance probabilities decay to zero over time, at a rate that is directly related to the size of the LIR. Hence the size of the positive LIR is usually directly related to the drop in bar heights on the left of figure 3. To produce the LIR, we replace each monthly net return $R$ in (3) with its continuously compounded counterpart

$\log 1+R$ before computing it, yielding

$$LIR = \frac{1.35\% - 1.10\%}{7.83\%} = 3.2\% > 0. \tag{4}$$

which is smaller than (3) but still positive, again favoring the fund over the index. Comparing the LIR (4) to the ordinary IR (3), we see that the relatively small size of (4) is heavily influenced by the fund's historical average *log gross* return (1.4%) being much smaller than its ordinary historical average return (1.99%). This is due to the fund's relatively high volatility, which causes the average log gross return to be much smaller than the average return itself. The small size of the LIR (4) implies that the fund's underperformance probabilities will decay slowly as the holding period lengthens, as seen in the slow decay of bar heights on the left of Figure 3.[10]

This property, developed in the Appendix, is summarized below:

**Proposition 2:** *Computing the Information Ratio (3) using the logarithms of fund and index gross returns, i.e. using continuously compounded returns, is a historical estimate of a performance measure (dubbed LIR) that when positive, is usually directly related to the probability that the fund's cumulative return will exceed that of the index. The higher the positive value of LIR, the faster the fund's underperformance probabilities decay toward zero.*

How can Proposition 2 be used to interpret the conventional Sharpe Ratio (1), which shows that the fund's historically estimated Sharpe Ratio is *lower* than the Sharpe Ratio of the index? Substitution of log gross (i.e. continuously compounded) returns $\log 1+R$ for the net returns $R$ in (1) and (2) produces the following log-modified Sharpe Ratio (LSR) comparison:

$$LSR^{fund} = 7.1\% < 12.4\% = LSR^{index}. \tag{5}$$

The fund and the index both have positive historically estimated LSRs. If the positivity of those

estimates is indicative of the the unobserved counterparts that they estimate, Proposition 2 implies that both the fund's and the index' cumulative returns will beat the cumulative *T-Bill* return over long holding periods. The higher positive LSR of the index indicates that its probabilities of underperforming the T-Bill decay to zero at a faster rate than the fund's do as the holding period lengthens. This is seen in Figure 4, which shows the faster decline of the index' probabilities of underperforming T-Bills as the holding period lengthens from 1 to 10 years.

So those who are mainly interested in outperforming a cumulative investment in one-month T-Bills should rank the index above the fund. But this does *not* imply that the index will outperform the *fund*, as verified in Figure 3, because the relevant performance measure for that comparison is the LIR (3), rather than the LSR (5).

### 3.2 Morningstar's Risk Adjusted Return

During 2002, Morningstar, Inc. changed its well-known star ratings procedure in several ways [15]. Funds are now assigned to one of 48 equity and bond fund categories differentiated by size, style, sector, geographic locale, bond term, etc. Within each category, they use a performance measure to rank order the funds in that category against each other. Let us now focus attention on the performance measure they use to rank funds within any particular one of their 48 categories. This Morningstar Risk Adjusted Return (MRAR) performance measure is [15, p.13]

$$MRAR(\gamma) = \left[ \frac{1}{T} \sum_{t=1}^{T} \left( \frac{1 + R_t^{fund}}{1 + R_t^b} \right)^{-\gamma} \right]^{-\frac{12}{\gamma}} - 1 \tag{6}$$

where $1 + R_t^{fund}$ denotes its (gross) monthly total return in month $t$ out of the past $T$ months, $1 + R_t^b$ is an analogous benchmark return. Like the Sharpe Ratio (1), Morningstar uses the one-month T-Bill return as the benchmark return $R^b$, but it will later prove useful to consider the consequences

of using other benchmarks. Similarly, $\gamma$ is a number that Morningstar sets at a constant equal to 2 for all funds, because "Morningstar's U.S. fund analysts have concluded that $\gamma = 2$ results in fund rankings that are consistent with the risk tolerances of typical investors."[15, p.13].

For the purpose of understanding the nature of (4), let us start by ranking our example fund and index over the same historical period of $T = 258$ months used earlier. Substituting the monthly fund and index returns into (4) yields their MRAR(2) ratings:

$$MRAR(2)^{fund} = -9.1\% < 4.3\% = MRAR(2)^{index} \tag{7}$$

which like the Sharpe Ratio comparison (1) - (2), ranks the fund below the index. The ranking would have been reversed had Morningstar used a lower $\gamma$ coefficient, in fact any $\gamma$ lower than 0.42. So its choice of $\gamma$ is critical to the ranking. One could also substitute the index (or another benchmark) return $R^b$ in (6) for Morningstar's T-Bill return , as done when changing the Sharpe Ratio (1) into the Information Ratio (3). But substituting the index for the T-Bill in this example yields a similar conclusion, because

$$\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1+R_t^{fund}}{1+R_t^{index}}\right)^{-\gamma}\right]^{-\frac{12}{\gamma}} - 1 = -5.5\% \tag{8}$$

is still a negative number, indicating that the fund would still be ranked below the index. So the different conclusions reached by the outperformance probability analysis of section 2 and the Morningstar MRAR($\gamma = 2$) performance measure are not just due to the latter's use of the T-Bill as a benchmark instead of a broad-based index.

In order to systematically relate the MRAR($\gamma$) measure to the outperformance probability analysis, it helps to transform the former into a simpler performance measure that produces the

same rank ordering of funds. The Appendix shows that (6) provides the same rank ordering as the following historical estimate of an expected power utility index:

$$U(\gamma) = \frac{1}{T}\sum_{t=1}^{T} - \left(\frac{1+R_t^{fund}}{1+R_t^b}\right)^{-\gamma} \tag{9}$$

The following Proposition 3, derived in the Appendix, provides a useful approximation of the MRAR($\gamma = 2$) performance measure.

**Proposition 3:** *For any constant level of $\gamma > 0$ and benchmark with return $R^b$, Morningstar's Risk-Adjusted Return measure MRAR($\gamma$) in (6) rates one fund higher than another when and only when the expected power utility $U(\gamma)$ in (9) does. A useful approximation of the ranking that is produced by Morningstar's use of the constant $\gamma = 2$ can be obtained by ranking funds in accord with the difference between the average excess continuously compounded return and its variance (i.e.* **squared** *standard deviation).*

In our example, Proposition 3's alternative performance measure for the fund is $0.8\% - (11.5\%)^2 = -0.5\%$, which indeed is lower than the corresponding $0.6\% - (4.6\%)^2 = +0.4\%$ value for the index, and hence in agreement with the MRAR(2) ranking (7). More generally, using the approximation to re-rank the approximately 1300 large growth and large value funds tracked by Morningstar results in an almost identical fund ranking; the rank correlation coefficient is 99.999%.

But how is the Morningstar measure related to the risk of underperformance analyzed in section 2? The following Proposition 4, developed in the Appendix, describes a simple modification of (9) that produces a performance measure consistent with the outperformance probability analysis conducted in section 2.

**Proposition 4:** *Associate each fund with its own different level of $\gamma$, called $\gamma_{max}$, that maxi-*

*mizes the size of its $U(\gamma)$ in (9). If a fund's $\gamma_{max}$ is positive, then the higher the value of $U(\gamma_{max})$, the faster the fund's underperformance probabilities decay toward zero. This is because the underperformance probabilities eventually decay at a rate of $-\log[-U(\gamma_{max})]$, which is higher when $U(\gamma_{max})$ is.*

In our example, the broad-based index was used for the benchmark return $R^b$. Substituting it into the function (9), and using the Excel "Solver" routine to maximize it over $\gamma$, showed that the fund is associated with a maximizing $\gamma_{max} = 0.39$. According to Proposition 4, the fund's underperformance probability will decay to zero as the holding period increases. But $-U(\gamma_{max}) = 0.9995$, so $-\log[.9995] = 0.049\%$ is quite small. So according to Proposition 4, the fund's underperformance probabilities will decay to zero at a very slow rate as the holding period lengthens. As a result, its probabilities of underperforming the index will persist for a very long time, as witnessed by the very slow decrease in bar heights on the left hand side of Figure 3.

### 3.3 Lipper's Preservation Measure

Lipper, Inc. uses two different performance measures when identifying their most preferred funds. They are intended for different investor groups. A preferred fund for "the most risk-averse fund investor" group is based on Lipper's assumption that "Investors perceive risk in terms of the frequency of losing money and also the extent (depth) of losses." [23, p.1].[11] The Lipper Preservation performance measure is the nonpositive number:

$$LP = \sum_{t=1}^{T} \min[0, R_t^{fund}] \tag{10}$$

which is just the sum of the fund's *negative* historical returns. The top 20% of such funds, i.e. the ones whose values of (10) are closest to the maximum possible value of 0, are designated "Lipper Leaders for Preservation." Lipper uses a $T = 36$ month fund history to produce this ranking, even when a fund has a longer history. The pros and cons and doing of so will be discussed in the following section 4. More generally,(10) can be divided by whatever value of $T$ is used, to produce an equivalent ranking measure:

$$ALP = \frac{1}{T}\sum_{t=1}^{T} \min[0, R_t^{fund}] \tag{11}$$

which is (minus one times) the fund's historical *average* loss. The highest possible value of (11) is still 0, which would be attained by one-month T-Bills that always have nonnegative yields. You just can't beat T-Bills for preservation of invested funds, albeit unadjusted for inflation! Applying (11) to the fund and index used in our example yields:

$$ALP^{fund} = -3.1\% < -1.2\% = ALP^{index} \tag{12}$$

so the Lipper Preservation measure would rank the index above the fund, as did the Sharpe Ratio comparison (1)- (2).

To better understand the Lipper Preservation measure, first note that

$$\sum_{t=1}^{T} \min[0, R^{fund}] = \sum_{t:R_t^{fund}=0} 0 + \sum_{t:R_t^{fund}<0} R_t^{fund}. \tag{13}$$

Denoting the number of negative historical returns (i.e. the ones used to form the second term in (13)) by $T^-$, multiply and divide the two terms in (13) by it to obtain the following expression equivalent to (11):

$$ALP = \frac{T^-}{T} \times \frac{\sum_{t:R_t^{fund}<0} R_t^{fund}}{T^-}. \tag{14}$$

Hence the Lipper Preservation measure decreases as either the fraction of months where losses occur increases, or the average value of those months' losses (i.e. minus one times the negative return) increases.

In the example, both the fund and the index earned a negative return (i.e. a loss) in close to 37% of the months between 1980-2001, although they weren't always earned during the same months. Hence the first term in (14) is nearly identical for the fund and index. But the second term, which the average of those negative returns, was much worse for the fund than the index (-8.3% vs. -3.3%), resulting in (12) when the two terms in (14) are multiplied together.

The following proposition, developed in the Appendix, provides a useful approximation of the ranking produced by the Lipper Preservation measure.

**Proposition 5:** *A useful approximation of the ranking that would be produced by the Lipper Preservation measure can be obtained by ranking funds in accord with the difference between their average returns and their return standard deviations.*

Figure 6 provides some graphical evidence of this equivalence, based on the derivation in the Appendix. Note that in our example, Proposition 5's alternative performance measure for the fund is $2.0\% - 10.8\% = -8.8\%$, which indeed is lower than the corresponding $1.2\% - 4.5\% = -3.3\%$ value for the index, and hence in agreement with the Lipper Preservation measure's ranking (see the equivalent ALP ranking (12)).[12] The following quote indicates that this approximation works quite well in practice:

> We have run some tests using our fund performance data and indeed found a close correlation between the Lipper preservation measure with the measure you suggested.[13]

# 4 Implementation Issues

## 4.1 The Use of Historical Returns: Average Returns Are the Culprit

The typical stability of fund rankings to the historical period used can be illustrated by examining the sensitivity of the Sharpe Ratio, which is perhaps the most widely used performance measure in both academia and industry. Figure 7 shows what would have happened had we ranked the fund versus the index via their most recent 10 year past historical Sharpe Ratios, starting in 1991 and continuing to re-rank them in every successive month through 2001. The ranking reversed, i.e. the fund would have been ranked higher than the index, only toward the end of the period. Figure 8 shows the rankings that would have resulted had only the most recent 3 year historical Sharpe Ratios been compared. We see that an earlier reversal in ranking also occurred, lasting throughout 1994. Comparing the vertical axes of the two figures shows that the range of historical Sharpe Ratios is far greater when only 3 years of historical returns are used than when 10 years are used. Because it is a ratio of historical average (excess of T-Bill) returns to historical return standard deviation, it is interesting to investigate which of the two contributes most to the fluctuation of the ratio. For this purpose, a good indicator of fluctation is the average absolute percentage change from month to month. For example, if a number goes up by 10% in one month (say, from 2 to 2.2) and down by 12% over the next month, its average absolute percentage change over the two months is 11%. It turns out that the rolling 3 year average historical returns, i.e. the numerators of the fund's rolling 3 year Sharpe Ratios in Figure 8, had an average absolute percentage change of 65%! But the corresponding rolling standard deviations used in the denominators had an average absolute percentage change of only 2.4%. Note that if both had always gone up or down by the same percentage in every month, the Sharpe Ratios would not have fluctuated at all. One might

think that the problem would be generally eliminated by using 10 years of historical data rather than just 3 years, but the fund's rolling 10 year historical average returns still fluctuate quite a bit relative to the rolling 10 year historical standard deviations; the former had an average absolute percentage change of 9% versus only 0.71% for the latter. In our example, ranking reversals weren't common, but this vast difference indicates that they could be more common in other examples. Moreover, it is well-known that the difficulty of estimating long-run (i.e. expected) returns from historical averages is not eliminated by measuring returns more frequently, e.g. daily or weekly.[13, pp.214-217], although this will generally improve volatility estimates.

Propositions 3 and 5 show that both the Morningstar and Lipper Preservation measures will also be heavily influenced by these problematic historical average returns. The traditional statistical foundation for assessing the accuracy of historical average returns as estimators of the unobserved expected return is the assumption that the longer the historical period used, the more accurate the estimate is, and the more stable rolling averages will be. The instability of historical averages may persist even when an unusually long historical period is used to form them. From this perspective, use of longer historical record of returns is always better.[14] But there has been a huge secular increase in the number of funds, implying that many won't have long return histories. We will examine a potential way to cope with this problem in section 4.2.

Moreover, neither the Sharpe, Morningstar, nor Lipper Preservation measures use benchmarks that are closely correlated with the equity funds that they rank. But simple algebra in the Appendix indicates that the *difference* between a fund's and a benchmark's historical averages will fluctate less about the true difference in means than the separate average of the fund, when the fund and benchmark returns are reasonably positively correlated and the benchmark is no more volatile than the fund.[15] The intuition for this finding comes when one considers a hypothetical fund

whose returns are just a constant $\alpha$ plus its benchmark's returns. The fund's and benchmark's returns are perfectly correlated, and they both have the same volatility. The difference of the two returns is just $\alpha$, which would be discovered using any historical period. This consideration favors adoption of performance measures like the IR (3), and the LIR (4) that is analyzed in Proposition 2, that depend on the average difference of fund returns from a closely related but less volatile benchmark; perhaps an index tailored to the fund's category. The practical consequences of all this are summarized in Proposition 6 below:

**Proposition 6:** *Either explicitly or implicitly, performance measures depend on historical average returns that require a high number of years to stably estimate. As a result, relative fund ratings will fluctuate more when shorter historical periods are used in comparisons, limiting performance measures' practical value to funds with relatively long histories. The problem may be at least partly alleviated by use of performance measures depending explicitly or implicitly on the average difference in returns between a fund and a closely related, less volatile benchmark. This consideration favors adoption of equity (bond) benchmarks when ranking equity (bond) funds, instead of a T-Bill benchmark.*

An illustration of the benefit associated with using a closely related benchmark is seen in our example, when analyzing the ability of the most recent 10 year historical LIR to predict whether or not the fund's subsequent 3 year cumulative return will exceed that of the index. A positive historical LIR is a prediction that the fund's future cumulative return will exceed that of the index, while a negative historical LIR predicts the opposite. These predictions were correct in 80% of the time periods over which our data permits comparison. Using the sign of the most recent 3 year historical LIR, the predictions were correct in only 65% of those time periods, due to the higher

instability of its numerator.

Because Proposition 6 does not support the case for a T-Bill benchmark (except when evaluating money market or short term bond funds), it is important to re-examine the conventional mean-variance portfolio choice results that underlie use of a T-Bill (i.e. riskless asset) benchmark in the Sharpe Ratio. In that theory, investors are just *assumed* to want higher mean wealth and lower wealth variance ("risk"), rather than some more intuitive want (e.g. beating a broad-based benchmark with high probability.) As noted by Fabozzi, Gupta, and Markowitz [7], the mean-variance theory is intended to *better inform* investor's decisions rather than to *explain* their current decisions. We have seen that seemingly different performance measures are all consistent with its prescription for high average returns and low return variance, although the funds are (from this perspective) undesirably ranked in isolation from other assets the investor holds. But why should the two statistics be weighted according to their (Sharpe) ratio, computed relative to its T-Bill benchmark? In this vein, Sharpe [19, p.31] argues that the Sharpe Ratio was developed "for situations in which an investor can use borrowing or lending to achieve his or her desired level of risk". Specifically, they must be able to invest *and borrow* at the riskless T-Bill rate in each period. This assumption is critical to enshrining the maximum Sharpe Ratio portfolio as the best mean-variance efficient portfolio of risky assets. Although bond traders can often borrow at close to T-Bill rates (via repurchase agreements), other less-heavily collateralized investors have an awfully hard time doing so.

Moreover, the conventional case for the Sharpe Ratio requires a less-well understood but equally important cross-correlation assumption, as described in Sharpe's own words:

> When choosing one from a among a group of funds of a particular type for inclusion in

> a larger set of holdings, the one with largest predicted excess return Sharpe Ratio may reasonably be chosen, if it can be assumed that all the funds in the set have similar correlations with the other holdings. If this condition is not met, some account should be taken of the differential levels of such correlations.[18, p.56]

This condition arises to ensure that when the high Sharpe Ratio fund is added to an investor's pre-existing portfolio of other assets, it will again be possible to use riskless investing and/or borrowing to raise the mean return above its pre-existing value, and/or lower the return variance below its pre-existing value. In other words, the above condition is needed to ensure systemic consistency with one-period, mean-variance theory.[16] Conditions like this one would also arise in other portfolio choice theories grounded in the alternative performance measures analyzed here. As noted in the introduction, this paper is primarily concerned with describing these alternative performance measures, and relating them to the probability of beating the benchmark used to construct them. So development of analogous consistency conditions is a good topic in need of future research.

A search of the Bloomberg database failed to uncover even a single published benchmark index for a blended portfolio representative of the stocks, bonds, bills, and bank CDs, and home equity held by representative individuals. In fact, the search failed to uncover a published blended index of just domestic stocks and bonds.[17] This failure provides some evidence that neither fund managers nor investors use benchmarks in a one-step, systemic way. But one must still choose whether or not to use a broad equity (bond) index to evaluate all equity (bond) funds, or to use an index specific to the fund's category. There is no shortage of published indexes narrowly tailored to funds size, style, and sector characteristics; in fact, it should be possible to develop separate benchmark

indices for each of Morningstar's 48 fund categories, or the categories used by Standard and Poors' Select Funds system. [18] Such category-specific indexes are best used by investors who (rightly or wrongly) either do not want to be well-diversified, or who want to become well-diversified by picking a fund or two (or more) from many of the numerous possible categories. They hope that each of their picks will outperform the benchmark in its category, so that their portfolio will outperform a diversified portfolio comprised of the various categories' benchmarks. See Belden and Waring [3] for a debate on the topic of broad versus narrow benchmarks.

### 4.2 Ranking Both Older and Newer Funds Together

The Lipper Preservation measure is based on the past 3 years of historical returns, regardless of whether or not a fund has a longer historical return record. So are the Standard and Poors' SelectFunds ratings as well as the system used by Charles Schwab. Morningstar has a more complex procedure. The performance of funds that have between 3 and 5 year records are ranked over the latest 3 year period, analogous to Lipper's procedure, and assigned a rating between five and one stars based on their relative rankings. They give their coveted five star rating to the top 10% of them, and their feared one star rating to the lowest 10% of them. 22.5% of funds receive four stars, and another 22.5% of funds receive two stars. The middle 35% of funds (i.e. the rest ) receive a middling three stars. But when measuring the performance of funds that have between 5 and 10 year records, 60% weight is given to its star rating based on the latest 5 year period's ranking, while 40% weight is given to its star rating based on the latest 3 year period's ranking. Funds that have more than 10 year records have 50% weight assigned to its 10 year star rating, 30% weight assigned to its 5 year star rating, and 20% weight to its 3 year star rating. The resulting resulting weighted averages produce fractional number of stars that are rounded up or down as appropriate.

The previous section documented the problems created by use of historical averages over 3 year periods. Proposition 5 shows that Lipper's use of just the most recent 3 years of historical returns will not generally be immune to those problems. While Morningstar's weighted average procedure utilizes histories up to 10 years when available, findings in Morey [14] indicate that its weighted averaging procedure is at least partly responsible for the higher ratings associated with the relatively older funds. To understand the phenomenon, consider what will generally happen as time passes as new returns roll in. As documented in the previous section, the new returns will generally have more impact on 3 year rolling average returns than on 10 year rolling average returns; the former are far less stable than the latter. Proposition 3 shows that those rolling averages will play a very important role in the performance measure that Morningstar uses to rank its funds, so a fund's 3 year star rating will likely be less stable than its 10 year star rating (if it is old enough to have one) as new returns roll in. Hence a 3 or 4 year old fund that attains a high overall rating, which is just its 3 year rating, is less likely to see it stay high for long than a fund older than 10 years would, because an older fund's overall rating is only 20% influenced by its equally unstable 3 year rating.

A promising alternative to using either short histories or weighted averages of longer histories was proposed by Stambaugh [20]. He shows that there might be a way around this apparent limit, by substituting *estimates* of returns that the relatively recently started funds would likely have had before they came into being! Before you reject this *filtering* approach out of hand, let us use our example to see how it could work.

In our example, let us make believe that the fund didn't exist for the full 1980-2001 period that the index did. Assume the fund only existed since 1997, so that at the end of 2001, we would only have 5 years of monthly fund returns to work with. While in the land of make believe, let us also assume that after reading sections 2 and 3, an advisory firm decided to adopt the LIR in Proposition

2 as its performance measure. Its numerator is the average of $\log(1+R^{fund})-\log(1+R^{index})$, which is just the difference of the two averages. Over the 5 years, the monthly average of $\log(1+R^{fund})$ was 1.09%, while the average of $\log(1+R^{index})$ was only 0.77% per month, so the difference was 0.31% per month. But after reading section 4.1, the advisory firm decides that there is too much uncertainty in 5 year averages to use this information in its ranking measure. In desperation, one of their analysts plots the 60 monthly observations of $\log(1+R^{fund})$ and $\log(1+R^{index})$ on the same graph, and notices that while the fund series is much more volatile, the two series appear to be highly correlated (see Figure 9). The analyst reports that the correlation coefficient of the two monthly return series is 86%. The analyst thinks that it is reasonable to assume that this high correlation will continue into the future past 2001. Then she had a flash of insight; the kind that marked her for future promotion. She reasoned that *had* the fund existed prior to 1997, its (log gross) returns would still have been highly correlated with the index ( log gross) returns, albeit probably more volatile. She posits a linear relationship between the two series in Figure 9, and uses a spreadsheet linear regression tool to find that the following regression line (t-stats in parenthesis):

$$\log(1+R_t^{fund}) = \overset{(-.464)}{-0.00375} + \overset{(12.73)}{1.90} * \log(1+R_t^{index}) \quad R^2 = 74\% \tag{15}$$

She notes that the constant $-0.004$ has a very low t-stat, and could probably be ignored, but the slope 1.90 cannot. Glancing back at Figure 9, she sees how multiplying the index series by 1.9 will produce a series whose upturns and downtowns are much more severe, just like the fund's are. Even though the regression fit is far from 100% ($R^2 = 74\%$), she decides to plug prior months index log gross returns into (15) in order to "backcast" projections of would-be fund returns those prior months. Grafting those 60 projected log gross fund returns onto the observed most recent 60 returns produces a 10 year monthly series, whose average will be compared with the observed average of

the 10 year index log gross monthly returns. In doing so, she hoped that the benefit of higher stability (using a 10 year average instead of the observed 10 year average) would outweigh the cost associated with projecting fund returns prior to its formal existence. After all, she reasoned, the funds managers certainly existed prior to the fund's existence. Why wouldn't they have managed it during the five prior years in the same way (philosophy, style, etc.) that they managed it during the subsequent five years?

She reported her results to the firm, who already knew that the fund's log gross return averaged 1.09% per month between 1997-2001. What the firm didn't yet have was a reasonable projection of what the fund's average might have been over the 1992-1996 period, which was needed to calculate the second term in the following weighted average formula for the fund's projected 10 year average:

$$\frac{1}{120}\sum_{t=1}^{120}\log(1+R_t^{fund}) = \frac{60}{120}\sum_{t=1}^{60}\log(1+R_t^{fund}) + \frac{60}{120}\sum_{t=61}^{120}\log(1+R_t^{fund}) \tag{16}$$

In (16), the weights are equal (i.e. each is just one-half) because both the observation period and the projection period were the same length (5 years). Pro-weighted weights are used in other cases. The analyst reported that her regression technique yielded a projected average of 1.83% per month for the fund between 1992-1996. Plugging this into (16) yielded a projected 10 year fund average of $60/120 * 1.09\% + 60/120 * 1.83\% = 1.46\%$ per month. Subtracting the 10 year index average of 0.96%, she told the firm that her estimate for the 10 average difference in log gross fund and index returns was 0.50% per month.

We can help the analyst and her firm decide what to do with this information, because we have the benefit of hindsight: we *know* how the fund actually performed between 1992-1996. Her projected 1992-1996 fund average of 1.83% was lower than its actual average of 2.65% over those years. As a result, her 10 year projection of 1.46% was also lower than the fund's actual 10 year

average of 1.87%. The index actual 10 year average was 1.09%, so the *actual* 10 year average difference in log gross fund and index returns was $1.87\% - 1.09\% = 0.78\%$, which is higher than the *projected* difference of 0.50% per month. While not perfect, the projected difference is closer to the actual 10 year difference than the observed 1997-2001 5 year difference of $1.09\% - 0.77\% = 0.32\%$ that would otherwise have been used by the firm.

Stambaugh [20] also shows how to use the longer index series to project the 10 year fund return variance and its covariance with the index, which could be used to project the 10 year standard deviation of the difference in log gross fund and index returns.[19] Dividing this into the 0.50% per month would provide the desired 10 year projected LIR.

Projections work best when the statistical model used to link the series is well-specified, i.e. $R^2$ is quite high, and stable over time. It is well-known that large-cap funds' returns often have $R^2$s in excess of 90% when regressed on a large-cap index like the S&P 500, because their holdings substantially overlap. So it appears that projections like this would be ideal for ranking large-cap funds of varying ages relative to a respected large-cap benchmark with a very long history, like the S&P 500. The technique should be similarly valuable when ranking funds in *any* category relative to a benchmark tailored for that category.

### 4.3 Do the Differences Make a Difference?

Propositions 3 and 5 showed that seemingly different performance measures (de-facto) reward high averages and penalize high standard deviations, utilizing different weightings of the two statistics. We also reported empirical evidence that the approximations in Propositions 3 and 5 worked quite well in practice. Similar mean-variance approximations of other seemingly different expected utility measures are also derivable. Those approximations will probably also work well in practice as long as

the absolute third moment of returns is finite and vanishes more rapidly than the first two moments do as returns are measured more frequently. Specifically, under this and some other regularity conditions, Ohlson [16] proved that expected utility rankings will converge to expected quadratic utility (i.e. mean-variance based) rankings as the interval between return measurements decreases to zero. Figure 10 uses the example fund 's log gross returns to illustrate the phenomenon. It shows that the 3rd central moment is quite small when annual data is used, especially when compared to the first two moments (i.e. the mean and variance) and shrinks to a near-zero number when monthly data is used. Hence the mean-variance approximation should be even better when daily or weekly fund returns are used, rather than the commonly used monthly returns. The differences and similarities of the performance measures' mean-variance approximations are summarized in the table below.

**Summary of De-Facto Weightings of Averages and Standard Deviations**

| *Performance Measure* | *Benchmark* | *Return Average* | *Standard Deviation* |
|---|---|---|---|
| Sharpe Ratio | T-Bill | Net | Divided Into |
| Morningstar | T-Bill | Log Gross | Squared and Subtracted From |
| Lipper Preservation | Zero | Net | Subtracted From |
| Log Information Ratio | User-Selected | Log Gross | Divided Into |

Will the above differences in performance measure construction cause big differences in fund ratings? Some evidence that they may not was provided in Sharpe's [19] study of the different, more complex performance measure used by Morningstar prior to 2002. This measure was essentially the sum of a fund's historical cumulative return in excess of the T-Bill cumulative return, and the negative value of the second term in (14). While it appears on the surface to be very different

from the conventional Sharpe Ratio, Sharpe [18, Figure 9] used historical fund returns from 1994-1996 to compare the two rankings, concluding that the correlation coefficient between the funds' percentile rankings under the two measures was 98.6%!However, Sharpe [19, pp.30-31] conjectures that this high correlation might break down when the historical return period is a bear market. Because Morningstar replaced the measure Sharpe studied with the MRAR(2) studied here, there isn't much cause to do that analysis. Instead, I used the three years of monthly returns on 1307 large cap funds from 1999-2001, which includes the bear market of 2001, to contrast the different fund rankings produced by Morningstar's MRAR(2) and Lipper's Preservation measure. Despite the differences described in the above table, the rank correlation coefficient for the two seemingly disparate rankings is 87.4%. The agreement seems quite marked at the bottom of the fund rankings, but a bit less-so at the top. The major source of the difference appears to be MRAR(2)'s de-facto subtraction of the squared volatility, rather than the volatility itself. As a result, a small increase in volatility decreases the Lipper (de-facto) measure by an equal amount, while the same increase in volatility has decreases the MRAR(2) (de-facto) measure by twice that amount, times the fund's monthly volatility. Because monthly fund volatilities are well below 50%, the Morningstar measure is less sensitive to volatility than the Lipper Preservation measure is, i.e. it weights volatility less when ranking the funds.

In light of these empirical findings and the above theoretical arguments establishing their plausibility, I conclude with a provocative conjecture:

**Conjecture:** *In practice, seemingly different performance measures may produce surprisingly similar fund rankings, as long as they incorporate similar benchmarks.*

# 5 Conclusions

Fund managers and/or their investor clients who are seeking to beat a benchmark should realize that even if they will beat the benchmark over the long-run, significant probabilities of underperforming it often persist for a surprisingly long time. Those desiring funds with rapidly shrinking underperformance probabilities should use a performance measure produced by replacing the net returns used in the well-known Information Ratio measure with log gross (i.e. continuously compounded) returns. The popular Sharpe Ratio is inappropriate for this purpose, in part because it uses ordinary net rather than log gross returns, but also because it uses a T-Bill benchmark. The Morningstar, Inc. performance measure used to produce its well-known "star ratings" also uses a T-Bill benchmark. But after substituting both the benchmark to be beaten, and a fund-dependent, optimized value of its curvature parameter $\gamma$ in place of its fixed level $\gamma = 2$, it will usually produce a ranking directly related to the outperformance probability, like the aforementioned log-modified Information Ratio does. It does not appear to be as simple to transform Lipper's Preservation measure into a measure that ranks funds in accord with their probabilities of beating the benchmark.

However, Lipper's Preservation performance measure does have something in common with all the alternative measures mentioned above: they all reward high average fund returns in excess of some benchmark, and penalize high return standard deviations (i.e. volatilities). They differ with respect to how these two statistics are weighted, and by whether ordinary net or log gross (i.e. continuously compounded) returns are used in their computation. Nevertheless, empirical evidence indicates that in practice, the different performance measures will produce fairly similar ratings for most funds, i.e. their rankings will be highly correlated.

The main difficulty impeding the reliability of performance measures is the instability of histori-

cal average returns. The problem is not eliminated by measuring returns more frequently, e.g. daily or weekly, rather than monthly. However, the historical average *difference* of reasonably closely correlated fund and benchmark returns will generally be more stable, so accuracy considerations favor the ranking of funds relative to a benchmark reflecting their category (e.g. ranking large-cap funds in accord with their probabilities of beating the S&P 500, rather than in accord with their probabilities of beating T-Bills.) By doing so, the very practical concern of how to rank old funds with a long track record against young funds, with say, a 3 year track record can be partly eased by using filtering techniques to "backcast" returns (relative to the highly correlated benchmark) that the young funds might have earned had they been in business earlier.

# Appendix

The purpose of this appendix is to more formally develop the paper's propositions. To develop Proposition 1, let us start by comparing the cumulative return $W_T^{fund}$ from a fund with period return $R_t^{fund}$ at period $t$, to a benchmark's $W_T^b$ with period return $R_t^b$ (e.g. the broad-based index return used in the example). The following equivalent inequalities:

$$W_T^{fund} \equiv \prod_{t=1}^{T}\left(1+R_t^{fund}\right) < \prod_{t=1}^{T}\left(1+R_t^{b}\right) \equiv W_T^b \;\; iff$$

$$\log\frac{W_T^{fund}}{T} = \frac{\sum_{t=1}^{T}\log\left(1+R_t^{fund}\right)}{T} < \frac{\sum_{t=1}^{T}\log\left(1+R_t^{b}\right)}{T} = \log\frac{W_T^b}{T} \tag{1}$$

show that the underperformance probability

$$Prob\left[W_T^{fund} < W_T^b\right] = Prob\left[\frac{\sum_{t=1}^{T}\log\left(1+R_t^{fund}\right)}{T} < \frac{\sum_{t=1}^{T}\log\left(1+R_t^{b}\right)}{T}\right]. \tag{2}$$

Note that a log gross return $\log(1+R_t)$ is a continously compounded period return, denoted $r_t$, because $e^{r_t} = 1+R_t$. Hence (2) shows that the fund's underperformance probability over a horizon of $T$ periods is the probability that the time average of its continuously compounded period return is less than the benchmark's time averaged continuously compounded return. Laws of large numbers appropriate to the return processes imply that

$$\begin{aligned}\lim_{T\to\infty}\sum_{t=1}^{T} r_t^{fund} &= E[r^{fund}] \\ \lim_{T\to\infty}\sum_{t=1}^{T} r_t^{b} &= E[r^{b}]\end{aligned} \tag{3}$$

where in the case of non-IID ergodic processes, the expectation operator is interpreted as the ergodic mean. Hence an implication of (2) - (3) is

$$\lim_{T\to\infty} Prob\left[W_T^{fund} < W_T^b\right] = 0 \;\; iff \;\; E[r^{fund}] > E[r^b]. \tag{4}$$

So the fund's underperformance probability will decay to zero as the holding period lengthens precisely when its expected log gross (i.e. continuously compounded) return exceeds the benchmark's. This finding can be used to explain the decay to zero of the bootstrap simulated underperformance probabilities in Figure 4 and on the left hand side of Figure 3. The bootstrap used here randomly resamples the past history of continuously compounded returns, i.e. it samples from the *empirical* distribution.[20] Because Figure 1 shows that the fund's historical cumulative return exceeded the benchmark's (either the broad-based index or the T-Bill), so did its historical time averaged continously compounded return (see (1)), which is the *expected* value of this empirical distribution. Hence the right hand side of (4) characterizes the *empirical* distribution, so its left hand side must also characterize it, i.e. the bootstrapped underperformance probabilities in Figure 4 and and on the left hand side of Figure 3 must asymptotically approach zero as the holding period approaches infinity. But the underperformance probabilities could decay to zero at a very slow rate of decay, as seen on the left hand side of Figure 3.

But what determines the decay rate of the underperformance probabiliities? Proposition 2 attempts to answer this harder question. To develop it, the underperformance probabilities must actually decay to zero, so w.l.o.g. use (4) to require that $E[r^{fund}] > E[r^b]]$. For the moment, let us also assume that the distribution of the process generating the period returns $r_t^{fund} - r_t^b$ is IID normal with expected value $E[r^{fund} - r^b]$ and variance $Var[r^{fund} - r^b]$.[21] From (2), the underperformance probability for horizon length $T$ is the probability that the corresponding future time averaged value of $r_t^{fund} - r_t^b$ is less than zero. Because of the normality assumption, the distribution of this time average is also normally distributed, with expected value $E[r^{fund} - r^b]$ and variance $Var[r^{fund} - r^b]/T$. As such, transformation to the standard normal variate $Z$ shows that the underperformance probability is:

$$Prob[W_T^{fund} < W_T^b] = Prob\left[Z \ < \frac{0 - E[r^{fund} - r^b]}{\sqrt{Var[r^{fund} - r^b]/T}}\right] = Prob\left[Z > LIR\,\sqrt{T}\right] \qquad (5)$$

where $LIR$ denotes the log-modified Information Ratio $\frac{E[r^{fund}-r^b]}{\sqrt{Var[r^{fund}-r^b]}}$ described in Proposition 2. From (5), we see that at least in the IID normally distributed case, the LIR performance measure rank orders funds inversely to their probabilities of underperforming the same benchmark, for *any* fixed value of the holding period $T$. By (4), the underperformance probability asymptotically decays to zero if and only if the LIR is positive. A consequence of the standard normal distribution's right hand tail is that the higher the LIR, the faster the underperformance probability decays to zero as $T \to \infty$, i.e. the faster the bar heights decline in Figure 4 or on the left hand side of Figure 3.

But Proposition 2 did not require the normality assumption. Can the above calculations be at least partly generalized to non-normal processes? A Central Limit Theorem [10] appropriate to the process generating the differential return $r_t^{fund} - r_t^b$ implies that the time average of this differential return is approximately normally distributed for suitably large $T$, i.e. the distribution of the time average is asymptotically normally distributed. Note that there are CLTs that don't require that the differential return process by independent, nor that it be identically distributed. In fact, there are CLTs for the kind of weak dependence assumptions that are required for the applicability of the time series analysis tools that are widely used in financial econometrics (e.g. GMM estimation). Hence the general validity of Proposition 2 depends on how good the CLT normal approximation is for holding periods of interest to investors, with the form of the mainly determining the nature of the estimator needed for the variance in the LIR denominator (e.g. a weakly dependent process may require a Bartlett kernel, Newey-West type variance estimator). The reliance on the CLT is the reason for use of the word "generally" in Proposition 2.

Some evidence that the CLT does provide a good approximation in practice is provided in the paper. Adopting the one-month T-Bill return as the benchmark, we saw in the paper that the historical LIR (due to the T-Bill benchmark used in the isomorphic Sharpe Ratio, this was dubbed the LSR) of the index exceeded the fund's. Figure 4 is produced by resampling from the empirical distribution, yet it indicates that the index probabilities of underperforming T-Bills does indeed decay toward zero at a faster rate than fund's probabilities do, at least over the 1 -10 year range of holding periods plotted in Figure 4. This is true despite the non-normal levels of skewness (-1.8) and kurtosis (10.8) that characterize the empirical distribution that was resampled to produce Figure 4, indicating that the CLT approximation is good enough to make the LIR useful, even when the holding period is as low as 1 -10 years. In fact, results in Stutzer [22], [21] show that the underperformance probabilities will eventually (i.e. asymptotically) decay at a rate of $\frac{1}{2}(LIR)^2$ per period, although the decay may initially be more rapid.

Now let us develop Proposition 3. To do so, let us first transform Morningstar's Risk-Adjusted Return measure MRAR($\gamma$) into the simpler, more familiar expected power utility index that preserves its rank ordering of funds. To do so, consider the rank order of two funds, denoted $R^{fund}$ and $R^{index}$. The following chain of equivalent inequalities holds:

$$\begin{aligned}
MRAR^{fund}(\gamma) &< MRAR^{index}(\gamma) \ \ iff \\
\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1+R_t^{fund}}{1+R_t^b}\right)^{-\gamma}\right]^{-\frac{12}{\gamma}} - 1 &< \left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1+R_t^{index}}{1+R_t^b}\right)^{-\gamma}\right]^{-\frac{12}{\gamma}} - 1 \ \ iff \\
-\frac{12}{\gamma}\log\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1+R_t^{fund}}{1+R_t^b}\right)^{-\gamma}\right] &< -\frac{12}{\gamma}\log\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1+R_t^{index}}{1+R_t^b}\right)^{-\gamma}\right] \ \ iff \\
\log\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1+R_t^{fund}}{1+R_t^b}\right)^{-\gamma}\right] &> \log\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1+R_t^{index}}{1+R_t^b}\right)^{-\gamma}\right] \ \ iff
\end{aligned}$$

$$\frac{1}{T}\sum_{t=1}^{T} - \left(\frac{1+R_t^{fund}}{1+R_t^b}\right)^{-\gamma} \quad < \quad \frac{1}{T}\sum_{t=1}^{T} - \left(\frac{1+R_t^{index}}{1+R_t^b}\right)^{-\gamma} \tag{6}$$

Financial theorists will recognize (6) as the historical estimate of the expected power utility of the fund return *relative* to a benchmark return. This is the first claim in Proposition 3. This power utility has a degree of constant relative risk aversion equal to $1+\gamma$. Note that this is *not* the degree of risk aversion to fluctations in investor wealth; it is the degree of risk aversion to fluctations in the ratio of wealth earned by investment in the fund to wealth earned by investment in the benchmark. As such, one cannot appeal to the usual experimental or market evidence when specifying $\gamma$ for a representative investor; Morningstar must martial other evidence to support its assumption that $1+(\gamma=2)=3$ is a representative degree of investor risk aversion for the purpose of ranking funds.[22] Now suppose we substitute continously compounded returns $r \equiv \log 1+R$, for the gross returns $1+R$ in (6), producing the equivalent performance comparison:

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} - \left(\frac{e^{r_t^{fund}}}{e^{r_t^b}}\right)^{-\gamma} \quad &< \quad \frac{1}{T}\sum_{t=1}^{T} - \left(\frac{e^{r_t^{index}}}{e^{r_t^b}}\right)^{-\gamma} \quad iff \\ \frac{1}{T}\sum_{t=1}^{T} -e^{-\gamma(r_t^{fund}-r_t^b)} \quad &< \quad \frac{1}{T}\sum_{t=1}^{T} -e^{-\gamma(r_t^{index}-r_t^b)} \end{aligned} \tag{7}$$

Financial theorists will recognize (7) as the historical estimate of the expected exponential utility (of the fund's *excess* return) with constant *absolute* risk aversion equal to $\gamma > 0$, used in models of fund manager behavior, e.g. in Becker, Ferson, et.al.[2]. The approximate ranking statistic described at the end of Proposition 3 is established by using a normal approximation to the distribution of $r^{fund} - r^b$, i.e. by using a lognormal approximation to the distribution of the return relative $1+R^{fund}/1+R^b$. This approximation truncates all third and higher order cumulants. It is well known that the expected exponential utility of the normally distributed $r^{fund} - r^b$ is just:

$$E\left[-e^{-\gamma(r^{fund}-r^b)}\right] = E[r^{fund}-r^b] - \frac{\gamma}{2}Var[r^{fund}-r^b]. \tag{8}$$

Substituting Morningstar's value of $\gamma = 2$ into (8) yields the approximate ranking measure $E[r^{fund}-r^b] - \sqrt{Var[r^{fund}-r^b]}^2$, as claimed at the end of Proposition 3.

The conventional use of (7) as a ranking index uses a constant value of $\gamma > 0$ to rank funds, like Morningstar does. To establish Proposition 4, Stutzer [22] used Cramer's Large Deviation Theorem [4, Chap.1] for the sample average of (assumed) IID log return processes to show that the generalized expected exponential utility index

$$\max_{\gamma} E\left[-e^{-\gamma(r^{fund}-r^b)}\right] \tag{9}$$

whose historical estimate is:

$$\max_{\gamma} \frac{1}{T}\sum_{t=1}^{T} -e^{-\gamma(r_t^{fund}-r_t^b)} \tag{10}$$

has a positive maximizing value of $\gamma$ if and only if $E[r^{fund}] > E[r^b]$. Equation (4) shows that this happens if and only if the fund's underperformance probabilities approach zero as $T$ increases. This establishes the decay to zero of underperformance probabilites if and only if $\gamma_{max} > 0$, as claimed in Proposition 4. The large deviations theorem also establishes that the underperformance probabilities decay to zero at an asymptotic rate equal to $-\log E[e^{-\gamma(r^{fund}-r^b)}]$, which is historically estimated by $-\log[-(10)]$. The latter reduces to $-\log[U(\gamma_{max})]$ defined in Proposition 4, upon substitution of the defining $\log[1+R]$ for each continuously compounded return $r$ in (10). Hence the larger the fund's performance measure (10), the smaller (larger) the underperformance (outperformance) probability is. Hence (10) produces a rank ordering of funds that generally agrees with that produced by the LIR developed in Appendix 1, differences being attributable to cases where IID Central Limit and IID Large Deviations approximations for the tails of the distributions of

sample averages are far enough apart to cause differences in their rank orderings. Generalizations to weakly dependent, non-IID processes were developed in Stutzer [21] and in Foster and Stutzer [8].

Proposition 5 is developed by first noting that the performance ALP defined in the paper is an historical estimate of the expectation $E[\min[0, R^{fund}]]$. The paper's Equation (14) produces an historical estimate of $Prob[R^{fund} < 0] \times E[R^{fund} | R^{fund} < 0]$. Elementary probability theory implies that this product can also be written:

$$E[R^{fund}] - Prob[R^{fund} \geq 0] E[R^{Fund} | R^{fund} \geq 0] \tag{11}$$

To calculate the dependence of this expression on the first two moments of $R^{fund}$, let us approximate the distribution of $R^{fund}$ with a normal distribution that has the same mean and variance. This distribution has the same first two moments (and hence the same first two cumulants), but truncates the effects of all third and higher order cumulants. Using the standardizing transformation, calculate:

$$Prob[R^{fund} \geq 0] = Prob\left[Z \geq -\frac{E[R^{fund}]}{\sqrt{Var[R^{fund}]}}\right] = N\left(\frac{E[R^{fund}]}{\sqrt{Var[R^{fund}]}}\right) \tag{12}$$

where $N$ denotes the standard normal cumulative distribution function. The conditional expectation in (11) is an integral that is computed in standard logit models (e.g, see Amemiya [1, p.367]), yielding:

$$E[R^{Fund} | R^{fund} \geq 0] = E[R^{fund}] + \sqrt{Var[R^{fund}]} \frac{N'\left(\frac{E[R^{fund}]}{\sqrt{Var[R^{fund}]}}\right)}{N\left(\frac{E[R^{fund}]}{\sqrt{Var[R^{fund}]}}\right)} \tag{13}$$

where $N'$ denotes the standard normal probability *density* function. Substituting (12) and (13) into (11) and simplifying yields the desired closed form formula for the population value of the

historical estimate ALP:

$$E[\min[0, R^{fund}]] = E[R^{fund}]\left(1 - N\left(\frac{E[R^{fund}]}{\sqrt{Var[R^{fund}]}}\right)\right) - \sqrt{Var[R^{fund}]}N'\left(\frac{E[R^{fund}]}{\sqrt{Var[R^{fund}]}}\right) \tag{14}$$

Numerical evaluation of (14) indicates that it is well approximated by a linear function of $E[R^{fund}] - \sqrt{Var[R^{fund}]}$ e.g. see Figure 6, as claimed in Proposition 5.

Proposition 6 claims that the instability of performance measures that require historical average returns is partly alleviated by using performance measures that depend on the historical average *difference* of reasonably correlated assets' returns. To see this, examine the following equivalent chain of comparisons between the standard error of a fund's historical average and the differential average's standard error:

$$\begin{aligned}
\sqrt{Var[\frac{1}{T}\sum_{t=1}^{T}(R_t^{fund} - R_t^b)]} &< \sqrt{Var[\frac{1}{T}\sum_{t=1}^{T} R_t^{fund}]} \;\; iff \\
Var[\frac{1}{T}\sum_{t=1}^{T}(R_t^{fund} - R_t^b)] &< Var[\frac{1}{T}\sum_{t=1}^{T} R_t^{fund}] \;\; iff \\
\frac{1}{T}(Var[R^{fund}] + Var[R^b] - 2Cov[R^{fund}, R^b]) &< \frac{1}{T}Var[R^{fund}] \;\; iff \\
\frac{Var[R^{fund}] + Var[R^b] - 2Cov[R^{fund}, R^b]}{\sqrt{Var[R^{fund}]}\sqrt{Var[R^b]}} &< \frac{Var[R^{fund}]}{\sqrt{Var[R^{fund}]}\sqrt{Var[R^b]}} \;\; iff \\
\frac{Cov[R^{fund}, R^b]}{\sqrt{Var[R^{fund}]}\sqrt{Var[R^b]}} &> \frac{1}{2}\frac{\sqrt{Var[R^b]}}{\sqrt{Var[R^{fund}]}}
\end{aligned} \tag{15}$$

The left hand side of (15) is the correlation coefficient of the fund and benchmark returns. The example in the paper used the index as the benchmark, which is the 87% correlated with the fund. This is well above the right hand side of (15), which is only 21%. Note that the left hand side correlation will typically be high when the benchmark portfolio is representative of the fund's

category (e.g. asset class (stocks or bonds), style (growth or value), or size (large cap or small cap)). The right hand side will typically be low when the fund is more volatile than the benchmark. Both of these considerations are typically met when a broad-based equity (bond) index is used to benchmark the performance of equity (bond) funds, in an Information Ratio or an LIR performance measure (for the latter, just substitute continuously compounded returns $\log[1 + R]$ for the net returns $R$ in (15).

## Notes

[1]This paper was prepared for the IFID Centre (Prof. Moshe Milevsky, Director) conference at the Fields Institute, Toronto, CA. I acknowledge the assistance of Paul Kaplan at Morningstar, Inc. and Linbo Fan at Lipper, Inc. in establishing the empirical relevance of the approximations developed herein, and helpful comments that Prof. Milevsky made on a prior draft.

[2]The punch line is a paraphrase from one of Woody Allen's essays.

[3]The historical return data was kindly provided by Mark Hulbert. The index series used here is neither a mutual nor exchange-traded index fund, whose returns would differ somewhat from it, while the other series is associated with active management. The two series used here are sufficient to illustrate the issues arising in rating the performance of one investment relative to an alternative.

[4]There are also more complex ways to implement bootstrapping. Parametric estimates of return processes for both the fund and the index can be utilized if good parametric models can be found, or non-parametric "block" bootstrap procedures can be utilized if returns are significantly serially dependent; in our application, it is the difference of two funds' returns that would need to be serially dependent. The simpler procedure is sufficient for the pedagogical purpose of this paper.

[5]Of course, an optimist looking at Figure 3 might not be focused on its left hand side, which only show the fund's probabilities of underperforming the index. For example, the closest middle bar indicates that there is a 68% chance that the fund will fail to beat the index by a factor of more than two; but this means that there is a $100\% - 68\% = 32\%$ chance that the fund *will* more than double the index cumulative return after 10 years!

[6]As shown in the Appendix, this is due to the lower standard error associated with the *difference*

of averages drawn from closely correlated random variables.

[7]According to Schwab's website, "To make the Mutual Fund Select List, a fund had to have a high risk-adjusted return coupled with a high total return and low expenses." The risk-adjusted return arises from the Sharpe Ratio. The quantitative part of Standard and Poors evaluation is similarly motivated. Specifically , Standard and Poors averages a fund's absolute performance decile ranking over each of the past 3 years, does the same with its Sharpe Ratio, and then averages the resulting two numbers to determine a fund's combined ranking. Select funds are further evaluated by Standard and Poors staff, who consider "the management process and resources put in place to run a fund" (personal communication from Eleanore De Bar, Director, Standard and Poors Fund Services,Europe).

[8]Sharpe [18, p.50] calls this the "ex-post" ratio.

[9]Because the funds in a specific category have existed (or have reported returns) for different lengths of time, and because some analysts might feel the recent past is more indicative of the future to come, the choice of the number of past historical months $T$ to use when ranking them is problematic. Issues involving the choice of $T$ will be discussed in section 4.

[10]Of course, the bootstrap resamples from the historical distribution, so this assessment is actually based on the historical estimate of LIR, not its actual value. The Appendix shows that the fund's underperformance probabilities will decay to zero as the holding period lengthens when $E[\log R_{fund} - \log R_{index}] > 0$, because cumulative return is the produce of period returns, rather than the sum of them. When this condition holds, The Appendix also shows that for suitably long holding periods, the underperformance probability will be lower when the actual LIR is higher, i.e.

when $E[\log R_{fund} - \log R_{index}]/StdDev[\log R^{fund} - \log R^{index}]$ is lower.

[11]At www.lipperleaders.com, Lipper also ranks funds according to their total return (i.e. no risk-adjustment), and according to a more complex and harder to interpret hierarchical scheme. Funds with returns that have a Hurst Exponent higher than .5 are more "persistent" than a random walk, generally having positive autocorrelations at different lags (and motivating the term "consistent returns" [5]), slowly decaying autocorrelations as the lag length increases ("long memory" or "long-term dependence") , and infinite long run variance. Funds with with a Hurst exponent less than .5 are "antipersistent", with negative autocorrelations. Lipper uses three years of daily fund returns to sorts funds into high ($H > .55$), medium ($.45 < H < .55$) and low ($H < .45$) ranges based on their AR(1) residuals' estimated Hurst Exponents. A loss-aversion type of utility is then used to rank funds within each of these ranges. The top ranked quintile of funds in each of the three Hurst categories are dubbed "Lipper Leaders for Preservation". Lo and MacKinlay [12, Chap.6] discuss the formidable difficulties involved in properly estimating and interpreting the Hurst Exponent and the rescaled range statistics that it is based on. I have two conjectures. First, it will be very difficult in practice to discern whether or not the funds are sorted properly into their high, medium, and low Hurst categories using only three years of data. Second, it will be difficult to accurately estimate the loss-aversion utility with only three years of data, due to its inevitable (de-facto) dependence on average returrns.

[12]The fund and index had almost identical, modestly negative levels of historical monthly skewness (-.75 vs. -.77), so their third moments could not have contributed as much to their relative ranking as their first two moments did.

[13]Personal correspondence from Linbo Fan, Research Analyst, Lipper, Inc.

[14]Of course, the possibility of secular change in the return generating process calls this reasoning into question.

[15]This argument is made by Roll [17, p.20].

[16]It is important to note that the conventional Sharpe Ratio, which is defined using ordinary net rather than log gross returns, "is not a complete summary of the risks of a multiperiod investment strategy and should never be used as the sole criterion for making an investment decision" [11].

[17]However, Moshe Milevsky informed me that in Canada, the National Post publishes the FPX, a balanced index including both stocks and bonds.

[18]Private firms, e.g. the Frank Russell Co., publish a large number of narrowly tailored equity indices. Firms such as BARRA and Richards and Tierney produce custom benchmark indices used to evaluate the performance of fund managers charged with maintaining a fixed size, style, or sector orientation.

[19]Stambaugh's paper also shows how to generalize the idea to incorporate multiple correlated series starting at different times, and develops several alternative applications.

[20]However, the simple bootstrap implemented in the paper was not constructed to exploit any possible serial dependencies in the return processes.

[21]For example, the return difference for a period length $\Delta t$ could be generated by a continuous time lognormal process, with instantaneous mean $\mu$ and volatility $\sigma$. If so, $E[r^{fund} - r^b] = (\mu - \sigma^2/2)\Delta t$ while $Var[r^{fund} - r^b] = \sigma^2 \Delta t$.

[22]Morningstar's equivalent MRAR($\gamma = 2$) performance measure is intended to measure the an-

nualized certainty equivalent return implied by this utility.

## References


[1] Takeshi Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

[2] Connie Becker, Wayne Ferson, David H. Myers, and Michael J. Schill. Conditional market timing with benchmark investors. *Journal of Financial Economics*, 52(1):119–148, 1999.

[3] Susan Belden and M. Barton Waring. Compared to what? a debate on picking benchmarks. *Journal of Investing*, 10(4):66–72, 2001.

[4] James A. Bucklew. *Large Deviation Techniques in Decision, Simulation, and Estimation*. Wiley, 1990.

[5] Andrew Clark. White paper on consistent return. Lipper Leaders Technical Document, Lipper, Inc., 2001.

[6] Diane Del Guercio and Paula A. Tkac. Star power: The effect of Morningstar ratings on mutual fund flows. Working Paper 2001-15, Research Department, Federal Reserve Bank of Atlanta.

[7] Frank J. Fabozzi, Francis Gupta, and Harry M. Markowitz. The legacy of modern portfolio theory. *Journal of Investing*, 11(3):7–23, 2002.

[8] F. Douglas Foster and Michael Stutzer. Performance and risk aversion of funds with benchmarks: A large deviations approach. Working Paper, University of Colorado Finance Department, 2002.

[9] Thomas H. Goodwin. The Information Ratio. *Financial Analysts Journal*, 54(4):34–43, 1998.

[10] E.L. Lehmann. *Elements of Large Sample Theory*. Springer-Verlag, 1999.

[11] Andrew W. Lo. The statistics of Sharpe Ratios. *Financial Analysts Journal*, 58(4):36–52, 2002.

[12] Andrew W. Lo and A. Craig MacKinlay. *A Non-Random Walk Down Wall Street*. Princeton University Press, 1999.

[13] David G. Luenberger. *Investment Science*. Oxford University Press, New York, 1998.

[14] Matthew R. Morey. Mutual fund age and morningstar ratings. *Financial Analysts Journal*, 54:56–63, March/April 2002.

[15] Morningstar. The new Morningstar rating methodology. Morningstar Research Report, April 22, 2002, Morningstar, Inc., Chicago, IL.

[16] James A. Ohlson. The asymptotic validity of quadratic utility. In W.T. Ziemba and R.G. Vickson, editors, *Stochastic Optimization Models in Finance*. Academic Press, 1975.

[17] Richard Roll. A mean/variance analysis of tracking error. *Journal of Portfolio Management*, 18(4):13–22, 1992.

[18] William Sharpe. The Sharpe Ratio. *Journal of Portfolio Management*, 21(1):49–58, 1994.

[19] William Sharpe. Morningstar's risk-adjusted ratings. *Financial Analysts Journal*, 54(4):21–33, 1998.

[20] Robert F. Stambaugh. On the exclusion of assets from tests of the two-parameter model: A sensitivity analysis. *Journal of Financial Economics*, 10(3):237–268, 1982.

[21] Michael Stutzer. Portfolio choice with endogenous utility: A large deviations approach. Journal of Econometrics, 2002(forthcoming).

[22] Michael Stutzer. A portfolio performance index. *Financial Analysts Journal*, 56(3):52–61, 2000.

[23] Jeff Tjornehoj. White paper on preservation. Lipper Leaders Technical Document, Lipper, Inc., 2001.

**Figure 1: Volatile Fund Beats Market Index**

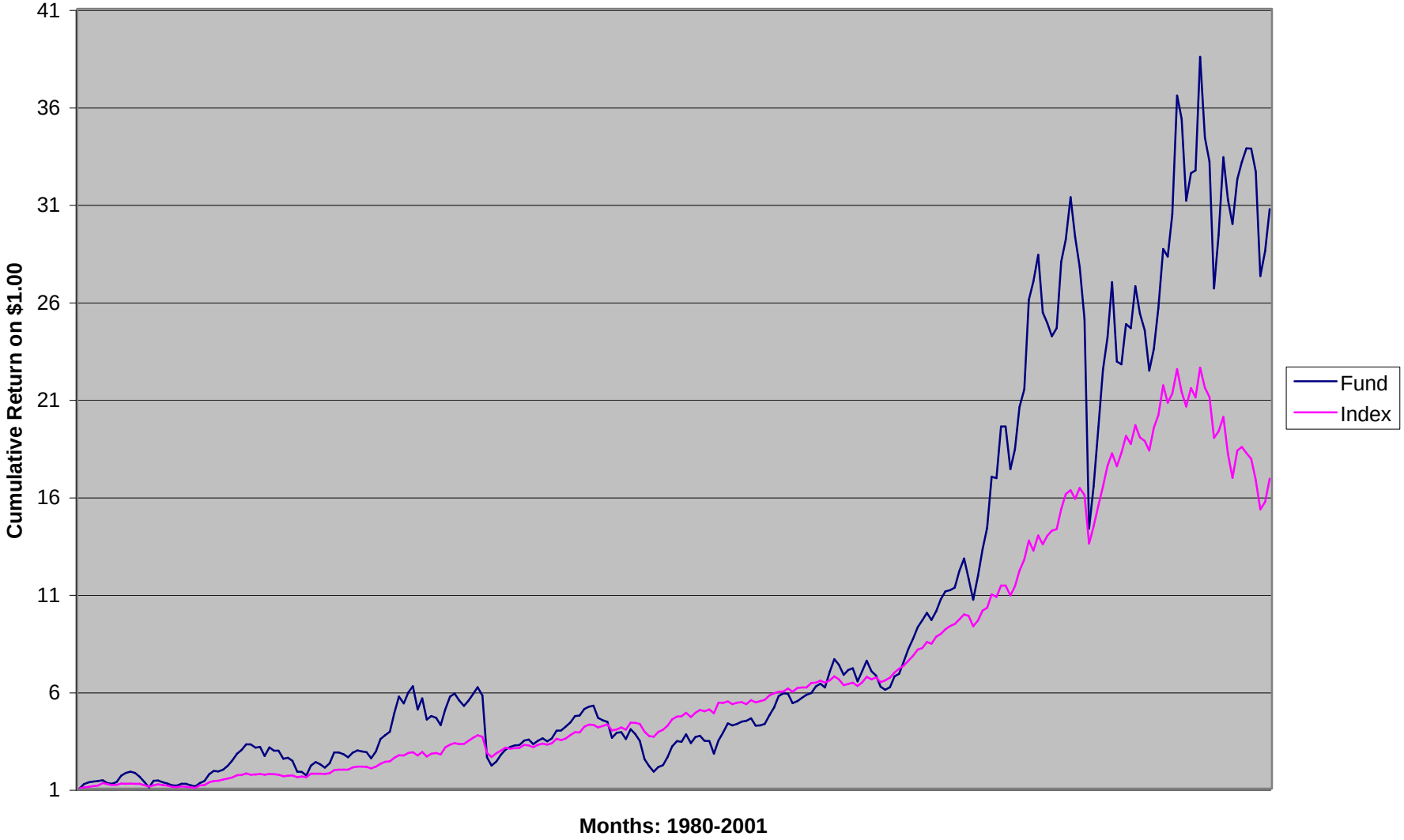


**Figure 2: Volatile Mutual Fund vs. Market Index**

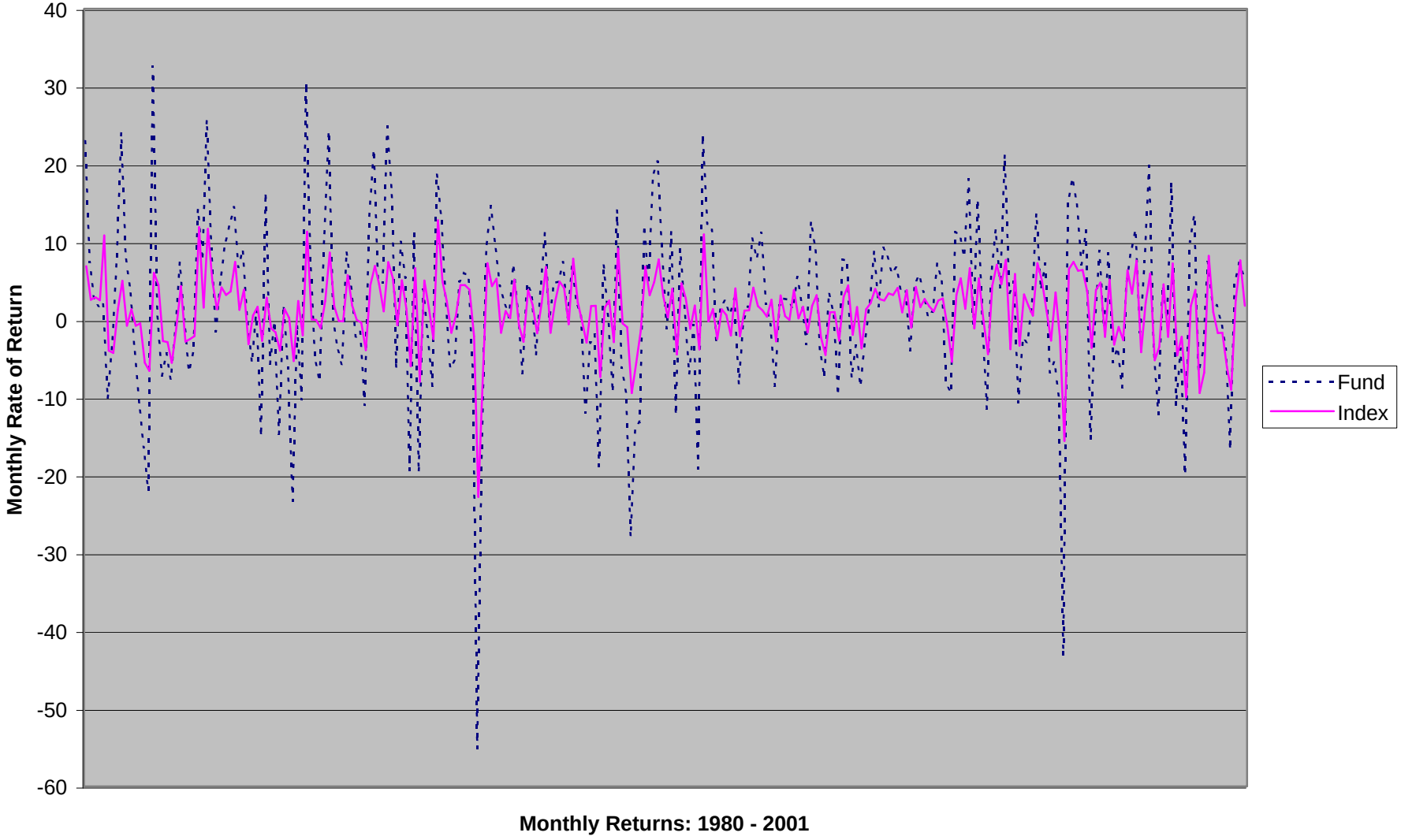


**Figure 3: Possible Future Ratios of Fund to Index Cumulative Returns**

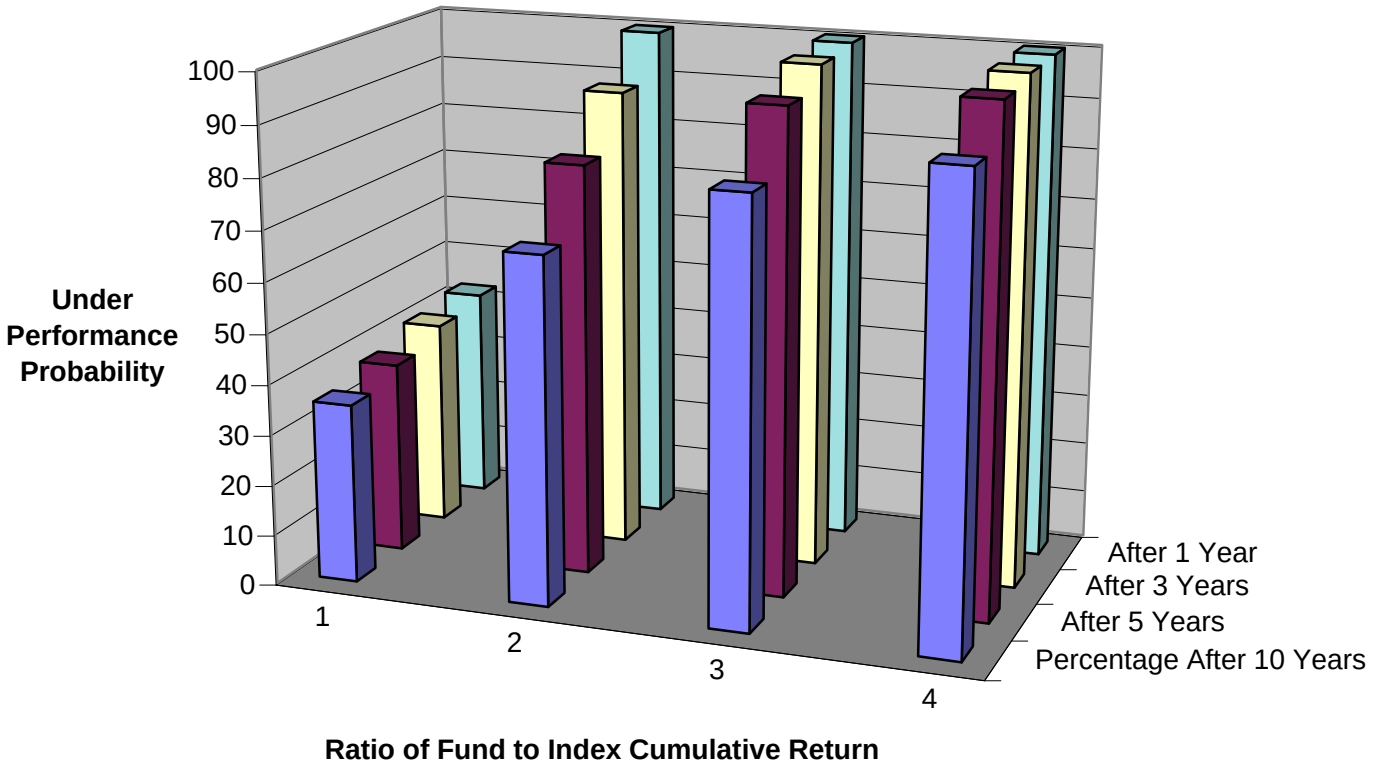


**Figure 4 : Probability of Underperforming T- Bills**

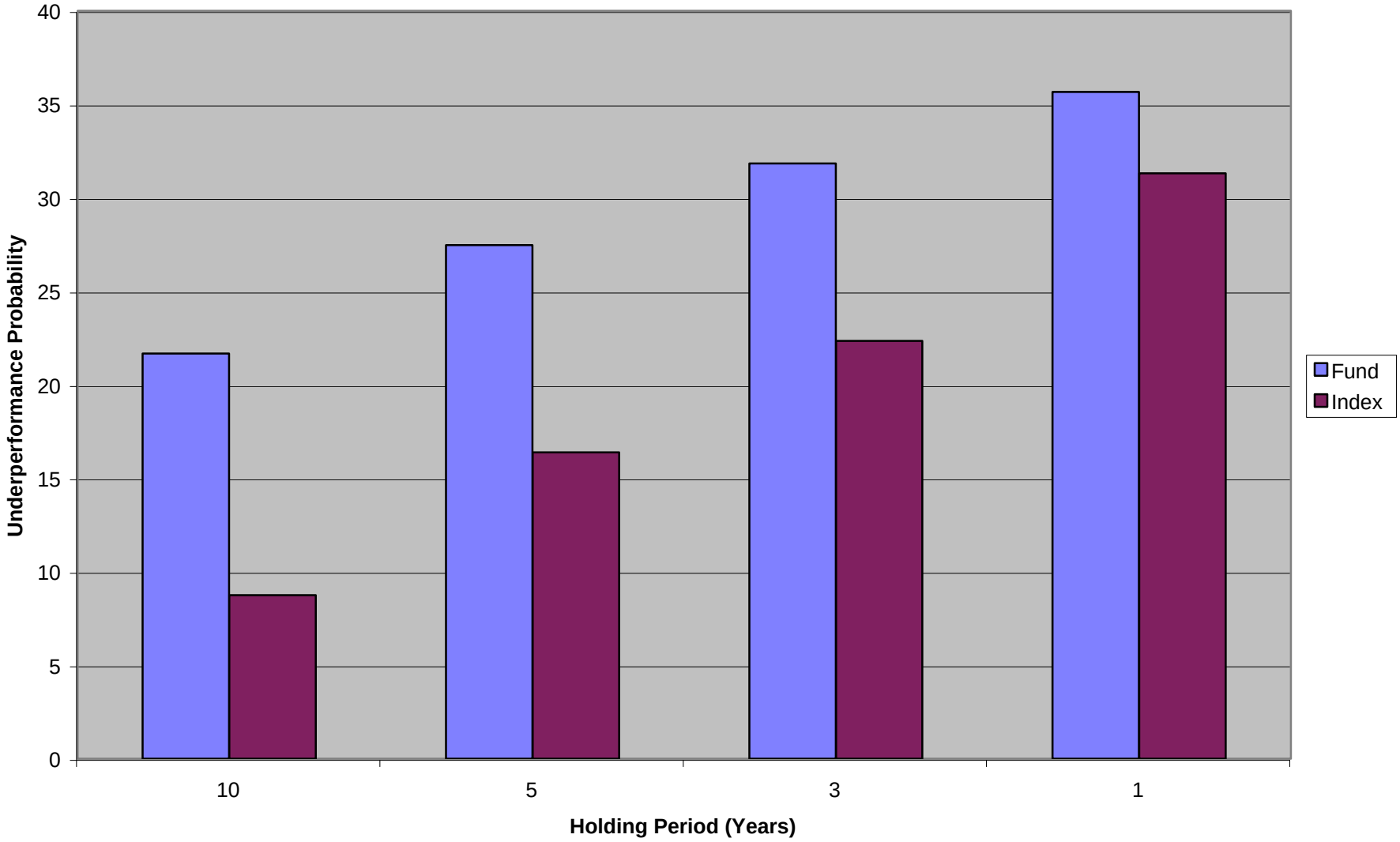
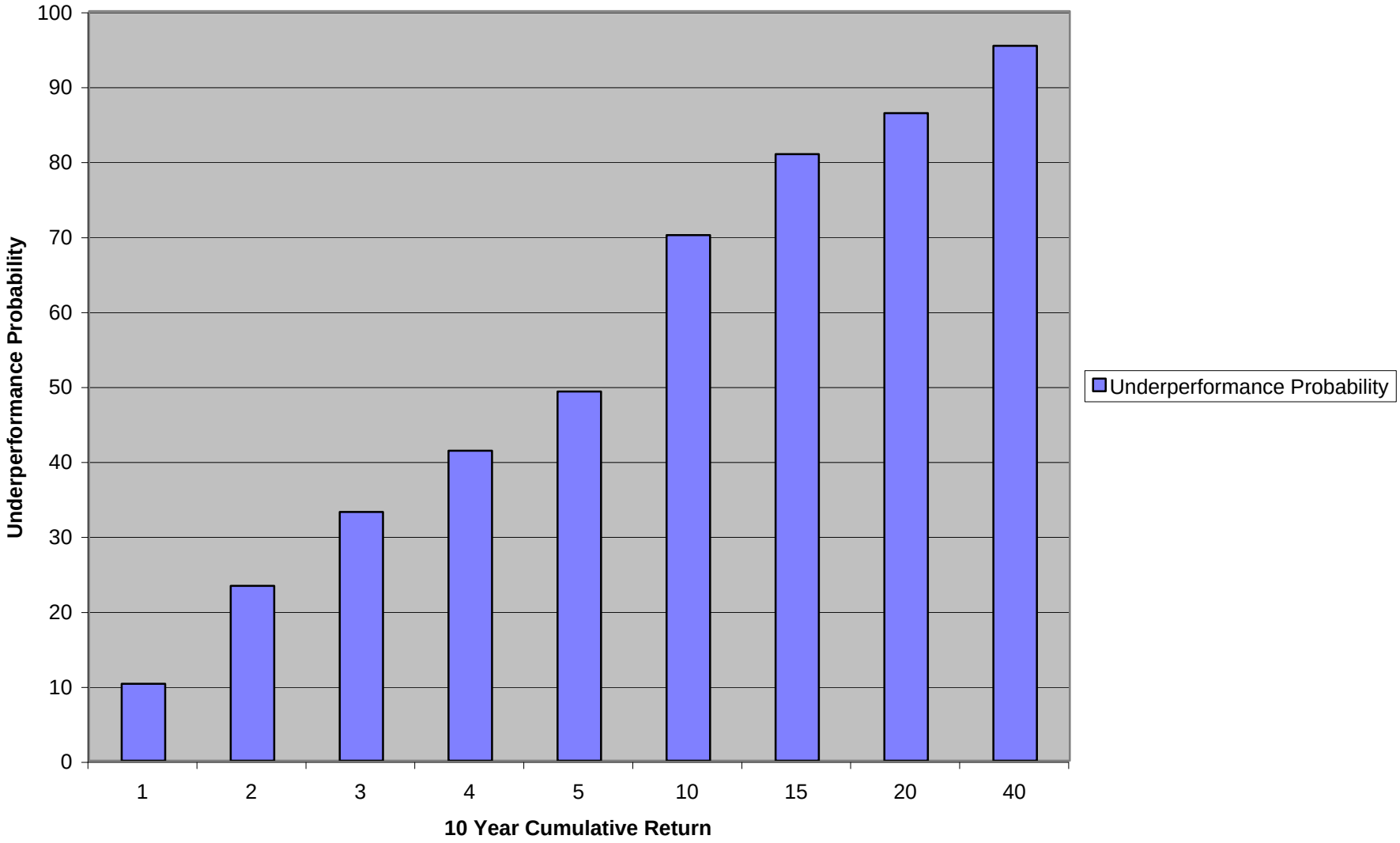


**Figure 5: Fund's 10 yr. Holding Period Return Possibilities**



## Figure 6: ALP with Bell-Shaped Returns

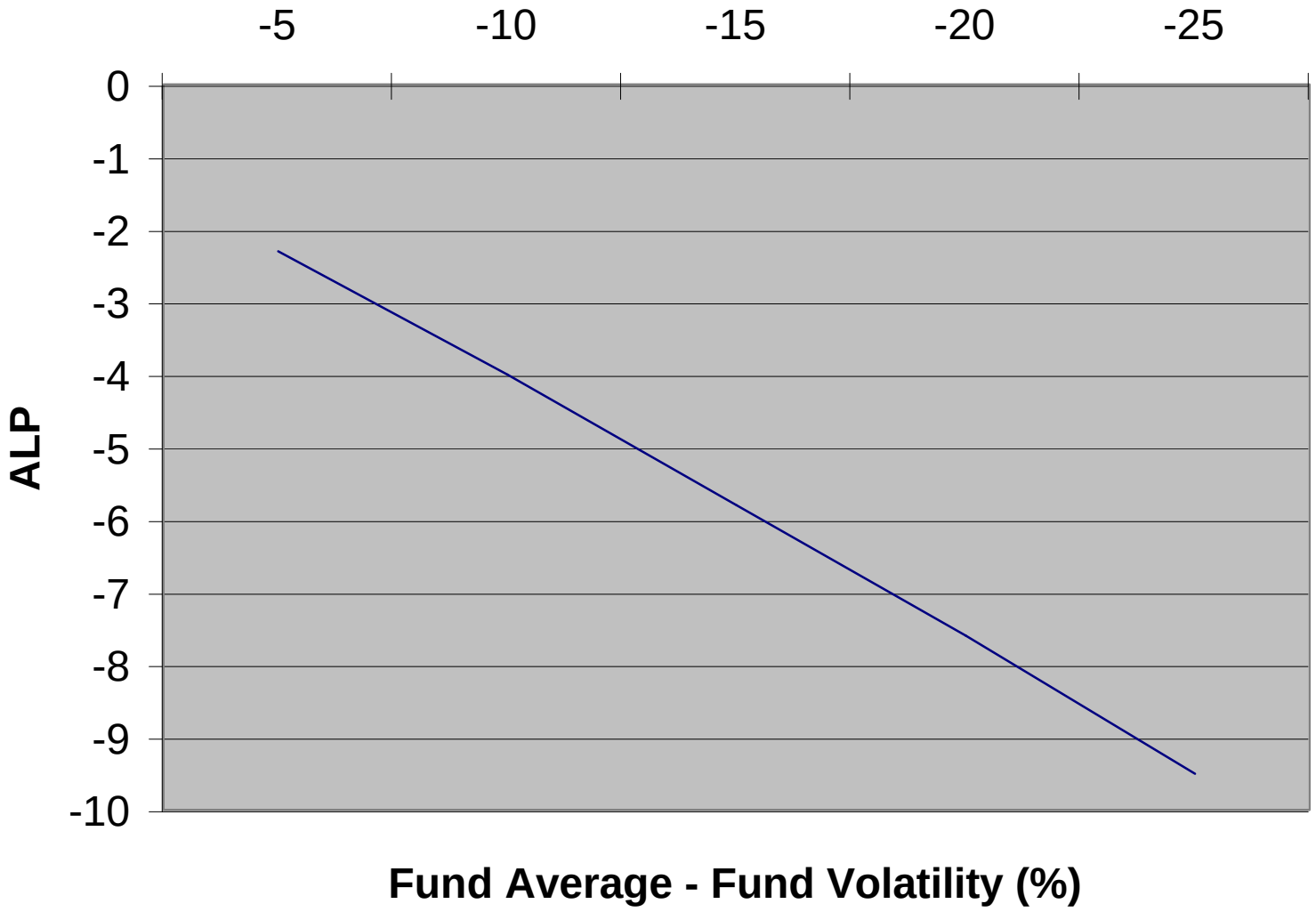


**Figure 7: Comparison of Rolling 10 Year Sharpe Ratios**

**Figure 8: Comparison of 3 Year Rolling Sharpe Ratios**

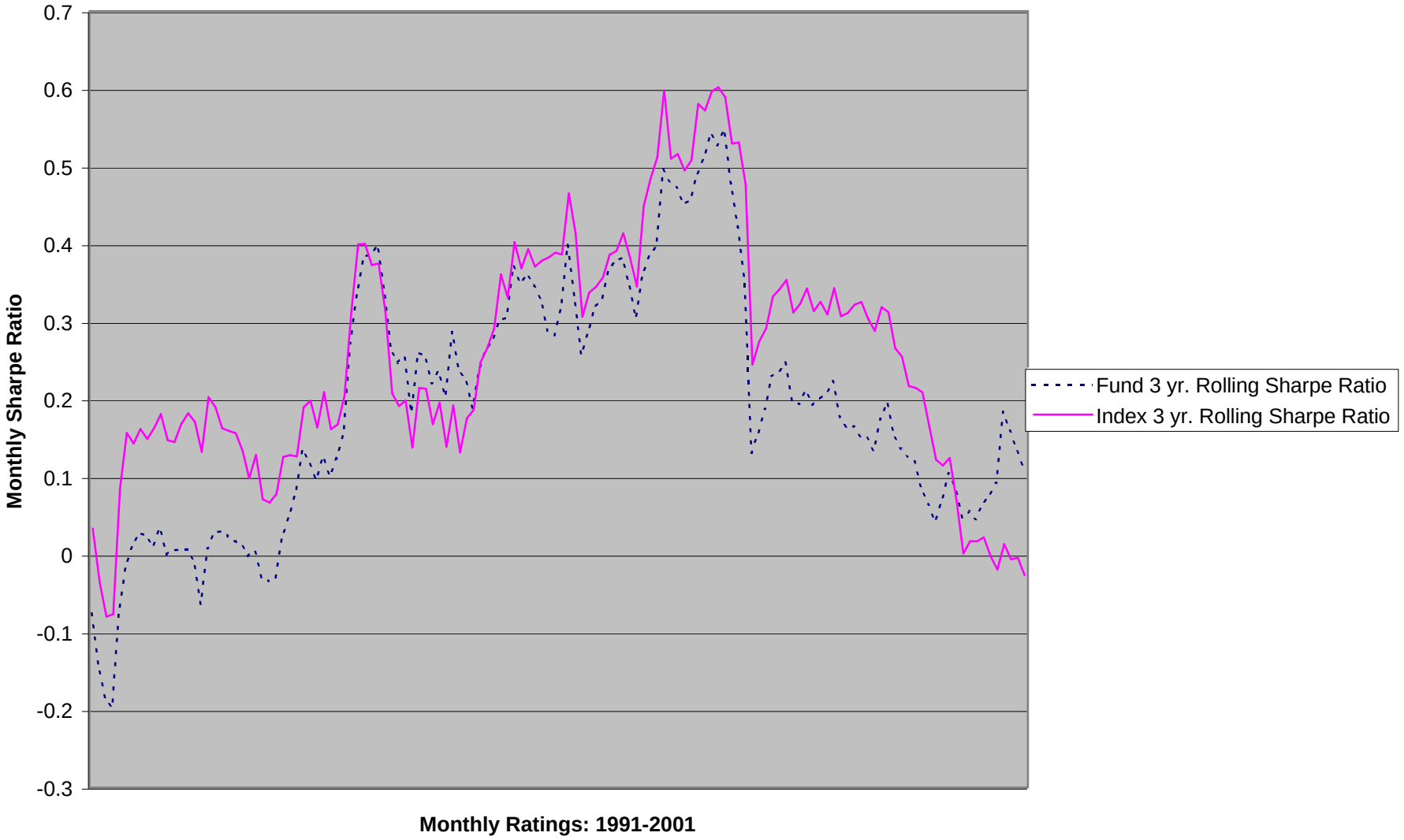


**Figure 9: Fund and Index Log Gross Returns**

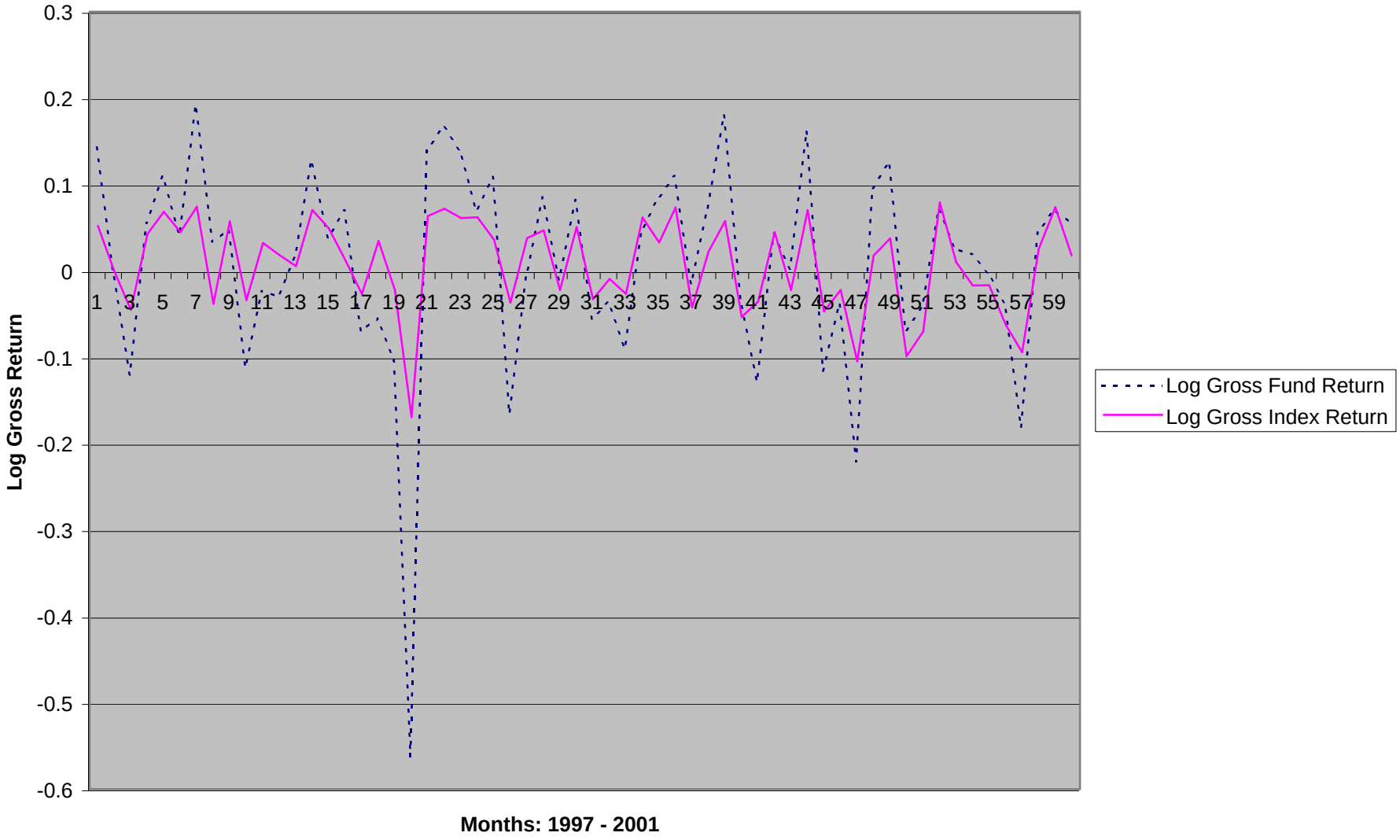
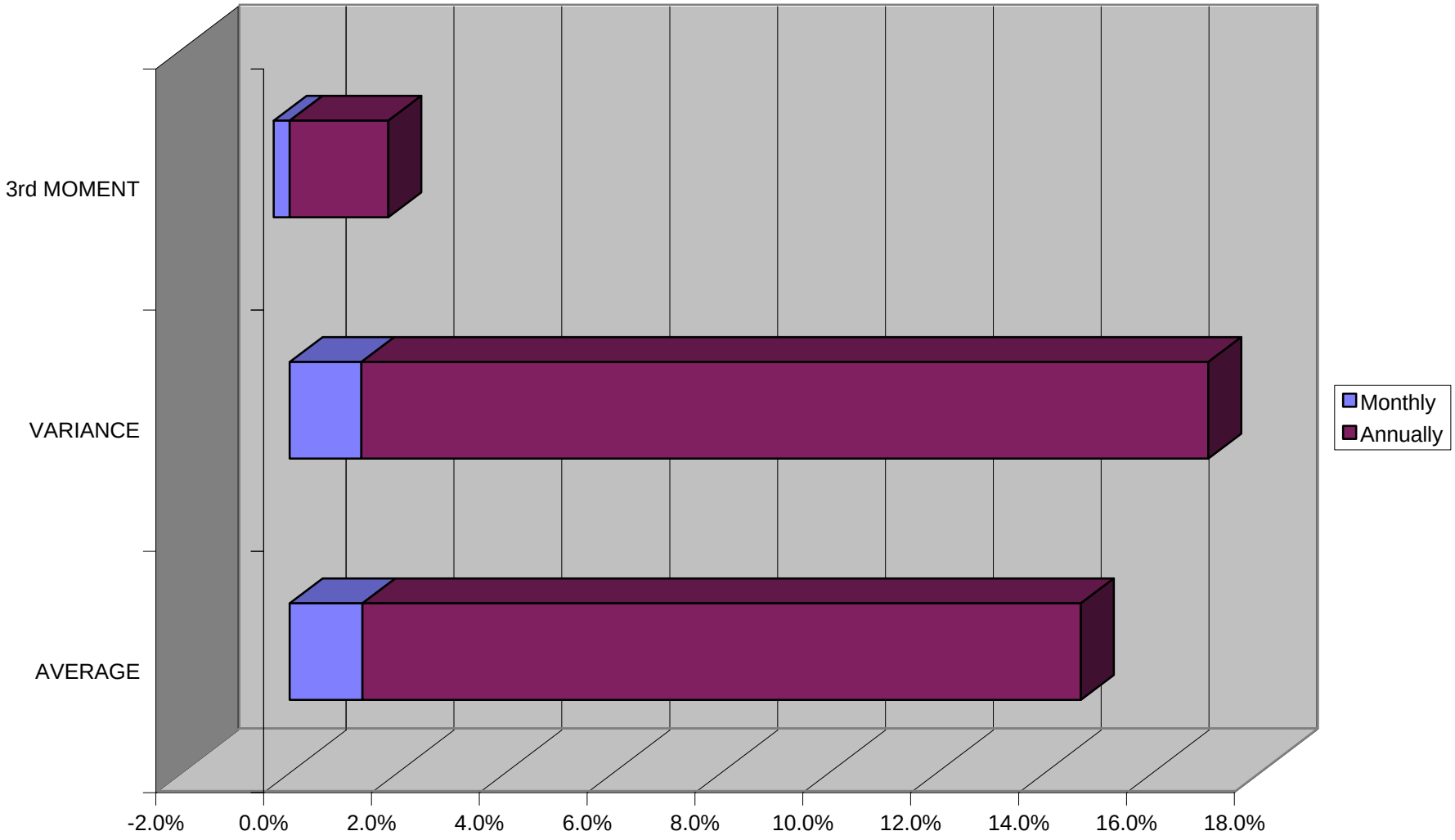


**Figure 10: Fund Returns' Vanishing 3rd Moment**



## References


[1] Takeshi Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

[2] Connie Becker, Wayne Ferson, David H. Myers, and Michael J. Schill. Conditional market timing with benchmark investors. *Journal of Financial Economics*, 52(1):119–148, 1999.

[3] Susan Belden and M. Barton Waring. Compared to what? a debate on picking benchmarks. *Journal of Investing*, 10(4):66–72, 2001.

[4] James A. Bucklew. *Large Deviation Techniques in Decision, Simulation, and Estimation*. Wiley, 1990.

[5] Andrew Clark. White paper on consistent return. Lipper Leaders Technical Document, Lipper, Inc., 2001.

[6] Diane Del Guercio and Paula A. Tkac. Star power: The effect of Morningstar ratings on mutual fund flows. Working Paper 2001-15, Research Department, Federal Reserve Bank of Atlanta.

[7] Frank J. Fabozzi, Francis Gupta, and Harry M. Markowitz. The legacy of modern portfolio theory. *Journal of Investing*, 11(3):7–23, 2002.

[8] F. Douglas Foster and Michael Stutzer. Performance and risk aversion of funds with benchmarks: A large deviations approach. Working Paper, University of Colorado Finance Department, 2002.

[9] Thomas H. Goodwin. The Information Ratio. *Financial Analysts Journal*, 54(4):34–43, 1998.

[10] E.L. Lehmann. *Elements of Large Sample Theory*. Springer-Verlag, 1999.

[11] Andrew W. Lo. The statistics of sharpe ratios. *Financial Analysts Journal*, 58(4):36–52, 2002.

[12] Andrew W. Lo and A. Craig MacKinlay. *A Non-Random Walk Down Wall Street*. Princeton University Press, 1999.

[13] David G. Luenberger. *Investment Science*. Oxford University Press, New York, 1998.

[14] Matthew R. Morey. Mutual fund age and morningstar ratings. *Financial Analysts Journal*, 54:56–63, March/April 2002.

[15] Morningstar. The new Morningstar rating methodology. Morningstar Research Report, April 22, 2002, Morningstar, Inc., Chicago, IL.

[16] James A. Ohlson. The asymptotic validity of quadratic utility. In W.T. Ziemba and R.G. Vickson, editors, *Stochastic Optimization Models in Finance*. Academic Press, 1975.

[17] Richard Roll. A mean/variance analysis of tracking error. *Journal of Portfolio Management*, 18(4):13–22, 1992.

[18] William Sharpe. The Sharpe Ratio. *Journal of Portfolio Management*, 21(1):49–58, 1994.

[19] William Sharpe. Morningstar's risk-adjusted ratings. *Financial Analysts Journal*, 54(4):21–33, 1998.

[20] Robert F. Stambaugh. On the exclusion of assets from tests of the two-parameter model: A sensitivity analysis. *Journal of Financial Economics*, 10(3):237–268, 1982.

[21] Michael Stutzer. Portfolio choice with endogenous utility: A large deviations approach. Journal of Econometrics, 2002(forthcoming).

[22] Michael Stutzer. A portfolio performance index. *Financial Analysts Journal*, 56(3):52–61, 2000.

[23] Jeff Tjornehoj. White paper on preservation. Lipper Leaders Technical Document, Lipper, Inc., 2001.