Performance and Risk Aversion of Funds with Benchmarks: A Large Deviations Approach [1]

F. Douglas Foster and Michael Stutzer

Professor of Finance, Australian Graduate School of Management, Sydney, AU.

Professor of Finance, University of Iowa, USA.

ABSTRACT

Mutual fund performance is often measured relative to a designated benchmark portfolio. This paper provides a simple method of ranking portfolios' probabilities of outperforming a benchmark portfolio. Ranking fund performance in this way is identical to ranking each fund's portfolio with an expected generalized power utility index, that uses a fund and benchmark-specific risk aversion parameter implied by the fund's portfolio choice. When the difference between funds' log returns and those of the benchmark are from different Gaussian processes, we derive different modifications of the *selection* Sharpe ratio (1994) associated with Roll's (1992) Tracking Error Variance (TEV)-efficiency notion. We develop feasible nonparametric and parametric estimators of a fund's performance index value, and the implied degree of risk aversion of the equivalent expected generalized power utility. We apply these estimators to rank the small fraction of mutual funds that (from the results of an hypothesis test) could outperform the S&P 500 index in the long run, and to estimate the implied degrees of risk aversion of their managers. Our procedure produces more plausible and precise estimates of managerial risk aversion than other recent estimates.

1 Introduction

As noted by Roll [23, p.13], " Today's professional money manager is often judged by total return performance relative to a prespecified benchmark, usually a broadly diversified index of assets." He argues that "This is a sensible approach because the sponsor's most direct alternative to an active manager is an index fund matching the benchmark." A typical example, of more than just professional interest to academic readers, is the following statement by the TIAA-CREF Trust Company:

> Different accounts have different benchmarks based on the client's overall objectives...Accounts for clients who have growth objectives with an emphasis on equities will be benchmarked heavily toward the appropriate equity index – typically the S&P 500 index – whereas an account for a client whose main objective is income and safety of principal will be measured against a more balanced weighting of the S&P 500 and the Lehman Corporate/Government Bond Index.[30, p.3]

How should plan sponsors and the investors they represent evaluate the performance of a fund like this? William Sharpe [27, p.32] asserts that

> The key information an investor needs to evaluate a mutual fund is (i) the fund's likely future exposures to movements in major asset classes, (ii) the likely added (or subtracted) return over and above a benchmark with similar exposures, and (iii) the likely risk vis-á-vis the benchmark.

Procedures for implementing this recommendation will differ, depending on the quantitative framework used for measuring "return over and above a benchmark" in (ii) and "risk vis-á-vis the benchmark" in (iii). Once these notions are defined precisely, they can be combined into a performance measure used to rank fund portfolios. *When returns are multivariate normally distributed*, the most common measure used is the *selection* Sharpe ratio [26, pp.50-1], [27, p.32], defined as:

$$E[R_p - R_b]/\sqrt{Var[R_p - R_b]} \tag{1}$$

where $R_p - R_b$ is the difference between the portfolio and benchmark's gross returns, measuring the "return over and above a benchmark" in (ii). The denominator in (1) measures the "risk vis-á-vis the benchmark" in (iii). Note that (1) reduces to the *conventional* Sharpe ratio when the benchmark portfolio is a constant return riskfree asset.[2] In the presence of a riskfree asset used to define the conventional Sharpe ratio, the risky asset portfolio that maximizes the ratio (i.e. the mean-variance efficient *tangency* portfolio) is the only risky asset portfolio held by mean-variance efficient investors. This provides a popular rationale for using it to rank the performance of risky asset portfolios. The natural generalization of mean-variance efficiency *relative to a benchmark the investor wants to beat* is Roll's [23] Tracking Error Variance (TEV)-efficiency, resulting from minimization of the tracking error variance $Var[R_p - R_b]$ subject to a constraint on the desired size of $E[R_p - R_b] > 0$. So in the presence of a riskfree asset, it is not surprising that the risky asset portfolio that maximizes the *selection* Sharpe ratio (1) is the only one held by TEV-efficient investors. When there isn't a riskfree asset, a TEV-efficient investor's risky asset portfolio depends on his/her degree of absolute risk aversion γ used in the following maximization of expected exponential utility:

$$\max_p E\left[-e^{-\gamma(R_p - R_b)}\right]. \tag{2}$$

These results have motivated Brennan [6], Gomez and Zapatero [13] and Becker, Ferson, et al. [3] to assume that (2) is a fund manager's criterion when ranking normally distributed portfolios. But several questions arise when considering the general legitimacy of using (2) to rank portfolios relative to a designated benchmark:

- The TEV logic underlying (2) is single-period. Is (2) an appropriate criterion for those trying to beat the designated benchmark over a long horizon?
- Returns aren't necessarily normally distributed. How should (2) be modified in light of this?
- An advantage of (1) is the absence of critical, exogenous preference parameters, like γ in (2). In the absence of a riskfree asset used to rationalize (1) via TEV-efficiency, can an index be found that obviates the need to know exogenous preference parameters?

This paper argues that the first bulleted question can be answered by ranking portfolios in accord with an index of the probability that they will outperform the benchmark over typical long-term investors' time horizons. Section 2 starts by using the Gärtner-Ellis Large Deviations Theorem [9, Chap.2] to show that the appropriate index is the following function:

$$D_p \equiv \max_{\gamma} -\lim_{T\to\infty} \frac{1}{T} \log E\left[e^{-\gamma\left(\sum_{t=1}^{T}(\log R_{pt} - \log R_{bt})\right)}\right]. \tag{3}$$

Section 2 also answers the second and third bulleted questions, by showing that ranking a portfolio p in accord with (3) is equivalent to ranking it in accord with the asymptotic expected *power* utility of the ratio of wealth invested in the portfolio to wealth that would be earned by alternatively investing in the benchmark. This utility has relative risk aversion equal to $1 + \gamma_p$, where γ_p denotes the maximized value of γ in (3), and hence does not need to be exogenously specified. However, when

approximation of time averaged log gross returns by arithmetic averaged net returns is reasonable (e.g. when $Var[R_p - R_b]$ is small) and when $R_p - R_b$ is IID, section 2.2 shows that the single-period exponential, TEV-based criterion (2) does arise from our criterion *without* assuming normality, by substituting our maximizing, endogenous γ_p from (3) for the γ in (2). We also show that that under the additional assumption of normality, this substitution of γ_p for γ in (2) reduces to the selection Sharpe ratio (1), *without* assuming the presence of a riskfree asset. *In this sense, the outperformance probability hypothesis nests the better-known criteria (1) and (2) as special cases, and hence is not subject to critiques commonly made of different probability-based criteria, e.g. expected utility maximization subject to "safety-first", Value-At-Risk (VAR) constraints.*[2]. Section 2.1 contains the appropriate modifications of the selection Sharpe ratio that arise under non-IID normality, e.g. in the popularly assumed GARCH(1,1) case. Then, we develop some historical returns-based estimators of the general performance index in section 3. Section 3.1.1 applies them to rank the relatively few mutual fund portfolios that, according to standard hypothesis tests, could outperform the S&P 500 in the long run. Section 3.1.2 shows that those funds' choices imply degrees of risk aversion γ_p for their managers that are both plausible and reasonably precise, in contrast to the implausibly large and imprecise managerial risk aversion estimates that Becker, Ferson, et al.[3] made by using (2) over the same data period. We identify features of their model that contributed to those results. Section 4 develops some consequences of the auxiliary hypothesis that fund managers act (either now or eventually) as-if they maximized our performance index, including a Lucas critique [20] of econometric specifications that, like Becker, Ferson, et al.(op.cit), treat the risk aversion coefficient as an exogenous parameter. Section 5 concludes with several future research topics that are directly suggested by our findings.

While some other connections to the portfolio choice and asset pricing literatures are made in the following section, the most closely related papers use other outperformance probability criteria, and are now summarized. Unlike our approach, Stutzer [28] is not based on the probability that the fund's cumulative return should exceed the benchmark's, and did not contain an empirical ranking of mutual funds. But its method is currently used by Morningstar, Inc. to produce its "Global Star Ratings" [16] of mutual funds, so our alternative approach should be of interest to performance analysts like them, as well as the fund managers they rank. Our approach also permits a stochastic benchmark, generalizing the constant growth rate benchmark used in the constantly rebalanced portfolio choice model of Stutzer [29]. The framework of that portfolio choice paper was used by Pham [21], to model optimal dynamic portfolio choice when a risky asset's returns are generated by the process adopted in Bielecki, Pliska, and Sherris [4]. Finally, Browne [8, Sec. 4] formulated a related, but more specific criterion for optimal dynamic portfolio choice. After imposing restrictive portfolio and benchmark parametric price process restrictions, he characterized the portfolio that maximized "the probability of beating the benchmark by some predetermined percentage, before going below it by some other predetermined percentage." In contrast, our outperformance probability criterion differs from his, by not assuming that managers are constrained to use such floors and ceilings, specific time horizons, nor specific parametric return processes. Moreover, we do not study optimal portfolio choice, but instead use our large deviations approach to derive the relationships between the outperformance probability performance criterion and more widely used, but (heretofore) seemingly unrelated, moment-based and expected utility-based performance criteria, and to derive and empirically apply estimators of a fund's performance and the degree of risk aversion implied by a fund's chosen portfolio.

2 An Index of Outperformance Probability

Ex-ante, wealth at some future time T arising from initial wealth W_0 invested in some portfolio strategy p will be $W_T^p = W_0 \prod_{t=1}^{T} R_{pt}$, where R_{pt} denotes the random gross return from the strategy between times $t-1$ and t. Note that the validity of this expression does not depend on the length of the time inteval between $t-1$ and t, nor the particular times t at which the random gross returns are measured. Similarly, an alternative investment of W_0 in a different portfolio b, dubbed the "benchmark", yields $W_T^b = W_0 \prod_{t=1}^{T} R_{bt}$. Taking logs and subtracting shows that:

$$\log W_T^p - \log W_T^b = \log \frac{W_T^p}{W_T^b} = \sum_{t=1}^{T} \log R_{pt} - \sum_{t=1}^{T} \log R_{bt}. \tag{4}$$

From (4), we see that the portfolio strategy p outperforms the benchmark b when and only when the sum of its log gross returns exceeds the benchmark's. Dividing both sides of (4) by T yields the following expression for the difference of the two continuously compounded growth rates of wealth:

$$\log W_T^p/T - \log W_T^b/T = \frac{1}{T} \log \frac{W_T^p}{W_T^b} = \frac{1}{T} \sum_{t=1}^{T} (\log R_{pt} - \log R_{bt}). \tag{5}$$

Suppose one wants to rank portfolios according to the rank ordering of their respective probabilities for the event that $W_T^p > W_T^b$. Using (5), this *outperformance event* is the event that (5) is greater than zero, i.e. that the portfolio's continuously compounded growth rate of wealth exceeds that of the benchmark. Hence one desires a rank ordering of the probabilities

$$Prob\left[\frac{1}{T} \sum_{t=1}^{T} (\log R_{pt} - \log R_{bt}) > 0\right]. \tag{6}$$

which is equivalent to ordering the complimentary probabilities from lowest to highest, i.e.*we seek to rank a portfolio strategy inversely to its underperformance probability*:

$$(7) \qquad Prob\left[\frac{1}{T}\sum_{t=1}^{T}(\log R_{pt} - \log R_{bt}) \leq 0\right].$$

Of course, the rank ordering of portfolio strategies via (7) could depend on the exact value of the investor's horizon T. Because it is difficult for performance analysts to determine an exact value of an investor's horizon (when one exists), and because short horizon investors may have different portfolio rankings than long horizon investors, let us try to develop a ranking of (7) that will be valid for all T greater than a suitably large T. This is similar in spirit to the motivation behind choice of an infinite horizon for the investor's objective in most consumption-based asset pricing models (for a survey, see Kocherlakota [19]), or in many portfolio choice models (e.g., in Grossman and Vila [14]). *Supporting evidence in Stutzer [29] shows that portfolios with relatively low underperformance probabilities (7) for suitably large T often also have relatively low underperformance probabilities for small T (or even* all *T) as well. Hence shorter term investors may also make use of the results.*

To produce this ranking, we use the following two step procedure. First, reasonably assume that investors are not interested in portfolio strategies that (almost surely) will not even beat the benchmark when given an infinite amount of time to do so; in that event, they would prefer investing in the benchmark to investing in the portfolio. Hence, one should restrict the ranking to portfolios for which the underperformance probability (7) approaches zero as $T \to \infty$. More formally, one need only rank portfolio strategies p for which

$$(8) \qquad \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}(\log R_{pt} - \log R_{bt}) > 0.$$

Inequality (8) requires that the so-called *ergodic mean* of the log portfolio gross returns exceeds the benchmark's (a.e.). In the familiar special case of IID returns, (8) requires that the portfolio's expected log gross return exceeds the benchmark's. In fact, if one weren't worried about the probability of underperforming at finite horizons T, applying the law of large numbers to the limit in (8) shows that the highest performing fund would be the one with highest expected log return, i.e. the "growth optimal" fund that maximizes the expected log utility of wealth. It will almost surely *asymptotically* generate more wealth than the benchmark and all other funds will. But this "time diversification" argument in favor of log utility is irrelevant, because as Rubinstein [25] so effectively demonstrated, significant underperformance probabilities persist over time spans well in excess of typical investors' retirement horizons. As he summarized: "The long run may be long indeed"!

So we will now quantify this downside, and formulate alternative utility-based formulations that reflect it. To do so, the second step of our procedure seeks a rank ordering index for the underperformance probability (7) of portfolios satisfying (8), i.e. those portfolios whose underperformance probabilities decay to zero as $T \to \infty$. Fortunately, the powerful, yet simply stated Gärtner-Ellis Large Deviations Theorem [9, Chap.2] is tailor-made for this purpose. This theorem shows that those portfolios' underperformance probabilities (7) will decay to zero as $T \to \infty$, at a portfolio-dependent exponential rate. As a result, the underperformance probability of a portfolio with a higher decay rate will approach zero faster as $T \to \infty$, and hence its complementary outperformance probability will approach 1 faster, so the portfolio will have a higher probability of outperforming for suitably large T. Again, it is important to emphasize that while this is formally an asymptotic criterion, in practice it will produce a ranking that applies to much shorter investor horizons T as well (see Stutzer [29] for substantial evidence establishing this). In summary, *the*

underperformance probability's rate of decay to zero as $T \to \infty$ is our proposed ranking index for portfolios. A portfolio whose underperformance probability decays to zero at a higher rate will be ranked higher than a portfolio with a lower decay rate.

Direct application of the Gärtner-Ellis Large Deviations Theorem [9, Chap.2] shows that the decay rate of the underperformance probability (7) is

$$D_p \equiv \max_{\theta} -\lim_{T\to\infty} \frac{1}{T} \log E\left[e^{\theta \sum_{t=1}^{T} (\log R_{pt} - \log R_{bt})}\right] \tag{9}$$

Under the restriction (8), the maximizing θ in (9) will be negative (see Stutzer [29]), so without loss of generality one may substitute a value $-\gamma$, where $\gamma > 0$. Hence the rank ordering index is the decay rate:

$$D_p = \max_{\gamma>0} -\lim_{T\to\infty} \frac{1}{T} \log E\left[e^{-\gamma \sum_{t=1}^{T} (\log R_{pt} - \log R_{bt})}\right] \tag{10}$$

An expected utility interpretation of (10) is found by first substituting (4) into it and simplifying to yield:

$$D_p = \max_{\gamma>0} \lim_{T\to\infty} -\frac{1}{T} \log E\left[\left(\frac{W_T^p}{W_T^b}\right)^{-\gamma}\right]. \tag{11}$$

Now multiply (11) by $-T$, exponentiate both sides, and multiply both sides by -1 to obtain the following large T approximation:

$$-e^{-D_p T} \approx E\left[-\left(\frac{W_T^p}{W_T^b}\right)^{-\gamma_p}\right] \tag{12}$$

where γ_p in (12) denotes the maximizing γ from (11). Note that the left hand side of (12) is increasing in D_p, so using D_p to rank order portfolios produces the same rank order as the expected generalized power utility on the right hand side of (12) for suitably large T.

There are two differences between the right hand side of (12) and a power utility of wealth with exogenous degree of relative risk aversion $1+\gamma$. First, the argument of the power function in (12) is not the wealth invested in the portfolio strategy, but instead is the ratio of it to the wealth that would have been earned if invested in the benchmark. This ratio is the state variable in the formulations of Browne [8]. It is analogous to the argument in the period utility of Abel's [1] "external habit formation", consumption-based criterion. Instead of our ratio of individual wealth to wealth created from an exogenous benchmark investment, its argument is the ratio of individual consumption to an exogenous benchmark function of aggregate (past) consumption ("keeping up with the Joneses"). In fact, when generalizing this argument to model other forms of this consumption externality, Gali [12, Footnote 2] noted that "such a hypothesis may be given an alternative interpretation: agents in the model can be thought of as professional "portfolio managers" whose performance is evaluated in terms of the return on their portfolio relative to the rest of managers and/or the market." While our infinite horizon criterion for pure investment (11) is perhaps more reminscent of the asymptotic growth of expected utility criterion $J_p = \lim_{T\to\infty} \frac{1}{T} \log E[(W_T^p)^\alpha]^{1/\alpha}$ used by Grossman and Zhou [15] and Bielecki, Pliska and Sherris [4] [3], it should be possible to adapt the argument leading to (11) to analyze the consumption/investment problem with consumption externalities. The meaning of the degree of risk aversion in all these models is not the usual aversion to mean preserving spreads of distributions of wealth or consumption, but rather of distributions of wealth or consumption *relative to a benchmark.*

The second and more unusual difference between our criterion and all these others is that the curvature parameter γ_p on the right hand side of (12) is determined by maximization of (10), and is hence dependent on the stochastic process for $\log R_{pt} - \log R_{bt}$. Hence for the purpose of outperformance probability analysis, analysts should not make the assumption that the degree of

risk aversion $1+\gamma_p$ is a constant, independent of the benchmark or investment opportunity set that funds face.[4]

Applicability of the Gärtner-Ellis Theorem requires that one maintain the assumptions that the limit in (10) exists (possibly as the extended real number $+\infty$) for all $\gamma > 0$, and is differentiable at any γ yielding a finite limit. Many log return processes adopted by analysts will satisfy these hypotheses, as will now be demonstrated by example.

2.1 Time Varying Gaussian Log Returns

In order to both illustrate the calculation (10) and to relate it to the selection Sharpe ratio, suppose that for each time t, $\log R_{pt} - \log R_{bt}$ is normally distributed, so that each partial sum $\sum_{t=1}^{T}(\log R_{pt} - \log R_{bt})$ in (10) is a normally distributed random variable with mean $\sum_t E[\log R_{pt} - \log R_{bt}]$ and variance $Var[\sum_t(\log R_{pt} - \log R_{bt})]$. But (10) is just -1 times the time average of the log moment generating functions of these normally distributed random variables, evaluated at the maximizing $-\gamma_p$. Hence in the Gaussian case, (10) is just the quadratic function

$$D_p = \max_{\gamma} \lim_{T\to\infty} \frac{\sum_{t=1}^{T} E[\log R_{pt} - \log R_{bt}]}{T}\gamma - \frac{1}{2}\frac{Var[\sum_{t=1}^{T}(\log R_{pt} - \log R_{bt})]}{T}\gamma^2. \tag{13}$$

This first term in (13) will exist whenever the ergodic mean (8) exists, while the second term will exist whenever the analogous ergodic variance exists. These are standard assumptions to make in econometric estimation. Setting the first derivative of (13) equal to zero and solving for the maximizing γ yields:

$$\gamma_p = \frac{\lim_{T\to\infty}\sum_{t=1}^{T} E[\log R_{pt} - \log R_{bt}]/T}{\lim_{T\to\infty} Var[\sum_{t=1}^{T}(\log R_{pt} - \log R_{bt})]/T} \tag{14}$$

which from assumption (8) is positive, as asserted earlier. Now substitute (14) back into (13) and rearrange to obtain the following underperformance probability decay rate in the Gaussian case:

$$D_p = \frac{1}{2}\left(\frac{\lim_{T\to\infty}\sum_{t=1}^{T} E[\log R_{pt} - \log R_{bt}]/T}{\sqrt{\lim_{T\to\infty} Var[\sum_{t=1}^{T} \log R_{pt} - \log R_{bt}]/T}}\right)^2. \tag{15}$$

Hence the Gaussian performance index (15) depends on the ratio of the long run mean excess log return to its long run standard deviation, and is hence a generalization of the usual Sharpe Ratio. This differs from the Gaussian (i.e. second order) approximation of the performance criterion in Grossman and Zhou [15] and Bielecki, Pliska, and Sherris [4], which (the latter paper shows) is the long run mean minus an exogenous risk aversion parameter times the long run variance.

In the important special case where $\log R_{pt} - \log R_{bt}$ process is the widely-assumed GARCH(1,1) process with Gaussian errors [5], i.e. $\log R_{pt} - \log R_{bt} = \mu + e_t$; $e_t \sim N(0, \sigma_t^2)$; and $\sigma_t^2 = \sigma^2 + \alpha e_{t-1}^2 + \beta\sigma_{t-1}^2$, the index (15) simplifies to:

$$D_p = \frac{1}{2}\left(\frac{\mu}{\sqrt{\frac{\sigma^2}{1-(\alpha+\beta)}}}\right)^2. \tag{16}$$

which exists when the process is stationary, i.e. $\alpha + \beta < 1$. In this case, the denominator in (15) exists and is equal to the unconditional variance $\frac{\sigma^2}{1-(\alpha+\beta)}$ in (16). Funds satisfying (8) will be those with $\mu > 0$, which can then be ranked in accord with (16). In section 3.1.1, we report the results of estimating (16) for an appropriate sample of mutual funds described there.

2.2 Familiar Performance Measures as Approximations

The single period, exponential ranking index (2) has been rationalized by its consistency with Roll's (op.cit.) TEV-efficiency when the difference in returns $R_{pt} - R_{bt}$ is IID normal. To obtain something akin to (2) by an approximation of our index (10), first approximate a log gross return $\log R$ by its net return $R - 1$, i.e. substitute $R_{pt} - R_{bt}$ for $\log R_{pt} - \log R_{bt}$ in (10). Now under the restriction

that the difference in the time series of equity returns is produced by a serially independent process, one obtains the following index:

$$\max_{\gamma} - \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \log E\left[e^{-\gamma(R_{pt}-R_{bt})}\right] \tag{17}$$

which under the additional restriction that the independent distributions are identically distributed reduces to the single-period index

$$-\log E\left[e^{-\gamma_p(R_p-R_b)}\right] \tag{18}$$

that rank orders portfolios in the same way as the single-period expected exponential utility:

$$E\left[-e^{-\gamma_p(R_p-R_b)}\right]. \tag{19}$$

Note that the index (19) is based on the exponential function used in the TEV-based index (2), despite the fact that the argument for it made no use of normality. But (19) is not the same as (2), because (19) uses the portfolio-dependent, maximizing γ_p when ranking portfolio p, rather than some constant value of γ used for all p. However, if we impose the additional restriction that is used to rationalize the TEV hypothesis, i.e. that $R_p - R_b$ is normally distributed, we can substitute the Gaussian (quadratic) log moment generating function into the equivalent problem (18) and solve to yield the following maximizing γ_p

$$\gamma_p = E[R_p - R_b]/Var[R_p - R_b]. \tag{20}$$

Substituting (20) back into that log moment generating function and rearranging yields

$$\frac{1}{2}\left(\frac{E[R_p - R_b]}{\sqrt{Var[R_p - R_b]}}\right)^2 \tag{21}$$

which is half the squared selection Sharpe ratio (1). Hence we see that under the log return approximation, and the IID normal process restriction used to rationalize the TEV hypothesis, use of the fund-specific maximizing γ_p (20) transforms the exponential utility index (2) into a parameter-free ranking index (21), that ranks the funds in the same order as the selection Sharpe ratio (1)! Hence our general index (10) or (11) may be viewed as a generalization of the TEV-based indices (1) and (2), to be used when the approximation of log returns by net returns, the IID assumption, and the normality assumption are unwarranted. Moreover, the log-modified, selection Sharpe ratio-based indices (15) and (16) that arise under different lognormal dependency assumptions are *not* subject to the two most common criticisms of the mean-variance analysis used to establish the usual Sharpe ratios. To quote Rubinstein [24], "Mean-variance results in discrete time are justified by either multivariate normality of security returns (which is inconsistent with limited liability) or quadratic utility (which is inconsistent with non-satiation beyond some level of wealth and implies increasing absolute risk aversion). Both of these assumptions have serious problems for some purposes."

Finally, to see what happens when we maintain the IID assumption without either the approximation of log returns by net returns nor the normality assumption, one can apply the IID restriction directly to the difference of log gross returns in (10), producing the following index:

$$\max_{\gamma} -\log E\left[\left(\frac{R_p}{R_b}\right)^{-\gamma}\right] \tag{22}$$

which yields the same rank ordering of portfolios as the expected power utility of the return ratio

$$E\left[-\left(\frac{R_p}{R_b}\right)^{-\gamma_p}\right], \tag{23}$$

instead of the expected exponential utility of the return difference (19).

3 Estimation of the Performance Index

The simplest estimator of the performance index (10) arises when one makes the additional assumption that the *excess log return process* $X_{pt} \equiv \log R_{pt} - \log R_{bt}$ is serially independent, but *not* necessarily identically distributed. As argued earlier, one need only rank the fund portfolios p that almost surely will outperform the benchmark, i.e. that satisfy (8). Under the additional identically distributed assumption, (8) requires that $E[X_p] > 0$, and (10) reduces to

$$D_p = \max_{\gamma>0} -\log E\left[e^{-\gamma X_p}\right] \tag{24}$$

Given an observed time series of past excess (relative to the benchmark) log returns, denoted $X_p(t)$, $t = 0, \ldots T$, one can consistently estimate (24) by substituting its sample analog, which is just:

$$\hat{D}_p = \max_{\gamma>0} -\log \frac{1}{T}\sum_{t=1}^{T} e^{-\gamma X_p(t)}. \tag{25}$$

It is important to note that (25) is still a sensible nonparametric estimator when X_{pt} is merely independent, rather than IID, because while we found that excess log returns of funds over the S&P 500's have low or zero autocorrelations, the assumption that they are identically distributed is less defensible. But if the analyst was truly confident that a GARCH model, or some other specific parametric stochastic process, generated $X_p(t)$, (10) could be calculated (like it was for Gaussian processes in section 2.1) and parametrically estimated for each fund p. This could be repeated for all funds to produce the ranking. Because of the popularity of the GARCH(1,1) model, the following section will contrast the estimates from (25) with estimates of (16) produced by estimating the GARCH(1,1) model preceding (16).

3.1 Empirical Results

It is useful to compare our fund performance and risk aversion estimates with results from other recent studies. To foster this comparison, we examined mutual funds during the 228 months starting in January 1976 and ending in December 1994. This coincides with the estimation period used in Becker, Ferson, et al. (op.cit), and is almost identical to the estimation period used in the mutual fund performance study of Wermers [31]. We also followed Wermers in using the CRSP Survival-Bias Free US Mutual Fund Database, originated by Mark Carhart [10]. After describing our results and quantifying the sampling error present in them, we will use them to re-examine some of the claims in Wermers, which of course were based on different performance criteria. We will then compare our findings about managerial risk aversion to the implausibly large managerial risk aversion parameter estimates that Becker, Ferson, et al. obtained, and identify features of both their model and estimation strategy that contributed to those poor estimates. We chose what is arguably the most common cited benchmark, the S&P 500 index.[5]

3.1.1 Fund Performance

In accord with the rationale presented in section 2, one should only rank funds (if any) that would asymptotically outperform the S&P 500, i.e. one should only rank funds that satisfy (8). We attempted to identify those funds in ways consistent with each of our proposed estimators. When the estimator (25) is used to rank the funds, the S&P 500 benchmark monthly returns $\log R_b(t)$ were subtracted from the corresponding monthly returns of each of equity mutual funds returns $\log R_p(t)$ to produce the historical time series $X_p(t)$, and used to conduct the usual *one-way* paired difference of means tests of the null hypothesis $H_0 : E[X_p] = 0$ versus the alternative hypothesis $H_1 : E[X_p] > 0$. The test statistic is $\overline{X_p} \equiv \sum_{t=1}^{228} X_p(t)/228$.

Performance analysts will not be surprised to find out that the null hypothesis was rejected in favor of the alternative ($t > 1.65$) for only 32 of the 347 CRSP mutual funds whose returns persisted from January 1976 to December 1994. Now if one were testing an hypothesis about whether or not a *typical* fund beat the S&P 500, a survival bias would result from examining only those funds that survived over that entire period. But that hypothesis is *not* being tested here. Even though the procedure here examines only those 347 funds that were skillful or lucky enough to survive those 228 months, a standard hypothesis test concludes that only 32 of the 347 could asymptotically outperform the S&P 500, strengthening our conclusion that relatively few funds will do so.

Our performance index is now used to rank those 32 funds. The test statistic and the performance index estimator (25) might be problematic for funds whose X_{pt} are serially correlated. But a standard test of the hypothesis that the 1st through 6 th autocorrelation correlation coefficients are all zero was rejected at the 5% level for only 8 of the 32 funds, and even those had low autocorrelation coefficients.

Summary statistics and performance rankings for those 32 funds are reported in Table 1, ranked in order of their estimated performance index values $\hat{D}_p\%$ from (25). A (naive) analyst who is only concerned with the terminal wealth in the funds at the end of the ranking period, i.e. the cumulative return over the ranking period, would *not* rank the funds by $\hat{D}_p\%$. Instead, that analyst would rank the funds in order of the average difference in log returns over the rating period (5), listed in the second column of Table 1 as $\overline{X}_p\%$. But $\overline{X}_p\%$ ranks the funds very differently than $\hat{D}_p\%$, which is used by analysts concerned with eventual underperformance that didn't happen during the ranking period. However, the two rankings do agree on the top-ranked fund 36450. Perhaps not surprisingly, this is the Fidelity Magellan fund. Our estimates show that its portfolio strategy resulted in the least probability (7) of underperforming the S&P 500 benchmark, because

the probability of underperforming it decays to zero as $T \to \infty$ at the highest estimated rate $\hat{D}_p = 4.95\%$ per month. Using the well-known, compounding "Rule of 72" approximation, the underperformance probability (7) will eventually be cut in half about every $72/4.95 \approx 15$ months. While the bottom-ranked fund 18050 should also eventually outperform the S&P 500 (due to rejection of its null $E[X_p] = 0$ with $t > 1.65$), its probability of underperformance is estimated to die off much slower, i.e. it will eventually halve only every $72/0.60 = 120$ months.

Table 1 also reports the comparable corresponding GARCH(1,1) parametric estimates of (16), for the six funds p for which the GARCH stationarity condition $\alpha_p + \beta_p < 1$ was satisfied in sample, the parametric hypothesis outperformance screen $\mu_p > 0$ at the level $t > 1.65$ was satisfied, and the GARCH error normality χ^2 screen was passed. Table 1 shows that the nonparametric and GARCH(1,1) estimates are close in four of those six cases, but not close for funds 37600 and 16820. The sample stationarity condition value for fund 37600 was .99, suggesting that it might not be stationary. Another fund that didn't pass the nonparametric difference in means test for inclusion in Table 1 had a relatively low performance index value (16) of 0.65. Given the poor fits of the GARCH(1,1) model, we focus on the nonparametric estimator in what follows.

We designed and conducted a bootstrap resampling study, to study the likely impact of sampling error on the stability of our nonparametrically estimated rankings. We resampled the 228 months (with replacement) 10,000 times, to construct alternative possible $X_p(t)$ series for each of the 32 funds that could (on the basis of the nonparametric hypothesis test) outperform the S&P 500. After each of the 10000 replications, we estimated the 32 funds performance index values (25) and ranked them. In each replication, we followed Brown and Goetzmann's [7] in classifying fund performance as either above ("high") or below ("low") the median performance for that replication. Figure 1 shows the nonparametric results for the funds listed in Table 1. The first panel of Figure 1 clearly

shows a high degree of stability where it is most needed, i.e. at the top end of the ranking. The highest rated fund (Fidelity Magellan) stayed above the median in virtually all replications. More detail is provided in the second panel of Figure 1, which lists each fund's four transition probabilities of moving from above (below) the median to above (below) the median over successive replications. The panel confirms that it is more common for low funds to stay low and high funds to stay high in successive replications; as expected, the top and bottom ranked funds experience the most stability in this sense.

It is fruitful to re-examine one of the findings highlighted in Wermers' [31] study of mutual fund performance, conducted over an almost identical period. Wermers (op.cit, p. 1686) asked the question: "Do higher levels of mutual fund trading result in higher levels of performance?" Wermers attempted to answer this question by constructing a hypothetical portfolio implemented by annually shifting money into funds that had relatively high turnover during the previous year. He concludes (op.cit, p. 1690): "Although these high-turnover funds have negative (but insignificant) characteristic-adjusted net returns, their average unadjusted net return over our sample period significantly beats that of the Vanguard Index 500 fund."

There are two questions left unanswered by Wermers' conclusion. First, we have shown that fund investors, who want their invested wealth to exceed that which would have accrued in the S&P 500 stocks, must not use the average net return as a performance measure. Rather, they must use the average log (1 + net return), akin to the geometric average, which can be significantly lower when returns are volatile. Second, Wermers examined a *strategy* of annually moving money from low turnover funds to high turnover funds. Significant load payments may pile up while doing this, but even if they did not, investors may also want to know whether or not it pays to buy-and-hold individual mutual funds that have relatively high average turnover.

As Wermers notes, annual turnover rates are reported for each fund. So we compiled the median of the 19 years' turnover rates for each of the 32 funds that could (on the basis of the standard hypothesis test) asymptotically outperform the S&P 500. These ranged from a high of 181% to a low of 15%; half of these 32 funds had median annual turnover below 65%. Half of the 347 funds that had the full 228 monthly returns reported for our data period had median annual turnover below 43%, as reported in Table 2. So it is fair to say that the 32 outperforming funds had somewhat higher turnover than the rest.

Also, half of the 32 funds in Table 2 had median annual stock allocations that exceeded 85%, compared to 80% for the 347 funds. Moreover, the 32 funds had expense ratios that were similar to the 347 funds; half of each group had a median annual expense ratio below 89 basis points. *Hence, the outperformance of the 32 funds does not appear to be associated with atypical stock allocations or expense ratios.*

3.1.2 Risk Aversion Estimates

The final column of Table 1 lists the estimates of the fund managers' implied degrees of risk aversion $1+\gamma_p$ from (25), and their bootstrapped standard errors (in parentheses). No fund's implied degree of risk aversion is implausibly large, nor are the estimates imprecise, i.e. the corresponding standard errors are considerably smaller than the estimates. Table 1 shows that there is positive, but not perfect, correlation between a fund manager's degree of risk aversion and performance ranking. To help see why, let us approximate the index by substituting the net return difference for the log return difference, and assume that the net return difference is IID Gaussian. Then, (20) shows that the manager's degree of risk aversion will be high whenever the ratio $E[R_p - R_b]/Var[R_p - R_b]$ is high. In this case, the approximated performance index is (21), i.e. half the squared selection Sharpe

ratio, which is half the ratio $E[R_p - R_b]^2/Var[R_p - R_b]$. So in this case, γ_p differs from D_p only because its numerator $E[R_p - R_b]$ is not squared. So there will be a *tendency* for performance and risk aversion to be directly related, but the relationship is not perfect. In other words, a manager's choice of a portfolio attaining relatively low underperformance probabilities (i.e. a relatively high decay rate of the probability) is a somewhat noisy signal that he/she acts *as-if* he/she is relatively highly averse to the risk of underperforming the benchmark. The most extreme outperformance would occur when $R_p = R_b + c$ with $c > 0$, i.e. when the portfolio return is perfectly correlated with the benchmark, but is always higher by a constant amount. In this case, $Var[R_p - R_b] = 0$, so the degree of risk aversion is infinite, and shorting the benchmark to purchase the portfolio would then provide an arbitrage opportunity – the ultimate in outperformance! The top-ranked fund 36450 (Fidelity Magellan) certainly did not choose a portfolio that was that good, but its outstanding performance is a noisy signal that its management acted as-if it was more risk averse ($1+\gamma_p = 13.5$) to the possibility of underperforming the S&P 500 than other funds were.

Our estimates of implied fund risk aversion in Table 1 are in stark contrast to the much larger and less precisely estimated values reported by Becker, Ferson, et al. (op.cit), despite the fact that they also examined fund returns over the same period we did. To understand the differences between our results and theirs, we need to highlight four major differences between their model and our model. They first assume that at each time t, each fund uses its own exogenous risk aversion coefficient γ to evaluate the expected exponential utility index (2) of the excess return in each period t:

$$E[-e^{-\gamma(R_{pt} - R_{bt})}] \tag{26}$$

Second, they assume that $R_{pt} - R_{bt}$ are conditionally Gaussian, in which case the expectation in

(26) specializes to:

$$E[R_{pt} - R_{bt}] - \frac{\gamma}{2} Var[R_{pt} - R_{bt}]. \tag{27}$$

which their manager maximizes period by period. In their framework, a manager's γ is fixed, so an equivalent index for their manager is $\gamma * (27)$. So under their time-varying Gaussian process assumption, comparison of their $\gamma * (27)$ and our (13) reveals that we substitute differences in moments of the *log* returns, because they are the relevant returns for earned wealth comparisons (i.e. the cumulative returns, as shown in section 2), and that we substitute the maximizing γ_p from (14) rather than their assumed exogenous γ. The result of our substitutions is the natural generalization (15) of the selection Sharpe ratio (1) appropriate to their time-varying Gaussian case.

Third, because of their desire to build and test a model of *market timing* funds, they additionally assumed that portfolio choices were restricted to choosing time-varying weights of the S&P 500 and the riskfree T-Bill, so that the portfolio's return is

$$R_{pt} = x_t R_{S\&Pt} + (1 - x_t) R_{ft}, \tag{28}$$

and they also assumed that the fund's benchmark portfolio is some unobservable, fixed weighted average of the S&P 500 and the T-Bill, i.e.

$$R_{bt} = h R_{S\&Pt} + (1 - h) R_{ft}. \tag{29}$$

Combining (28) and (29) shows that the argument of their ranking index (27) is:

$$R_{pt} - R_{bt} = (x_t - h)(R_{S\&Pt} - R_{ft}). \tag{30}$$

Fourth and finally, Becker, Ferson, et al. assume the manager substitutes the (conditional) mean and variance of (30) into (27), and then chooses the equity allocation fraction x_t to maximize (27)

at each point in time t, i.e.

$$x_t = h + \frac{E[R_{S\&Pt}] - R_{ft}}{\gamma Var[R_{S\&Pt}]}. \tag{31}$$

But we will now argue that their model's implication (31) will counterfactually lead to equity allocation fractions x_t that will be greater than 100% much of the time, unless the manager has implausibly high risk aversion γ and/or the fund's benchmark places an implausibly high weight $1-h$ on the T-Bill. Such funds would have to have been levered frequently, often using short positions in T-Bills to further augment the investors' funds totally placed in the index. Such leverage would be expected to occur occasionally in some funds, and often in a few funds, but in most funds it is unlikely to *typically* occur. To illustrate this counterfactual implication of (31), we will use Becker, Ferson, et al.'s own estimates for these parameters, for the "asset allocation" category of funds that they (reasonably) believed to behave most in accord with their model. Their Table 5 (op.cit, p. 139) reports summary statistics for individual asset allocator funds' estimated γ and h. The mean value of these fund managers' risk aversion coefficient estimates is an implausibly high $\gamma = 93.6$, while the mean value of the managers' respective *benchmarks'* T-Bill weights was a more sensible $1-h = 15\%$. Using their own claim that the unconditional moment ratio $E[R_{S\&P} - R_f]/Var[R_{S\&P}]$ has "a typical magnitude of approximately equal to two" (op.cit, p. 123), (31) predicts that a fund with these parameters would *on average* allocate a plausible fraction $x_t = 85\% + 2/95 = 87\%$ to equities. But note that had the risk aversion coefficient been any lower than $\gamma = 13$, their predicted equity allocation fraction would have exceeded 100% on average. The only way to avoid this counterfactual implication is for the fund's benchmark to place a higher weight $1-h$ on T-Bills than is plausible. For example, to be consistent with a plausible risk aversion value of $\gamma = 6$, the fund's benchmark would have to place a $1 - h = 2/6 = 33\%$ weight on the T-Bill return, in order

for the fund to avoid being leveraged on-average. How many actively managed funds would be allowed to have an easy benchmark like this?

The problem is analogous to the "equity premium puzzle" plaguing the standard CRRA, representative agent consumption-based asset pricing model [19]. In order for that model to be consistent with both the expected stock and T-Bill returns, its representative agent's risk aversion parameter must have an implausibly high value. Similarly, in order for the Becker, Ferson, et al. model to be consistent with a typical fund's observed stock weight and plausible T-Bill weighting in their benchmarks, its risk aversion parameter must also be implausibly high. Both models impose the ad-hoc restriction that the risk aversion parameter does not depend endogenously on the investment opportunity set. Our outperformance probability criterion necessitates elimination of that restriction.

Moreover, before trying to rank the *amount* by which funds outperformed the benchmark, we first tested whether they *could* asymptotically outperform the benchmark, i.e. was $E[\log R_p - \log R_b] > 0$? From (27), we immediately see that the analogous test in the Becker, Ferson, et al. model is $E[R_{pt} - R_{bt}] > 0$ for all t, because for those t where this didn't happen, their criterion function (27) shows that this manager should have chosen to invest in the benchmark rather than his/her portfolio, regardless of his/her degree of risk aversion. For plausibly low values of the benchmark's T-Bill asset weight $1 - h$ in (29), they surely would have discovered (like we did for $1 - h = 0$) that many managers failed the unconditional implication that $E[R_p - R_b] > 0$.

Further evidence of the Becker, Ferson, et al. model's misspecification is presented in their extremely low reported value of $R^2 = .04$ for the individual asset allocator fund models' fits. They report similarly problematic fits in every style subset of funds examined, which they acknowledge would not be expected to fit their market timing specification as closely as the asset allocator funds,

e.g. they might depend on style or stock selection to beat their benchmarks. They accurately conclude that "...the risk aversion estimates are imprecise and the power of the tests for timing ability seem low (op.cit, p. 145)." This conclusion was due to their frequent findings of standard errors that were not less than the parameter γ estimates themselves, as well as rejections of their models' GMM moment restrictions by Hansen's J-statistic.

In contrast, the last column in Table 1 shows that our estimates of rational managers' degrees of implied fund manager risk aversion $1+\gamma_p$ are neither very large nor imprecisely estimated, even though we used the same data period . Moreover, these results occurred for funds that are of most interest to both investors and performance analysts, i.e. those expected to asymptotically outperform a tough benchmark, i.e. an all S&P 500 benchmark with a T-Bill weight of $1-h=0$, as assumed in Roll's [23] original TEV hypothesis used to rationalize their period utility (27).

4 Outperformance Probability as a Fund Manager Criterion

Some differences in managers' risk aversion estimates in Table 1 are attributable to sampling error. But suppose fund managers *all* try to choose portfolios that maximize the outperformance probability. *If* (i) they all have the *same* benchmark and (ii) they all utilize the *same* investment opportunity set, then they would all choose the *same* portfolio $p_{\max} \equiv \arg\max_p D_p$, and would all have the *same* implied degree of risk aversion $\gamma_{p_{\max}}$. But some differences will undoubtedly be caused by violations of (i) or (ii). Condition (ii) is particularly unlikely, as it requires that the funds face the same restrictions (if any) on trading, and that these funds agree on all the conditional moments of each portfolio's log returns in excess of the benchmark's at each point in time, i.e. they must agree on the forms and parameters of all portfolios' excess log return processes. This

is unlikely to be true in practice, due to differences in managers' opinions about them.[6]

Of course, the same phenomenon arises when interpreting the standard investment textbooks' proposition that in the presence of a riskfree asset, all mean-variance investors will choose to hold risky assets in the same proportions as the Sharpe ratio maximizing "tangency" portfolio does, i.e. in the same proportions as the market portfolio in the CAPM. It is "as-if" everyone adopted the riskfree asset as their benchmark portfolio, and (ignoring the effects of multi-period accumulation of wealth and the consequent need to use log gross returns) ranked risky asset portfolios in accord with the Sharpe ratio. To paraphrase Sharpe [27, p.26] (for our purpose rather than his), such choices are clearly inconsistent with the observed behavior of the vast majority of investors, who do not hold risky assets in the same proportions as each other. Surely some of these differences are due to investors' different perceptions about the investment opportunity set.

Moreover, an implicit maintained hypothesis in estimating exogenous, differing γ models like Becker, Ferson, et al. is that all managers agree on the performance-relevant aspects of the investment opportunity set. One sees from (27) that *any* particular TEV-efficient portfolio is a rationalizable choice for *some* particular manager with a different level of γ, even if all managers make the same estimates of all portfolios' excess return expectations and variances. This makes it far more difficult for these models to *test* the hypothesis that managers make common estimates of those expectations and variances. More generally, how would one test specific hypotheses about differences in managers' opinions about particular investment styles, with models that permit attribution of their different portfolio choices to differences in their exogenous γ "parameters"? In contrast, an outperformance probability maximizing manager will always choose the same portfolio as another outperformance probability maximizing manager who shares his/her views about the investment opportunity set and is similarly constrained by rules or regulations. Our hypothesis

makes a sharper prediction in this dimension, and hence is more potentially falsifiable, just as the hypothesis of Sharpe ratio maximization makes sharper portfolio choice predictions than the general hypothesis of expected utility maximization. Subsequent to Popper's seminal work [22], scientists have considered a more potentially falsifiable hypothesis to be better than a less potentially falsifiable one, unless empirical evidence clearly favors the latter. Section 3.1.2 shows that the Becker, Ferson, et al. empirical evidence clearly does not favor their hypothesis over ours, so it appears (at this point) that Popper's logic favors our hypothesis over theirs.

Moreover, our outperformance probability analysis shows that the assumption of a manager-specific, exogenous γ is subject to the critique made in a justly celebrated paper by Robert E. Lucas [20]. He criticized econometric analyses that incorrectly hold "parameters" of an optimizing agent's decision rule fixed in the face of policy-induced changes in the decision making environment. He showed that these "parameters" would change when agents optimized more thoroughly.

Econometric analyses that fix γ as a managerial preference "parameter", like that in Becker, Ferson, et al., are subject to a similar critique. Our hypothesis implies that the optimal portfolio decision rule depends on this "parameter", which in turn depends on the decision making environment, i.e. the benchmark portfolio and investment opportunity set, via maximization of the outperformance probability index (10) or (11). Plan sponsors and/or their investors, who designate a fund benchmark for management to beat, are analogous to Lucas' policymakers. Should they designate a tougher benchmark, they should anticipate that the outperformance probability maximizing manager must take more (underperformance) risk in order to beat it, i.e. he/she will act as-if he/she had a lower degree of aversion to underperformance risk.

This critique is most starkly illustrated in the case of a single manager contracted to run two separate funds: one for a group of conservative investors who designated the 3 month T-Bill as the

benchmark portfolio, and the other for a group of investors who designated the S&P 500 as the benchmark portfolio.[7] The manager would quickly surmise that it is much easier to outperform a T-Bill benchmark than an S&P 500 benchmark, and that he/she should choose a much more conservative portfolio when managing the former portfolio in order to maximize (minimize) the probability of outperforming (underperforming) it over finite investor horizons. A priori, there is no reason for theorists to rule out the possibility that the manager acted as-if he/she used a higher coefficient of risk aversion when managing the former fund than he/she did when managing the latter. Stutzer [29] shows that this is precisely what will happen when the easy and hard benchmarks are deterministic, and this continues to be true here. While this explanation for the different choices is unusual, it follows from our power utility criterion (11), which was *derived* from the deeper hypothesis of outperformance maximization. The usual explanation *assumes rather than derives* a power utility, and then imposes the as yet empirically and experimentally unsupported *ad-hoc restriction* that its constant degree of risk aversion is completely exogenous to the manager's benchmark and the manager's views about the investment opportunity set available to beat it.

5 Conclusions

Mutual fund performance is often measured relative to a designated benchmark portfolio. This paper provides performance analysts with a simple index of funds' probabilities of outperforming a benchmark portfolio. We showed how our index reduces to use of the *selection* Sharpe (1994) ratio, consistent with Roll's (1992) TEV-efficiency hypothesis, under a restrictive IID Gaussian process approximation. We also showed how to modify the selection ratio to index the outperformance probability when funds' excess log returns (over the benchmark's) are generated by time-varying

Gaussian processes.

Without imposing approximations nor process restrictions, the outperformance probability index is equivalent to using an asymptotic expected generalized power utility, which differs in two ways from the familiar power utility of wealth. First, the argument of the utility function is the ratio of wealth earned in the fund to what would have otherwise been earned from investing in the benchmark. Second, the power utility's risk aversion "parameter" must, like portfolio choice itself, be determined by maximization. In this way, a fund manager's portfolio choice implies the fund's degree of aversion to the risk (quantified by the probability) that the fund will underperform the fund's benchmark over finite horizons. This surprising finding implies that if a fund manager seeks to maximize the outperformance probability, performance analysts must not assume that this manager's implied degree of risk aversion will be independent of the benchmark designated by the fund sponsor and/or investors, nor will it be independent of the manager's perceived investment opportunity set. Failure to account for this endogeneity of fund manager risk aversion exposes analysts to the celebrated Lucas Critique [20] of econometric policy evaluation.

In order to illustrate the feasibility and plausibility of this approach, we derived simple parametric and nonparametric estimators for the performance ranking index and its implied degree of risk aversion, and applied them to rank the performance of mutual funds that (based on standard hypothesis tests) could asymptotically outperform the S&P 500. We concluded that only 32 out of 347 funds will be able to asymptotically outperform the S&P 500, even though those 347 funds managed to survive the 19 year test period. Those that outperformed had overall equity allocations and expense ratios that were similar to those that didn't. The implied coefficients of risk aversion of the 32 outperforming funds ranged between 5.6 and 13.5. Not surprisingly, the highest ranked fund is the Fidelity Magellan fund, which also had the highest implied degree of risk aversion. Its

choice of a portfolio with the smallest probabilities of underperforming the S&P 500 (over finite horizons) is a signal that it acted as-if it had the highest aversion to the risk of underperforming it. We contrasted our funds' risk aversion estimates to the implausibly high and imprecise risk aversion estimates reported by Becker, Ferson, et al. (1999), and identified reasons for the stark differences between our results and theirs.

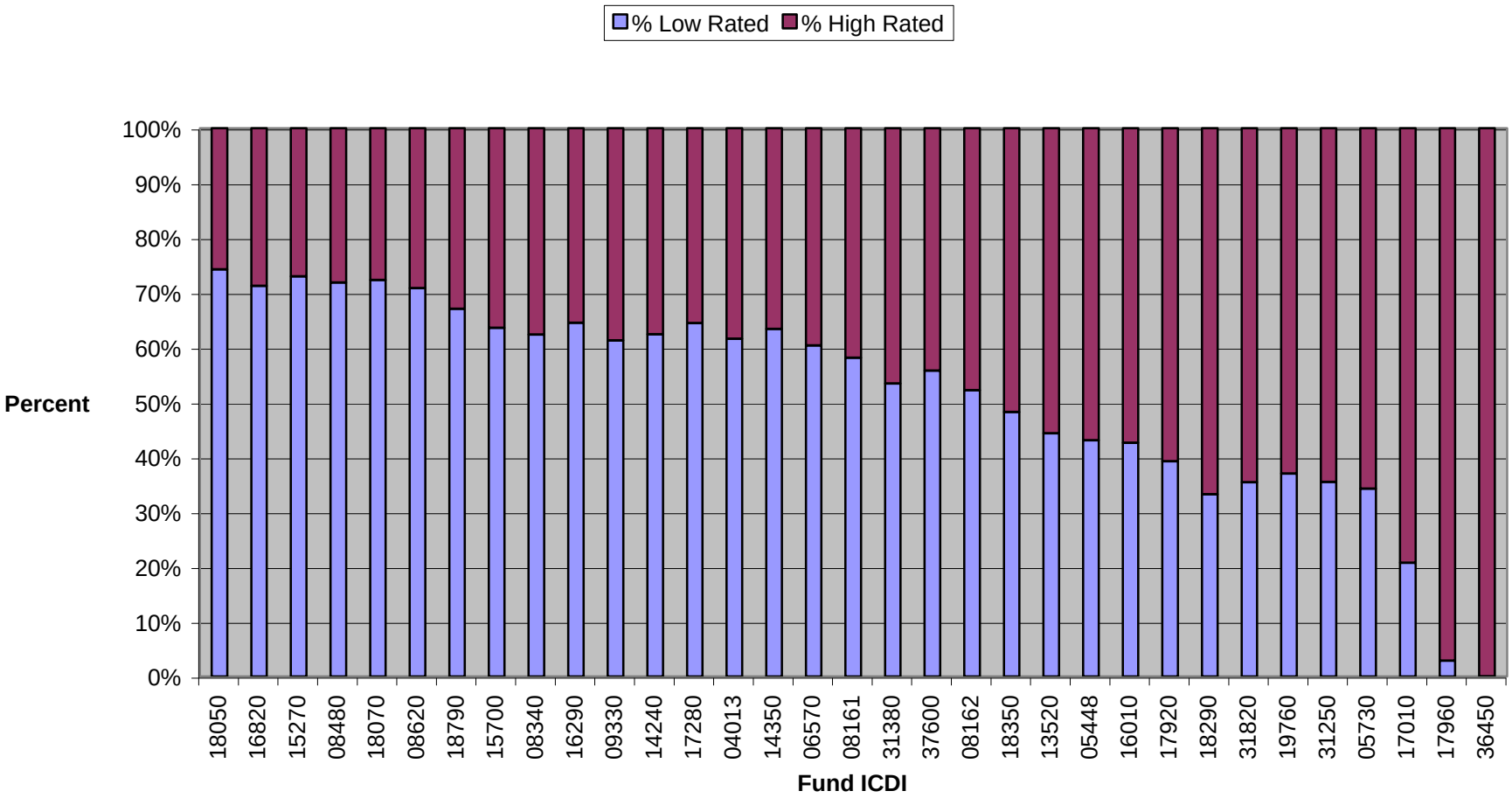
But there are many other problems left for future research. The most straightforward extension would be to adapt the estimation criterion in Kitamura and Stutzer [17] to permit nonparametric estimation of our performance index, in applications where log fund returns in excess of the benchmark's log returns are weakly dependent. Indeed, later analysis of this estimation criterion contained in Kitamura and Stutzer [18] showed that it is a decay rate criterion similar to our performance index. Estimation of weakly dependent processes may be more important when ranking fixed income funds relative to their benchmarks.

Our analysis also suggests some good theoretical topics. Under the auxiliary hypothesis that fund managers try to maximize our performance index, it might be possible to derive a non-IID Gaussian generalization of Brennan's (1993) altered CAPM-like risk-return relationship for an economy with benchmark investors. Another good topic is to formulate an analogous outperformance probability index for the more general consumption/investment problem, and eventually, an external habit formation-like consumption-based asset pricing relation [1] predicated on agents' solution of it. More realistic pricing is starting to emerge from such models, when agents are allowed to have (unexplained) heterogeneous degrees of risk aversion. Such heterogeneity arises naturally here, under the innocuous assumption that investors have different benchmarks, and/or under the assumption that investors perceive different investment opportunity sets, either through heterogeneous trading restrictions (e.g. short-sales) or views about assets' return process parameters.

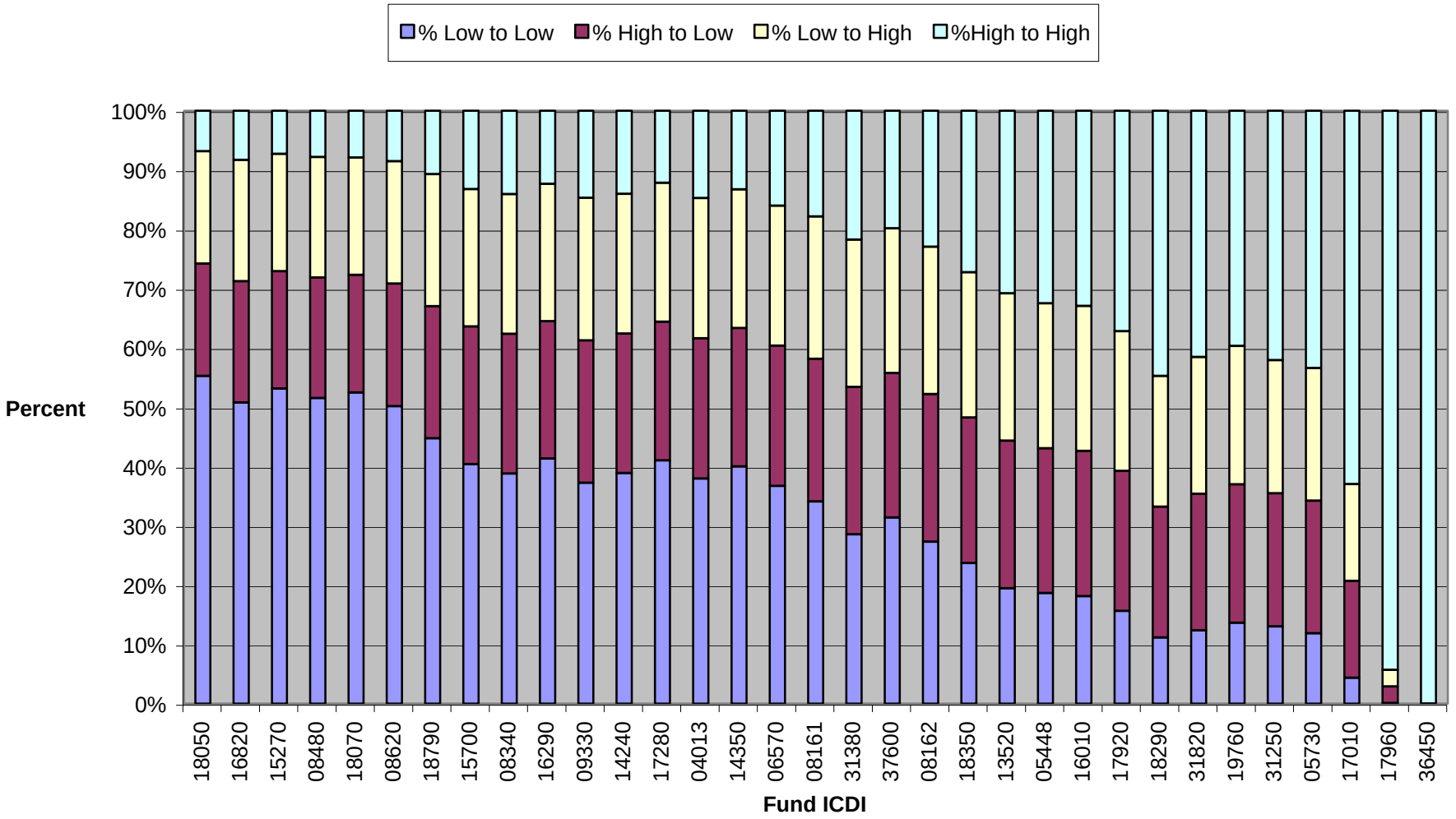
Mutual Funds with $E[X_p] > 0$							
CRSP #	$\overline{X_p}\%$	SDev %	Skew	Kurt	$\hat{D}_p\%$ (25)	$\hat{D}_p\%$ (16)	$1+\hat{\gamma}_p$
36450	.80	2.4	-.66	5.94	4.95		13.5 (5.5)
17960	.50	2.1	-.42	3.62	2.77		12.6 (4.1)
17010	.33	1.7	-.26	0.64	1.81	1.55	12.8 (4.2)
05730	.41	2.4	-.16	1.00	1.41		8.9 (2.9)
31250	.47	2.9	.36	1.51	1.38		7.9 (2.5)
19760	.42	2.6	.17	1.76	1.37		8.5 (2.8)
31820	.29	1.8	-.09	0.71	1.34	1.35	10.9 (3.8)
18290	.43	2.7	-.01	2.26	1.28		8.0 (2.7)
17920	.35	2.3	.33	3.25	1.23		9.0 (3.4)
16010	.25	1.6	-.02	2.31	1.19		11.4 (4.6)
05448	.25	1.6	-.03	2.31	1.19		11.4 (4.6)
13520	.26	1.7	.06	0.64	1.15		10.9 (4.1)
18350	.27	1.9	.92	4.15	1.07		10.2 (3.9)
08162	.40	2.8	-.065	5.11	0.97		5.8 (2.8)
37600	.33	2.4	.16	.80	0.95	0.62	7.8 (2.9)
31380	.37	2.7	-.79	6.43	0.94		6.9 (3.1)
08161	.55	4.1	-.72	5.40	0.86		5.1 (1.9)
06570	.27	2.1	-.09	1.34	0.83		8.2 (3.4)
14350	.26	2.0	-.37	2.30	0.82		8.3 (3.6)
04013	.44	3.5	-.01	0.86	0.81		5.7 (2.0)
17280	.03	2.3	-.30	1.63	0.81		7.4 (3.0)
14240	.38	3.0	-.74	4.04	0.80		6.1 (2.5)
09330	.47	3.7	-.51	1.52	0.78	0.88	5.3 (1.9)
16290	.27	2.2	-.39	2.19	0.76		7.6 (3.3)
08340	.24	2.0	.01	1.29	0.76		8.2 (3.5)
15700	.23	1.9	-.57	5.07	0.74		8.4 (4.1)
18790	.30	2.5	-.26	4.02	0.72		6.8 (3.0)
08620	.29	2.4	-.45	2.35	0.72		6.9 (3.0)
18070	.35	3.1	.19	3.45	0.66		5.8 (2.4)
08480	.35	3.1	-.30	.40	0.65	0.67	5.7 (2.2)
15270	.23	2.1	.42	3.35	0.64		8.3 (3.4)
16820	.30	2.7	-.28	0.97	0.63	1.06	6.2 (2.6)
18050	.33	3.0	-.12	1.27	0.60		5.6 (2.3)

Table 1: The 32 out of 347 CRSP mutual funds that could (on the basis of a standard hypothesis test) statistically significantly outperform the S&P 500 index during January 1976-December 1994. Rankings are in order of the nonparametrically estimated decay rate $\hat{D}_p$ from (25). No fund's implied degree of relative risk aversion $1+\hat{\gamma}_p$ from (25) is implausibly high, and all have standard errors (in parentheses) that are generally much lower than the estimates themselves, unlike the risk aversion estimates reported in Becker, Ferson, et al. (1999).

Fund Categorization



Persistence in Mutual Fund Rankings



Mutual Funds with $E[X_p] > 0$						
CRSP #	Fund	Turnover	Expense %	Stock %	Bond %	Cash %
36450	Fidelity Magellan	1.26	1.08	92	2	1
17960	Fidelity Destiny I	0.75	0.68	NA	NA	NA
17010	NY Venture	0.59	1.02	91	2	5
05730	Mutual Shares	0.67	0.74	67	21	13
31250	Lindner Growth	0.21	0.89	80	6	14
19760	Sequoia	0.32	1.00	66	0	27.2
31820	Guardian Park Ave.	0.48	0.72	82	10	7
18290	Acorn Inv. Trust	0.32	0.82	85	3	9
17920	Van Kamp-Amer Pace/A	0.45	1.01	86	0	14
16010	Neub &Berm Partners Fund	1.81	0.95	73	2	25
05448	Neub & Berm Partners Trust	NA	NA	NA	NA	NA
13520	Fidelity Equity Income	1.07	0.72	68	26	5
18350	Pioneer II	0.26	0.81	90	1	8
08162	20th Cent.Select	0.95	1.01	99	0	1
37600	SteinRoe Special	0.61	0.96	84	1	9
31380	Evergreen Fund/Y	0.48	1.13	92	0	8
08161	20th Cent.Growth	1.05	1.01	99	0	1
06570	Templeton Growth/I	0.15	0.75	87	4	10
14350	AMCAP	0.18	0.73	84	0	14
04013	Spectra	NA	NA	NA	NA	NA
17280	Nicholas	0.27	0.86	89	0	11
14240	Weingarten Equity	1.09	1.19	96	0	4
09330	Loomis-Sayles Cap. Dev.	1.29	0.79	98.6	0	1.4
16290	IDS New Dimensions/A	0.70	0.74	87	0	13
08340	Windsor	0.34	0.53	84	2	10
15700	Van Kamp-Amer Comstock/A	0.61	0.79	85	0	11
18790	Janus	1.63	1.02	71	1	27
08620	Growth Fund of America	0.22	0.76	81	0	16
18070	Value Line Leveraged Growth	1.13	0.91	95	0	3
08480	St. Paul Growth	0.84	1.00	84	0	16
15270	Charter	1.06	1.25	77	0	23
16820	Putnam Voyager/A	0.65	1.10	93	0	6
18050	Van Kamp-Amer Emerg.Grwth/A	0.93	1.05	91	0	8.30
	Median: All 347 Funds	**0.43**	**0.89**	**80**	**0**	**6**

Table 2: Comparison of the 32 outperforming funds' characteristics to the median characteristics of all 347 CRSP mutual funds over the period January 1976-December 1994. The 32 outperforming funds had slightly higher turnover and fractional allocation to equities than the 347 funds did, and similar expense ratios. Hence, the outperformance of the 32 funds does not appear to be associated with atypical stock allocations or expense ratios.

Notes

[1]The authors wish to acknowledge helpful comments from Tom Smith, Richard Heaney, Juan-Pedro Gomez, and seminar participants at the University of Minnesota and University of Colorado.

[2]Note that subtracting gross returns in (1) gives the same number as the more common subtraction of net returns.

[3]By algebra similar to that establishing (12), we derive its large T approximation to be $e^{\alpha J_p T}/\alpha \approx E[(W_T^p)^\alpha/\alpha]$, where $\alpha < 1$ and nonzero. Compared to the right hand side of (12), their utility also has CRRA (i.e. $1-\alpha$). But unlike the right hand side of (12), their utility is an unbounded power form, its curvature parameter α is assumed to be exogenous and independent of the portfolio return process, it has no benchmark in its argument, and it is a purely subjective criterion that has no objective outperformance probability interpretation.

[4]It is not obvious that use of this criterion as a ranking index over distributions of wealth relative to a benchmark violates one or more of the Von-Neumann/Morganstern axioms that imply expected utility representations of choices over lotteries. But even if it does, the frequent publication of papers utilizing Epstein and Zin's, Frisch Award-winning non-expected utility representation [11] indicates that this should not be a litmus test for publication.

[5]To obtain a total return benchmark, we used the CRSP value-weighted index, which includes distributions from the S&P 500 stocks.

[6]Readers doubting this point should watch a randomly chosen episode of the public television show "Wall Street Week".

[7]This is not unrealistic; many fund managers run more than one fund.

References

[1] Andrew Abel. Asset prices under habit formation and keeping up with the Joneses. *American Economic Review*, 80:38–42, 1990.

[2] Suleiman Basak and Alex Shapiro. Value-At-Risk based risk management: Optimal policies and asset prices. *Review of Financial Studies*, 14(2):371–405, 2001.

[3] 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.

[4] T. R. Bielecki, S.R. Pliska, and M. Sherris. Risk sensitive asset allocation. *Journal of Economic Dynamics and Control*, 24:1145–1177, 2000.

[5] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.

[6] Michael Brennan. Agency and asset pricing. Finance Working Paper no.6-93, Anderson Graduate School of Management, University of California, Los Angeles, 1993.

[7] Stephen J. Brown and William N. Goetzmann. Performance persistence. *Journal of Finance*, 50(2):679–698, 1995.

[8] Sid Browne. Beating a moving target: Optimal portfolio strategies for outperforming a stochastic benchmark. *Finance and Stochastics*, 3:275–294, 1999.

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

[10] Mark Carhart. On persistence in mutual fund performance. *Journal of Finance*, 52(1):57–82, 1997.

[11] Larry G. Epstein and Stanley E. Zin. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica*, 57(4):937–969, 1989.

[12] Jordi Gali. Keeping up with the Joneses: Consumption externalities, portfolio choice, and asset prices. *Journal of Money, Credit, and Banking*, 26(1):1–8, 1994.

[13] Juan-Pedro Gómez and Fernando Zapatero. Asset pricing implications of benchmarking: A two-factor CAPM. Working Paper, Finance Dept., Univ. of Southern California, March 1999.

[14] S. Grossman and J-L. Vila. Optimal dynamic trading with leverage constraints. *Journal of Financial and Quantitative Analysis*, 27(2):151–168, 1992.

[15] S. Grossman and Z. Zhou. Optimal investment strategies for controlling drawdowns. *Mathematical Finance*, 3(3):241–276, 1993.

[16] Paul Kaplan and Jim Knowles. The Stutzer performance index: Summary of mathematics, rationale, and behavior. Quantitative Research Dept., Morningstar, Inc., Chicago, IL, March 2001.

[17] Yuichi Kitamura and Michael Stutzer. An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65(4):861–874, 1997.

[18] Yuichi Kitamura and Michael Stutzer. Connections between entropic and linear projections in asset pricing estimation. *Journal of Econometrics*, 107(1):159–174, 2002.

[19] Narayanan R. Kocherlakota. The equity premium: It's still a puzzle. *Journal of Economic Literature*, 34(1):42–71, 1996.

[20] Robert E. Lucas. Econometric policy evaluation: A critique. In Karl Brunner and Allan Meltzer, editors, *The Phillips Curve and Labor Markets: Volume 1, Carnegie-Rochester Conference Series on Public Policy*. North-Holland, 1976.

[21] Huyen Pham. A large deviations approach to optimal long term investment. CNRS-UMR Working Paper No. 633, Laboratoire de Probabilités et Modèles Aléatoires, Universités de Paris, January 2001.

[22] Karl Popper. *The Logic of Scientific Discovery*. Routledge, 1977 (14th Printing).

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

[24] Mark Rubinstein. CRRA portfolio theory. Dept. of Finance, University of California at Berkeley.

[25] Mark Rubinstein. Continuously rebalanced investment strategies. *Journal of Portfolio Management*, 18(1):78–81, 1991.

[26] William Sharpe. Mutual fund performance. *Journal of Business*, 39(1):119–138, 1966.

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

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

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

[30] TIAA-CREF. TIAA-CREF Trust Company's personal touch. TIAA-CREF Investment Forum, September 2000.

[31] Russ Wermers. Mutual fund performance: An empirical decomposition into stock-picking talent, style, transactions costs, and expenses. *Journal of Finance*, 55(4):1655–1703, 2000.