Peer Reviewed Articles Showing Lack of Skill of Investment Managers

False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas

Laurent Barras, O. Scaillet, Russ Wermers

The Selection and Termination of Investment Management Firms by Plan Sponsors

Amit Goyal, Sunil Wahal

Luck versus Skill in the Cross Section of Mutual Fund Returns

Eugene F. Fama, Kenneth R. French

University of Maryland
Robert H. Smith School
RESEARCH PAPER NO. RHS 06-043
-

Swiss Finance Institute
RESEARCH PAPER NO. 08-18

False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas

Laurent Barras
Swiss Finance Institute
Imperial College London – Tanaka Business School

O. Scaillet
University of Geneva – HEC
Swiss Finance Institute

Russ R. Wermers
University of Maryland- Robert H. Smith School of Business

May 1, 2008

False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas*

Laurent Barras†, Olivier Scaillet‡, and Russ Wermers§

First version, September 2005; This version, May 2008

JEL Classification: G11, G23, C12

Keywords: Mutual Fund Performance, Multiple-Hypothesis Test, Luck, False Discovery Rate

*We are grateful to S. Brown, B. Dumas, M. Huson, A. Metrick, L. Pedersen, E. Ronchetti, R. Stulz, M.-P. Victoria-Feser, M. Wolf, as well as seminar participants at Banque Cantonale de Genève, BNP Paribas, Bilgi University, CREST, Greqam, INSEAD, London School of Economics, Maastricht University, MIT, Princeton University, Queen Mary, Solvay Business School, NYU (Stern School), Universita della Svizzera Italiana, University of Geneva, University of Georgia, University of Missouri, University of Notre-Dame, University of Pennsylvania, University of Virginia (Darden), the Imperial College Risk Management Workshop (2005), the Swiss Doctoral Workshop (2005), the Research and Knowledge Transfer Conference (2006), the Zeuthen Financial Econometrics Workshop (2006), the Professional Asset Management Conference at RSM Erasmus University (2008), the Joint University of Alberta/University of Calgary Finance Conference (2008), the annual meetings of EC^2 (2005), ESEM (2006), EURO XXI (2006), ICA (2006), AFFI (2006), SGF (2006), and WHU Campus for Finance (2007) for their helpful comments. We also thank C. Harvey (the Editor), the Associate Editor and the Referee (both anonymous) for numerous helpful insights. The first and second authors acknowledge financial support by the National Centre of Competence in Research "Financial Valuation and Risk Management" (NCCR FINRISK). Part of this research was done while the second author was visiting the Centre Emile Bernheim (ULB).

†Swiss Finance Institute and Imperial College, Tanaka Business School, London SW7 2AZ, UK. Tel: +442075949766. E-mail: l.barras@ic.ac.uk

‡Swiss Finance Institute at HEC-University of Geneva, Boulevard du Pont d'Arve 40, 1211 Geneva 4, Switzerland. Tel: +41223798816. E-mail: scaillet@hec.unige.ch

§University of Maryland - Robert H. Smith School of Business, Department of Finance, College Park, MD 20742-1815, Tel: +13014050572. E-mail: wermers@umd.edu

ABSTRACT

This paper uses a new approach to determine the fraction of truly skilled managers among the universe of U.S. domestic-equity mutual funds over the 1975 to 2006 period. We develop a simple technique that properly accounts for "false discoveries," or mutual funds which exhibit significant alphas by luck alone. We use this technique to precisely separate actively managed funds into those having (1) unskilled, (2) zero-alpha, and (3) skilled fund managers, net of expenses, even with cross-fund dependencies in estimated alphas. This separation into skill groups allows several new insights. First, we find that the majority of funds (75.4%) pick stocks well enough to cover their trading costs and other expenses, producing a zero alpha, consistent with the equilibrium model of Berk and Green (2004). Further, we find a significant proportion of skilled (positive alpha) funds prior to 1995, but almost none by 2006, accompanied by a large increase in unskilled (negative alpha) fund managers–due both to a large reduction in the proportion of fund managers with stockpicking skills and to a persistent level of expenses that exceed the value generated by these managers. Finally, we show that controlling for false discoveries substantially improves the ability to find funds with persistent performance.

Investors and academic researchers have long searched for outperforming mutual fund managers. Although several researchers document negative average fund alphas, net of expenses and trading costs (e.g., Jensen (1968), Lehman and Modest (1987), Elton et al. (1993), and Carhart (1997)), recent papers show that some fund managers have stock-selection skills. For instance, Kosowski, Timmermann, Wermers, and White (2006; KTWW) use a bootstrap technique to document outperformance by some funds, while Baks, Metrick, and Wachter (2001), Pastor and Stambaugh (2002b), and Avramov and Wermers (2006) illustrate the benefits of investing in actively-managed funds from a Bayesian perspective. While these papers are useful in uncovering whether, on the margin, outperforming mutual funds exist, they are not particularly informative regarding their prevalence in the entire fund population. For instance, it is natural to wonder how many fund managers possess true stockpicking skills, and where these funds are located in the cross-sectional *estimated* alpha distribution. From an investment perspective, precisely locating skilled funds maximizes our chances of achieving persistent outperformance.[1]

Of course, we cannot observe the *true* alpha of each fund in the population. Therefore, a seemingly reasonable way to estimate the prevalence of skilled fund managers is to simply count the number of funds with sufficiently high estimated alphas, $\widehat{\alpha}$. In implementing such a procedure, we are actually conducting a multiple (hypothesis) test, since we simultaneously examine the performance of several funds in the population (instead of just one fund).[2] However, it is clear that this simple count of significant-alpha funds does not properly adjust for luck in such a multiple test setting–many of the funds have significant estimated alphas by luck alone (i.e., their true alphas are zero). To illustrate, consider a population of funds with skills just sufficient to cover trading costs and expenses (zero-alpha funds). With the usual chosen significance level of 5%, we should expect that 5% of these zero-alpha funds will have significant estimated alphas–some of them will be unlucky ($\widehat{\alpha} < 0$) while others are lucky ($\widehat{\alpha} > 0$), but all will be "false discoveries"–funds with significant *estimated* alphas, but zero *true* alphas.

This paper implements a new approach to controlling for false discoveries in such a multiple fund setting. Our approach much more accurately estimates (1) the proportions of unskilled and skilled funds in the population (those with *truly* negative and positive

[1] From an investor perspective, "skill" is manager talent in selecting stocks sufficient to generate a positive alpha, net of trading costs and fund expenses.

[2] This multiple test should not be confused with the joint hypothesis test with the null hypothesis that all fund alphas are equal to zero in a sample. This test, which is employed by several papers (e.g., Grinblatt and Titman (1989, 1993)), addresses only whether at least one fund has a non-zero alpha among several funds, but is silent on the prevalence of these non-zero alpha funds.

alphas, respectively), and (2) their respective locations in the left and right tails of the cross-sectional *estimated* alpha (or estimated alpha t-statistic) distribution. One main virtue of our approach is its simplicity–to determine the proportions of unlucky and lucky funds, the only parameter needed is the proportion of zero-alpha funds in the population, π_0. Rather than arbitrarily imposing a prior assumption on π_0, our approach estimates it with a straightforward computation that uses the p-values of individual fund estimated alphas–no further econometric tests are necessary. A second advantage of our approach is its accuracy. Using a simple Monte-Carlo experiment, we demonstrate that our approach provides a much more accurate partition of the universe of mutual funds into zero-alpha, unskilled, and skilled funds than previous approaches that impose an a priori assumption about the proportion of zero-alpha funds in the population.[3]

Another important advantage of our approach to multiple testing is its robustness to cross-sectional dependencies among fund estimated alphas. Prior literature has indicated that such dependencies, which exist due to herding and other correlated trading behaviors (e.g., Wermers (1999)), greatly complicate performance measurement in a group setting. However, Monte Carlo simulations show that our simple approach, which requires only the (alpha) p-value for each fund in the population–and not the estimation of the cross-fund covariance matrix–is quite robust to such dependencies.

We apply our novel approach to the monthly returns of 2,076 actively managed U.S. open-end, domestic-equity mutual funds that exist at any time between 1975 and 2006 (inclusive), and revisit several important themes examined in the previous literature. We start with an examination of the long-term (lifetime) performance (net of trading costs and expenses) of these funds. Our decomposition of the population reveals that 75.4% are zero-alpha funds–funds having managers with some stockpicking abilities, but that extract all of the rents generated by these abilities through fees. Among remaining funds, only 0.6% are skilled (true $\alpha > 0$), while 24.0% are unskilled (true $\alpha < 0$). While our empirical finding that the majority are zero-alpha funds is supportive of the long-run equilibrium theory of Berk and Green (2004), it is surprising that we find so many truly negative-alpha funds–those that overcharge relative to the skills of their managers. Indeed, we find that such unskilled funds underperform for long time periods, indicating that investors have had some time to evaluate and identify them as underperformers.

We also find some notable time trends in our study. Examining the evolution of

[3]The reader should note the difference between our approach and that of KTWW (2006). Our approach simultaneously estimates the prevalence and location of outperforming funds in a group, while KTWW test for the skills of a single fund that is chosen from the universe of alpha-ranked funds. As such, our approach examines fund performance from a more general perspective, with a richer set of information about active fund manager skills.

each skill group between 1990 and 2006, we observe that the proportion of skilled funds dramatically decreases from 14.4% to 0.6%, while the proportion of unskilled funds increases sharply from 9.2% to 24.0%. Thus, although the number of actively managed funds has dramatically increased, skilled managers (those capable of picking stocks well enough to overcome their trading costs and expenses) have become increasingly rare.

Motivated by the possibility that funds may outperform over the short-run, before investors compete away their performance with inflows, we conduct further tests over five-year subintervals–treating each five-year fund record as a separate "fund." Here, we find that the proportion of skilled funds equals 2.4%, implying that a small number of managers have "hot hands" over short time periods. These skilled funds are located in the extreme right tail of the cross-sectional estimated alpha distribution, which indicates that a *very* low p-value is an accurate signal of short-run fund manager skill (relative to pure luck). Across the investment subgroups, Aggressive Growth funds have the highest proportion of managers with short-term skills, while Growth & Income funds exhibit no skills.

The concentration of skilled funds in the extreme right tail of the estimated alpha distribution suggests a natural way to choose funds in seeking out-of-sample persistent performance. Specifically, we form portfolios of right-tail funds that condition on the frequency of "false discoveries"–during years when our tests indicate higher proportions of lucky, zero-alpha funds in the right tail, we move further to the extreme tail to decrease false discoveries. Forming such a false discovery controlled portfolio at the beginning of each year from January 1980 to 2006, we find a four-factor alpha of 1.45% per year, which is statistically significant. Notably, we show that this luck-controlled strategy outperforms prior persistence strategies used by Carhart (1997) and others, where constant top-decile portfolios of funds are chosen with no control for luck.

Our final tests examine the performance of fund managers before expenses (but after trading costs) are subtracted. Specifically, while fund managers may be able to pick stocks well enough to cover their trading costs, they usually do not exert direct control over the level of fund expenses and fees–management companies set these expenses, with the approval of fund directors. We find evidence that indicates a very large impact of fund fees and other expenses. Specifically, on a pre-expense basis, we find a much higher incidence of funds with positive alphas–9.6%, compared to our above-mentioned finding of 0.6% after expenses. Thus, almost all outperforming funds appear to capture (or waste through operational inefficiencies) the entire surplus created by their portfolio managers. It is also noteworthy that the proportion of skilled managers (before expenses) declines substantially over time, again indicating that portfolio managers with skills have become increasingly rare. We also observe a large reduction in the proportion of unskilled funds when we move from net alphas to pre-expense alphas (from 24.0% to 4.5%), indicating a big role for excessive fees (relative to manager stockpicking skills) in underperforming funds. Although industry sources argue that competition among funds has reduced fees and expenses substantially since 1980 (Rea and Reid (1998)), our study indicates that a large subgroup of investors appear to either be unaware that they are being overcharged (Christoffersen and Musto (2002)), or are constrained to invest in high-expense funds (Elton, Gruber, and Blake (2007)).

The remainder of the paper is as follows. The next section explains our approach to separating luck from skill in measuring the performance of asset managers. Section 2 presents the performance measures, and describes the mutual fund data. Section 3 contains the results of the paper, while Section 4 concludes.

I The Impact of Luck on Mutual Fund Performance

A Overview of the Approach

A.1 Luck in a Multiple Fund Setting

Our objective is to develop a framework to precisely estimate the fraction of mutual funds in a large group that truly outperform their benchmarks. To begin, suppose that a population of M actively managed mutual funds is composed of three distinct performance categories, where performance is due to stock-selection skills. We define such performance as the ability of fund managers to generate superior model alphas, net of trading costs as well as all fees and other expenses (except loads and taxes). Our performance categories are defined as follows:

- **Unskilled funds**: funds having managers with stockpicking skills insufficient to recover their trading costs and expenses, creating an "alpha shortfall" ($\alpha < 0$),
- **Zero-alpha funds**: funds having managers with stockpicking skills sufficient to just recover trading costs and expenses ($\alpha = 0$),
- **Skilled funds**: funds having managers with stockpicking skills sufficient to provide an "alpha surplus," beyond simply recovering trading costs and expenses ($\alpha > 0$).

Note that our above definition of skill is one that is relative to expenses, and not in an absolute sense. This definition is driven by the idea that consumers look for mutual funds that deliver surplus alpha, net of all expenses.[4]

[4] However, perhaps a manager exhibits skill sufficient to more than compensate for trading costs, but the fund management company overcharges fees or inefficiently generates other services (such as administrative services, e.g., record-keeping)–costs that the manager usually has little control over. In

Of course, we cannot observe the true alphas of each fund in the population. Therefore, how do we best infer the prevalence of each of the above skill groups from performance estimates for individual funds? First, we use the t-statistic, $\widehat{t}_i = \widehat{\alpha}_i/\widehat{\sigma}_{\widehat{\alpha}_i}$, as our performance measure, where $\widehat{\alpha}_i$ is the estimated alpha for fund i, and $\widehat{\sigma}_{\widehat{\alpha}_i}$ is its estimated standard deviation–KTWW (2006) show that the t-statistic has superior properties relative to alpha, since alpha estimates have differing precision across funds with varying lives and portfolio volatilities. Second, after choosing a significance level, γ (e.g., 10%), we observe whether $\widehat{t}_i$ lies outside the thresholds implied by γ (denoted by t_γ^- and t_γ^+), and label it "significant" if it is such an outlier. This procedure, simultaneously applied across all funds, is a multiple-hypothesis test:

$$
\begin{aligned}
H_{0,1} &: \alpha_1 = 0, \quad H_{A,1} : \alpha_1 \neq 0, \\
... &: ... \\
H_{0,M} &: \alpha_M = 0, \quad H_{A,M} : \alpha_M \neq 0. \qquad (1)
\end{aligned}
$$

To illustrate the difficulty of controlling for luck in this multiple test setting, Figure 1 presents a simplified hypothetical example that borrows from our empirical findings (to be presented later) over the last five years of our sample period. In Panel A, individual funds within the three skill groups–unskilled, zero alpha, and skilled–are assumed to have true annual four-factor alphas of -3.2%, 0%, and 3.8%, respectively (the choice of these values is explained in Appendix B).[5] The individual fund t-statistic distributions shown in the panel are assumed to be normal for simplicity, and are centered at -2.5, 0, and 3.0 (which correspond to the prior-mentioned assumed true alphas; see Appendix B).[6] The t-distribution shown in Panel B is the cross-section that (hypothetically) would be observed by a researcher. This distribution is a mixture of the three skill-group distributions in Panel A, where the weight on each distribution is equal to the proportion of zero-alpha, unskilled, and skilled funds, respectively, in the population of mutual funds (specifically, $\pi_0 = 75\%$, $\pi_A^- = 23\%$, and $\pi_A^+ = 2\%$; see Appendix B).

Please insert Figure 1 here

a later section (III.D.1), we redefine stockpicking skill in an absolute sense (net of trading costs only) and revisit some of our basic tests to be described.

[5]Individual funds within a given skill group are assumed to have identical true alphas in this illustration. In our empirical section, our approach makes no such assumption.

[6]The actual t-statistic distributions for individual funds are non-normal for most U.S. domestic equity funds (KTWW (2006)). Accordingly, in our empirical section, we use a bootstrap approach to more accurately estimate the distribution of t-statistics for each fund (and their associated p-values).

To illustrate further, suppose that we choose a significance level, γ, of 10% (corresponding to $t_\gamma^- = -1.65$ and $t_\gamma^+ = 1.65$). With the test shown in Equation (1), the researcher would expect to find 5.4% of funds with a positive and significant t-statistic.[7] This proportion, denoted by $E(S_\gamma^+)$, is represented by the shaded region in the right tail of the cross-sectional t-distribution (Panel B). Does this area consist merely of skilled funds, as defined above? Clearly not, because some funds are just lucky; as shown in the shaded region of the right tail in Panel A, zero-alpha funds can exhibit positive and significant estimated t-statistics. By the same token, the proportion of funds with a negative and significant t-statistic (the shaded region in the left-tail of Panel B) overestimates the proportion of unskilled funds, because it includes some unlucky zero-alpha funds (the shaded region in the left tail in Panel A). Note that we have not considered the possibility that skilled funds could be *very* unlucky, and exhibit a negative and significant t-statistic. In our example of Figure 1, the probability that the estimated t-statistic of a skilled fund is lower than $t_\gamma^- = -1.65$ is less than 0.001%. This probability is negligible, so we ignore this pathological case. The same applies to unskilled funds that are very lucky.

The message conveyed by Figure 1 is that we measure performance with a limited sample of data, therefore, unskilled and skilled funds cannot easily be distinguished from zero-alpha funds. This problem can be worse if the cross-section of actual skill levels has a complex distribution (and not all fixed at the same levels, as assumed by our simplified example), and is further compounded if a substantial proportion of skilled fund managers have low levels of skill, relative to the error in estimating their t-statistics. To proceed, we must employ a procedure that is able to precisely account for "false discoveries," i.e., funds that falsely exhibit significant estimated alphas (i.e., their true alphas are zero) in the face of these complexities.

A.2 Measuring Luck

How do we measure the frequency of "false discoveries" in the tails of the cross-sectional (alpha) t-distribution? At a given significance level γ, it is clear that the probability that a zero-alpha fund (as defined in the last section) exhibits luck equals $\gamma/2$ (shown as the dark shaded region in Panel A of Figure 1)). If the proportion of zero-alpha funds in the population is π_0, the expected proportion of "lucky funds" (zero-alpha funds with

[7] From Panel A, the probability that the observed t-statistic is greater than $t_\gamma^+ = 1.65$ equals 5% for a zero-alpha fund and 84% for a skilled fund. Multiplying these two probabilities by the respective proportions represented by their categories (π_A^- and π_A^+) gives 5.4%.

positive and significant t-statistics) equals

$$E(F_\gamma^+) = \pi_0 \cdot \gamma/2. \tag{2}$$

Now, to determine the expected proportion of skilled funds, $E(T_\gamma^+)$, we simply adjust $E(S_\gamma^+)$ for the presence of these lucky funds:

$$E(T_\gamma^+) = E(S_\gamma^+) - E(F_\gamma^+) = E(S_\gamma^+) - \pi_0 \cdot \gamma/2. \tag{3}$$

Since the probability of a zero-alpha fund being unlucky is also equal to $\gamma/2$ (i.e., the grey and black areas in Panel A of Figure 1 are identical), $E(F_\gamma^-)$, the expected proportion of "unlucky funds," is equal to $E(F_\gamma^+)$. As a result, the expected proportion of unskilled funds, $E(T_\gamma^-)$, is similarly given by

$$E(T_\gamma^-) = E(S_\gamma^-) - E(F_\gamma^-) = E(S_\gamma^-) - \pi_0 \cdot \gamma/2. \tag{4}$$

What is the role played by the significance level, γ, chosen by the researcher? By defining the significance thresholds t_γ^- and t_γ^+, γ determines the portion of the right (or left) tail which is examined for lucky vs. skilled funds (or unlucky vs. unskilled funds), as described by Equations (3) and (4). By varying γ, we can determine the location of skilled (or unskilled) funds–by measuring the proportion of such funds in any segment of the cross-section.

This flexibility in choosing γ provides us with opportunities to make important insights into the merits of active fund management. First, by choosing a larger γ (i.e., lower t_γ^- and t_γ^+, in absolute value), we can estimate the proportions of unskilled and skilled funds in a larger portion of the left and right tails of the cross-sectional t-distribution, respectively–thus, giving us an appreciation of the proportions of unskilled and skilled funds in the entire population, π_A^- and π_A^+. That is, as we increase γ, $E(T_\gamma^-)$ and $E(T_\gamma^+)$ converge to π_A^- and π_A^+, thus minimizing Type II error (failing to locate truly unskilled or skilled funds). Alternatively, by reducing γ, we can determine the precise location of unskilled or skilled funds in the extreme tails of the t-distribution. For instance, choosing a very low γ (i.e., very large t_γ^- and t_γ^+, in absolute value) allows us to determine whether extreme tail funds are skilled or simply lucky (unskilled or simply unlucky)–information that is quite useful to investors trying to locate skilled (or avoid unskilled) managers.

A.3 Estimation Procedure

The key to our approach to measuring luck in a group setting, as shown in Equation (2), is the estimator of the proportion, π_0, of zero-alpha funds in the population. Here, we turn to a recent estimation approach developed by Storey (2002)–called the "False Discovery Rate" (FDR) approach. The FDR approach is very straightforward, as its sole inputs are the (two-sided) p-values associated with the (alpha) t-statistics of each of the M funds. By definition, zero-alpha funds satisfy the null hypothesis, $H_{0,i} : \alpha_i = 0$, and, therefore, have p-values that are uniformly distributed over the interval $[0, 1]$.[8] On the other hand, p-values of unskilled and skilled funds tend to be very small because their estimated t-statistics tend to be far from zero (see Panel A of Figure 1). We can exploit this information to estimate π_0 without knowing the exact distribution of the p-values of the unskilled and skilled funds.

To explain further, a key intuition of the FDR approach is that it uses information from the center of the cross-sectional t-distribution (which is dominated by zero-alpha funds) to correct for luck in the tails. To illustrate the FDR procedure, suppose we randomly draw 2,076 t-statistics (the number of funds in our study), each from one of the three t-distributions in Panel A of Figure 1. Each t-statistic is drawn from a given distribution with probability according to our estimates of the proportion of unskilled, zero-alpha, and skilled funds in the population, $\pi_0 = 75\%$, $\pi_A^- = 23\%$, and $\pi_A^+ = 2\%$, respectively. Thus, our draw of t-statistics comes from a known frequency of each type (23%, 75%, and 2%). Next, we apply the FDR technique to estimate these frequencies–from the sampled t-statistics, we compute two-sided p-values, $\widehat{p}_i$, for each of the 2,076 funds, then plot them in Figure 2.

Please insert Figure 2 here

The darkest grey zone near zero captures the majority of p-values of unskilled and skilled funds ($\pi_A^- + \pi_A^+$ =25%). The area below the horizontal line at 0.075 represents the true (but unknown to the researcher) proportion, π_0, of zero-alpha funds (75%), since zero-alpha funds have uniformly distributed p-values. The researcher estimates π_0 from the histogram of observed p-values as follows. If we take a sufficiently high threshold λ^* (e.g., $\lambda^* = 0.6$), we know that the vast majority of p-values higher than λ^* come from

[8] To see this, let us denote by t and p the t-statistic and p-value of the zero-alpha fund. We have $p = 1 - F(|t|)$, where $F(t) = prob(\left|\widehat{t}_i\right| < |t| \,|\, \alpha_i = 0)$. The p-value is uniformly distributed over $[0, 1]$ since its cdf, $G(p) = prob(\widehat{p}_i < p) = prob(1 - F(|t_{\widehat{p}_i}|) < p) = prob(|t_{\widehat{p}_i}| > F^{-1}(1-p)) = 1 - F\left(F^{-1}(1-p)\right) = p$.

zero-alpha funds. Thus, we first measure the proportion of the total area that is covered by the four lightest grey bars on the right of λ^*, $\widehat{W}(\lambda^*)/M$ (where $\widehat{W}(\lambda^*)$ denotes the number of funds having p-values exceeding λ^*). Then, we extrapolate this area over the entire interval $[0,1]$ by multiplying by $1/(1-\lambda^*)$ (e.g., if $\lambda^* = 0.6$, the area is multiplied by 2.5):[9]

$$\widehat{\pi}_0(\lambda^*) = \frac{\widehat{W}(\lambda^*)}{M} \cdot \frac{1}{(1-\lambda^*)}. \tag{5}$$

To select λ^*, we use the simple data-driven approach suggested by Storey (2002) and explained in detail in Appendix A.

Substituting the estimate $\widehat{\pi}_0$ in Equations (2), (3), and replacing $E(S_\gamma^+)$ with the observed proportion of significant funds in the right tail, $\widehat{S}_\gamma^+$, we can easily estimate $E(F_\gamma^+)$ and $E(T_\gamma^+)$ corresponding to any chosen significance level, γ. The same approach can be used in the left tail by replacing $E(S_\gamma^-)$ in Equation (4) with the observed proportion of significant funds in the left tail, $\widehat{S}_\gamma^-$. This implies the following estimates of the proportions of unlucky and lucky funds:

$$\widehat{F}_\gamma^- = \widehat{F}_\gamma^+ = \widehat{\pi}_0 \cdot \gamma/2. \tag{6}$$

Using Equation (6), the estimated proportions of unskilled and skilled funds (at the chosen significance level, γ) are, respectively, equal to

$$\begin{aligned} \widehat{T}_\gamma^- &= \widehat{S}_\gamma^- - \widehat{F}_\gamma^- = \widehat{S}_\gamma^- - \widehat{\pi}_0 \cdot \gamma/2, \\ \widehat{T}_\gamma^+ &= \widehat{S}_\gamma^+ - \widehat{F}_\gamma^+ = \widehat{S}_\gamma^+ - \widehat{\pi}_0 \cdot \gamma/2. \end{aligned} \tag{7}$$

Finally, we estimate the proportions of unskilled and skilled funds in the entire population as

$$\widehat{\pi}_A^- = \widehat{T}_{\gamma^*}^-, \qquad \widehat{\pi}_A^+ = \widehat{T}_{\gamma^*}^+, \tag{8}$$

where γ^* is a sufficiently high significance level–we choose γ^* with a simple data-driven method explained in Appendix A.

B Comparison of Our Approach with Existing Methods

The previous literature has followed two alternative approaches when estimating the proportions of unskilled and skilled funds. The "full luck" approach proposed by Jensen

[9] This estimation procedure cannot be used in a one-sided multiple test, since the null hypothesis is tested under the least favorable configuration (LFC). For instance, consider the following null hypothesis $H_{0,i} : \alpha_i \leq 0$. Under the LFC, it is replaced with $H_{0,i} : \alpha_i = 0$. Therefore, all funds with $\alpha_i \leq 0$ (i.e., drawn from the null) have inflated p-values which are not uniformly distributed over $[0,1]$.

(1968) and Ferson and Qian (2004) assumes, a priori, that all funds in the population have zero alphas, $\pi_0 = 1$. Thus, for a given significance level, γ, this approach implies an estimate of the proportions of unlucky and lucky funds equal to $\gamma/2$.[10] At the other extreme, the "no luck" approach reports the observed number of significant funds (for instance, Ferson and Schadt (1996)) without making a correction for luck.

What are the errors introduced by assuming, a priori, that π_0 equals 0 or 1, when it does not accurately describe the population? To address this question, we compare the bias produced by these two approaches relative to our FDR approach across different possible values for π_0 ($\pi_0 \in [0, 1]$) using our simple framework of Figure 1. Our procedure consists of three steps. First, for a chosen value of π_0, we create a simulated sample of 2,076 fund t-statistics (corresponding to our fund sample size) by randomly drawing from the three distributions in Panel A of Figure 1 in the proportions π_0, π_A^-, and π_A^+. For each π_0, the ratio π_A^-/π_A^+ is held fixed to 11.5 (0.23/0.02), as in Figure 1, to assure that the proportion of skilled funds remains low compared to the unskilled funds. Second, we use these sampled t-statistics to estimate the proportion of unlucky ($\alpha = 0$, $\widehat{\alpha} < 0$), lucky ($\alpha = 0$, $\widehat{\alpha} > 0$), unskilled ($\alpha < 0$, $\widehat{\alpha} < 0$), and skilled ($\alpha > 0$, $\widehat{\alpha} > 0$) funds under each of the three approaches–the "no luck," "full luck," and FDR techniques.[11] Third, under each approach, we repeat these first two steps 1,000 times to compare the average value of each estimator with its true population value.

Please insert Figure 3 here

Specifically, Panel A of Figure 3 compares the three estimators of the expected proportion of unlucky funds. The true population value, $E(F_\gamma^-)$, is an increasing function of π_0 by construction, as shown by Equation (2). While the average value of the FDR estimator closely tracks $E(F_\gamma^-)$, this is not the case for the other two approaches. Note that, by assuming that $\pi_0 = 0$, the "no luck" approach consistently underestimates $E(F_\gamma^-)$ when the true proportion of zero-alpha funds is higher ($\pi_0 > 0$). Conversely, the "full luck" approach, which assumes that $\pi_0 = 1$, overestimates $E(F_\gamma^-)$ when $\pi_0 < 1$. To illustrate the extent of the bias, consider the case where $\pi_0 = 75\%$. While the "no luck" approach substantially underestimates $E(F_\gamma^-)$ (0% instead of its true value of 7.5%), the "full luck" approach overestimates $E(F_\gamma^-)$ (10% instead of its true 7.5%). The biases for estimates of lucky funds $E(F_\gamma^+)$ shown in Panel B are exactly the same,

[10] Jensen (1968) summarizes the "full luck" approach as follows: "...if all the funds had a true α equal to zero, we would expect (merely by random chance) to find 5% of them having t values 'significant' at the 5% level."

[11] We choose $\gamma = 0.20$ to examine a large portion of the tails of the cross-sectional t-distribution, although other values for γ provide similar results.

since $E(F_\gamma^+) = E(F_\gamma^-)$.

Estimates of the expected proportions of unskilled and skilled funds ($E(T_\gamma^-)$ and $E(T_\gamma^+)$) provided by the three approaches are shown in Panels C and D, respectively. As we move to higher true proportions of zero-alpha funds (a higher value of π_0), the true proportions of unskilled and skilled funds, $E(T_\gamma^-)$ and $E(T_\gamma^+)$, decrease by construction. In both panels, our FDR estimator accurately captures this feature, while the other approaches do not fare well due to their fallacious assumptions about the prevalence of luck. For instance, when $\pi_0 = 75\%$, the "no luck" approach exhibits a large upward bias in its estimates of the total proportion of unskilled and skilled funds, $E(T_\gamma^-) + E(T_\gamma^+)$ (37.3% rather than the correct value of 22.3%). At the other extreme, the "full luck" approach underestimates $E(T_\gamma^-) + E(T_\gamma^+)$ (17.8% instead of 22.3%).

Panel D reveals that the "no luck" and "full luck" approaches also exhibit a nonsensical positive relation between π_0 and $E(T_\gamma^+)$. This result is a consequence of the low proportion of skilled funds in the population. First, as π_0 rises, the additional lucky funds drive the proportion of significant funds up, making the "no-luck" approach wrongly believe that more skilled funds are present. Second, the few skilled funds in the population cannot offset the excessive luck adjustment made by the "full luck" approach, which actually produces negative estimates of $E(T_\gamma^+)$.

In addition to the bias properties exhibited by our FDR estimators, their variability is low because of the large cross-section of funds ($M = 2,076$). To understand this, consider our main estimator $\widehat{\pi}_0$ (the same arguments apply to the other estimators). Since $\widehat{\pi}_0$ is a proportion estimator that depends on the proportion of $\widehat{p}_i > \lambda^*$, the Law of Large Numbers drives it close to its true value with our large sample size. For instance, taking $\lambda^* = 0.6$ and $\pi_0 = 75\%$, σ_{π_0} is as low as 2.5% with independent p-values ($1/30^{th}$ the magnitude of π_0).[12] In the appendix, we provide further evidence of the remarkable accuracy of our estimators using Monte-Carlo simulations.

C Estimation under Cross-Sectional Dependence among Funds

Mutual funds can have correlated residuals if they "herd" in their stockholdings (Wermers (1999)) or hold similar industry allocations. In general, cross-sectional dependence in fund estimated alphas greatly complicates performance measurement. Any inference test with dependencies becomes quickly intractable as M rises, since this requires the

[12] Specifically, $\widehat{\pi}_0 = (1-\lambda^*)^{-1} \cdot 1/M \sum_{i=1}^{M} x_i$, where x_i follows a binomial distribution with probability of success $P_{\lambda^*} = prob(\widehat{p}_i > \lambda^*) = 0.30$ (i.e., the rectangle area delimited by the horizontal black line and the vertical line at $\lambda^* = 0.6$ in Figure 2). Therefore, we have $\sigma_x = (P_{\lambda^*}(1-P_{\lambda^*}))^{\frac{1}{2}} = 0.46$, and $\sigma_{\pi_0} = (1-\lambda^*)^{-1} \cdot \sigma_x/\sqrt{M} = 2.5\%$.

estimation and inversion of an $M \times M$ residual covariance matrix. In a Bayesian framework, Jones and Shanken (2005) show that performance measurement requires intensive numerical methods when investor prior beliefs about fund alphas include cross-fund dependencies. Further, KTWW (2006) show that a complicated bootstrap is necessary to test the significance of performance of a fund located at a particular alpha rank, since this test depends on the joint distribution of all fund estimated alphas–cross-correlated fund residuals must be bootstrapped simultaneously.

An important advantage of our approach is that we estimate the p-value of each fund in isolation–avoiding the complications that arise because of the dependence structure of fund residuals. However, high cross-sectional dependencies could potentially bias our estimators. To illustrate this point with an extreme case, suppose that all funds produce zero alphas ($\pi_0 = 100\%$), and that fund residuals are perfectly correlated (perfect herding). In this case, all fund p-values would be the same, and the p-value histogram would not converge to the uniform distribution, as shown in Figure 2. Clearly, we would make serious errors no matter where we set λ^*.

In our sample, we are not overly concerned with dependencies, since we find that the average correlation between four-factor model residuals of pairs of funds is only 0.08. Further, many of our funds do not have highly overlapping return data, thus, ruling out highly correlated residuals by construction. Specifically, we find that 15% of the funds pairs do not have a single monthly return observation in common; on average, only 55% of the return observations of fund pairs is overlapping. As a result, we believe that cross-sectional dependencies are sufficiently low to allow consistent estimators (i.e., mutual fund residuals satisfy the ergodicity conditions discussed in Storey, Taylor, and Siegmund (2004)).

However, in order to explicitly verify the properties of our estimators, we run a Monte-Carlo simulation. In order to closely reproduce the actual pairwise correlations between funds in our dataset, we estimate the residual covariance matrix directly from the data, then use these dependencies in our simulations. In further simulations, we impose other types of dependencies, such as residual block correlations or residual factor dependencies, as in Jones and Shanken (2005). In all simulations, we find both that average estimates (for all of our estimators) are very close to their true values, and that confidence intervals for estimates are comparable to those that result from simulations where independent residuals are assumed. These results, as well as further details on the simulation experiment are discussed in Appendix B.

II Performance Measurement and Data Description

A Asset Pricing Models

To compute fund performance, our baseline asset pricing model is the four-factor model proposed by Carhart (1997):

$$r_{i,t} = \alpha_i + b_i \cdot r_{m,t} + s_i \cdot r_{smb,t} + h_i \cdot r_{hml,t} + m_i \cdot r_{mom,t} + \varepsilon_{i,t}, \tag{9}$$

where $r_{i,t}$ is the month t excess return of fund i over the riskfree rate (proxied by the monthly T-bill rate); $r_{m,t}$ is the month t excess return on the value-weighted market portfolio; and $r_{smb,t}$, $r_{hml,t}$, and $r_{mom,t}$ are the month t returns on zero-investment factor-mimicking portfolios for size, book-to-market, and momentum obtained from Kenneth French's website.

We also implement a conditional four-factor model to account for time-varying exposure to the market portfolio (Ferson and Schadt (1996)),

$$r_{i,t} = \alpha_i + b_i \cdot r_{m,t} + s_i \cdot r_{smb,t} + h_i \cdot r_{hml,t} + m_i \cdot r_{mom,t} + B^{'} \left(z_{t-1} \cdot r_{m,t} \right) + \varepsilon_{i,t}, \tag{10}$$

where z_{t-1} denotes the $J \times 1$ vector of centered predictive variables, and B is the $J \times 1$ vector of coefficients. The four predictive variables are the one-month T-bill rate; the dividend yield of the CRSP value-weighted NYSE/AMEX stock index; the term spread, proxied by the difference between yields on 10-year Treasurys and three-month T-bills; and the default spread, proxied by the yield difference between Moody's Baa-rated and Aaa-rated corporate bonds. We have also computed fund alphas using the CAPM and the Fama-French (1993) models. These results are summarized in Section III.D.2.

To compute each fund t-statistic, we use the Newey-West (1987) heteroscedasticity and autocorrelation consistent estimator of the standard deviation, $\widehat{\sigma}_{\widehat{\alpha}_i}$. Further, KTWW (2006) find that the finite-sample distribution of $\widehat{t}$ is non-normal for approximately half of the funds. Therefore, we use a bootstrap procedure (instead of asymptotic theory) to compute fund p-values. In order to estimate the distribution of $\widehat{t}_i$ for each fund i under the null hypothesis $\alpha_i = 0$, we use a residual-only bootstrap procedure, which draws with replacement from the regression estimated residuals $\{\widehat{\varepsilon}_{i,t}\}$.[13] For each

[13] To determine whether assuming homoscedasticity and temporal independence in individual fund residuals is appropriate, we have checked for heteroscedasticity (White test), autocorrelation (Ljung-Box test), and Arch effects (Engle test). We have found that only a few funds present such regularities. We have also implemented a block bootstrap methodology with a block length equal to $T^{\frac{1}{5}}$ (proposed by Hall, Horowitz, and Jing (1995)), where T denotes the length of the fund return time-series. All of our results to be presented remain unchanged.

fund, we implement 1,000 bootstrap replications. The reader is referred to KTWW (2006) for details on this bootstrap procedure.

B Mutual Fund Data

We use monthly mutual fund return data provided by the Center for Research in Security Prices (CRSP) between January 1975 and December 2006 to estimate fund alphas. Each monthly fund return is computed by weighting the net return of its component shareclasses by their beginning-of-month total net asset values. The CRSP database is matched with the Thomson/CDA database using the MFLINKs product of Wharton Research Data Services (WRDS) in order to use Thomson fund investment-objective information, which is more consistent over time. Wermers (2000) provides a description of how an earlier version of MFLINKS was created. Our original sample is free of survivorship bias, but we further select only funds having at least 60 monthly return observations in order to obtain precise four-factor alpha estimates. These monthly returns need not be contiguous. However, when we observe a missing return, we delete the following-month return, since CRSP fills this with the cumulated return since the last non-missing return. In unreported results, we find that reducing the minimum fund return requirement to 36 months has no material impact on our main results, thus, we believe that any biases introduced from the 60-month requirement are minimal.

Our final universe has 2,076 open-end, domestic equity mutual funds existing for at least 60 months between 1975 and 2006. Funds are classified into three investment categories: Growth (1,304 funds), Aggressive Growth (388 funds), and Growth & Income (384 funds). If an investment objective is missing, the prior non-missing objective is carried forward. A fund is included in a given investment category if its objective corresponds to the investment category for at least 60 months.

Table I shows the estimated annualized alpha as well as factor loadings of equally-weighted portfolios within each category of funds. The portfolio is rebalanced each month to include all funds existing at the beginning of that month. Results using the unconditional and conditional four-factor models are shown in Panels A and B, respectively.

Please insert Table I here

Similar to results previously documented in the literature, we find that unconditional estimated alphas for each category is negative, ranging from -0.45% to -0.60% per annum. Aggressive Growth funds tilt towards small capitalization, low book-to-market, and momentum stocks, while the opposite holds for Growth & Income funds. Introducing time-varying market betas provides similar results (Panel B). In tests available upon request from the authors, we find that all results to be discussed in the next section are qualitatively similar whether we use the unconditional or conditional version of the four-factor model. For brevity, we present only results from the unconditional four-factor model.

III Empirical Results

A Impact of Luck on Long-Term Performance

We begin our empirical analysis by measuring the impact of luck on long-term mutual fund performance, measured as the lifetime performance of each fund (over the period 1975-2006) using the four-factor model of Equation (9). Panel A of Table II shows estimated proportions of zero-alpha, unskilled, and skilled funds in the population ($\widehat{\pi}_0$, $\widehat{\pi}_A^-$, and $\widehat{\pi}_A^+$), as defined in Section I.A.1, with standard deviations of estimates in parentheses. These point estimates are computed using the procedure described in Section I.A.3, while standard deviations are computed using the method of Genovese and Wasserman (2004)–which is described in the appendix.

Please insert Table II here

Among the 2,076 funds, we estimate that the majority–75.4%–are zero-alpha funds. Managers of these funds exhibit stockpicking skills just sufficient to cover their trading costs and other expenses (including fees). These funds, therefore, capture all of the economic rents that they generate–consistent with the long-run prediction of Berk and Green (2004).

Further, it is quite surprising that the estimated proportion of skilled funds is statistically indistinguishable from zero (see "Skilled" column). This result may seem surprising in light of some prior studies, such as Ferson and Schadt (1996), which find that a small group of top mutual fund managers appear to outperform their benchmarks, net of costs. However, a closer examination–in Panel B–shows that our adjustment for luck is key in understanding the difference between our study and prior research.

To be specific, Panel B shows the proportion of significant alpha funds in the left and right tails ($\widehat{S}_\gamma^-$ and $\widehat{S}_\gamma^+$, respectively) at four different significance levels ($\gamma = 0.05$, 0.10, 0.15, 0.20). Similar to past research, there are many significant alpha funds in the right tail–$\widehat{S}_\gamma^+$ peaks at 8.2% of the total population (170 funds) when $\gamma = 0.20$. However, of course, "significant alpha" does not always mean "skilled fund manager."

Illustrating this point, the right side of Panel B decomposes these significant funds into the proportions of lucky zero-alpha funds and skilled funds ($\widehat{F}_{\gamma}^{+}$ and $\widehat{T}_{\gamma}^{+}$, respectively). Clearly, we cannot reject that all of the right tail funds are merely lucky outcomes among the large number of zero-alpha funds (1,565), and that none of these right-tail funds have truly skilled fund managers.

It is interesting (Panel A) that 24% of the population (499 funds) are truly unskilled fund managers–unable to pick stocks well enough to recover their trading costs and other expenses.[14] In untabulated results, we find that left-tail funds, which are overwhelmingly comprised of unskilled (and not merely unlucky) funds, have a relatively long fund life–12.7 years, on average. And, these funds generally perform poorly over their entire lives, making their survival puzzling. Perhaps, as discussed by Elton, Gruber, and Busse (2003), such funds exist if they are able to attract a sufficient number of unsophisticated investors, who are also charged higher fees (Christoffersen and Musto (2002)).

The bottom of Panel B presents characteristics of the average fund in each segment of the tails. Although the average estimated alpha of right-tail funds is somewhat high (between 4.8% and 6.5% per year), this is simply due to very lucky outcomes for a small proportion of the 1,565 zero-alpha funds in the population. It is also interesting that expense ratios are higher for left-tail funds, which likely explains some of the underperformance of these funds (we will revisit this issue when we examine pre-expense returns in a later section). Turnover does not vary systematically among the various tail segments, but left-tail funds are much smaller than right-tail funds, presumably due to the combined effects of outflows and poor investment returns. Results for the three investment-objective subgroups (Aggressive Growth, Growth, and Growth & Income) are similar–these results are available upon request from the authors.

As mentioned earlier, the universe of U.S. domestic equity mutual funds has expanded substantially since 1990. Accordingly, we next examine the evolution of the proportions of unskilled and skilled funds over time. To accomplish this, at the end of each year from 1989 to 2006, we estimate the proportions of unskilled and skilled funds using the entire return history for each fund up to that point in time–this would correspond to the entire history of fund returns (starting in 1975) observed by a researcher for the universe of domestic equity funds at that point in time. For instance, our initial estimates, on December 31, 1989, cover the first 15 years of the sample, 1975-89, while our final estimates, on December 31, 2006, are based on the entire 32 years of the sample,

[14] This minority of funds is the driving force explaining the negative average estimated alpha that is widely documented in the literature (e.g., Jensen (1968), Carhart (1997), Elton et al. (1993), and Pastor and Stambaugh (2002a)).

1975-2006 (i.e., these are the estimates shown in Panel A of Table II).[15] The results in Panel A of Figure 4 show that the proportion of funds with non-zero alphas (equal to the sum of the proportions of skilled and unskilled funds) remains fairly constant over time. However, there are dramatic changes in the relative proportions of unskilled and skilled funds: from 1989 to 2006. Specifically, the proportion of skilled funds declines from 14.4% to 0.6%, while the proportion of unskilled funds rises from 9.2% to 24.0% of the entire universe of funds. These changes are also reflected in the population average estimated alpha (shown in Panel B), which drops from 0.16% to -0.97% per year over the same period.

Please insert Figure 4 here

Panel B also displays the yearly count of funds included in the estimated proportions of Panel A. From 1996 to 2005, there are more than 100 additional actively managed domestic-equity mutual funds per year.[16] Interestingly, this coincides with the time trend in unskilled and skilled funds shown in Panel A–the huge increase in numbers of actively managed mutual funds has resulted in a much larger proportion of unskilled funds, at the expense of skilled funds. Either the growth of the fund industry has coincided with greater levels of stock market efficiency, making stockpicking a more difficult and costly endeavor, or the large number of new managers simply have inadequate skills. It is also interesting that, during our period of analysis, many fund managers with good track records left the sample to manage hedge funds (as shown by Kostovetski (2007)), and that indexed investing increased substantially.

B Impact of Luck on Short-Term Performance

Our above results indicate that funds do not achieve superior long-term alphas, perhaps because flows compete away any alpha surplus. However, we might find evidence of funds with superior short-term alphas, before investors become fully aware of such outperformers due to search costs.

To test for short-run mutual fund performance, we partition our data into six non-overlapping subperiods of five years, beginning with 1977-1981 and ending with 2002-2006. For each subperiod, we include all funds having 60 monthly return observations, then compute their respective alpha p-values–in other words, we treat each fund during

[15] To be included at the end of a given year, a fund must have at least 60 monthly return observations before that date, although these observations need not be contiguous.

[16] Since we require 60 monthly observations to measure fund performance, this rise reflects the massive entry of new funds over the period 1993-2001.

each five-year period as a separate "fund."[17] We pool these five-year records together across all time periods to represent the average experience of an investor in a randomly chosen fund during a randomly chosen five-year period. After pooling, we obtain a total of 3,311 p-values from which we compute our different estimators. Results for the entire population (All Funds) are shown in Table III, while results for Growth, Aggressive Growth, and Growth & Income funds are displayed in Panels A, B, and C of Table IV, respectively.

Please insert Table III here

First, Panel A of Table III shows that a small fraction of funds (2.4% of the population) exhibit skill over the short-run (with a standard deviation of 0.7%). Thus, short-term superior performance is rare, but does exist, as opposed to long-term performance. Second, these skilled funds are located in the extreme right tail of the cross-sectional t-distribution. Panel B of Table III shows that, with a γ of only 10%, we capture almost all skilled funds, as $\widehat{T}_{\gamma}^{+}$ reaches 2.3% (close to its maximum value of 2.4%). Proceeding toward the center of the distribution (by increasing γ to 0.10 and 0.20) produces almost no additional skilled funds and almost entirely additional zero-alpha funds that are lucky ($\widehat{F}_{\gamma}^{+}$). Thus, skilled fund managers, while rare, may be somewhat easy to find, since they have extremely high t-statistics (extremely low p-values)–we will use this finding in our next section, where we attempt to find funds with out-of-sample skills. It is notable that we find evidence of short-term outperformance of some funds here, but no evidence of long-term outperformance in the prior section of this paper. This is consistent with Berk and Green (2004), where outperforming funds exist only until investors are successfully able to locate them.

In the left tail, we observe that the great majority of funds are unskilled, and not merely unlucky zero-alpha funds. For instance, in the extreme left tail (at $\gamma = 0.05$), the proportion of unskilled funds, $\widehat{T}_{\gamma}^{-}$, is roughly five times the proportion of unlucky funds, $\widehat{F}_{\gamma}^{-}$ (9.4% versus 1.8%). Here, the short-term results are similar to the prior-discussed long-term results–the great majority of left-tail funds are truly unskilled. It is also interesting that true skills seem to be inversely related to turnover, as indicated by the substantially higher levels of turnover of left-tail funds (which are mainly unskilled funds). Unskilled managers apparently trade frequently to appear skilled, which ultimately hurts their performance. Perhaps poor governance of some funds explains why they end up in the left tail (net of expenses), in the short-run–they overexpend on both

[17] Note that reducing the number of observations comes at a cost: it increases the standard deviation of the estimated alphas, making the p-values of non-zero alpha funds harder to distinguish from those of zero-alpha funds.

trading costs (through high turnover) and other expenses relative to their skills.

Table IV shows results for investment-objective subgroups. Panel A shows that the proportions of skilled Growth funds in various segments of the right tail are similar to those of the entire universe (from Table III). However, Aggressive-Growth funds (Panel B) exhibit somewhat higher skills. For instance, at $\gamma = 0.05$, 73% of significant Aggressive-Growth funds are truly skilled (3.1/4.9). On the other hand, Panel C shows that no Growth & Income funds are truly skilled, but that a substantial proportion of them are unskilled. The long-term existence of this category of actively-managed funds, which includes "value funds" and "core funds" is remarkable in light of these poor results.

Please insert Table IV here

C Performance Persistence

Our previous analysis reveals that only 2.4% of the funds are skilled over the short-term. Can we detect these skilled funds over time, in order to capture their superior alphas? Ideally, we would like to form a portfolio containing only the truly skilled funds in the right tail; however, since we only know which segment of the tails in which they lie, but not their identities, such an approach is not feasible.

Nonetheless, the reader should recall from the last section that skilled funds are located in the extreme right tail. By forming portfolios containing all funds in this extreme tail, we have a greater chance of capturing the superior alphas of the truly skilled ones. For instance, Panel B of Table III shows that, at $\gamma = 0.05$, the proportion of skilled funds among all significant funds, $\widehat{T}_\gamma^+/\widehat{S}_\gamma^+$, is about 50%, which is much higher than the proportion of skilled funds in the entire universe, 2.4%.

To select a portfolio of funds, we use the False Discovery Rate in the right tail, FDR^+. At a given significance level, γ, the FDR^+ is defined as the expected proportion of lucky funds among all significant funds in this tail:

$$FDR_\gamma^+ = E\left(\frac{F_\gamma^+}{S_\gamma^+}\right). \tag{11}$$

The FDR_γ^+ provides a simple portfolio formation rule.[18] When we set a low FDR^+ target, we allow only a small proportion of lucky funds ("false discoveries") in the chosen

[18] Our new measure, FDR_γ^+, is an extension of the traditional FDR introduced in the statistical literature (e.g., Benjamini and Hochberg (1995), Storey (2002)), since the latter does not distinguish between bad and good luck. The traditional measure is $FDR_\gamma = E\left(F_\gamma/S_\gamma\right)$, where $F_\gamma = F_\gamma^+ + F_\gamma^-$, $S_\gamma = S_\gamma^+ + S_\gamma^-$.

portfolio. Specifically, we set a sufficiently low significance level, γ, so as to include skilled funds along with a small number of zero-alpha funds that are extremely lucky. Conversely, increasing the FDR^+ target has two opposing effects on a portfolio. First, it decreases the portfolio's expected future performance, since the proportion of lucky funds in the portfolio is higher. However, it also increases its diversification, since more funds are selected–reducing the volatility of the portfolio's out-of-sample performance. Accordingly, we examine five FDR^+ target levels in our persistence test: 10%, 30%, 50%, 70%, and 90%.

The construction of the portfolios proceeds as follows. At the end of each year, we estimate the alpha p-values of each existing fund using the previous five-year period. Using these p-values, we estimate the FDR^+_γ over a range of chosen significance levels (γ =0.01, 0.02,..., 0.60). Following Storey (2002) and Storey and Tibshirani (2003), we implement the following straightforward estimator of the FDR^+_γ :

$$\widehat{FDR}^+_\gamma = \frac{\widehat{F}^+_\gamma}{\widehat{S}^+_\gamma} = \frac{\widehat{\pi}_0 \cdot \gamma/2}{\widehat{S}^+_\gamma}, \tag{12}$$

where $\widehat{\pi}_0$ is the estimator of the proportion of zero-alpha funds described in Section I.A.3. For each FDR^+ target level, we determine the significance level, γ^P, that provides an $\widehat{FDR}^+_{\gamma^P}$ as close as possible to this target. Then, only funds with p-values smaller than γ^P are included in an equally-weighted portfolio. This portfolio is held for one year, after which the selection procedure is repeated. If a selected fund does not survive after a given month during the holding period, its weight is reallocated to the remaining funds during the rest of the year to mitigate survival bias. The first portfolio formation date is December 31, 1979 (after five years of returns have been observed), while the last is December 31, 2005.

In Panel A of Table V, we show the FDR level ($\widehat{FDR}^+_{\gamma^P}$) of the five portfolios, as well as the proportion of funds in the population that they include ($\widehat{S}^+_{\gamma^P}$) during the five-year formation period, averaged over the 27 formation periods (ending from 1979 to 2005)–and, their respective distributions. First, we observe (as expected) that the achieved FDR increases with the FDR target assigned to a portfolio. However, the average $\widehat{FDR}^+_{\gamma^P}$ does not always match its target. For instance, FDR10% achieves an average of 41.5%, instead of the targeted 10%–during several formation periods, the proportion of skilled funds in the population is too low to achieve a 10% FDR target.[19]

[19] For instance, the minimum achievable FDR at the end of 2003 and 2004 is equal to 47.0% and 39.1%, respectively. If we look at the $\widehat{FDR}^+_{\gamma^P}$ distribution for the portfolio FDR 10% in Panel A,we observe that in 6 years out of 27, the $\widehat{FDR}^+_{\gamma^P}$ is higher than 70%.

Of course, a higher FDR target means an increase in the proportion of funds included in a portfolio–as shown in the rightmost columns of Panel A–since our selection rule becomes less restrictive.

In Panel B, we present the average out-of-sample performance (during the following year) of these five false discovery controlled portfolios, starting January 1, 1980 and ending December 31, 2006. We compute the estimated annualized alpha, $\widehat{\alpha}$, along with its bootstrapped p-value; annualized residual standard deviation, $\widehat{\sigma}_{\varepsilon}$; information ratio, IR= $\widehat{\alpha}/\widehat{\sigma}_{\varepsilon}$; four-factor model loadings; annualized mean return (minus T-bills); and annualized time-series standard deviation of monthly returns. The results reveal that our FDR portfolios successfully detect funds with short-term skills. For example, the portfolios FDR10% and 30% produce out-of-sample alphas (net of expenses) of 1.45% and 1.15% per year (significant at the 5% level). As the FDR target rises to 90%, the proportion of funds in the portfolio increases, which improves diversification ($\widehat{\sigma}_{\varepsilon}$ falls from 4.0% to 2.7%). However, we also observe a sharp decrease in the alpha (from 1.45% to 0.39%), reflecting the large proportion of lucky funds contained in the FDR90% portfolio.

Please insert Table V here

Panel C examines portfolio turnover–we determine the proportion of funds which are still selected using a given false discovery rule 1, 2, 3, 4, and 5 years after their initial inclusion. The results sharply illustrate the short-term nature of truly outperforming funds. After 1 year, 40% or fewer funds remain in portfolios FDR10% and 30%, while after 3 years, these percentages drop below 6%.

Finally, we examine, in Figure 5, how the estimated alpha of the portfolio FDR10% evolves over time using expanding windows. The initial value, on December 31, 1989, is the yearly out-of-sample alpha, averaged over the period 1980 to 1989, while the final value, on December 31, 2006, is the yearly out-of-sample alpha, averaged over the entire period 1980-2006 (i.e., this is the estimated alpha shown in Panel B of Table V). Again, these are the entire history of persistence results that would be observed by a researcher at the end of each year. The similarity with Figure 4 is striking. While the alpha accruing to the FDR10% portfolio is impressive at the beginning of the 1990s, it consistently declines thereafter. As the proportion, π_A^+, of skilled funds falls, the FDR approach moves much further to the extreme right tail of the cross-sectional t-distribution (from 5.7% of all funds in 1990 to 0.9% in 2006) in search of skilled managers. However, this change is not sufficient to prevent the performance of FDR10% from dropping substantially.

Please insert Figure 5 here

It is important to note the differences between our approach to persistence and that of the previous literature (e.g., Hendricks, Patel, and Zeckhauser (1993), Elton, Gruber, and Blake (1996), Carhart (1997)). These prior papers generally classify funds into fractile portfolios based on their past performance (past returns, estimated alpha, or alpha t-statistic) over a previous ranking period (one to three years). The size of fractile portfolios (e.g., deciles) are held fixed, with no regard to the changing proportion of lucky funds within these fixed fractiles. As a result, the signal used to form portfolios is likely to be noisier than our FDR approach. To compare these approaches with ours, Figure 5 displays the performance evolution of two top decile portfolios which are formed based on ranking funds by their alpha t-statistic, estimated over the previous one and three years, respectively.[20] Over most years, the FDR approach performs much better, consistent with the idea that it much more precisely detects skilled funds. However, this performance advantage declines during later years, when the proportion of skilled funds decreases substantially, making them much tougher to locate. Therefore, we find that the superior performance of the FDR portfolio is tightly linked to the prevalence of skilled funds in the population.

D Additional Results

D.1 Performance Measured with Pre-Expense Returns

In our baseline framework described previously, we define a fund as skilled if it generates a positive alpha net of trading costs, fees, and other expenses. Alternatively, skill could be defined, in an absolute sense, as the manager's ability to produce a positive alpha before expenses are deducted. Measuring performance on a pre-expense basis allows one to disentangle the manager's stockpicking skills, net of trading costs, from the fund's expense policy–which may be out of the control of the fund manager. To address this issue, we add monthly expenses (1/12 times the most recent reported annual expense ratio) to net returns for each fund, then revisit the long-term performance of the mutual fund industry.[21]

Panel A of Table VI contains the estimated proportions of zero-alpha, unskilled, and skilled funds in the population ($\widehat{\pi}_0$, $\widehat{\pi}_A^-$, and $\widehat{\pi}_A^+$), on a pre-expense basis. Comparing

[20] We use the t-statistic to be consistent with the rest of our paper, but the results are qualitatively similar when we rank on the estimated alpha.

[21] We discard funds which do not have at least 60 pre-expense return observations over the period 1975-2006. This leads to a small reduction in our sample from 2,076 to 1,836 funds.

these estimates with those shown in Table II, we observe a striking reduction in the proportion of unskilled funds–from 24.0% to 4.5%. This result indicates that only a small fraction of fund managers have stockpicking skills that are insufficient to at least compensate for their trading costs. Instead, mutual funds produce negative net-of-expense alphas chiefly because they charge excessive fees, in relation to the selection abilities of their managers. In Panel B, we further find that the average expense ratio across funds in the left tail is lower when performance is measured prior to expenses (1.3% versus 1.5% per year), indicating that high fees (potentially charged to unsophisticated investors) are a chief reason why funds end up in the extreme left tail, net of expenses. In addition, turnover seems to have no relation to pre-expense performance, as with the long-term net-of-expense results of Table II.

Please insert Table VI here

In the right tail, we find that 9.6% of fund managers have stockpicking skills sufficient to more than compensate for trading costs (Panel A). Consistent with Berk and Green (2004), the rents stemming from their skills are extracted through fees and other expenses, driving the proportion of net-expense skilled funds to zero.

Since 75.4% of funds produce zero net-expense alphas, it seems surprising that that we do not find more pre-expense skilled funds. However, this is due to the relatively small impact of expense ratios on fund performance in the center of the cross-sectional t-distribution. Adding back these expenses leads only to a marginal increase in the alpha t-statistic, making the power of the tests rather low.[22]

Finally, in untabulated tests, we find that the proportion of skilled funds in the population decreases from 27.5% to 10% between 1996 and 2006. This implies that the decline in net-expense skills noted in Figure 4 is mostly driven by a reduction in stockpicking skills over time (as opposed to an increase in expenses for (pre-expense) skilled funds).

On the contrary, the proportion of pre-expense unskilled funds remains equal to zero until the end of 2003. Thus, poor stockpicking skills (net of trading costs) cannot explain the large increase in the proportion of unskilled funds (net of both trading costs and expenses) from 1996 onwards. This increase is likely to be due to rising expenses

[22] The average expense ratio across funds with $|\widehat{\alpha}_i| < 1\%$ is approximately 10 bp per month. Adding back these expenses to a fund with zero net-expense alpha only increases its t-statistic mean from 0 to 0.9 (based on $T^{\frac{1}{2}}\alpha_A/\sigma_\varepsilon$, with $T = 384$, and $\sigma_\varepsilon = 0.021$). It implies that the null and alternative t-statistic distributions are extremely difficult to distinguish (for a fund with a (pre-expense) t-statistic mean of 0.9, the probability of observing a negative (pre-expense) t-statistic is equal to 18%!).

charged by funds with weak stock-selection abilities, or the introduction of new funds with high expense ratios and little stockpicking skills.

D.2 Performance Measured with Other Asset Pricing Models

Our estimation of the proportions of unskilled and skilled funds, $\widehat{\pi}_A^-$ and $\widehat{\pi}_A^+$, obviously depends on the choice of the asset pricing model. To examine the sensitivity of our results, we repeat the long-term (net of expense) performance analysis using the (unconditional) CAPM and Fama-French models. Based on the CAPM, we find that $\widehat{\pi}_A^-$ and $\widehat{\pi}_A^+$ are equal to 14.3% and 8.6% respectively, which is much more supportive of active management skills, compared to Section III.A.1. However, this result may be due to the omission of the size, book-to-market, and momentum factors. This conjecture is confirmed in Panel A of Table VII: the funds located in the right tail (according to the CAPM) have substantial loadings on the size and the book-to-market factors, which carry positive risk premia over our sample period (3.7% and 5.4% per year, respectively).

Please insert Table VII here

Turning to the Fama-French (1993) model, we find that $\widehat{\pi}_A^-$ and $\widehat{\pi}_A^+$ amount to 25.0% and 1.7%, respectively. These proportions are very close to those obtained with the four-factor model, since only one factor is omitted. As expected, the 1.1% difference in the estimated proportion of skilled funds between the two models (1.7%-0.6%) can be explained by the momentum factor. As shown in Panel B, the funds located in the right tail (according to the Fama-French model) have substantial loadings on the momentum factor, which carries a positive risk premium over the period (9.4% per year).

D.3 Bayesian Interpretation

Although we operate in a classical frequentist framework, our new FDR measure, FDR^+, also has a natural Bayesian interpretation.[23] To see this, we denote, by H_i, a random variable which takes the value of -1 if fund i is unskilled, 0 if it has zero alpha, and +1 if it is skilled. The prior probabilities for the three possible values (-1, 0, +1) are given by the proportion of each skill group in the population, π_A^-, π_0, and π_A^+. The Bayesian version of our FDR^+ measure, denoted by fdr_γ^+, is defined as the posterior probability that fund i has a zero alpha given that its t-statistic, $\widehat{t}_i$, is positive and significant: $fdr_\gamma^+ = prob\left(H_i = 0 | \, \widehat{t}_i \in \Gamma_A^+(\gamma)\right)$, where $\Gamma_A^+(\gamma) = \left(t_\gamma^+, +\infty\right)$. Using

[23] Our demonstration follows from the arguments used by Efron and Tibshirani (2002) and Storey (2003) for the traditional FDR, defined as $FDR_\gamma = E\left(F_\gamma / S_\gamma\right)$, where $F_\gamma = F_\gamma^+ + F_\gamma^-$, $S_\gamma = S_\gamma^+ + S_\gamma^-$.

Bayes theorem, we have:

$$fdr_{\gamma}^{+} = \frac{prob\left(\widehat{t}_i \in \Gamma_A^{+}(\gamma)\middle|\, H_i = 0\right) \cdot prob(H_i = 0)}{prob\left(\widehat{t}_i \in \Gamma_A^{+}(\gamma)\right)} = \frac{\gamma/2 \cdot \pi_0}{E(S_{\gamma}^{+})}. \tag{13}$$

Stated differently, the fdr_{γ}^{+} indicates how the investor changes his prior probability that fund i has a zero alpha ($H_i = 0$) after observing that its t-statistic is significant. In light of Equation (13), our estimator $\widehat{FDR}_{\gamma}^{+} = (\gamma/2 \cdot \widehat{\pi}_0)/\widehat{S}_{\gamma}^{+}$ can therefore be interpreted as an empirical Bayes estimator of fdr_{γ}^{+}, where π_0 and $E(S_{\gamma}^{+})$ are directly estimated from the data.[24]

In the recent Bayesian literature on mutual fund performance (e.g., Baks, Metrick, and Wachter (2001) and Pastor and Stambaugh (2002a)), attention is given to the posterior distribution of the fund alpha, α_i, as opposed to the posterior distribution of H_i. Interestingly, our approach also provides some relevant information for modeling the fund alpha prior distribution in an empirical Bayes setting. The parameters of the prior can be specified based on the relative frequency of the three fund skill groups (zero-alpha, unskilled, and skilled funds). In light of our estimates, an empirically-based alpha prior distribution is characterized by a point mass at $\alpha = 0$, reflecting the fact that 75.4% of the funds yield zero alphas, net of expenses. Since $\widehat{\pi}_A^{-}$ is higher than $\widehat{\pi}_A^{+}$, the prior probability of observing a negative alpha is higher than that of observing a positive alpha. These empirical constraints yield an asymmetric prior distribution. A tractable way to model the left and right parts of this distribution is to exploit two truncated normal distributions in the same spirit as in Baks, Metrick, and Wachter (2001). Further, we estimate that 9.6% of the funds have an alpha greater than zero, before expenses. While Baks, Metrick, and Wachter (2001) set this probability to 1% in order to examine the portfolio decision made by a skeptical investor, our analysis reveals that this level represents an overly skeptical belief.

IV Conclusion

In this paper, we apply a new method for measuring the skills of fund managers in a group setting. Specifically, the "False Discovery Rate" (FDR) approach provides a simple and straightforward method to estimate the proportion of funds within a population

[24] A full Bayesian estimation of fdr_{γ}^{+} requires to posit prior distributions for the proportions π_0, π_A^{-}, and π_A^{+}, and for the distribution parameters of $\widehat{t}_i$ for each skill group. This method based on additional assumptions (including independent p-values) as well as intensive numerical methods is illustrated by Tang, Ghosal, and Roy (2007) in the case of the traditional FDR.

that have stockpicking skills. In Monte Carlo simulations, we show that our novel approach gives very accurate estimates of the proportion of skilled funds (those providing a positive alpha, net of trading costs and expenses), zero-alpha funds, and unskilled funds (those providing a negative alpha) in the entire population. Further, we can use these estimates to provide accurate counts of skilled funds within various intervals in the right tail of the cross-sectional (estimated) alpha distribution, as well as unskilled funds within segments of the left tail.

We also apply the FDR technique to show that the proportion of skilled fund managers has diminished rapidly over the past 20 years. On the contrary, unskilled fund managers have increased substantially in the population over this period. Further analysis of pre-expense alphas reveals that the increase in unskilled fund managers (net of expenses) is due to an increase in the number of funds who charge high fees while possessing no particular stockpicking skills.

Our paper focuses the long-standing puzzle of actively managed mutual fund underperformance on the minority of truly underperforming funds. Most actively managed funds provide either positive or zero net-of-expense alphas, putting them at least on par with passive funds. Still, it is puzzling why investors seem to increasingly tolerate the existence of a large minority of funds that produce negative alphas, when an increasing array of passively managed funds have become available (such as ETFs). Perhaps a class of unsophisticated or inattentive investors remain shareholders in funds after they have clearly demonstrated (over time) their inferior returns. Or, as Elton, Gruber, and Blake (2007) discuss, maybe investors are forced to make constrained rational decisions–since these authors document that many 401(k) plans offer inefficient choices of mutual funds.

While our paper focuses on mutual fund performance, our approach has potentially wide applications in finance. It can be used in any setting in which a multiple hypothesis test is run and a large sample is available. We list two illustrative examples. First, technical trading can be implemented with a myriad of trading rules (e.g., Sullivan, Timmermann, and White (1999)). Our estimators can be used to determine the impact of luck on the performance of all these trading rules simultaneously. Second, testing the presence of commonality in liquidity boils down to regressing an individual stock liquidity measure on the market liquidity measure (e.g., Chordia, Roll, and Subrahmanyam (2000)). Since this regression is run for each individual stock, we are dealing with multiple testing. As a result, a correct measurement of commonality in liquidity necessitates a proper adjustment for luck. Because our approach only requires the estimation of π_0, controlling for luck in multiple testing is trivial: the only input required is a vector of p-values, one for each stock.

V Appendix

A Estimation Procedure

A.1 Determining the Value for λ^* from the Data

We use the bootstrap procedure proposed by Storey (2002) and Storey, Taylor, and Siegmund (2004). This resampling approach chooses λ from the data such that an estimate of the Mean-Squared Error (MSE) of $\widehat{\pi}_0(\lambda)$ is minimized. First, we compute $\widehat{\pi}_0(\lambda)$ using Equation (5) across a range of λ values ($\lambda = 0.30, 0.35, ..., 0.70$). Second, for each possible value of λ, we form 1,000 bootstrap replications of $\widehat{\pi}_0(\lambda)$ by drawing with replacement from the $M \times 1$ vector of fund p-values. These are denoted by $\widehat{\pi}_0^b(\lambda)$, for $b = 1, ..., 1,000$. Third, we compute the estimated MSE for each possible value of λ:

$$\widehat{MSE}(\lambda) = \frac{1}{1,000} \sum_{b=1}^{1,000} \left[\widehat{\pi}_0^b(\lambda) - \min_{\lambda} \widehat{\pi}_0(\lambda) \right]^2. \tag{14}$$

We choose λ^* such that $\lambda^* = \arg\min_\lambda \widehat{MSE}(\lambda)$. In unreported results (available upon request), we find that fixing λ to 0.5 or 0.6 yields similar results to those obtained with the bootstrap procedure (see also Storey (2002)). Still, the main advantage of the bootstrap approach is that it is entirely data-driven.

A.2 Determining the Value for γ^* from the Data

Similar to the approach used to determine λ^*, we use a bootstrap procedure which minimizes the estimated MSE of $\widehat{\pi}_A^-(\gamma)$ and $\widehat{\pi}_A^+(\gamma)$. First, we compute $\widehat{\pi}_A^-(\gamma)$ using Equation (8) across a range of γ ($\gamma = 0.10, 0.15, ..., 0.50$). Second, we form 1,000 bootstrap replications of $\widehat{\pi}_A^-(\gamma)$ for each possible value of γ. These are denoted by $\widehat{\pi}_A^{b-}(\gamma)$, for $b = 1, ..., 1,000$. Third, we compute the estimated MSE for each possible value of γ:

$$\widehat{MSE}^-(\gamma) = \frac{1}{1,000} \sum_{b=1}^{1,000} \left[\widehat{\pi}_A^{b-}(\gamma) - \max_{\gamma} \widehat{\pi}_A^-(\gamma) \right]^2. \tag{15}$$

We choose γ^- such that $\gamma^- = \arg\min_\gamma \widehat{MSE}^-(\gamma)$. We use the same data-driven procedure for $\widehat{\pi}_A^+(\gamma)$ to determine $\gamma^+ = \arg\min_\gamma \widehat{MSE}^+(\gamma)$. If $\min_\gamma \widehat{MSE}^-(\gamma) < \min_\gamma \widehat{MSE}^+(\gamma)$, we set $\widehat{\pi}_A^-(\gamma^*) = \widehat{\pi}_A^-(\gamma^-)$. To preserve the equality $1 = \pi_0 + \pi_A^+ + \pi_A^-$, we set $\widehat{\pi}_A^+(\gamma^*) = 1 - \widehat{\pi}_0 - \widehat{\pi}_A^-(\gamma^*)$. Otherwise, we set $\widehat{\pi}_A^+(\gamma^*) = \widehat{\pi}_A^+(\gamma^+)$ and $\widehat{\pi}_A^-(\gamma^*) = 1 - \widehat{\pi}_0 - \widehat{\pi}_A^+(\gamma^*)$.

A.3 Determining the Standard Deviation of the Estimators

We rely on the large-sample theory proposed by Genovese and Wasserman (2004). The essential idea is to recognize that the estimators $\widehat{\pi}_0(\lambda^*)$, $\widehat{S}_\gamma^+$, $\widehat{F}_\gamma^+$, $\widehat{T}_\gamma^+$, $\widehat{S}_\gamma^-$, $\widehat{F}_\gamma^-$, and $\widehat{T}_\gamma^-$ are all stochastic processes indexed by λ^* or γ which converge to a Gaussian process when the number of funds, M, goes to infinity. Proposition 3.2 of Genovese and Wasserman (2004) shows that $\widehat{\pi}_0(\lambda^*)$ is asymptotically normally distributed when $M \to \infty$, with standard deviation $\widehat{\sigma}_{\widehat{\pi}_0(\lambda^*)} = \left(\frac{\widehat{W}(\lambda^*)(M-\widehat{W}(\lambda^*))}{M^3(1-\lambda^*)^2}\right)^{\frac{1}{2}}$, where $\widehat{W}(\lambda^*)$ denotes the number of funds having p-values exceeding λ^*. Similarly, we have $\widehat{\sigma}_{\widehat{F}_\gamma^+} = (\gamma/2)\,\widehat{\sigma}_{\widehat{\pi}_0(\lambda^*)}$, $\widehat{\sigma}_{\widehat{S}_\gamma^+} = \left(\frac{\widehat{S}_\gamma^+(1-\widehat{S}_\gamma^+)}{M}\right)^{\frac{1}{2}}$, and $\widehat{\sigma}_{\widehat{T}_\gamma^+} = \left(\widehat{\sigma}^2_{\widehat{S}_\gamma^+} + (\gamma/2)^2\,\widehat{\sigma}^2_{\widehat{\pi}_0(\lambda^*)} + 2\frac{(\gamma/2)}{1-\lambda^*}\widehat{S}_\gamma^+\frac{\widehat{W}(\lambda^*)}{M}\right)^{\frac{1}{2}}$ (using the equality $\widehat{S}_\gamma^+ = \widehat{F}_\gamma^+ + \widehat{T}_\gamma^+$). Standard deviations for the estimators in the left tail $\left(\widehat{S}_\gamma^-,\ \widehat{F}_\gamma^-,\ \widehat{T}_\gamma^-\right)$ are obtained by simply replacing $\widehat{S}_\gamma^+$ with $\widehat{S}_\gamma^-$ in the above formulas.

Finally, if $\gamma^* = \gamma^+$, the standard deviation of $\widehat{\pi}_A^+$ and $\widehat{\pi}_A^-$ are respectively given by $\widehat{\sigma}_{\widehat{\pi}_A^+} = \widehat{\sigma}_{\widehat{T}_{\gamma^*}^+}$, and $\widehat{\sigma}_{\widehat{\pi}_A^-} = \left(\widehat{\sigma}^2_{\widehat{\pi}_A^+} + \widehat{\sigma}^2_{\widehat{\pi}_0(\lambda^*)} - 2\left(\frac{1}{1-\lambda^*}\right)\widehat{S}_{\gamma^*}^+\frac{\widehat{W}(\lambda^*)}{M} - 2\,(\gamma^*/2)\,\widehat{\sigma}^2_{\widehat{\pi}_0(\lambda^*)}\right)^{\frac{1}{2}}$ (using the equality $\widehat{\pi}_A^+ = 1 - \widehat{\pi}_0^+ - \widehat{\pi}_A^-$). Otherwise if $\gamma^* = \gamma^-$, we just reverse the superscripts $+/-$ in the two formulas above.

B Monte-Carlo Analysis

B.1 Under Cross-Sectional Independence

We use Monte-Carlo simulations to examine the performance of all estimators used in the paper: $\widehat{\pi}_0$, $\widehat{\pi}_A^-$, $\widehat{\pi}_A^+$, $\widehat{S}_\gamma^-$, $\widehat{F}_\gamma^-$, $\widehat{T}_\gamma^-$, and $\widehat{S}_\gamma^+$, $\widehat{F}_\gamma^+$, $\widehat{T}_\gamma^+$. We generate the $M \times 1$ vector of fund monthly excess returns, r_t, according to the four-factor model (market, size, book-to-market, and momentum factors):

$$\begin{aligned} r_t &= \alpha + \beta F_t + \varepsilon_t, \qquad t = 1, ..., T, \\ F_t &\sim N(0, \Sigma_F), \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2 I), \end{aligned} \tag{16}$$

where α denotes the $M \times 1$ vector of fund alphas, and β is the $M \times 4$ matrix of factor loadings. The 4×1 vector of factor excess returns, F_t, is normally distributed with covariance matrix Σ_F. ε_t is the $M \times 1$ vector of normally distributed residuals. We initially assume that the residuals are cross-sectionally independent and have the same variance σ_ε^2, so that the covariance matrix of ε_t can simply be written as $\sigma_\varepsilon^2 I$, where I is the $M \times M$ identity matrix.

Our estimators are compared with their respective true population values defined as follows. The parameters π_0, π_A^-, and π_A^+ denote the true proportions of zero-alpha, unskilled, and skilled funds. The expected proportions of unlucky and lucky funds, $E(F_\gamma^-)$ and $E(F_\gamma^+)$, are both equal to $\pi_0 \cdot \gamma/2$. To determine the expected proportions of unskilled and skilled funds, $E(T_\gamma^-)$ and $E(T_\gamma^+)$, we use the fact that, under the alternative hypothesis $\alpha_i \neq 0$, the fund t-statistic follows a non-central student distribution with $T-5$ degrees of freedom and a noncentrality parameter equal to $T^{\frac{1}{2}}\alpha_A/\sigma_\varepsilon$ (Davidson and MacKinnon (2004), p. 169):

$$
\begin{aligned}
E(T_\gamma^-) &= \pi_A^- \cdot prob\left(t < t_{T-5,\gamma/2}\,|H_A, \alpha_A < 0\right), \\
E(T_\gamma^+) &= \pi_A^+ \cdot prob\left(t > t_{T-5,1-\gamma/2}\,|H_A, \alpha_A > 0\right),
\end{aligned} \tag{17}
$$

where $t_{T-5,\gamma/2}$ and $t_{T-5,1-\gamma/2}$ denote the quantiles of probability level $\gamma/2$ and $1-\gamma/2$, respectively (these quantiles correspond to the thresholds t_γ^- and t_γ^+ used in the text). Finally, we have $E(S_\gamma^-) = E(F_\gamma^-) + E(T_\gamma^-)$, and $E(S_\gamma^+) = E(F_\gamma^+) + E(T_\gamma^+)$.

To compute these population values, we need to set values for the (true) proportions π_0, π_A^-, π_A^+, as well as for the means of the non-central student distributions (required to compute Equation (17)). In order to set realistic values, we estimate π_0, π_A^-, and π_A^+ at the end of each of the final five years of our sample (2002-2006) using the entire return history for each fund up to that point in time. These estimates are then averaged to produce values that reflect the recent trend observed in Figure 4: $\pi_0 = 75\%$, $\pi_A^- = 23\%$, and $\pi_A^+ = 2\%$. To determine the means of the t-statistic distributions of the unskilled and skilled funds, we use a simple calibration method. We first compute the average $\widehat{T}_\gamma^-$ and $\widehat{T}_\gamma^+$ (at $\gamma = 0.20$) over the final 5 years of our sample (2002-2006). Inserting these values along with $\pi_A^- = 23\%$ and $\pi_A^+ = 2\%$ in Equation (17), we can determine what are the means of the distributions which satisfy both equalities. The resulting values are -2.5 and 3, and correspond to an annual four-factor alpha of -3.2% and 3.8%, respectively (using the equality $t_A = T^{\frac{1}{2}}\alpha_A/\sigma_\varepsilon$).

The total number of funds, M, used in the simulation is equal to 1,400.[25] The input for β is equal to the empirical loadings of a random draw of 1,400 funds (among the total population of 2,076 funds). Consistent with our database, we set $T = 384$ (months), $\sigma_\varepsilon = 0.021$ (equal to the empirical average across the 1,400 funds), and proxy Σ_F by its empirical counterpart. To build the vector of fund alphas, α, we need to determine the identity of the unskilled and skilled funds. This is done by randomly choosing 322 funds (i.e., 23% of the entire population) to which we assign a negative alpha (-3.2% per

[25] We use this sample size to allow for comparison with the dependence case (described hereafter), which uses a sample of 1,400 correlated fund returns. Since our original sample of funds is larger than 1,400 ($M = 2,076$), our assessment of the precision of the estimators in this section is conservative.

year), and 28 funds (2% of the population), to which we assign a positive alpha (3.8% per year).

After randomly drawing F_t and ε_t ($t = 1, ..., 384$), we construct the fund return time-series according to Equation (14), and compute their t-statistics by regressing the fund returns on the four-factor model. To determine the alpha p-values, we use the fact that the fund t-statistic follows a Student distribution with $T - 5$ degrees of freedom under the null hypothesis $\alpha_i = 0$. Then, we compute $\widehat{\pi}_0$, $\widehat{\pi}_A^-$, and $\widehat{\pi}_A^+$ using Equations (5) and (8). $\widehat{S}_\gamma^-$ and $\widehat{S}_\gamma^+$ correspond to the observed number of significant funds with negative and positive alphas, respectively. $\widehat{F}_\gamma^-$ and $\widehat{F}_\gamma^+$ are computed with Equation (6). $\widehat{T}_\gamma^-$ and $\widehat{T}_\gamma^+$ are given in Equation (7). We repeat this procedure 1,000 times.

In Table AI, we compare the average value of each estimator (over the 1,000 replications) with the true values. The figures in parentheses denote the lower and upper bounds of the estimator 90%-confidence interval. We set γ equal to 0.05 and 0.20. In all cases, the simulation results reveal that the average values of our estimators closely match the true values, and that their 90%-confidence intervals are narrow. This result is not surprising in light of the large cross-section of funds available in our sample.

Please insert Table AI here

B.2 Under Cross-Sectional Dependence

The return-generating process is the same as the one shown in Equation (16), except that the fund residuals are cross-correlated:

$$\varepsilon_t \sim N(0, \Sigma), \tag{18}$$

where Σ denotes the $M \times M$ residual covariance matrix. The main constraint imposed on Σ is that it must be positive semi-definite. To achieve this, we select all funds with 60 valid return observations over the final five years (2002-2006), which is the period over which we have the largest possible cross-section of funds existing simultaneously–898 funds, whose covariance matrix, Σ_1, is directly estimated from the data.[26] To assess the precision of our estimators, we also need to account for the non-overlapping returns observed in the long-term fund data due to funds that do not exist at the same time. To address this issue, we introduce 502 uncorrelated funds, and write the covariance matrix

[26] The 25%, 50%, and 75% pairwise correlation quantiles are -0.09, 0.05, and 0.19, respectively.

for the resulting 1,400 funds as follows:[27]

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \sigma_\varepsilon^2 I \end{pmatrix}. \tag{19}$$

An an input for β, we use the empirical factor loadings of the 898 funds, along with the loadings of a random draw of the 502 remaining funds. The vector of fund alphas, α, is built by randomly choosing the identity of the unskilled and skilled funds, as in the independence case. The results in Table AII indicate that all estimators remain nearly unbiased ($\widehat{\pi}_0$, $\widehat{\pi}_A^-$, and $\widehat{\pi}_A^+$ exhibit small biases). Looking at the 90% confidence intervals, we logically observe that the dispersion of the estimators widens under cross-sectional dependence. However, the performance of the estimators is still very good.

Please insert Table AII here

Apart from this baseline dependence scenario, we also examine three other cases (the results are available upon request). First, we introduce correlation by block among each skill group (zero-alpha, unskilled, and skilled funds) to account for their possible similar bets. Inside each block (representing 10% of each skill group), we set the pairwise correlation equal to 0.15 or 0.30. Second, we use the residual factor specification proposed by Jones and Shanken (2005) in order to capture the role of non-priced factors. We assume that all fund residuals depend on a common residual factor, and that the unskilled and skilled funds are affected by specific residual factors. In the two cases, the results show that the precision of the estimators remain very close to those obtained under the independence case. Finally, we consider the extreme dependence case where the fund population only consists of the 898 correlated funds. We find that all estimators remain unbiased as shown in Tables AI and AII. But unsurprisingly, the confidence intervals widen slightly compared to those shown in Table AII (on average 2% are added on each side of the interval).

[27] The total number of fund pairs, P, is given by $M(M-1)/2$, where $M = 1,400$. If there are X uncorrelated funds in the population, the total number of uncorrelated fund pairs, I, equals $X \cdot (M - X) + X(X-1)/2$. In our data, 15% of the funds pairs do not have any return observations in common, and 55% of the observations are common to the remaining pairs (85%). Therefore, we estimate that the proportion of uncorrelated pairs is equal to 53% ($15\% + 85\% \cdot 45\%$). With 502 uncorrelated funds, I/P amounts to 58%, and is very close to the ratio observed in the data.

Table AI

Monte-Carlo Analysis under Cross-Sectional Independence

We examine the average value and the 90%-confidence interval (in parentheses) of the different estimators based on 1,000 replications. For each replication, we generate monthly fund returns for 1,400 funds and 384 periods using the four-factor model (market, size, book-to-market, and momentum factors). Fund residuals are independent from one another. The true parameter values for the proportions of zero-alpha, unskilled, and skilled funds (π_0, π_A^-, and π_A^+) are set to 75%, 23%, and 2%. We set the true four-factor annual alpha equal to -3.2% for the unskilled funds and +3.8% for the skilled ones. In each tail (left and right), we assess the precision of the different estimators at two significance levels (γ=0.05 and 0.20).

Fund Proportion	True	Estimator (90% interval)
Zero-alpha funds (π_0)	75.0	75.1 (71.7,78.6)
Unskilled funds (π_A^-)	23.0	22.9 (19.7,25.9)
Skilled funds (π_A^+)	2.0	2.0 (0.3,3.8)

	Significance level $\gamma = 0.05$		Significance level $\gamma = 0.20$	
Left Tail	True	Estimator (90% interval)	True	Estimator (90% interval)
Significant funds $E\left(S_\gamma^-\right)$	18.1	18.1 (16.4,19.7)	27.9	27.9 (26.1,30.0)
Unlucky funds $E\left(F_\gamma^-\right)$	1.8	1.8 (1.8,1.9)	7.5	7.5 (7.1,7.9)
Unskilled funds $E\left(T_\gamma^-\right)$	16.2	16.2 (14.6,17.9)	20.4	20.4 (18.2,22.7)

	Significance level $\gamma = 0.05$		Significance level $\gamma = 0.20$	
Right Tail	True	Estimator (90% interval)	True	Estimator (90% interval)
Significant funds $E\left(S_\gamma^+\right)$	3.6	3.6 (2.8,4.4)	9.4	9.4 (8.2,10.8)
Lucky funds $E\left(F_\gamma^+\right)$	1.8	1.8 (1.8,1.9)	7.5	7.5 (7.1,7.9)
Skilled funds $E\left(T_\gamma^+\right)$	1.7	1.7 (0.9,2.5)	1.9	1.9 (0.5,3.3)

Table AII

Monte-Carlo Analysis under Cross-Sectional Dependence

We examine the average value and the 90%-confidence interval (in parentheses) of the different estimators based on 1,000 replications. For each replication, we generate monthly fund returns for 1,400 funds and 384 periods using the four-factor model (market, size, book-to-market, and momentum factors). We assume that funds are cross-sectionally correlated and use the emprical covariance matrix of the fund residuals as the true covariance matrix. The true parameter values for the proportions of zero-alpha, unskilled, and skilled funds (π_0, π_A^-, and π_A^+) are set to 75%, 23%, and 2%. We set the true four-factor annual alpha equal to -3.2% for the unskilled funds and +3.8% for the skilled ones. In each tail (left and right), we assess the precision of the different estimators at two significance levels (γ=0.05 and 0.20).

Fund Proportion	True	Estimator (90% interval)
Zero-alpha funds (π_0)	75.0	75.2 (69.5,80.8)
Unskilled funds (π_A^-)	23.0	22.8 (17.0,28.9)
Skilled funds (π_A^+)	2.0	1.9 (0.0,6.5)

	Significance level $\gamma = 0.05$		Significance level $\gamma = 0.20$	
Left Tail	True	Estimator (90% interval)	True	Estimator (90% interval)
Significant funds $E\left(S_\gamma^-\right)$	18.1	18.1 (15.3,20.7)	27.9	27.9 (24.3,32.3)
Unlucky funds $E\left(F_\gamma^-\right)$	1.8	1.8 (1.6,2.1)	7.5	7.6 (6.6,8.3)
Unskilled funds $E\left(T_\gamma^-\right)$	16.2	16.2 (13.4,19.1)	20.4	20.4 (16.3,24.6)

	Significance level $\gamma = 0.05$		Significance level $\gamma = 0.20$	
Right Tail	True	Estimator (90% interval)	True	Estimator (90% interval)
Significant funds $E\left(S_\gamma^+\right)$	3.5	3.6 (2.4,5.3)	9.4	9.4 (6.6,12.6)
Lucky funds $E\left(F_\gamma^+\right)$	1.8	1.8 (1.6,2.1)	7.5	7.6 (6.6,8.3)
Skilled funds $E\left(T_\gamma^+\right)$	1.7	1.7 (0.5,3.8)	1.9	1.9 (0.1,5.6)

References

[1] Avramov, Doron, and Russ Wermers, 2006, Investing in Mutual Funds When Return are Predictable, Journal of Financial Economics 81, 339-377.

[2] Baks, Klaas P., Andrew Metrick, and Jessica Wachter, 2001, Should Investors Avoid all Actively Managed Mutual Funds? A Study in Bayesian Performance Evaluation, Journal of Finance 56, 45-85.

[3] Benjamini, Yoav, and Yosef Hochberg, 1995, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society 57, 289-300.

[4] Berk, Jonathan B., and Richard C. Green, 2004, Mutual Fund Flows and Performance in Rational Markets, Journal of Political Economy 112, 1269-1295.

[5] Carhart, Mark M., 1997, On Persistence in Mutual Fund Performance, Journal of Finance 52, 57-82.

[6] Chordia, Tarun, Richard Roll, and Avanidhar Subrahmanyam, 2000, Commonality in Liquidity, Journal of Financial Economics 56, 3-28.

[7] Christoffersen, Susan E. K., and David K. Musto, 2002, Demand Curves and the Pricing of Money Management, Review of Financial Studies 15, 1495-1524.

[8] Davidson, Russell, and James G. MacKinnon, 2004, Econometric Theory and Methods, Oxford University Press, New-York.

[9] Efron, Bradley, and Robert Tibshirani, 2002, Empirical Bayes Methods and False Discovery Rates for Microarrays, Genetic Epidemiology 23, 70-86.

[10] Elton, Edwin J., Martin J. Gruber, Sanjiv Das, and Matthew Hlavka, 1993, Efficiency with Costly Information: A Reinterpretation of Evidence from Managed Portfolios, Review of Financial Studies 6, 1-22.

[11] Elton, Edwin J., Martin J. Gruber, Christopher R. Blake, 1996, The Persistence of Risk-Adjusted Mutual Fund Performance, Journal of Business 69, 133-157.

[12] Elton, Edwin J., Martin J. Gruber, and Jeffrey Busse, 2004, Are Investors Rational? Choices among Index Funds, Journal of Finance 59, 261-288.

[13] Elton, Edwin J., Martin J. Gruber, and Christopher R. Blake, 2007, Participant Reaction and the Performance of Funds offered by 401(k) Plans, Journal of Financial Intermediation 16, 249-271.

[14] Fama, Eugene F., and Kenneth R. French, 1993, Common Risk Factors in the Returns on Stocks and Bonds, Journal of Financial Economics 33, 3-56.

[15] Ferson, Wayne E., and Rudi W. Schadt, 1996, Measuring Fund Strategy and Performance in Changing Economic Conditions, Journal of Finance 51, 425-461.

[16] Ferson, Wayne E., and Meijun Qian, 2004, Conditional Performance Evaluation Revisited, in Research Foundation Monograph of the CFA Institute.

[17] Genovese, Christopher, and Larry Wasserman, 2004, A Stochastic Process Approach to False Discovery Control, Annals of Statistics, 32, 1035-1061.

[18] Grinblatt, Mark, and Sheridan Titman, 1989, Mutual Fund Performance: An Analysis of Quarterly Portfolio Holdings, Journal of Business 62, 393-416.

[19] Grinblatt, Mark, and Sheridan Titman, 1993, Performance Measurement without Benchmarks: An Examination of Mutual Fund Returns, Journal of Business 66, 47-68.

[20] Hall Peter, Joel L. Horowitz, and Bing-Yi Jing, 1995, On Blocking Rules for the Bootstrap with Dependent Data, Biometrika 82, 561-574.

[21] Hendricks, Darryll, Jayendu Patel, and Richard Zeckhauser, 1993, Hot Hands in Mutual Funds: The Persistence of Performance, 1974-88, Journal of Finance 48, 93-130.

[22] Jensen, Michael C., 1968, The Performance of Mutual Funds in the Period 1945-1964, Journal of Finance 23, 389-416.

[23] Jones, Christopher S., and Jay Shanken, 2005, Mutual Fund Performance with Learning across Funds, Journal of Financial Economics 78, 507-552.

[24] Kosowski, Robert, Allan Timmermann, Halbert White, and Russ Wermers, 2006, Can Mutual Fund "Stars" Really Pick Stocks? New Evidence from a Bootstrap Analysis, Journal of Finance 61, 2551-2595.

[25] Kostovetsky, Leonard, 2007, Brain Drain: Are Mutual Funds Losing their Best Minds? Working Paper, Princeton University.

[26] Lehmann, Bruce N., and David M. Modest, 1987, Mutual Fund Performance Evaluation: A Comparison of Benchmarks and Benchmark Comparisons, Journal of Finance 42, 233-265.

[27] Newey, Whitney K., and Kenneth D. West, 1987, A Simple, Positive Semi-Definite, Heteroscedasticity and Autocorrelation Consistent Covariance Matrix, Econometrica 55, 703-708.

[28] Pastor, Lubos, and Robert F. Stambaugh, 2002a, Mutual Fund Performance and Seemingly Unrelated Assets, Journal of Financial Economics 63, 315-349.

[29] Pastor, Lubos, and Robert F. Stambaugh, 2002b, Investing in Equity Mutual Funds, Journal of Financial Economics 63, 351-380.

[30] Rea, John D., and Brian K. Reid, 1998, Trends in the Ownership Cost of Equity Mutual Funds, Investment Company Institute Perspective, November.

[31] Storey, John D., 2002, A Direct Approach to False Discovery Rates, Journal of the Royal Statistical Society 64, 479–498.

[32] Storey, John D., 2003, The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value, Annals of Statistics 31, 2013-2035.

[33] Storey, John D., and Robert Tibshirani, 2003, Statistical Significance for Genomewide Studies, Proceedings of the National Academy of Sciences 100, 9440-9445.

[34] Storey, John D., Jonathan E. Taylor, and David Siegmund, 2004, Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach, Journal of the Royal Statistical Society 66, 187-205.

[35] Sullivan Ryan, Allan Timmermann, and Halbert White, 1999, Data-Snooping, Technical Trading Rule Performance and the Bootstrap, Journal of Finance 54, 1647-1691.

[36] Tang, Yongqiang, Subhashis Ghosal, and Anindya Roy, 2007, Nonparametric Bayesian Estimation of Positive False Discovery Rates, Biometrics 63, 1126-1134.

[37] Wermers, Russ, 1999, Mutual Fund Herding and the Impact on Stock Prices, Journal of Finance 55, 581-622

[38] Wermers, Russ, 2000, Mutual Fund Performance: An Empirical Decomposition into Stock-Picking Talent, Style, Transaction Costs, and Expenses, Journal of Finance 55, 1655-1695.

Table I

Performance of the Equally-Weighted Portfolio of Funds

Results for the unconditional and conditional four-factor models are shown in Panels A and B for the entire fund population (All funds), as well as Growth, Aggressive Growth, and Growth & Income funds. The regressions are based on monthly data between January 1975 and December 2006. Each Panel contains the estimated annualized alpha ($\widehat{\alpha}$), the estimated exposures to the market ($\widehat{b}_m$), size ($\widehat{b}_{smb}$), book-to-market ($\widehat{b}_{hml}$), and momentum factors ($\widehat{b}_{mom}$), as well as the adjusted R^2 of an equally-weighted portfolio that includes all funds that exist at the beginning of each month. Figures in parentheses denote the Newey-West (1987) heteroskedasticity and autocorrelation consistent estimates of p-values, under the null hypothesis that the regression parameters are equal to zero.

Panel A Unconditional Four-Factor Model

	$\widehat{\alpha}$	$\widehat{b}_m$	$\widehat{b}_{smb}$	$\widehat{b}_{hml}$	$\widehat{b}_{mom}$	R^2
All (2,076)	-0.48%	0.95	0.17	-0.01	0.02	98.0%
	(0.12)	(0.00)	(0.00)	(0.38)	(0.09)	
Growth (1,304)	-0.45%	0.95	0.16	-0.03	0.02	98.0%
	(0.16)	(0.00)	(0.00)	(0.15)	(0.07)	
Aggressive	-0.53%	1.04	0.43	-0.17	0.09	95.8%
Growth (388)	(0.22)	(0.00)	(0.00)	(0.00)	(0.00)	
Growth &	-0.47%	0.87	-0.04	0.17	-0.03	98.2%
Income (384)	(0.09)	(0.00)	(0.02)	(0.00)	(0.01)	

Panel B Conditional Four-Factor Model

	$\widehat{\alpha}$	$\widehat{b}_m$	$\widehat{b}_{smb}$	$\widehat{b}_{hml}$	$\widehat{b}_{mom}$	R^2
All (2,076)	-0.60%	0.96	0.17	-0.02	0.02	98.2%
	(0.09)	(0.00)	(0.00)	(0.23)	(0.08)	
Growth (1,304)	-0.59%	0.96	0.16	-0.03	0.03	98.2%
	(0.10)	(0.00)	(0.00)	(0.08)	(0.05)	
Aggressive	-0.49%	1.05	0.43	-0.19	0.08	96.2%
Growth (388)	(0.24)	(0.00)	(0.00)	(0.00)	(0.00)	
Growth &	-0.58%	0.87	-0.04	0.16	-0.03	98.3%
Income (384)	(0.05)	(0.00)	(0.02)	(0.00)	(0.02)	

Table II

Impact of Luck on Long-Term Performance

Panel A displays the estimated proportions of zero-alpha, unskilled, and skilled funds in the entire fund population (2,076 funds). We measure fund performance with the unconditional four-factor model over the entire period 1975-2006. Panel B counts the proportions of significant funds in the left and right tails of the cross-sectional t-statistic distribution ($\widehat{S}_\gamma^-$, $\widehat{S}_\gamma^+$) at four significance levels (γ=0.05, 0.10, 0.15, 0.20). In the leftmost columns, the significant group in the left tail, $\widehat{S}_\gamma^-$, is decomposed into unlucky and unskilled funds ($\widehat{F}_\gamma^-$, $\widehat{T}_\gamma^-$). In the rightmost columns , the significant group in the right tail, $\widehat{S}_\gamma^+$, is decomposed into lucky and skilled funds ($\widehat{F}_\gamma^+$, $\widehat{T}_\gamma^+$). Figures in parentheses denote the standard deviation of the different estimators. The bottom of Panel B also presents the characteristics of each significant group ($\widehat{S}_\gamma^-$, $\widehat{S}_\gamma^+$): the average estimated alpha (% per year), expense ratio (% per year), turnover (% per year), and median size measured by the total net asset under management (millions USD).

Panel A Proportion of Unskilled and Skilled Funds

	Zero alpha($\widehat{\pi}_0$)	Non-zero alpha	Unskilled($\widehat{\pi}_A^-$)	Skilled($\widehat{\pi}_A^+$)
Proportion	75.4 (2.5)	24.6	24.0 (2.3)	0.6 (0.8)
Number	1,565	511	499	12

Panel B Impact of Luck in the Left and Right Tails

	Left Tail				Right Tail				
Signif. level(γ)	0.05	0.10	0.15	0.20	0.20	0.15	0.10	0.05	Signif. level(γ)
Signif. $\widehat{S}_\gamma^-$(%)	11.6	17.2	21.5	25.4	8.2	6.0	4.2	2.2	Signif. $\widehat{S}_\gamma^+$(%)
	(0.7)	(0.8)	(0.9)	(0.9)	(0.6)	(0.5)	(0.4)	(0.3)	
Unlucky $\widehat{F}_\gamma^-$(%)	1.9	3.8	5.6	7.6	7.6	5.6	3.8	1.9	Lucky $\widehat{F}_\gamma^+$(%)
	(0.0)	(0.1)	(0.2)	(0.3)	(0.3)	(0.2)	(0.1)	(0.0)	
Unskilled $\widehat{T}_\gamma^-$(%)	9.8	13.6	16.1	18.2	0.6	0.4	0.4	0.3	Skilled $\widehat{T}_\gamma^+$(%)
	(0.7)	(0.9)	(1.0)	(1.1)	(0.7)	(0.6)	(0.5)	(0.3)	
Alpha(% year)	-5.5	-5.0	-4.7	-4.6	4.8	5.2	5.6	6.5	Alpha(% year)
Exp.(% year)	1.6	1.5	1.5	1.5	1.2	1.2	1.2	1.2	Exp.(% year)
Turn.(% year)	100	99	98	96	126	95	95	105	Turn.(% year)
Size(million \$)	81	88	86	84	731	985	888	745	Size(million \$)

Table III

Impact of Luck on Short-Term Performance

Panel A displays the estimated proportions of zero-alpha, unskilled, and skilled funds in the entire fund population (3,311 funds). We measure fund performance with the unconditional four-factor model over non-overlapping 5-year periods between 1977-2006. Panel B counts the proportions of significant funds in the left and right tails of the cross-sectional t-statistic distribution ($\widehat{S}_{\gamma}^{-}$, $\widehat{S}_{\gamma}^{+}$) at four significance levels (γ=0.05, 0.10, 0.15, 0.20). In the leftmost columns, the significant group in the left tail, $\widehat{S}_{\gamma}^{-}$, is decomposed into unlucky and unskilled funds ($\widehat{F}_{\gamma}^{-}$, $\widehat{T}_{\gamma}^{-}$). In the rightmost columns , the significant group in the right tail, $\widehat{S}_{\gamma}^{+}$, is decomposed into lucky and skilled funds ($\widehat{F}_{\gamma}^{+}$, $\widehat{T}_{\gamma}^{+}$). Figures in parentheses denote the standard deviation of the different estimators. The bottom of Panel B also presents the characteristics of each significant group ($\widehat{S}_{\gamma}^{-}$, $\widehat{S}_{\gamma}^{+}$): the average estimated alpha (% per year), expense ratio (% per year), turnover (% per year), and median size measured by the total net asset under management (millions USD).

Panel A Proportion of Unskilled and Skilled Funds

	Zero alpha($\widehat{\pi}_0$)	Non-zero alpha	Unskilled($\widehat{\pi}_A^{-}$)	Skilled($\widehat{\pi}_A^{+}$)
Proportion	72.2 (2.0)	27.8	25.4 (1.7)	2.4 (0.7)
Number	2,390	921	841	80

Panel B Impact of Luck in the Left and Right Tails

	Left Tail				Right Tail				
Signif. level(γ)	0.05	0.10	0.15	0.20	0.20	0.15	0.10	0.05	Signif. level(γ)
Signif. $\widehat{S}_{\gamma}^{-}$(%)	11.2	16.8	21.4	24.9	9.6	7.8	5.9	3.5	Signif. $\widehat{S}_{\gamma}^{+}$(%)
	(0.5)	(0.6)	(0.7)	(0.8)	(0.5)	(0.5)	(0.4)	(0.3)	
Unlucky $\widehat{F}_{\gamma}^{-}$(%)	1.8	3.6	5.4	7.2	7.2	5.4	3.6	1.8	Lucky $\widehat{F}_{\gamma}^{+}$(%)
	(0.0)	(0.0)	(0.1)	(0.2)	(0.2)	(0.1)	(0.0)	(0.0)	
Unskilled $\widehat{T}_{\gamma}^{-}$(%)	9.4	13.2	16.0	17.7	2.4	2.4	2.3	1.7	Skilled $\widehat{T}_{\gamma}^{+}$(%)
	(0.6)	(0.7)	(0.8)	(0.8)	(0.6)	(0.5)	(0.4)	(0.3)	
Alpha(% year)	-6.5	-5.9	-5.5	-5.3	6.7	7.0	7.2	7.5	Alpha(% year)
Exp.(% year)	1.4	1.3	1.3	1.3	1.2	1.2	1.2	1.2	Exp.(% year)
Turn.(% year)	98	105	100	99	80	81	83	81	Turn.(% year)
Size(million \$)	242	244	252	244	623	628	664	749	Size(million \$)

Table IV

Short-Term Performance across Investment Categories

The impact of luck in the left and right tails for three investment categories (Growth, Aggressive Growth, and Growth & Income funds) is presented in Panels A, B, and C, respectively. We measure fund performance with the unconditional four-factor model over non-overlapping 5-year periods between 1977-2006. For each panel, we count the proportions of significant funds in the left and right tails of the cross-sectional t-statistic distribution ($\widehat{S}_{\gamma}^{-}$, $\widehat{S}_{\gamma}^{+}$) at four significance levels (γ=0.05, 0.10, 0.15, 0.20). In the leftmost columns, the significant group in the left tail, $\widehat{S}_{\gamma}^{-}$, is decomposed into unlucky and unskilled funds ($\widehat{F}_{\gamma}^{-}$, $\widehat{T}_{\gamma}^{-}$). In the rightmost columns , the significant group in the right tail, $\widehat{S}_{\gamma}^{+}$, is decomposed into lucky and skilled funds ($\widehat{F}_{\gamma}^{+}$, $\widehat{T}_{\gamma}^{+}$). Figures in parentheses denote the standard deviation of the different estimators. The bottom of Panel B also presents the characteristics of each significant group ($\widehat{S}_{\gamma}^{-}$, $\widehat{S}_{\gamma}^{+}$): the average estimated alpha (% per year), expense ratio (% per year), turnover (% per year), and median size measured by the total net asset under management (millions USD).

Panel A Growth funds

	Left Tail				Right Tail				
Signif. level(γ)	0.05	0.10	0.15	0.20	0.20	0.15	0.10	0.05	Signif. level(γ)
Signif. $\widehat{S}_{\gamma}^{-}$(%)	11.3	16.6	21.4	25.2	9.9	8.1	5.9	3.5	Signif. $\widehat{S}_{\gamma}^{+}$(%)
	(0.7)	(0.8)	(0.9)	(1.0)	(0.7)	(0.6)	(0.5)	(0.4)	
Unlucky $\widehat{F}_{\gamma}^{-}$(%)	1.8	3.6	5.5	7.3	7.3	5.5	3.6	1.8	Lucky $\widehat{F}_{\gamma}^{+}$ (%)
	(0.0)	(0.1)	(0.2)	(0.2)	(0.2)	(0.2)	(0.1)	(0.0)	
Unskilled $\widehat{T}_{\gamma}^{-}$(%)	9.5	13.0	15.9	17.9	2.6	2.6	2.3	1.7	Skilled $\widehat{T}_{\gamma}^{+}$ (%)
	(0.7)	(0.9)	(1.0)	(1.1)	(0.8)	(0.7)	(0.6)	(0.4)	
Alpha(% year)	-6.0	-5.6	-5.2	-5.1	6.8	6.8	6.8	7.3	Alpha(% year)
Exp.(% year)	1.4	1.3	1.3	1.3	1.2	1.2	1.2	1.2	Exp.(% year)
Turn.(% year)	101	113	110	107	79	79	79	75	Turn.(% year)
Size(million $)	253	253	253	252	589	593	621	580	Size(million $)

Table IV

Short-Term Performance across Investment Categories (Continued)

Panel B Aggressive Growth funds

	Left Tail				Right Tail				
Signif. level(γ)	0.05	0.10	0.15	0.20	0.20	0.15	0.10	0.05	Signif. level(γ)
Signif. $\widehat{S}_{\gamma}^{-}$(%)	12.0	16.0	19.4	22.2	11.2	9.4	7.1	4.9	Signif. $\widehat{S}_{\gamma}^{+}$(%)
	(1.1)	(1.4)	(1.6)	(1.8)	(1.3)	(1.1)	(1.0)	(0.8)	
Unlucky $\widehat{F}_{\gamma}^{-}$(%)	1.8	3.6	5.4	7.2	7.2	5.4	3.6	1.8	Lucky $\widehat{F}_{\gamma}^{+}$(%)
	(0.1)	(0.2)	(0.3)	(0.4)	(0.4)	(0.3)	(0.2)	(0.1)	
Unskilled $\widehat{T}_{\gamma}^{-}$(%)	10.2	12.3	14.0	15.0	4.0	4.0	3.5	3.1	Skilled $\widehat{T}_{\gamma}^{+}$(%)
	(1.3)	(1.6)	(1.7)	(1.9)	(1.4)	(1.2)	(1.1)	(0.9)	
Alpha(% year)	-9.3	-8.6	-8.1	-7.6	8.5	8.8	9.7	9.7	Alpha(% year)
Exp.(% year)	1.5	1.5	1.5	1.5	1.3	1.3	1.3	1.3	Exp.(% year)
Turn.(% year)	127	123	119	117	105	104	107	104	Turn.(% year)
Size(million $)	154	187	160	192	1,014	949	1,021	1,073	Size(million $)

Panel C Growth & Income funds

	Left Tail				Right Tail				
Signif. level(γ)	0.05	0.10	0.15	0.20	0.20	0.15	0.10	0.05	Signif. level(γ)
Signif. $\widehat{S}_{\gamma}^{-}$(%)	11.5	17.4	22.5	26.8	7.3	5.5	3.7	1.8	Signif. $\widehat{S}_{\gamma}^{+}$(%)
	(1.1)	(1.4)	(1.5)	(1.6)	(1.0)	(0.8)	(0.7)	(0.5)	
Unlucky $\widehat{F}_{\gamma}^{-}$ (%)	1.8	3.7	5.5	7.3	7.3	5.5	3.7	1.8	Lucky $\widehat{F}_{\gamma}^{+}$(%)
	(0.1)	(0.2)	(0.3)	(0.4)	(0.4)	(0.3)	(0.2)	(0.1)	
Unskilled $\widehat{T}_{\gamma}^{-}$ (%)	9.8	13.7	17.0	19.5	0.0	0.0	0.0	0.0	Skilled $\widehat{T}_{\gamma}^{+}$(%)
	(1.2)	(1.5)	(1.7)	(1.8)	(1.1)	(0.9)	(0.8)	(0.5)	
Alpha(% year)	-4.9	-4.5	-4.2	-4.0	4.9	5.3	5.1	4.9	Alpha(% year)
Exp.(% year)	1.3	1.2	1.2	1.2	1.1	1.1	1.0	1.0	Exp.(% year)
Turn.(% year)	69	68	66	64	56	57	54	45	Turn.(% year)
Size(million $)	295	346	348	337	492	482	473	1,787	Size(million $)

Table V

Performance Persistence Based on the False Discovery Rate

For each of the five FDR targets (10%, 30%, 50%, 70%, and 90%), Panel A contains descriptive statistics on the FDR level ($\widehat{FDR}^{+}_{\gamma^P}$) achieved by each portfolio, as well as the proportion of funds in the population that it includes ($\widehat{S}^{+}_{\gamma^P}$). The panel shows the average values of $\widehat{FDR}^{+}_{\gamma^P}$ and $\widehat{S}^{+}_{\gamma^P}$ over the 27 annual formation dates (from December 1979 to 2005), as well as their respective distributions. Panel B displays the performance of each portfolio over the period 1980-2006. We estimate the annual four-factor alpha ($\widehat{\alpha}$) with its bootstrap p-value, its annual residual standard deviation ($\widehat{\sigma}_{\varepsilon}$), its annual information ratio (IR=$\widehat{\alpha}/\widehat{\sigma}_{\varepsilon}$), its loadings on the market ($\widehat{b}_m$), size ($\widehat{b}_{smb}$), book-to-market ($\widehat{b}_{hml}$), and momentum factors ($\widehat{b}_{mom}$), and its annual excess mean, and standard deviation. In Panel C, we examine the turnover of each portfolio. We compute the proportion of funds that are still included in the portfolio 1, 2, 3, 4, and 5 years after their initial selection.

Panel A Portfolio Statistics

	Achieved False Discovery Rate $\left(\widehat{FDR}^{+}_{\gamma^P}\right)$					Included proportion of funds($\widehat{S}^{+}_{\gamma^P}$)				
	Mean	10-30	30-50	50-70	>70%	Mean	0-6	6-12	12-24	>24%
FDR10%	41.5%	14	6	1	6	3.0%	25	2	0	0
FDR30%	47.5%	8	12	1	6	8.2%	15	7	3	2
FDR50%	60.4%	0	14	7	6	20.9%	5	7	4	11
FDR70%	71.3%	0	4	12	11	29.7%	1	5	5	16
FDR90%	75.0%	0	4	9	14	33.7%	0	3	4	20

Panel B Performance Analysis

	$\widehat{\alpha}$(p-value)	$\widehat{\sigma}_{\varepsilon}$	IR	$\widehat{b}_m$	$\widehat{b}_{smb}$	$\widehat{b}_{hml}$	$\widehat{b}_{mom}$	Mean	Std dev
FDR10%	1.45%(0.04)	4.0%	0.36	0.93	0.16	-0.04	-0.02	8.3%	15.4%
FDR30%	1.15%(0.05)	3.3%	0.35	0.94	0.17	-0.02	-0.03	8.1%	15.4%
FDR50%	0.95%(0.10)	2.9%	0.33	0.96	0.20	-0.06	-0.01	8.1%	16.1%
FDR70%	0.68%(0.15)	2.7%	0.25	0.97	0.19	-0.06	-0.01	7.9%	16.1%
FDR90%	0.39%(0.30)	2.7%	0.14	0.97	0.19	-0.05	-0.00	7.8%	16.0%

Panel C Portfolio Turnover

	Proportion of funds remaining in the portfolio...				
	After 1 year	After 2 years	After 3 years	After 4 years	After 5 years
FDR10%	36.7	12.8	3.4	0.8	0.0
FDR30%	40.0	14.7	5.1	1.7	1.3
FDR50%	48.8	23.5	12.3	4.7	2.6
FDR70%	52.2	29.0	17.4	9.5	6.3
FDR90%	55.9	33.8	20.4	13.0	8.5

Table VI
Impact of Luck on Long-Term Pre-Expense Performance

Panel A displays the estimated proportions of zero-alpha, unskilled, and skilled funds in the entire fund population on a pre-expense basis (1,836 funds). We add the monthly expenses to net return of each fund, and measure performance with the unconditional four-factor model over the entire period 1975-2006. Panel B counts the proportions of significant funds in the left and right tails of the cross-sectional t-statistic distribution ($\widehat{S}_\gamma^-$, $\widehat{S}_\gamma^+$) at four significance levels (γ=0.05, 0.10, 0.15, 0.20). In the leftmost columns, the significant group in the left tail, $\widehat{S}_\gamma^-$, is decomposed into unlucky and unskilled funds ($\widehat{F}_\gamma^-$, $\widehat{T}_\gamma^-$). In the rightmost columns , the significant group in the right tail, $\widehat{S}_\gamma^+$, is decomposed into lucky and skilled funds ($\widehat{F}_\gamma^+$, $\widehat{T}_\gamma^+$). Figures in parentheses denote the standard deviation of the different estimators. The bottom of Panel B also presents the characteristics of each significant group ($\widehat{S}_\gamma^-$, $\widehat{S}_\gamma^+$): the average estimated alpha prior to expenses (in % per year), expense ratio (in % per year), turnover (in % per year), and median size measured by the total net asset under management (in millions of dollars).

Panel A Proportion of Unskilled and Skilled Funds

	Zero alpha($\widehat{\pi}_0$)	Non-zero alpha	Unskilled($\widehat{\pi}_A^-$)	Skilled($\widehat{\pi}_A^+$)
Proportion	85.9 (2.7)	14.1	4.5 (1.0)	9.6 (1.5)
Number	1,577	259	176	83

Panel B Impact of Luck in the Left and Right Tails

	Left Tail				Right Tail				
Signif. level(γ)	0.05	0.10	0.15	0.20	0.20	0.15	0.10	0.05	Signif. level(γ)
Signif. $\widehat{S}_\gamma^-$(%)	4.3	7.5	10.2	12.8	17.3	13.1	9.3	5.8	Signif. $\widehat{S}_\gamma^+$(%)
	(0.5)	(0.6)	(0.7)	(0.8)	(0.9)	(0.8)	(0.7)	(0.5)	
Unlucky $\widehat{F}_\gamma^-$(%)	2.1	4.3	6.4	8.6	8.6	6.4	4.3	2.1	Lucky $\widehat{F}_\gamma^+$(%)
	(0.0)	(0.1)	(0.1)	(0.2)	(0.2)	(0.1)	(0.1)	(0.0)	
Unskilled $\widehat{T}_\gamma^-$(%)	2.2	3.2	3.8	4.2	8.7	6.6	5.0	3.6	Skilled $\widehat{T}_\gamma^+$(%)
	(0.5)	(0.6)	(0.8)	(0.9)	(1.0)	(0.9)	(0.7)	(0.5)	
Pre Expense									Pre Expense
Alpha(% year)	-5.9	-5.2	-4.8	-4.5	4.4	4.8	5.0	5.3	Alpha(% year)
Exp.(% year)	1.5	1.4	1.3	1.3	1.3	1.3	1.3	1.3	Exp.(% year)
Turn.(% year)	105	108	108	108	106	111	122	84	Turn.(% year)
Size(million $)	81	90	93	90	430	578	948	1,000	Size(million $)

Table VII

Loadings on Omitted Factors

We determine the proportions of significant funds in the left and right tails ($\widehat{S}_{\gamma}^{-}$, $\widehat{S}_{\gamma}^{+}$) at four significance levels (γ=0.05, 0.10, 0.15, 0.20) according to each asset-pricing model over the period 1975-2006. For each of these significant groups, we compute their average loadings on the omitted factors from the four-factor model: size ($\widehat{b}_{smb}$), book-to-market ($\widehat{b}_{hml}$), and momentum ($\widehat{b}_{mom}$). Panel A shows the results obtained with the unconditional CAPM, while Panel B repeats the same procedure with the unconditional Fama-French model.

Panel A Unconditional CAPM

	Left Tail				Right Tail				
Signif. level(γ)	0.05	0.10	0.15	0.20	0.20	0.15	0.10	0.05	Signif. level(γ)
Size($\widehat{b}_{smb}$)	0.06	0.07	0.09	0.09	0.27	0.28	0.28	0.36	Size($\widehat{b}_{smb}$)
Book($\widehat{b}_{hml}$)	-0.14	-0.14	-0.13	-0.14	0.34	0.35	0.36	0.37	Book($\widehat{b}_{hml}$)
Mom.($\widehat{b}_{mom}$)	0.00	0.00	0.00	0.01	-0.01	-0.01	-0.02	-0.01	Mom.($\widehat{b}_{mom}$)

Panel B Unconditional Fama-French model

	Left Tail				Right Tail				
Signif. level(γ)	0.05	0.10	0.15	0.20	0.20	0.15	0.10	0.05	Signif. level(γ)
Mom.($\widehat{b}_{mom}$)	-0.02	-0.03	-0.02	-0.03	0.09	0.10	0.11	0.12	Mom.($\widehat{b}_{mom}$)

Figure 1

Outcome of the Multiple Performance Test

Panel A shows the distribution of the fund t-statistic across the three skill groups (zero-alpha, unskilled, and skilled funds). We set the true four-factor alpha equal to -3.2% and +3.8% per year for the unskilled and skilled funds (implying that the t-statistic distributions are centered at -2.5 and +3). Panel B displays the cross-sectional t-statistic distribution. It is a mixture of the three distributions in Panel A, where the weight on each distribution depends on the proportion of zero-alpha, unskilled, and skilled funds in the population (π_0, π_A^-, and π_A^+). In this example, we set $\pi_0 = 75\%$, $\pi_A^- = 23\%$, and $\pi_A^+ = 2\%$ to match our average estimated values over the final 5 years of our sample.

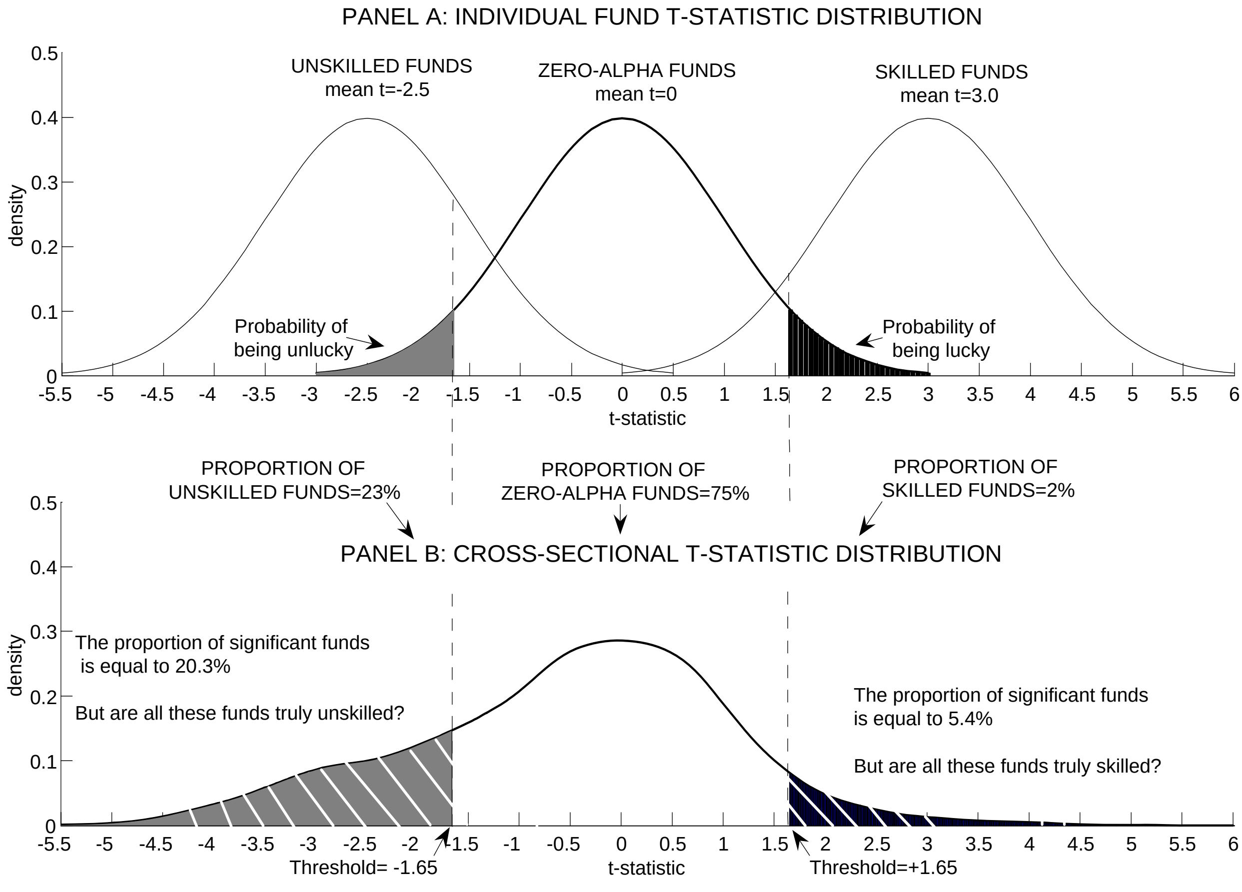


Figure 2
Histogram of Fund p-values

This figure represents the p-value histogram of M=2,076 funds (as in our database). For each fund, we draw its t-statistic from one of the distributions in Figure 1 (Panel A) according to the proportion of zero-alpha, unskilled, and skilled funds in the population (π_0, π_A^-, and π_A^+). In this example, we set $\pi_0 = 75\%$, $\pi_A^- = 23\%$, and $\pi_A^+ = 2\%$ to match our average estimated values over the final 5 years of our sample. Then, we compute the two-sided p-values of each fund from its respective t-statistic, and plot them in the histogram.

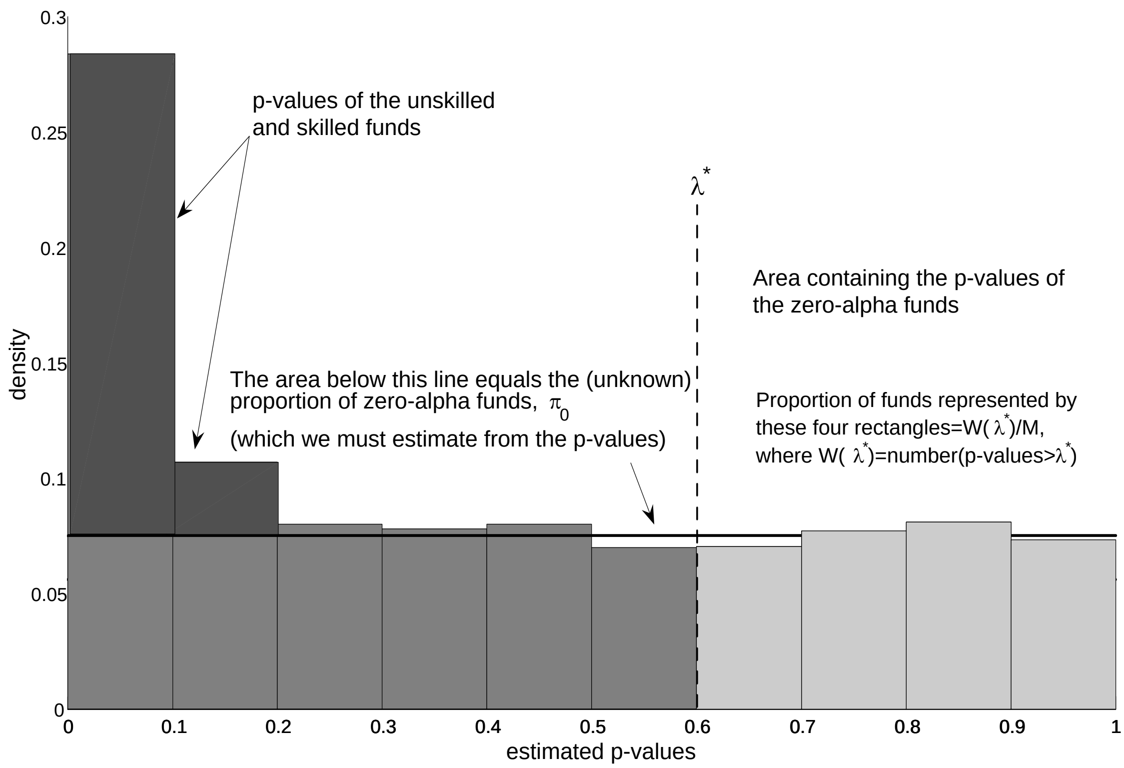
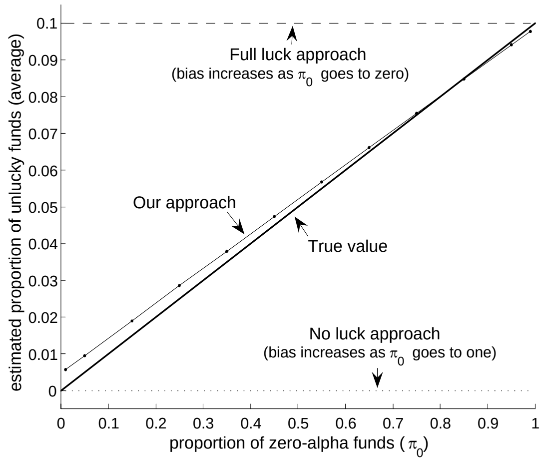


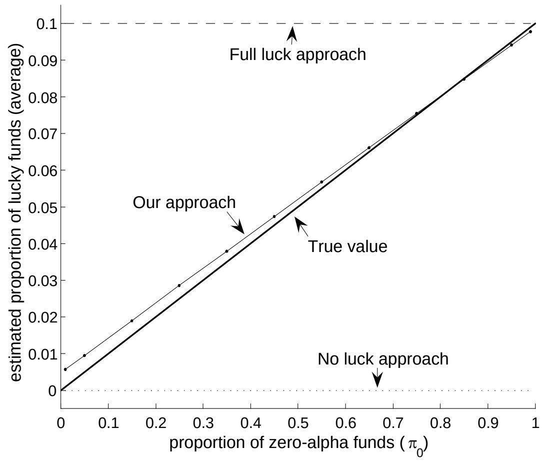
Figure 3

Measuring Luck: Comparison with Existing Approaches

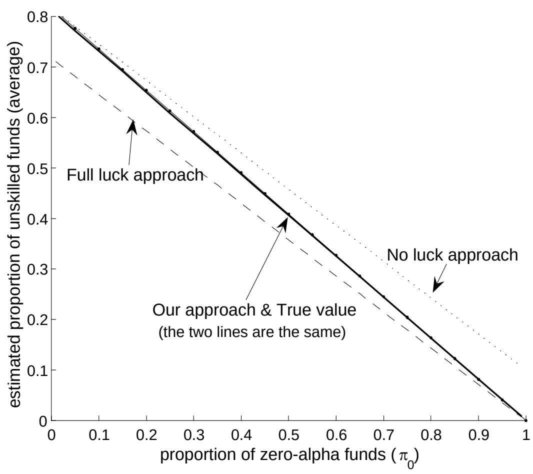
This figure examines the bias of different estimators produced by the three approaches ("no luck", "full luck", and "our approach") as a function of the proportion of zero-alpha funds, π_0. We examine the estimators of the proportions of unlucky, lucky, unskilled, and skilled funds in Panel A, B, C, and D, respectively. The "no luck" approach assumes that π_0=0, the "full luck" approach assumes that π_0=1, while "our approach" estimates π_0 directly from the data. For each approach, we compare the average estimator value (over 1,000 replications) with the true population value. For each replication, we draw the t-statistic for each fund i (i=1,...,2,076) from one of the distributions in Figure 1 (Panel A) according to the weights π_0, π_A^-, and π_A^+, and compute the different estimators at the significance level $\gamma = 0.20$. For each π_0, the ratio π_A^- over π_A^+ is held fixed to 11.5 (0.23/0.02) as in Figure 1.



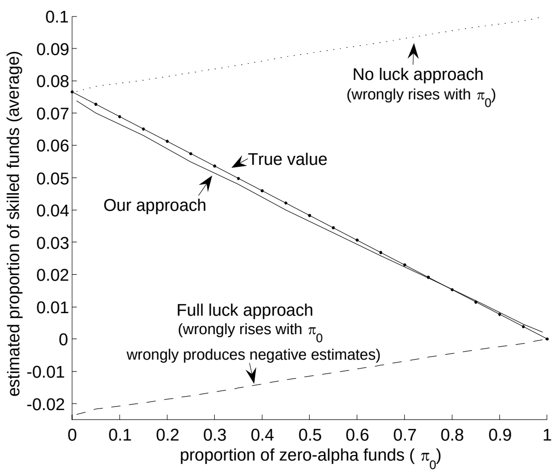
(A) Unlucky funds (left tail)



(B) Lucky funds (right tail)



(C) Unskilled funds (left tail)



(D) Skilled funds (right tail)

Figure 4
Evolution of Mutual Fund Performance over Time

Panel A plots the evolution of the estimated proportions of unskilled and skilled funds ($\widehat{\pi}_A^-$ and $\widehat{\pi}_A^+$) between 1989 and 2006. At the end of each year, we measure $\widehat{\pi}_A^-$ and $\widehat{\pi}_A^+$ using the entire fund return history up to that point. The initial estimates at the end of 1989 cover the period 1975-1989, while the last ones in 2006 use the period 1975-2006. The performance of each fund is measured with the unconditional four-factor model. Panel B displays the growth in the mutual fund industry (proxied by the total number of funds used to compute $\widehat{\pi}_A^-$ and $\widehat{\pi}_A^+$ over time), as well as its average alpha (in % per year).

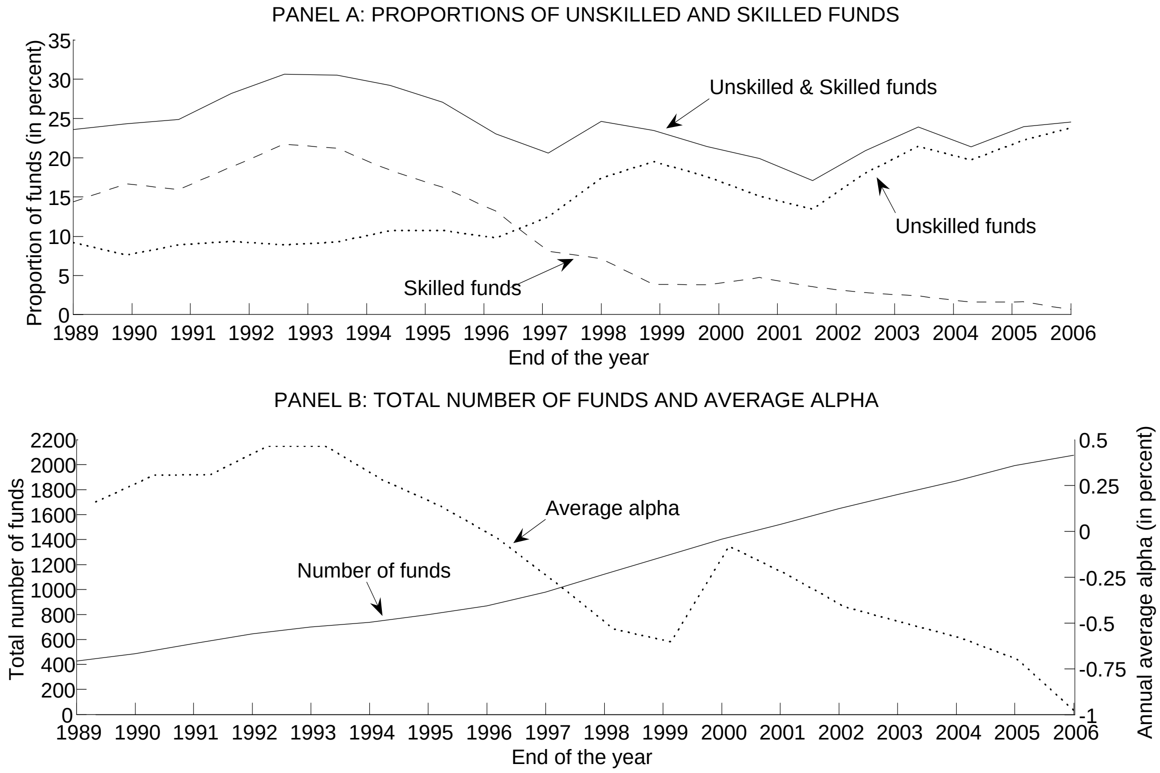
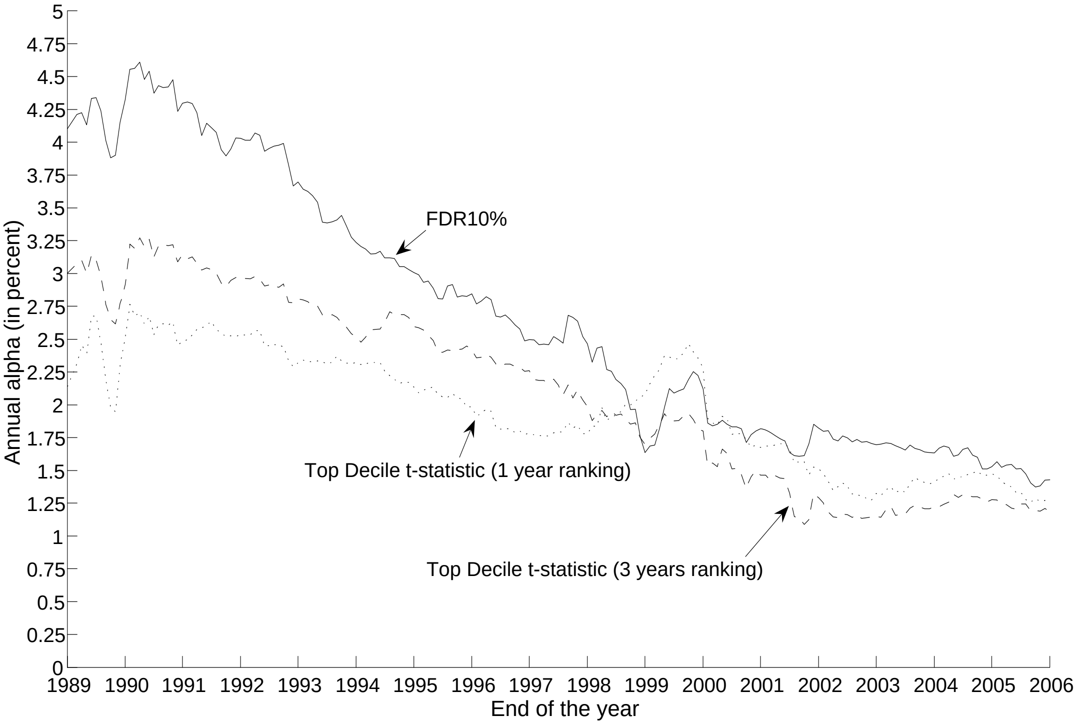


Figure 5

Performance of the Portfolios FDR10% over Time

The graph plots the evolution of the estimated annual four-factor alpha of the portfolio $FDR10\%$. This portfolio contains funds located in the right tail of the cross-sectional t-statistic distribution such that the targeted proportion of lucky funds included in the portfolio is equal to 10%. At the end of each year from 1989 to 2006, the portfolio's alpha is estimated using the portfolio return history up to that point. The initial estimates cover the period 1980-1989 (the first five years are used for the initial portfolio formation on December 31, 1979), while the last ones use the entire portfolio history from 1980 up to 2006. For comparison purposes, we also show the performance of top decile portfolios formed according to a t-statistic ranking, where the t-statistic is estimated over the prior one and three years, respectively.



THE JOURNAL OF FINANCE • VOL. LXIII, NO. 4 • AUGUST 2008

The Selection and Termination of Investment Management Firms by Plan Sponsors

AMIT GOYAL and SUNIL WAHAL*

ABSTRACT

We examine the selection and termination of investment management firms by 3,400 plan sponsors between 1994 and 2003. Plan sponsors hire investment managers after large positive excess returns but this return-chasing behavior does not deliver positive excess returns thereafter. Investment managers are terminated for a variety of reasons, including but not limited to underperformance. Excess returns after terminations are typically indistinguishable from zero but in some cases positive. In a sample of round-trip firing and hiring decisions, we find that if plan sponsors had stayed with fired investment managers, their excess returns would be no different from those delivered by newly hired managers. We uncover significant variation in pre- and post-hiring and firing returns that is related to plan sponsor characteristics.

ALLEN (2001) ARGUES THAT FINANCIAL INSTITUTIONS matter for asset pricing and laments the lack of attention to their behavior. Despite this clarion call, academic research has focused on two types of institutions, banks and mutual funds. There are good reasons for this. Banks have been a historically important component of the economy, and mutual funds are a relatively new but sizeable channel for retail investors to participate in capital markets. In addition, good data for both these types of institutions are widely available, permitting researchers to tackle issues with precision. However, another category of institutions, namely plan sponsors and institutional asset managers, is equally if not more important. At the end of 2003, there were 47,391 plan sponsors in the United States (corporate and public retirement plans, unions, endowments, and foundations), which were responsible for delegating investment of $6.3 trillion to institutional investment managers (Money Market Directory (2004)). At

*Goyal is at the Goizueta Business School, Emory University and Wahal is at the WP Carey School of Business, Arizona State University. We are indebted to Allison Howard and Nick Mencher at Invesco; David Baeckelandt, Maggie Griffin, and Robert Stein at Mercer Investment Consulting; Keith Arends at iisearches for assistance with data issues; and to the Goizueta Business School and the Q-Group for financial support. Jesse Ferlianto, Peter Left, Margaret Petri, Marko Svetina, and Fridge Vanzyp provided research assistance. We thank an anonymous referee, Roberto Barontini, George Benston, Mark Carhart, Chris Geczy, Charles Hadlock, Larry Harris, Narasimhan Jegadeesh, Kevin Johnson, Ananth Madhavan, and Ed Rice and the seminar participants at Hong Kong University of Science and Technology, the University of Washington, Goldman Sachs Asset Management, Arizona State University, University of California at Irvine, Barclays Global Investors, Boston College, Brown University, the European Finance Association Meetings in Moscow, the American Finance Association meetings in Boston, and the Mitsui Life Symposium at the University of Michigan for helpful comments and suggestions.

that time, there were 7,153 equity, bond, and hybrid mutual funds with total assets of $5.4 trillion (Investment Company Institute (2004)). The enormity of the assets under the jurisdiction of plan sponsors and their potential impact on asset prices are compelling reasons to examine their behavior.[1] Moreover, the fact that the assets managed by many plan sponsors fund the retirement incomes of their beneficiaries makes studying their behavior important from a personal and public policy perspective.

A comparison of institutional investment to the more widely studied retail marketplace provides some perspective. There are three basic streams to the retail investment/mutual fund literature: (1) investigations of performance, including persistence, (2) studies of the relationship between fund flows and returns, and (3) analyses of investment choices made by individual investors. The general conclusion that emerges from these streams is that the level of excess performance and the degree of persistence is weak and elusive, the relationship between flows and returns is convex, and retail investors make investment choices that can be construed as suboptimal by some and simply noisy by others.[2]

In the institutional realm, the streams are rivulets. Lakonishok, Shleifer, and Vishny (1992) provide the first investigation of performance and persistence. They persuasively argue that there are significant conflicts of interest in the money management industry and use proprietary data to examine the performance of 769 all-equity funds run by 341 investment managers. They paint a bleak picture of performance and argue, "[that] when all is said and done, we doubt that an industry that has added little if any value can continue to exist in its present form" p. 341. Coggin, Fabozzi, and Rahman (1993) also use proprietary data to study a sample of pension fund managers and find that they have limited skill in selecting stocks. Christopherson, Ferson, and Glassman (1998) find evidence of persistence among institutional equity managers using conditional methods and Busse, Goyal, and Wahal (2007) find that persistence exists in domestic equity and fixed income portfolios. Del Guercio and Tkac (2002) examine the relation between asset flows and returns and find that excess (as opposed to raw) returns are the relevant metric for the flow–performance relationship in the institutional arena. With one exception, the

[1] Institutions are more likely to be marginal traders than individual investors in most markets; consequently, their impact on asset pricing could be substantial. This is eloquently described by Cornell and Roll (2005 p. 59) who argue "... consumption decisions, whether to buy a television or take a vacation are made by consumers. The decision to buy IBM or Intel is delegated," and develop a simple yet elegant delegated-agent asset-pricing model.

[2] A partial list of contributions in the literature on performance and persistence includes Bollen and Busse (2005), Brown and Goetzmann (1995), Carhart (1997), Carhart et al. (2004), Daniel et al. (1997), Elton et al. (1992), Goetzmann and Ibbotson (1994), Grinblatt and Titman (1992), Grinblatt, Titman, and Wermers (1995), Hendricks, Patel, and Zeckhauser (1993), Ippolito (1989), Jensen (1968), Wermers (2000), and Zheng (1999). Fund flows and returns are studied by Chevalier and Ellison (1999), Gruber (1996), and Sirri and Tufano (1998). The third stream includes Barber and Odean (2000), Barber, Odean, and Zheng (2005), and Odean (1998, 1999). This list of citations is certainly not comprehensive. Omissions are not willful and we offer our apologies to authors not cited.

third stream, the actual investment choices by plan sponsors is dry. The exception is Heisler et al. (2007), who indirectly study why plan sponsors hire and fire investment managers by examining asset flows and accounts. Ex ante, one might expect that the level of expertise of plan sponsors in delegating assets to institutional investment management firms is higher than that of individual investors picking retail mutual funds. Whether this expertise generates excess returns or not is ultimately an empirical question. Our paper is the first to tackle this issue directly in the institutional marketplace.

Plan sponsors have certain investment goals and, working under self or externally imposed restrictions, allocate funds across asset classes in an attempt to achieve their goals. Within each asset class, mandates of specific dollar amounts are then delegated to investment management firms to be invested in a particular investment style. The raison d'être of a plan sponsor is then twofold: (1) to conduct asset allocation and (2) to hire managers to deliver benchmarked returns, monitor, and if necessary, fire investment managers.[3] It is this second task, that is, the hiring and firing of investment managers by plan sponsors, that we focus on in this paper.

We compile a unique database of 8,755 hiring decisions by 3,417 plan sponsors that delegate $627 billion in mandates between 1994 and 2003. We examine benchmark-adjusted cumulative excess returns, information ratios, and calendar-time alphas from factor models up to 3 years before and after hiring. All measurement methods show that for domestic equity and fixed income mandates, pre-hiring returns are positive, large, and statistically significant, but that post-hiring returns are statistically indistinguishable from zero. For international equity mandates, however, both pre- and post-hiring excess returns are positive and large.

Plan sponsors hire investment managers either because new inflows need to be invested or to replace terminated investment managers. Our sample of terminations consists of 869 firing decisions by 482 plan sponsors that withdraw almost $105 billion in mandates between 1996 and 2003. The number of terminations is substantially smaller than hiring decisions because data sources are geared toward assisting investment managers in obtaining new business and because there is a natural disinclination to report terminations. One obvious reason for terminating investment managers is underperformance. But we find that plan sponsors also terminate investment managers for a host of reasons unrelated to performance. Non-performance terminations are related to the plan sponsor (such as reallocations from one investment style to another or the merger of two plans) or events at the investment management firm (such as personnel turnover, the merger of two investment management firms, or regulatory actions). Excess returns prior to firing are negative for performance-based terminations but not for others. Post-firing excess returns for the entire sample are statistically indistinguishable from zero in the first 2 years after termination, but positive in the third

[3] Although we frequently refer to "investment managers," our unit of analysis throughout the paper is the investment management firm, not individuals at these firms.

year. Three-year post-firing returns are also positive for performance-based terminations.

To gauge the opportunity costs associated with both hiring and firing decisions, one has to compare post-hiring returns with the post-firing returns that would have been delivered by fired investment management firms. Since there are a multitude of complicated mechanisms by which firing and hiring decisions are coordinated, we build a sample of "round-trip" firing and hiring decisions manually. We identify 412 round-trip decisions between 1996 and 2003. For these decisions, the return difference between hired and fired managers prior to the round-trip is positive. After the round-trip the return differential is negative but with large standard errors.[4]

The aggregate results described above mask considerable variation in selection and termination. There are a number of different types of plan sponsors that run the gamut from defined benefit corporate plans to unions, foundations, public and private universities, and local- and state-level public plans. They vary in size from tiny multiemployer union plans like the Detroit Ironworkers Local 25 to behemoths such as the California Public Employees Retirement System. Size brings with it scale economies and perhaps expertise in selection and monitoring of investment managers. Consistent with this we find that larger plans are less likely to retain consultants to assist them in the selection process and have higher post-hiring excess returns than their smaller counterparts. Also important is the notion of "headline risk" in which some sponsors are sensitive to public scrutiny in the event of underperformance. We find that headline risk-sensitive sponsors are likely to chase investment styles with high returns in the past 3 years, to retain consultants to assist them in their hiring decisions, and to terminate managers for poor performance. But they have lower post-hiring returns than those that are headline risk-resistant or risk-neutral. Moreover, although consultants add value to hiring decisions on average (i.e., consultant-advised decisions have higher post-hiring returns), they destroy value in advising large plan sponsors. Lakonishok, Shleifer, and Vishny (1992) and Hart (1992) argue that overfunded corporate plans have little incentive to generate superior performance. Underfunding of plans, on the other hand, could generate large risk-taking incentives. For a limited sample of corporate and public plans for which we obtain funding ratios, we find that overfunded plans are less likely to engage in style-chasing and have lower post-hiring returns than underfunded plans. Underfunded plans are more likely to fire underperforming investment managers than overfunded plans. Finally, we also construct an asset allocation index that proxies for the lack of restrictions from investment policy statements and find that this index is positively correlated with post-hiring excess returns. The general picture that emerges from this cross-sectional analysis is that economic fundamentals such as size, the

[4] These results are similar to those of Odean (1998) for retail investors and Elton, Gruber, and Blake (2006) for 401(k) plans. Odean finds that the excess returns on winning stocks sold by individual investors are larger than the excess returns on loser stocks that could be, but are not, sold. Elton et al. (2006) find that administrators select funds that did well in the past but after the change, do no better than funds that were dropped.

potential for adverse publicity, restrictions, and funding demands "matter," in the sense that they influence various aspects of hiring and firing.

Notwithstanding this variation, the conclusion to be drawn from our broad results depends largely on one's view of performance persistence, and of the role of frictional costs. Since all of our hiring decisions are for active investment managers, they represent an unsuccessful attempt by plan sponsors to seek excess post-hiring returns. This lack of success could be because there is no persistence in investment manager returns. But Christopherson et al. (1998) and Busse et al. (2007) show that there is persistence in institutional portfolios over 1 to 2 years. The fact that there is some persistence justifies the plan sponsor's conditioning of hiring on returns, at least on an ex ante basis. Zero post-hiring excess returns indicate that, on average, plan sponsors have no timing ability.

For hiring decisions necessitated by the termination of incumbent investment managers, one has to judge the hired manager's returns against the returns that the fired manager would have delivered (i.e., the opportunity costs described above), as well as frictional costs in moving portfolios. Since the difference between pre-hiring and pre-firing returns is large, hiring and firing decisions can be justified ex ante by plan sponsors. Ex post, there are some opportunity losses. Addressing the issue of how much transaction costs add to these losses is more difficult because there are no publicly available data on the costs of moving portfolios. The process of moving assets from the legacy portfolio of the fired investment manager to the target portfolio of the hired manager is frequently outsourced to "transition management firms" that attempt to minimize the costs associated with the transition. Estimates of transition costs by practitioners in the public press suggest that average costs range between 2% and 5% of the portfolio, with a standard deviation of 1% (see, for example, Proszek (2002), Bollen (2004), and Werner (2001)). Private estimates of all-in transition costs provided to us by an anonymous large transition management firm vary between 1.0% and 2.0%. This firm also indicates that transition costs are much higher for international, fixed income, and small-cap transitions, and when the legacy and target portfolios are in different asset classes. Regardless of the actual magnitude, the size of this transition business, estimated by some observers to be almost $2 trillion annually, suggests that transaction costs are substantial.[5]

Given our results, a reader could reasonably ask why plan sponsors make decisions that, ex post, appear to be costly. There are three plausible explanations. One is the hubristic belief among plan sponsors than they can time the hiring and firing decisions successfully. We stress that this behavior is not necessarily irrational, especially since there is persistence in performance. A second explanation is job preservation; to quote Lakonishok et al. (1992, p. 342), "those

[5] If such frictions are important, then one would expect the return threshold for retention decisions (in which an incumbent manager is "rehired") to be lower than for brand new hiring decisions. Consistent with this, we find that pre-retention excess returns are positive but lower than pre-hiring excess returns.

in charge of the plan must show that they are doing some work to preserve their position." Simply put, if plan sponsors did not hire and fire, their raison d'être would be nonexistent. We find that elements of hiring and firing tendencies, pre-event return thresholds, and post-event performance are related to plan sponsor attributes that reflect these agency relationships; broadly, the cross-sectional evidence is closely tied to this possibility. A third possible explanation is that these decisions are not as costly as our evidence would indicate because we are unable to fully measure the benefits. For example, it may be that termination disciplines fired investment managers and cause them to improve returns in the future. Indeed, investment managers who lose a larger fraction of their assets have higher post-termination returns. It may also be that termination disciplines incumbent (not fired) as well as potential investment managers. Unfortunately, we have no way of measuring this potentially offsetting benefit. Thus, while our results shed light on the efficacy of hiring and firing, we cannot necessarily conclude that these decisions are inefficient. The above explanations are not mutually exclusive. It is quite likely that all three play some role in the process.[6]

Our paper proceeds as follows. In Section I, we provide a brief description of the institutional marketplace and investment process. In Section II, we describe data sources and sample construction procedures. We present results on the selection of investment managers in Section III, and the termination of investment managers in Section IV, and round-trips in Section V. Section VI concludes.

I. Institutional Details

In this section, we describe the institutional marketplace and the investment process followed by most plan sponsors. A more detailed description of the pension fund industry can be found in Fabozzi (1997), Lakonishok et al.(1992), Logue and Rader (1998), and Travers (2004).

A. The Institutional Marketplace

There are basically two types of plan sponsors, those that manage retirement assets and those that manage nonprofit assets. The former include corporate plans; public plans for employees at the city, county, or state level; single-employer plans; and Taft–Hartley multiemployer plans for organized labor.[7] The latter include foundations and endowments, including those for universities. Retirement plans can be set up as defined benefit plans, defined

[6] A fourth possible explanation is that plan sponsors are simply unaware of these costs. We deem this explanation implausible.

[7] Such plans are set up under Section 302(c) (5) of the Taft–Hartley Act, passed by Congress in 1947. Plan assets are jointly managed by a board of trustees representing labor and management. This is a sizeable market. Brull (2006) reports that 1,600 multiemployer plans had assets totaling $333 billion in 2002, and covered almost 10 million workers in 2005. He also reports that some 30,000 single-employer plans reported assets of $1.6 trillion in 2002 and covered 34.6 million workers.

contribution plans, or both. In a defined benefit plan, beneficiaries receive a fixed set of payments upon retirement. The trustees of the plan are responsible for investing the beneficiaries' contributions to ensure that future benefits can be paid. In defined contribution plans, beneficiaries receive variable payments upon retirement. The plan sponsor typically selects providers of various investment options (such as Vanguard or Fidelity) who then allow beneficiaries to directly invest their assets in various funds. Some firms offer both defined benefit and defined contribution plans.

All plan sponsors share one common feature: The trustees of the plan are charged with the task of managing assets in the best interests of their beneficiaries. However, organizational structure and incentives can generate tremendous variation in behavior across plan sponsors. In corporate defined benefit plans, if the plan is overfunded, the excess funds belong to the corporation. This creates incentives for the treasurer's office (the trustee) to generate superior performance. But, Lakonishok et al. (1992) argue that firms' implicit contracts with employees may be such that excess funds are effectively handed over to employees. Hart (1992) argues that even if the excess funds belong to the corporation, considering agency issues, there is little incentive for management to generate superior performance. If the plan is underfunded this might provide an incentive to invest in risky assets, in part because, in the event of bankruptcy, the Pension Benefit Guarantee Corporation (PBGC) insures the benefits (up to a statutory limit) if the corporation has insufficient assets to cover its obligations. Lakonishok et al. (1992) note that this structure produces a bias against passive investment management (since it reduces the potential power of the treasurer's office), and against internal investment management (since it is easier to blame another organization for poor performance). In federal, state, or local government pension plans, the residual claimant is the government authority (and ultimately the taxpayer), and the trustees of the plan are political appointees and/or bureaucrats. Similarly, the residual claimants at single-employer union plans are union members and the PBGC provides downside protection. Trustees are drawn from members. However, in multiemployer Taft–Hartley plans, if one employer files for bankruptcy, the shortfall is assumed by solvent companies remaining in the plan. Non-retirement plans such as endowments and foundations do not receive any protection from the PBGC and do not have a residual claimant per se. Cash outflows for endowments and foundations have more of a discretionary element to them than retirement plans. If a foundation's performance is weak, it can lower distributions and curtail charitable activity whereas a retirement system has to fulfill its cash outflow obligations. Incentives are also provided by the market for human capital. Superior performance in managing the investment process can increase salaries and generate improved external employment opportunities. This appears to be the case, especially for endowments, where even though the residual claimant is not well-defined, executives that manage the investment process effectively generate significant human capital.[8]

[8] Two well-known examples of this are David Swensen of the Yale University Endowment and Jack Meyer of the Harvard University Endowment.

B. The Investment Process

The above discussion suggests that the goals of a plan sponsor are influenced by the structure of claims and the nature of payouts. The investment process followed by plan sponsors is designed to achieve those goals. Typically this process begins with an investment policy statement drafted by the investment committee, often spearheaded by a chief investment officer. The investment policy statement describes the goals of the plan sponsor, the road map for reaching those goals, and any restrictions on the investment process. The restrictions originate from a desire to control risk and return profiles and can take a variety of forms, varying from broad strategic asset allocation decisions to tactical adjustments around strategic targets. They can influence the quantity and quality of asset classes available. For instance, certain asset classes (such as hedge funds or real estate) may be excluded or capped at a particular percentage of total assets. There may also be restrictions on specific securities to be held within qualified asset classes. Quality restrictions, for example, might involve excluding "sin" stocks or including only dividend-paying securities. Effectively, asset allocations can be thought of as one realization of the goals and restrictions in the investment policy statement.

Plan sponsor size also generates variation in the investment process across plan sponsors. Larger plan sponsors likely benefit from economies of scale in generating information and managing the investment process. In addition, large plan sponsors have an advantage in that they may be allowed preferential access to certain funds because they can provide large amounts of capital; most investment management firms have minimum investment requirements that small plan sponsors may not be able to meet.

C. The Hiring and Firing Process

Once broad asset allocations have been established, the search for managers begins. The plan sponsor puts out a request for proposals (RFP) and may retain a consultant to assist in the search. The process involves screening investment managers who provide investment products in the mandate stated by the plan sponsor. The mandate can be either broad (e.g., domestic equity) or narrow (e.g., small-cap equity value). The list of candidate managers is then culled based on relative performance. The list is further trimmed with written questionnaires and interviews, and the investment committee or trustees make a final choice.

For an investment manager, being part of the initial list of managers is a critical hurdle. As a result, most organizations voluntarily provide information to various databases that record performance and other characteristics. Such databases are produced by independent organizations, such as iisearches (affiliated with Institutional Investor publications) or Nelson's Directories (affiliated with Thomson Financial), as well as by pension consultants such as Mercer Investment Consulting. A list of common databases is contained in Travers (2004).

Since different plan sponsors conduct manager searches that are correlated in time and investment mandate, pension consultants can reap economies in gathering information. To the extent that larger plan sponsors make more hiring/firing decisions, they may be less likely to employ consultants. Plan sponsors may also employ a consultant to shield themselves from adverse publicity associated with negative outcomes from hiring decisions.

Once an investment management firm has been hired, its performance is generally monitored on a quarterly basis. If performance relative to a benchmark deteriorates over consecutive evaluation horizons, the firm may be put on a "watch list." If performance improves, the firm is removed from the watch list. Continued deterioration in performance may result in the firm's contract being terminated. If the firm is terminated, the assets are transferred to the newly hired investment manager's portfolio by a transition organization. Large investment houses, such as State Street Global Advisors and Barclays Global Investors, provide such transition management services, the aim of which is to minimize the frictional loss in transitioning between the legacy and target portfolios.

Aside from performance, there are other reasons why an investment management firm may be terminated. The plan sponsor may view the superior performance of the investment manager's portfolio as being directly attributable to a particular individual. If such an individual(s) leaves the firm, the plan sponsor may decide to terminate its relationship with the investment management firm. For example, in 1996 the two principal partners of Apodaca–Johnston Capital Management separated to start their own investment management firms. As a result, the Los Angeles County Employees Retirement Association terminated its contract with the firm. In addition to personnel turnover, mergers between investment management firms can also prompt terminations. Finally, reasons that are specific to the plan sponsor, rather than the investment management firm, can cause terminations. For instance, a reorganization of the sponsor (perhaps because two corporations merged) may cause the reorganized plan to fire some investment managers. Alternatively, if the plan sponsor decides to change asset classes or investment styles, it may terminate investment managers in mandates that are downsized.

Hiring of investment managers also takes place for several reasons. The replacement of a fired manager or an increase in asset allocation to a particular mandate can trigger hiring. Additionally, if the size of the plan sponsor's asset base increases, it may hire new investment managers rather than increase allocations to existing managers.

II. Data Sources and Sample Construction

A. Selection and Termination Data

We obtain data on the selection and termination of investment managers from three different sources: the "Tracker" database developed by Mercer Investment Consulting, the "iisearches" database created by *Institutional Investor*

Publications, and electronic searches of articles published in *Pensions and Investments* (*P&I*). The Tracker and iisearches databases are used by investment management firms to market their services to plan sponsors. These sources provide the name of the plan sponsor, the type of the plan sponsor, the name of the investment manager hired, the name of the consultant(s), the type and amount of the investment mandate, and a hiring date. Although similar in spirit, the two databases differ in three key ways. First, the Tracker database does not record the termination of investment managers. The iisearches database does record parallel information on investment managers that are fired, but the firing data are sparse and record only single-matching firing and hiring decisions. Therefore, round-trips cannot be extracted in a straightforward way from the database. Second, iisearches provides a column containing textual information about the hiring/firing that can help in identifying the reason for the termination. Here again, the data are sparse—only some records contain textual information. As a result, we use manual searches in trade journals to fill in the gaps. Third, the Tracker database contains data from 1994 through 2003, whereas the iisearches database starts in 1995.

We also perform electronic searches for articles in *P&I*, a widely used and respected source of weekly information for this industry. It reports on searches and terminations by major plan sponsors, often providing contextual information that is not recorded in the Tracker or iisearches databases. We perform keyword searches of all issues of *P&I* between 1996 and 2003 using the following phrases: "hiring," "firing," and "termination." We then read these articles and manually record the same data elements as Tracker and iisearches.

We remove all non-U.S. plan sponsors from each of these databases and discard observations where the hiring (or firing) concerns custodians or record keepers. We also remove observations for employee-directed (defined contribution) retirement plans. This results in 15,940 hiring observations from Tracker, 11,537 hiring observations from iisearches, and 1,184 observations from *P&I*.

We use these data sources to create as comprehensive a sample as possible and to cross-check information. To eliminate duplicates, we first create master files that uniquely identify different permutations and spellings of plan sponsor, investment manager, and consultant names. We then splice the data sets together, from which we identify duplicate observations as those in which the same plan sponsor hires/fires the same investment manager within 90 days of each other. When data sources disagree on other aspects of the hiring/firing, we use a reasonable algorithm to determine the final value for the field (for instance, taking the minimum value of the mandate amount). Where the data sources disagree on the investment mandate, we treat the mandate as unknown.

B. Plan Sponsor Information and Asset Allocation Data

We use Nelson's Directory of Plan Sponsors, the Money Market Directory of Investment Managers and Plan Sponsors, and internet searches to classify each plan sponsor into nine categories: corporate; endowments and foundations; local public plans that represent general retirement interests for cities and counties; state public plans that refer to statewide plans such as the California Public Employees Retirement System; miscellaneous public plans that include police, fire, and municipal employee retirement plans for cities and counties; unions (including Taft–Hartley plans); public universities; private universities; and a miscellaneous category that includes insurance plans, health and hospital plans, trusts, and anonymous plans.

For corporate plans, we calculate funding ratios for the year prior to hiring/firing based on the procedure outlined in Franzoni and Marín (2006), except that rather than scaling by market capitalization, we use the ratio of fair value of plan assets to the projected benefit obligation. For public plans, we manually collect funding ratios from plan sponsor websites, relying especially on the public retirement systems website (www.prism-assoc.org). Not surprisingly, there is a reporting bias: Only large plans report this information. Since the obligations of nonretirement plans are largely discretionary, the notion of a funding ratio is not well-defined. Therefore, our funding ratio tests are only for corporate and public plans.

We obtain information on asset allocations for plan sponsors from two sources. *P&I* surveys the largest 1,000 corporate and public retirement plans in each year and records information on broad asset allocations in the following general categories: domestic equity, domestic fixed income, international equity, international fixed income, cash, private equity, real estate, mortgages, and “others” (including distressed debt, oil and gas, timber, etc.). These data also contain the percentage of assets that are indexed and that are managed internally. There are several important qualifications to these data. First, they include only retirement plans and specifically exclude endowments, foundations, unions, and insurance plans. Second, prior to 1996, only the largest 200 plan sponsors are surveyed. Third, the asset class categories and gradations change over time. For example, in some years, only allocations to equity, rather than domestic and international equity, are recorded. Similarly, allocations to private equity are not recorded until later in the time series. We supplement these data with hand-collected information from Nelson’s Directory of Plan Sponsors (2005). Nelson’s coverage of plan sponsors is better in that it includes endowments, foundations, and union plans. However, its gradation of asset classes is not as fine as *P&I* and we only observe allocations at the end of our sample period.

C. Returns and Asset Size Data

We obtain return information from Mercer’s Manager Performance Analytics database. This database contains quarterly returns (gross of fees) on approximately 9,000 products offered by 1,200 investment managers for the period from 1981 to 2005. These are “composite” returns for unrestricted portfolios. The actual returns earned by a plan sponsor may differ slightly from these composite returns if the plan imposes significant restrictions on the portfolio. The returns data are self-reported by investment management firms. Given that a successful track record of returns is critical for hiring, it is possible that some investment management firms "amend" prior years' returns in updating return information. We ensure that this is not the case—Mercer informs us that investment managers provide each quarter's return soon after the end of the quarter and are not permitted to update prior returns. In addition, the investment management firms in our sample comply with the performance reporting standards established by the CFA Institute (see http://www.cfainstitute.org/centre/ips).

Another potential concern is one of survivorship bias. We perform three checks to determine if survivorship bias influences our results. First, we compute attrition rates of investment managers and ensure that return histories disappear over time. Tabulations of return histories show an attrition rate of approximately 4% per year in our sample (by comparison, Carhart et al. (2002) report an average annual attrition rate of 3.6% for mutual funds). Second, we calculate the number of instances where pre-firing returns are available but post-firing returns are not. We find that the loss in data is trivial (10 observations for a 1-year horizon), suggesting that post-firing returns do not disappear from the sample because the pre-firing returns are negative. Third, we reexamine the portion of our firing database for which we have no returns (either pre- or post-firing). A vast majority of firing decisions for which we have no returns are where the mandate is unknown or in an asset class not covered by our returns database (private equity, venture capital, real estate, etc.).

Mercer provides multiple benchmark return indices appropriate for each product category. For example, for the small-cap product category, Mercer provides 13 different benchmark indices. The correlation coefficients between these different indices are generally very high. Therefore, we select one index for each product category that we believe best describes the investment objective of that category. A list of each product category and the chosen index, along with a brief description, is provided in Table A1. We obtain asset information from the Money Market Directory of Investment Managers. This database contains the investment management firms' name and the total assets under management in each year from 1996 to 2003.

D. Sample Construction

We match the hiring/firing database with the return data in two steps. We first match the names of investment management firms across the two databases. We use *Nelson's Directory of Investment Managers* (2004), the *Money Market Directory of Investment Managers and Plan Sponsors* (2004), and Internet searches to ensure that acquisitions of investment management firms are correctly accounted for in both databases. Second, we match information on the investment mandate from the hiring/firing database to one of the products in the returns database. This process results in a loss of some data for three reasons. First, Mercer's return database may not have returns for a particular investment management firm. Second, Mercer's return database may not have returns for the mandate for which the investment manager was hired or fired. This is often the case for "alternative asset" mandates that include venture capital and private equity. Third, we remove passive mandates from our sample since investment managers for these mandates are selected for their ability to provide low-cost passive exposure rather than beating a particular benchmark.

Sometimes, mandate information in the hiring/firing database is available only at a broad level while the returns are available at a refined level. For instance, a hiring record may indicate that XYZ Investment Partners was hired for a large-cap equity mandate. Our returns database may record return information for XYZ Investment Partners for large-cap growth, large-cap value, and large-cap core products. In such situations, we use an equally weighted average return across all the relevant products and match it to the investment mandate. We perform all our tests without this averaging and note that it does not affect our conclusions.

The intersection of the two databases produces a sample of 8,755 hiring decisions by 3,417 plan sponsors. These hiring decisions involve 602 investment managers hired to manage a total of $627 billion between 1994 and 2003. The firing database consists of 869 decisions by 482 plan sponsors between 1996 and 2003. These decisions involve the withdrawal of $105 billion from 247 investment managers.

E. Performance Measurement

We identify quarter zero as the quarter in which the hiring/firing takes place and then measure performance in several different ways. We calculate cumulative excess returns for the mandate (portfolio) of the investment manager as

$$CER_i(t, H) = \sum_{s=t}^{t+H-1} (R_{i,s} - R_{b,s}), \tag{1}$$

where $R_{i,s}$ is the return on the mandate type by the investment manager i in quarter s, and $R_{b,s}$ is the return on the benchmark in quarter s. We calculate *CERs* for 1, 2, and 3 years before and after an event, but we focus our discussion on the 3-year horizon because shorter period returns are noisy. In addition to *CERs,* we also report information ratios since they are widely used in the practitioner community, and calculate them as

$$IR_i(t, H) = \frac{\overline{CER}}{\sigma_{ER}}, \tag{2}$$

where $\overline{CER}$ is the mean excess return over the appropriate horizon and σ_{ER} is the standard deviation of the excess return.

The assessment of the statistical significance for *CERs* is a nontrivial matter. In our data, plan sponsors and investment managers can appear multiple times for different decisions. This repetition, in combination with overlapping periods in long-horizon returns, introduces cross-sectional and time-series dependencies that render typical standard errors unreliable. We follow Jegadeesh and Karceski (2004) and calculate conservative standard errors based on a calendar-time procedure that accounts for cross-correlations, heteroskedasticity, and serial correlation. Details of the calculations of standard errors are contained in the Appendix.

Benchmark adjustments are not risk adjustments. One alternative is to estimate factor models in the spirit of the mutual fund literature (e.g., Elton, Gruber, and Blake (1996) or Carhart (1997)). Ideally, we would want to estimate alphas from a factor model before and after each event. However, the short time series, in addition to the fact that our returns are quarterly, limits our ability to do so. To get around this problem, we follow a calendar-time portfolio approach to estimating factor models. This allows us to estimate alphas for each year before and after the event. The disadvantage is that since we do not obtain alphas for each decision, we cannot examine cross-sectional variation in performance measured by alphas.

We calculate separate calendar-time portfolio returns for 3 years to 1 year before and after hiring/firing decisions (in other words, we calculate six separate calendar-time portfolios for each asset class). For instance, a hiring decision in December 1998 is included in the 3-year pre-hiring calendar-time portfolio from December 1996 to November 1998. We then estimate alphas from factor models with the following specification for each of the calendar-time portfolios:

$$R_{p,t} = \alpha_p + \sum_{k=1}^{K} \beta_{p,k} f_{k,t} + \varepsilon_{p,t}, \tag{3}$$

where R_p is the excess return on portfolio p, and f_k is the k^{th} factor return. The models are estimated separately for domestic equity, fixed income, and international equity mandates. For domestic equity mandates, we follow Fama and French (1993) and use the market, size, and book-to-market factors obtained from Ken French's web site. For fixed income portfolios, we use the Lehman Brothers Aggregate Bond Index return, a term spread (computed as the difference between the long-term government bond return and the T-bill return), and a default spread return (computed as the difference between the corporate bond return and the long-term government bond). The default and term spread are obtained from Ibbotson Associates. For international equity mandates, we employ an international version of the three-factor model. We obtain the international market return and book-to-market factor from Ken French. The international size factor is computed as the difference between the S&P/Citigroup PMI World index return and the S&P/Citigroup EMI World index return, both of which exclude the United States (see http://www.globalindices.standardandpoors.com).

III. The Selection of Investment Managers

A. Sample Distribution

Panel A of Table I describes the distribution of hiring decisions. Of the 8,755 hiring decisions, 22% (1,927) originate from corporate plan sponsors. The

Table I

Distribution of Hiring Decisions by Plan Sponsors

Local public plans are those for cities and counties. State public plans are state-level retirement plans (such as Calpers). Misc. public plans include police, fire, municipal employee, and other such retirement plans at the city or county level. Unions include single and multiemployer unions and Taft–Hartley plans. The "miscellaneous" category includes anonymous corporate plans, insurance plans, health and hospital plans, and trusts. Headline risk-resistant plans are corporate plans, private universities, and miscellaneous plans. Headline risk-sensitive plans are local, state and miscellaneous public plans, unions, and public universities. Headline risk-neutral plans include nonuniversity endowments and foundations. Funding status for corporate pension plans is calculated as in Franzoni and Marín (2006). Funding ratios for public plans for the year prior to the hiring decision are obtained from the plan web sites.

		Plan Sponsor Size ($M)			Mandate Size ($M)		
	Number of Hirings	Mean	Median	*N*	Mean	Median	*N*
Panel A: Distribution by Type of Plan Sponsor							
Corporate	1,927	3,690	370	1,617	55	22	1,557
Endowments & foundations	729	1,080	190	532	25	12	625
Local public plans	1,655	7,952	500	1,601	98	25	1,545
State public plans	1,032	22,954	12,000	1,006	203	120	961
Misc. public plans	951	4,728	830	891	87	30	858
Unions	892	1,165	250	761	34	19	815
Public universities	351	1,297	200	324	36	12	317
Private universities	348	369	174	321	16	10	303
Miscellaneous	890	2,659	244	597	91	20	671
All	8,755	6,482	474	7,650	82	25	7,652
Panel B: Headline Risk							
Headline risk-sensitive	4,884	9,021	800	4,583	103	30	4,496
Headline risk-neutral	729	1,080	190	532	25	12	625
Headline risk-resistant	3,145	3,026	300	2,535	59	20	2,531
Panel C: Funding Status							
Corporate Plans							
Underfunded	330	1,952	375	307	49	21	242
Overfunded	355	1,959	447	338	54	25	297
Public Plans							
Underfunded	736	13,288	6,100	731	170	100	700
Overfunded	381	24,468	13,650	370	278	130	356

average size of such sponsors is $3.7 billion and the average mandate is for $55 million. State-level public plans are extremely large, averaging $22.9 billion in size and present mandates that are over $200 million. Local and miscellaneous public plans are considerably smaller. Endowments and foundations are smaller than corporate and state or local public plans with an average size of only $1 billion. Their average mandate size is also smaller ($25 million). Single and multiemployer union plans represent over 10% of the sample and their average mandate is for $34 million. The miscellaneous category includes 890 hiring decisions by insurance plans, trusts, and anonymous defined benefit plans.

In Panel B, we collapse these types of plans into three categories that reflect their sensitivity to adverse publicity in the event of poor performance. This categorization is based on the premise that sponsors whose boards of directors or investment committee members are political appointments are more likely to be subject to headline risk. In the spirit of Brickley, Lease, and Smith (1988), we categorize plans into headline risk-sensitive, risk-resistant, and risk-neutral groups. Headline risk-sensitive sponsors include local, state, and miscellaneous public plans, unions, and public universities. In such public institutions, appointments to boards are either direct placements by elected officials (e.g., in the case of gubernatorial appointments at state plans) or take place via a process that involves behind-the-scenes political maneuvering. Headline risk-neutral sponsors include non-university endowments and foundations, and headline risk-resistant sponsors are corporate plans, private universities, and miscellaneous plans. The objectives of the latter group are well-defined and the political influence in the board appointment process is not as large as for headline risk-sensitive sponsors. Headline risk-sensitive sponsors are larger, in part because they include the extremely large state public plans.

In Panel C, we report size and mandate statistics for plans that are over- or underfunded in the year prior to the hiring decision. Since the residual claimant and the nature of the guarantees (PBGC vs. taxpayers) are quite different for corporate versus public plans, we report separate statistics. Over- and underfunded corporate plans are quite similar in terms of size and mandate, but in the case of public plans, overfunded plans are significantly larger with bigger mandates.

Before RFPs can be issued and an investment management firm hired, a plan sponsor must create an asset allocation plan that incorporates its investment goals and restrictions. Unfortunately, to our knowledge, there is no database of restrictions and/or investment policy statements. Even though we cannot measure the restrictions imposed on a plan sponsor directly, we create a proxy by examining asset allocations. The idea is that plan sponsors that are relatively unrestricted are more likely to invest larger amounts in riskier asset classes; in effect, asset allocations represent a realization of constraints and investment policy statements. For instance, an endowment that allocates a large percentage of its assets to hedge funds is likely to be less restricted than one that is prohibited from such investments. To capture this idea, we create a simple allocation index that is the average of the allocation to equity (both domestic and international), alternative assets, nonindexed assets, and externally managed assets.[9] For plan sponsors without data on indexation or externally managed assets, the average is computed only from available data elements.[10]

Panel A of Table II shows average asset allocations for the different types of plan sponsors. Since our data sources provide different and not always consistent classifications of assets, we collapse all allocation information into five

[9] Although this allocation index measures the strategic aspect of investment policy restrictions, to the extent that strategic and tactical restrictions are correlated, it is a proxy for both.

[10] As a spot check, we check the value of this index for a handful of plan sponsors for which we obtain direct information on investment restrictions. We find that index values are indeed lower for plan sponsors that have quality and/or quantity restrictions on asset allocations.

Table II

Asset Allocations and Consultant Use

Alternative assets include buyout funds, venture capital, and hedge funds. Other assets include balanced, GICs, cash, real estate, timber, oil and gas. The number of observations across asset classes and allocation attributes are not equal because of data collection procedures and as a result, the sum of allocations is not equal to 100%. The allocation index is the average of the allocation to equity (both domestic and international), alternative assets, nonindexed assets, and externally managed assets. For plan sponsors without data on indexation and externally managed assets, the average is computed from the equity and alternative asset allocation. For probit regressions predicting the use of consultants, standard errors (in parentheses) account for clustering in observations where the investment manager is hired for a mandate in the same style and period by different plan sponsors.

	Asset Allocation Information											
	Asset Classes Allocations					Allocation Attributes			Number of Hirings (%)			
	Dom. Eq.	Fixed Inc.	Intl. Eq.	Alt. Assets	Other Assets	Indexed	Internal Mgd.	Allocation Index	Dom. Eq.	Fixed Inc.	Intl. Eq	Consultant Use (%)
Panel A: Plan Sponsor Type												
Corporate	48.5	26.8	10.6	11.9	9.5	8.5	3.3	0.65	53	20	20	50
Endow. & found.	48.6	29.7	7.5	6.9	6.3	–	–	0.34	60	19	13	58
Local public plans	46.8	35.4	9.9	1.9	6.3	17.3	10.3	0.45	45	23	21	82
State public plans	42.4	33.6	13.3	4.4	8.5	25.0	19.2	0.54	41	23	25	68
Misc. public plans	45.9	34.8	10.6	2.8	7.3	20.1	6.6	0.50	49	24	19	73
Unions	41.5	37.6	2.4	10.9	12.6	8.1	0.2	0.45	61	24	4	67
Public universities	47.5	26.3	11.3	8.4	4.6	–	–	0.35	52	25	16	64
Private universities	55.3	21.5	6.7	7.1	6.2	–	–	0.35	60	18	17	61
Miscellaneous	49.7	24.7	4.9	14.7	5.9	–	–	0.39	50	30	14	41
Panel B: Headline Risk												
Sensitive	45.2	34.6	10.1	3.9	7.5	20.7	12.6	0.48	48	23	18	73
Neutral	48.6	29.7	7.5	6.9	6.3	–	–	0.34	60	20	13	58
Resistant	49.4	26.1	9.1	12.1	8.7	8.7	4.3	0.59	53	23	18	49

(*continued*)

Table II—*Continued*

	Asset Allocation Information											
	Asset Classes Allocations					Allocation Attributes			Number of Hirings (%)			
	Dom. Eq.	Fixed Inc.	Intl. Eq.	Alt. Assets	Other Assets	Indexed	Internal Mgd.	Allocation Index	Dom. Eq.	Fixed Inc.	Intl. Eq	Consultant Use (%)
Panel C: Funding Status												
Public: underfunded	42.5	33.0	13.7	4.5	7.5	24.4	12.7	0.59	323	171	168	75
Public: overfunded	45.7	31.7	14.6	3.8	6.2	28.4	20.6	0.56	142	97	105	74
Corp.: underfunded	45.8	26.2	13.4	9.3	10.8	7.6	4.9	0.66	183	65	63	50
Corp.: overfunded	49.4	26.3	10.9	9.7	8.0	10.5	5.5	0.66	193	63	86	60

Panel D: Probit Regressions Predicting Consultant Use

	Intercept	Plan Size	Portfolio Age	Headline Resistant	Headline Sensitive	Domestic Equity	Fixed Income	Intnl. Equity	Funding Indicator	Sample Size
Full sample	0.40	−0.05	0.02	−0.26	0.42	0.29	0.13	0.26	–	7,328
	(0.14)	(0.01)	(0.03)	(0.06)	(0.06)	(0.07)	(0.08)	(0.08)		
Public plans	0.76	−0.09	0.12	–	–	0.49	0.42	0.38	0.07	1,060
	(0.36)	(0.03)	(0.06)			(0.14)	(0.15)	(0.15)	(0.09)	
Corporate plans	0.83	−0.19	0.01	–	–	0.36	0.30	0.48	0.26	615
	(0.42)	(0.03)	(0.06)			(0.27)	(0.29)	(0.29)	(0.10)	

asset classes: domestic equity, fixed income, international equity, alternative assets (buyout funds, venture capital, and hedge funds), and other assets (balanced, GICs, cash, real estate, timber, oil and gas, etc.).

Allocations to fixed income generate a more predictable stream of cash flows than those to equity. Therefore, plan sponsors that need to pay retirees might make higher allocations to fixed income than those whose outflows are more flexible. Consistent with this, public and union plans allocate between 33.6% and 37.6% of their assets to fixed income portfolios compared with endowments that only allocate 29.7%, and to public and private universities that allocate 26.3% and 21.5%, respectively. By this metric, allocations by corporate plans are relatively aggressive, allocating 48.5% of their assets toward domestic equity and only 26.8% to fixed income. Allocations to international equity portfolios are quite high from corporate and public plans (over 10%), particularly compared to unions that invest only 2% of their assets in international equity. Corporate plan and endowment allocations to alternative assets are also high, but surprisingly, allocations from union plans are also large.

Panel A also reports the percentage of assets that are indexed and managed internally. Since these data elements are only available from *P&I*, the sample does not match that for asset classes. In the available subsample, the data show that state public plans manage a significant proportion of their assets internally (19%) and also pursue indexation policies (25%), consistent with the increase in indexation reported by Lakonishok et al. (1992). In contrast, union plans rarely index and never manage their own assets.

The allocation index is highest for corporate plan sponsors (0.65). This is again consistent with the idea that corporate plan sponsors can be more aggressive in asset allocation because they are the residual claimant and because they are less constrained than other sponsors. Panel B shows asset allocations and the allocation index for plan sponsors classified by headline risk and Panel C shows the same data for public and corporate plans that are either over- or underfunded. Headline risk-resistant plan sponsors have higher allocations to domestic equity and alternative assets, and a significantly higher allocation index than for headline risk-sensitive plan sponsors. Interestingly, headline risk-neutral plan sponsors have the lowest allocation index. The correlation between funding status and asset allocation could reflect two opposing forces. It could be that plans with more restrictions become underfunded because these restrictions prevent them from constructing optimal portfolios. Or, it could be that plans with lower restrictions become underfunded because they unsuccessfully invested in riskier securities. Empirically, we find that funding status does not vary with asset allocations.

The last columns in Panels A, B, and C show variation in the use of consultants. For example, headline risk-sensitive sponsors are more likely to employ a consultant (73%) than headline risk-resistant sponsors (4%). But, such effects are likely correlated with other attributes such as the size of the plan sponsor or the asset class of the mandate. To provide a more complete description of this, we estimate multivariate probit models that predict the use of consultants in Panel D. The independent variables in these probit models proxy for the ideas discussed above. Plan sponsor size captures the notion that larger sponsors may have economies in hiring. We include the age of the portfolio managed by the investment management firm because consultants typically require a return history before recommending a portfolio to a sponsor. We also include indicator variables for headline risk-resistant and risk-sensitive plan sponsors, and allow the headline risk-neutral category to be picked up by the intercept. Since selection of investment managers in certain asset classes might require more expertise, we include indicator variables for domestic equity, international equity, and fixed income mandates.

Three versions of the probit model are reported in Panel D. Standard errors are reported in parentheses below the coefficients. The first model is estimated on the full sample and shows that headline risk-resistant (sensitive) plan sponsors are significantly less (more) likely to use a consultant. The implied probability changes from the coefficients are 10% for headline risk-resistant sponsors and 15% for headline risk-sensitive sponsors. The logarithm of plan sponsor size is negatively correlated with the use of consultants, consistent with our priors. Similar models augmented with an indicator variable for whether the plan is overfunded in the prior year for public (corporate) plans are also reported. The funding indicator is insignificant for public plans but positive for corporate plans.

B. Pre-hiring Performance

Plan sponsors hire investment managers to invest new asset inflows and to replace terminated investment managers. We examine pre-hiring performance in two ways. First, we modify the investment manager *CERs* described above to calculate style *CERs*. Our purpose is to determine the degree to which plan sponsors engage in style-chasing. Lakonishok et al. (1992) argue that the structure of this industry and the agency relationships within cause sponsors to allocate funds to different styles, rather than following a specific style or indexing. Barberis and Shleifer (2003) argue that style investing is particularly attractive to plan sponsors because style categorizations make it very easy to evaluate investment managers. Ideally, to detect style-chasing, we would like to directly examine shifts in asset classes and styles for each plan sponsor and correlate them with lagged market movements. Absent this information, we can provide some indirect evidence to bear on this issue by computing style excess returns and correlating them with hiring decisions. Specifically, we compute style *CERs* by cumulating the return of the investment style ($R_{b,s}$) minus the return of a broad index that reflects the return for that asset class. For example, to compute the style *CER* for small-cap growth, we cumulate the return difference between the small-cap growth benchmark (Russell 2500 Growth) and the Russell 3000 index. Second, we calculate investment manager *CERs* as described in Section II.E. Panel A of Table III shows style and investment manager excess returns 1, 2, and 3 years before hiring with standard errors in parentheses.

There is some evidence of style-chasing in domestic equity: The 3-year prehiring return is 1.20%, albeit with a standard error of 3.59%. In contrast, there is

Table III

Style and Investment Manager Excess Returns Prior to Hiring

Style excess returns are calculated by subtracting the average return for all styles in an asset class from the style return of the hiring decision. These excess returns are then cumulated over appropriate horizons. Style *CERs* are only shown for domestic equity mandates. Excess returns for investment managers are calculated by differencing the raw return for the manager in the hiring mandate from benchmark returns for the same mandate. Information on benchmarks is provided in Table A1. Heteroskedasticity, serial, and cross-correlation consistent standard errors are calculated using the procedure described in Jegadeesh and Karceski (2004). Panel B shows the results of regressions with style or investment manager excess returns. The return regression is $y_j = \beta x_j + \delta z_j + \varepsilon_j$, where y_j is the 3-year pre-hiring cumulative excess return, x_j is a vector of explanatory variables, and z_j is a dummy variable for whether a consultant was employed. The selection equation is $z_j^* = \gamma w_j + u_j$, where $z_j = \{ {}^{1,\ if\ z_j^* > 0}_{0,\ otherwise}$ and w_j is a vector of explanatory variables. The selectivity correction is identical to the first model in Panel D of Table II.

Panel A: Univariate Returns

	Style *CERs*			Investment Manager *CERs*		
	−3 to 0	−2 to 0	−1 to 0	−3 to 0	−2 to 0	−1 to 0
Domestic equity	1.20	0.95	0.49	12.21	8.54	4.21
	(3.59)	(2.62)	(1.17)	(2.50)	(2.27)	(1.52)
Fixed income	−0.43	−0.55	−0.26	3.55	2.28	1.15
	(1.01)	(0.70)	(0.33)	(0.27)	(0.29)	(0.22)
International equity	−0.30	−0.50	−0.58	17.05	11.80	5.70
	(1.47)	(0.85)	(0.67)	(3.61)	(2.66)	(1.37)

Panel B: Selectivity-Corrected Regressions Using 3-Year Pre-hiring Returns

	Style *CER* (−3 to 0)			Investment Manager *CER* (−3 to 0)		
Constant	−5.93	−1.45	−3.52	−7.49	7.79	6.23
	(2.75)	(1.56)	(4.08)	(0.78)	(2.21)	(1.32)
Headline-sensitive indicator	3.17	5.59	1.95	−1.35	−0.27	0.29
	(1.20)	(3.10)	(1.66)	(0.74)	(1.80)	(1.11)
Log (plan sponsor size)	0.17	0.63	0.11	0.37	0.24	0.23
	(0.11)	(0.43)	(0.16)	(0.10)	(0.29)	(0.14)
Consultant indicator	9.75	11.39	4.81	1.63	0.99	1.83
	(3.97)	(7.67)	(0.54)	(0.66)	(1.53)	(1.09)
Consultant * headline-sensitive	1.69	1.98	2.17	0.25	1.48	−0.28
	(0.74)	(1.55)	(1.12)	(0.91)	(2.01)	(1.33)
Overfunded indicator	–	−2.29	–	–	1.90	–
		(0.71)			(1.93)	
Allocation index	–	–	0.09	–	–	2.70
			(1.04)			(1.32)
Number of observations	7,594	1,746	4,444	6,648	1,544	3,898

no style-chasing in either fixed income or international equity. In terms of pre-hiring performance, the cumulative excess returns for investment managers are consistently positive across all horizons and for all asset classes. They are the largest for international equity with a 3-year pre-hiring return of 17.05% and smallest for fixed income with a 3-year pre-hiring excess return of 3.55%.

Clearly, and not surprisingly, plan sponsors condition their hiring decisions on the performance of investment managers.

In Panel B, we investigate how different attributes of plan sponsors are correlated with the return threshold at which investment managers are hired. The endogeneity of consultant use (see results in Panel D, Table II) necessitates a procedure that corrects for selectivity. We follow Madalla (1983) and estimate the following model:

$$y_j = \beta x_j + \delta z_j + \varepsilon_j, \tag{4}$$

where y_j represents 3-year pre-hiring cumulative style or investment manager excess return, x_j is a vector of explanatory variable, and z_j is a dummy variable for whether a consultant was employed. The selection equation is modeled as

$$z_j^* = \gamma w_j + u_j, \tag{5}$$

where $z_j = \{^{1,\ if\ z_j^*>0}_{0,\ otherwise}$ and w_j is a vector of explanatory variables. The regressions are estimated via a two-stage procedure and standard errors account for clustering, where an investment management firm is hired for a mandate in the same style and period by different plan sponsors.

The selection equation that we use is identical to the first model in Panel D of Table II and not shown in Table III. The independent variables (x_j) in the return regression measure plan sponsor attributes that, based on the discussion in Section I, we expect to be correlated with pre-hiring return thresholds. We present three regression models. The first model includes an indicator variable for headline-sensitive plan sponsors, the logarithm of plan sponsor size, a consultant indicator (from the first-stage regression), and an interaction effect between the consultant indicator and the headline-sensitive sponsor indicator. This base specification shows that sponsor size plays no role in style-chasing but that headline risk-sensitive plan sponsors engage in style-chasing. Sponsors that employ consultants also engage in more style-chasing than those that do not. An interaction effect between the two indicates that the presence of a consultant accentuates the style-chasing behavior in headline risk-sensitive plan sponsors rather than reducing it. In the second model, we add an indicator variable for whether the plan is overfunded. This drops the sample size since funding information is only available for a small sample of public and corporate plans. The overfunded indicator variable is significantly negative, indicating that overfunded plans do not engage in style-chasing, most likely because they have little incentive to do so. In the third model, we add the allocation index to the base model to see if our proxy for restrictions influences style returns. It does not.

We also study variation in investment manager pre-hiring returns using the same sets of models. The base model suggests that larger sponsors condition their hiring on larger investment manager returns. Similarly, the presence of consultants is positively correlated with pre-hiring investment manager returns. But neither funding levels nor the allocation index are related to pre-hiring investment manager returns. Overall, the data suggest that there is some style-chasing and that plan sponsors condition their hiring decisions on investment manager performance. The magnitudes of these effects are different for headline risk-sensitive plan sponsors and those that are advised by consultants. We turn now to an investigation of post-hiring performance.

C. Post-hiring Performance

Table IV shows cumulative excess returns (Panel A), information ratios (Panel B), and alphas from factor models (Panel C) 1, 2, and 3 years after hiring. For comparison purposes, we also show pre-hiring returns over the same horizons. To ensure that changing sample sizes between the pre- and post-period do not drive our results, we report excess returns for a balanced sample in which returns can be computed for matched horizons before and after hiring. In addition to the full sample, we also show separate results for domestic equity, fixed income, and international equity.

As before, pre-hiring performance is significantly positive using all three measures of excess returns. For the full sample, post-hiring performance is statistically flat. Cumulative excess returns 1, 2, and 3 years after hiring are 0.4%, 1.1%, and 1.8% with standard errors of 0.6%, 0.8%, and 1.1%, respectively. The only case in which post-hiring excess returns are positive and statistically significant is for international equity mandates. This effect for international equity appears to be quite robust for all performance measures.

Recall that the sample of hiring decisions is for active mandates in which, presumably, plan sponsors hope to earn future excess returns. Our results suggest that, on average, plan sponsors are unsuccessful in this endeavor. It could be that some plan sponsors are more successful than others because of differences in the nature of agency relationships and incentive structures. For example, the degree of headline risk faced by a plan sponsor could influence its ability to successfully pick managers that beat their benchmark. We study the degree to which such plan sponsor attributes result in superior post-hiring excess returns through selectivity-corrected return regressions analogous to those in Table IV. The dependent variable is the 3-year post-hiring cumulative excess return. The base regression model contains 3-year pre-hiring cumulative excess returns, plan sponsor size, consultant indicator, and headline risk-resistant, risk-sensitive, and risk-neutral indicators as explanatory variables. Since all the headline risk indicators are included, the model is estimated without an intercept. Fixed effects for detailed investment styles (not shown) allow for intercept shifts in post-hiring returns that are not picked up by the benchmark used to compute excess returns.[11]

[11] Although the dependent variable is an excess return (say, raw return of a small-cap value manager minus the return on a small-cap value index), there may still be heterogeneity in

Table IV

Investment Manager Excess Returns before and after Hiring

Panel A presents average cumulative excess returns computed by summing quarterly excess returns (raw minus benchmark return). Information on benchmarks is provided in Table A1. Heteroskedasticity, serial, and cross-correlation consistent standard errors are calculated using the procedure described in Jegadeesh and Karceski (2004). Panel B shows information ratios calculated by scaling the average excess return by its standard deviation. Panel C shows estimates of alphas from calendar-time regressions factor regressions with standard errors in parentheses. For domestic equity mandates, we use the Fama and French (1993) three-factor model with market, size, and book-to-market factors. For fixed income mandates, we employ a three-factor model with the Lehman Brothers Aggregate Bond Index return, a term spread (the difference between the long-term government bond return and the T-bill return), and a default spread (the difference between the corporate bond return and the long-term government bond return). For international equity mandates, we use international versions of the domestic equity three-factor models. In all pre- and post-return comparisons, we require a balanced sample (i.e., returns be available in matched pre- and post-hiring horizons).

	Pre-hiring Period (Years)			Post-hiring Period (Years)		
	−3 to 0	−2 to 0	−1 to 0	0 to 1	0 to 2	0 to 3
Panel A: Cumulative Excess Returns						
Full sample	10.39	7.04	3.42	0.42	1.12	1.88
	(1.87)	(1.45)	(0.97)	(0.61)	(0.85)	(1.11)
Domestic equity	12.54	8.72	4.25	−0.22	−0.07	0.77
	(2.85)	(2.31)	(1.52)	(0.85)	(1.31)	(1.86)
International equity	17.11	11.83	5.71	3.32	7.09	9.00
	(3.67)	(2.69)	(1.37)	(1.27)	(1.71)	(2.62)
Fixed income	3.72	2.32	1.16	0.30	0.65	0.80
	(0.24)	(0.29)	(0.23)	(0.23)	(0.42)	(0.55)
Panel B: Information Ratios						
Full sample	3.69	2.61	1.59	0.45	0.78	1.05
Domestic equity	3.14	2.31	1.34	−0.04	0.11	0.30
International equity	4.52	3.45	2.15	1.42	2.42	2.89
Fixed income	5.13	3.43	2.25	1.31	1.74	1.98
Panel C: Calendar-Time Alphas from Factor Regressions						
Domestic equity	1.10	1.09	1.06	−0.17	−0.13	−0.08
	(0.26)	(0.29)	(0.35)	(0.15)	(0.14)	(0.16)
International equity	1.47	1.54	1.31	0.77	0.68	0.61
	(0.45)	(0.53)	(0.55)	(0.33)	(0.32)	(0.27)
Fixed income	0.36	0.35	0.39	0.19	0.21	0.21
	(0.09)	(0.08)	(0.09)	(0.11)	(0.08)	(0.08)

The base regressions in Table V show strong evidence of return reversal. The negative coefficients on the pre-hiring return variable do not imply negative post-hiring returns, just that post-hiring returns are smaller than pre-hiring

investment manager returns within small-cap value asset class. For example, one manager might invest in micro-cap securities exclusively, even though its investment style is regarded as small-cap. These indicator variables account for such effects.

Table V

Post-hiring Selectivity-Corrected Excess Return Regressions

The return regression $y_j = \beta x_j + \delta z_j + \varepsilon_j$, where y_j is the 3-year post-hiring cumulative excess return, x_j is a vector of explanatory variables, and z_j is a dummy variable for whether a consultant was employed. The explanatory variables are computed as in earlier tables. The selection equation is $z_j^* = \gamma w_j + u_j$, where $z_j = \{ {}^{1,\ if\ z_j^* > 0}_{0,\ otherwise}$ and w_j is a vector of explanatory variables. The selectivity correction is done via a two-stage estimation procedure. The selection equations for the full sample, public plans, and corporate plans are as reported in Panel D of Table II and are not reported in this table. Standard errors (in parentheses) account for clustering in observations where the investment manager is hired for a mandate in the same style and period by different plan sponsors.

	All Plan Sponsors				Public Plans	Corp. Plans
Pre-hiring return	−0.17	−0.17	−0.24	−0.18	−0.17	−0.01
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.05)
Log (plan sponsor size)	0.61	0.99	0.37	0.25	0.32	1.04
	(0.09)	(0.14)	(0.14)	(0.13)	(0.31)	(0.44)
Headline risk-resistant indicator	−1.13	−2.87	1.01	−0.38	–	–
	(0.80)	(2.01)	(1.72)	(1.18)		
Headline risk-sensitive indicator	−1.70	−4.12	−0.22	−0.63	–	–
	(0.07)	(1.06)	(0.12)	(1.18)		
Headline risk-neutral indicator	0.26	2.28	1.17	1.06	–	–
	(0.95)	(1.22)	(1.19)	(1.05)		
Expected value of consultant	2.02	6.19	1.95	1.70	0.82	1.74
	(0.43)	(1.37)	(0.62)	(0.59)	(1.10)	(1.44)
Consultant ∗ plan sponsor size	–	−0.64	–	–	–	–
		(0.18)				
Log (mandate / assets$_{t-1}$)	–	–	−0.22	–	–	–
			(0.11)			
Allocation index	–	–	–	4.11	–	–
				(1.29)		
Underfunded indicator$_{t-1}$	–	–	–	–	1.51	–4.48
					(3.01)	(3.07)
Overfunded indicator$_{t-1}$	–	–	–	–	−1.62	−0.30
					(0.80)	(0.14)
Number of observations	6,170	6,170	3,184	3,633	921	513

returns. Larger plan sponsors appear to generate superior post-hiring performance, consistent with scale economies at the plan sponsor level. The sensitivity to headline risk could influence hiring decisions in two opposing ways. It could be that increased public scrutiny improves incentives and results in higher post-hiring performance. Alternatively, headline risk sensitivity could be a response to the lack of incentives for plan sponsors to generate superior performance. Consistent with the latter explanation, we find that the performance of headline risk-sensitive plan sponsors is generally negative, particularly when compared to sponsors that are neutral to such risk. Finally, post-hiring returns are higher for decisions in which a consultant was used in selecting the investment manager.

The above results indicate that smaller plan sponsors have lower post-hiring performance and that consultants add value. Since larger plan sponsors are less likely to employ consultants, it is also interesting to examine whether consultants add more or less value for them. In the second model, we find that the interaction effect between sponsor size and consultant use is negative. This suggests that consultants add value for smaller plan sponsors but are detrimental to the post-hiring performance of larger plan sponsors. This could be because consultants do not bring scale economies or expertise to larger plans and are instead used as a shield in the case of poor hiring decisions.

Scale diseconomies could be present for investment managers. Consider, for example, a small-cap growth manager that is at capacity with \$1 billion under management. If this manager then receives a \$200 million mandate from a state-level plan sponsor, its future returns could deteriorate because of higher trading costs. In the third model, we add the size of the mandate obtained by the investment manager, scaled by (lagged) assets under management. Mandate size scaled by assets is negatively related to post-hiring returns. In the fourth model, we augment the base regression with the asset allocation index. The regression shows a strong positive relation between post-hiring returns and the allocation index, suggesting that the imposition of restrictions is detrimental to performance. Finally, we would like to add the funding status of the plan in the year prior to the hiring decision to these regressions. But since these data are available only for a subset of public and corporate plans, we estimate such regressions separately for these sponsors (and accordingly drop the headline risk indicator variables). For both corporate and public plans, the overfunded plan indicator is negative and significant, consistent with Hart's (1992) argument that overfunded plans have little incentive to generate superior performance.

The economic magnitude of some of these effects is quite large. From the base specification, the average impact of a one-standard deviation increase in 3-year pre-hiring returns (with other variables evaluated at their mean) implies a decrease in 3-year post-hiring cumulative excess returns of 4.7%. Headline risk-sensitive sponsors have excess returns that are lower by 1.7% than their counterparts and the use of a consultant leads to an increase in 3-year post-hiring returns by over 2.0% depending on the specification. Lower performance for overfunded plans (compared to underfunded plans) varies from 1.6% for public plans to 0.3% for corporate plans.

D. Discussion

Our aggregate results show that plan sponsors condition their hiring decisions on superior performance. However, post-hiring performance is essentially flat. One way to think about these results is to consider the role of persistence in investment manager returns. If there is little or no persistence in the performance of investment managers in general, then on average, hiring decisions should produce zero excess returns. This does not necessarily mean that plan sponsors achieve their objectives, since they hire investment managers in our sample to deliver excess returns. However, it does imply no ex-post losses. A full-scale analysis of persistence is beyond the scope of our paper. However, Christopherson et al. (1998) and Busse et al. (2007) undertake such an analysis for institutional investment managers and find evidence of persistence among winners for up to 1 year, and in some cases, longer. Their persistence results indicate that plan sponsors *could* generate excess returns by appropriately timing hiring decisions but, apparently, they do not.

However, the aggregate results mask considerable cross-sectional variation, not only in elements of pre-hiring decisions (return thresholds, style-chasing, consultant use), but also in post-hiring performance. This variation is tied to plan sponsor attributes that reflect agency problems and incentive structures across plans.

IV. The Termination of Investment Managers

A. *Reasons for Termination*

Our firing sample consists of 869 termination decisions. The number of termination decisions captured by the data collection process is substantially smaller than hiring decisions for three reasons. First, the data sources that we use (which to our knowledge are the only publicly available sources) serve a marketing function, that is, they are designed to inform subscribers that a plan sponsor is searching for an investment manager in a particular asset class/mandate. These sources are not designed to track performance or to assign blame. As such, the emphasis is on new accounts and revenue. Second, termination decisions are generally viewed with some distaste and there is a natural disinclination to report terminations. Certainly, investment managers have no incentive to report their own terminations. Plan sponsors may choose not to publicize terminations because they may employ the same manager for another mandate, either currently, or in the future. Third, there has been an increase in the assets under the administration of plan sponsors over the sample period. Ergo, the number of hiring decisions in the population is likely to be larger than of firing decisions.

Panel A of Table VI shows the distribution of termination decisions by type of plan sponsor and within headline risk category. Also shown are statistics on plan sponsor and mandate size. All major categories of sponsors except private universities are represented in our data. The number of terminations by endowments and foundations (in the headline risk-neutral category) are quite small. The size and mandate statistics are similar to those reported for hiring decisions. Although we do not show the time-series distribution, the number of firing observations increases over time because our data sources do a better job of capturing such decisions in the later years.

We use the textual information in our data sources to manually categorize the reasons for the termination of the investment manager into six categories. Four of those categories are related to activities/events specific to the

Table VI

Distribution of Firing Decisions by Plan Sponsors

Definitions for variables in Panels A and C are the same as those reported in Table I. Panel B shows the distribution of firing decisions by reasons identified by the data sources. Investment manager mergers may be either before the termination or impending. Regulatory action against the investment manager is both announced and ongoing. Personnel turnover at the investment management firm may be forced or voluntary. Plan reorganizations occur when two plans have to be merged. Plan reallocation category refers to firings because the plan sponsor has decided to move away from the asset allocation / investment style offered by the investment manager. The "not reported" category includes terminations in which the plan sponsors were asked the reason for the termination but deliberately did not offer a reason. When no public document contains information about the termination, the reason for the determination is determined to be missing.

	Number of Firings	Plan Sponsor Size ($M)			Mandate Size ($M)		
		Mean	Median	N	Mean	Median	N
Panel A: Headline Risk and Plan Sponsor Type							
Headline Risk-Resistant							
Corporate	112	2,209	700	777	95	37	80
Private universities	29	176	150	27	16	13	19
Miscellaneous	47	4225	350	33	197	62	35
Headline Risk-Neutral							
Endowments & found.	29	6,899	722	24	31	35	13
Headline Risk-Sensitive							
Local public plans	238	5,716	650	197	104	50	213
State public plans	181	24,319	13,200	143	304	200	157
Misc. public plans	128	3,494	618	101	107	50	111
Unions	75	383	190	57	103	20	70
Public universities	30	273	200	26	21	10	23
Panel B: Distribution of Firing Decisions by Stated Reason							
Manager merger	22	5,951	1,100	19	142	55	15
Manager regulatory action	53	13,375	2,214	48	258	112	38
Manager personnel turnover	49	9,425	487	42	76	35	44
Manager performance	297	7,062	767	238	130	50	257
Plan reorganization	36	9,555	422	28	131	70	31
Plan reallocation	111	4,458	675	80	218	75	89
Not reported	104	8,181	433	88	108	38	94
Missing	197	9,081	870	142	144	55	153
Panel C: Distribution of Firing Decisions by Funding Status							
Corporate Plans							
Underfunded	22	4,198	1,200	19	198	83	16
Overfunded	20	1,494	950	13	36	30	11
Public Plans							
Underfunded	182	19,966	8,350	164	237	200	161
Overfunded	76	21,593	12,000	52	286	200	60

investment management firm: the merger of two investment management firms, regulatory action against the investment management firm, personnel turnover, and performance. Two of the categories are related to the plan sponsor itself: either a reorganization of the plan sponsor or a reallocation across asset classes.[12] If the text of the termination decision indicates that the plan sponsor executive willfully refused to provide the reason for the termination, we identify it as "not reported." This is different from "missing" because that category contains terminations for which we cannot find any information.

Only 34% (297 observations) of the total terminations (including those with unidentified reasons) are due to the performance of the investment manager. Activities and events at the investment manager firm that are unrelated to performance (mergers, regulatory action, and personnel turnover) account for another 14%. Plan sponsor changes (reorganizations and asset reallocations) are responsible for almost 17% of terminations.

There are two caveats associated with the termination reasons described above. First, the reasons are self-identified by the plan sponsor. Second, elements of current or future underperformance could creep into non-performance categories. An acquisition of one investment management by another might take place after underperformance. Alternatively, a plan sponsor may terminate an investment manager after the departure of key personnel because it believes that the departure will cause underperformance in the future.

Panel C shows the distribution of firing decisions, sponsor size, and mandates by the funding status of corporate and public sponsors. Out of the 112 terminations from corporate plan sponsors, we only have funding information for 42, which are roughly evenly split between under- and overfunded plans. The underfunded corporate plans are considerably larger than the overfunded plans. Of the 546 public plans in the termination sample, we have funding information for 258, and a significant majority of those are underfunded (70%).

In Table VII, we present a two-way frequency tabulation of the reasons for termination and plan sponsor attributes. As with hiring decisions, our purpose is to determine if headline risk, funding status, size, and consultant use influence the degree to which plan sponsors terminate investment managers for various reasons. Before presenting the results, we alert the reader to two important facts. First, some of the sample sizes for termination reasons are quite small. Although we report all cuts of the data, we only make inferences when sample sizes are reasonable. Second, our priors are well formed primarily for two termination reasons, performance and regulatory action. For example, we expect that headline risk-sensitive plan sponsors may be more likely to terminate managers for poor performance or regulatory action than headline risk-resistant sponsors. We cannot a priori make the same claim for plan sponsor reorganizations/reallocations or even for investment manager personnel turnover. Again, we make inferences only where we have sensible priors.

With those qualifications in mind, Table VII presents the frequency of termination decisions across subcategories of sponsors in Panels A through F for

[12] We also place some very low-frequency reasons in the above categories. Terminations because the consultant drops coverage of an investment manager (4 observations) or the plan sponsor is consolidating the number of managers to cut costs (22 observations), or the plan sponsor has funding needs (5 observations) are placed in the plan reorganization category. Three observations in which investment managers are terminated for style drift are included in the performance category.

Table VII

Two-Way Frequency Distribution of Firing Decisions

The table shows the number of firing decisions for each identified reason and subgroup (panel), as well as the percentage of observations in that column and category. For example, of the 297 terminations identified as due to poor performance, 79.1% originated from sponsors that are sensitive to headline risk. Frequency distributions are not shown for the "not reported" and missing categories. Frequencies are also not reported from intermediate groups (i.e., headline risk-neutral plan sponsors, medium-size plan sponsors, and sponsors with allocations indices in the middle group). Low and high cutoffs for the allocation index are based on the bottom and top quartiles. Similarly, small and large cutoffs for sponsor size are based on the bottom and top quartiles.

	Investment Manager Reasons				Plan Sponsor Reasons		
	Merger	Regulatory Action	Turnover	Performance	Reorganization	Reallocation	Total
Panel A: Headline Risk							
Headline risk-resistant	9.1	24.5	14.3	18.8	25.0	19.8	21.6
Headline risk-sensitive	90.9	67.6	81.6	79.1	75.0	73.9	75.0
Number of observations	22	53	49	297	36	111	869
Panel B: Public Plan Funding Status							
Underfunded plans	66.7	90.5	77.8	75.6	84.6	53.3	70.5
Overfunded plans	33.3	9.5	22.2	24.4	15.4	46.7	29.5
Number of observations	6	21	18	90	13	30	258
Panel C: Corporate Plan Funding Status							
Underfunded plans	0.0	100	33.3	64.7	0.0	20	52.4
Overfunded plans	0.0	0.0	66.7	35.3	100	80	47.6
Number of observations	0	4	3	9	3	1	42
Panel D: Allocation Index							
Low allocation index	35.0	25.0	35.7	30.8	30.0	37.8	32.1
High allocation index	20.0	12.5	16.7	22.3	6.7	16.3	20.9
Number of observations	20	48	42	247	30	98	708
Panel E: Consultant Use							
No consultant	22.7	15.1	20.4	22.7	30.6	21.6	22.9
Consultant	77.3	84.9	79.6	77.3	69.4	78.4	77.1
Number of observations	22	53	49	297	36	111	869
Panel F: Plan Sponsor Size							
Small plan sponsors	20	11.1	21.1	31.2	37.9	39.2	31.7
Large plan sponsors	25	37.8	26.3	21.7	27.6	20.6	21.8
Number of observations	20	45	38	263	29	97	757

each termination reason. Correct interpretation of these frequencies requires one to compare the frequency distribution across a subcategory and reason with the unconditional distribution across that subcategory (reported in the last column). For example, to determine if headline risk-sensitive plan sponsors are more likely to terminate for underperformance than headline risk-resistant sponsors, we compare their frequency distribution (79% vs. 18.8%) to that for all terminations (75% vs. 21%). Consistent with our expectations, headline risk-sensitive sponsors are more likely to terminate investment managers for poor performance (79%) than headline risk-resistant sponsors (18%); the p-value for this difference is 0.00. Overfunded plans may be less likely to terminate underperforming managers because they have some slack. Alternatively, they may be more likely to terminate for poor performance if they achieved overfunding via good firing decisions. We find that overfunded plans are less likely to terminate for poor performance than their counterparts, suggesting that the first effect dominates. Consultant-advised plans may be more likely to terminate underperforming managers because consultants want to distance themselves from the poor performance of investment managers. But we find that consultant-advised plans are no more likely to terminate investment managers for poor performance (and regulatory action) than those without consultants.

B. Pre- and Post-firing Performance

In Table VIII, we show average cumulative excess returns for investment managers prior to the termination. Panel A shows the excess returns and standard errors for all terminations as well as by the reason for termination. The average excess return for all terminations is not different from zero: The 3-year (1-year) excess return is 0.33% (−0.72%) with a standard error of 1.27% (0.68%). This reflects the heterogeneity in the reasons for termination. The excess returns prior to performance-based firing are significantly negative (−4.1% over 3 years with a standard error of 1.2%). In fact, poor performance and regulatory action are the only termination reasons that have negative pre-firing returns, although returns for the latter are not statistically significant. Excess returns prior to terminations due to mergers are positive but returns for the other termination reasons are statistically indistinguishable from zero. In Panel B we investigate whether headline risk, funding status, sponsor size, the allocation index, and consultant use are related to pre-firing returns using selectivity-corrected regressions similar to those employed earlier. These regressions are estimated for performance-based terminations only because that is where we expect such effects to be important. None of the variables that were important for pre- and post-hiring returns are important here, although it is entirely possible that the small size limits the ability of the regression to detect meaningful differences.

In Table IX, we show cumulative excess returns (Panel A), information ratios (Panel B), and calendar-time alphas from factor regressions (Panel C) after termination. To allow for easy comparisons, we also show pre-firing results in the same table and break up the results for domestic equity, fixed income, and

Table VIII

Pre-firing Investment Manager Excess Returns

The table shows pre-firing cumulative excess returns for investment management firms. Panel A shows returns for terminations due to each of the stated reasons. Panel B shows the results of regressions with investment manager excess returns. The return regression is $y_j = \beta x_j + \delta z_j + \varepsilon_j$, where y_j is the 3-year pre-hiring cumulative excess return, x_j is a vector of explanatory variables, and z_j is a dummy variable for whether a consultant was employed. The selection equation is $z_j^* = \gamma w_j + u_j$, where $z_j = \{ {}^{1,\ if\ z_j^* > 0}_{0,\ otherwise}$ and w_j is a vector of explanatory variables. The selectivity correction is identical to the first model in Panel D of Table II. Heteroskedasticity, serial, and cross-correlation consistent standard errors are in parentheses and are calculated using the procedure described in Jegadeesh and Karceski (2004).

Panel A: Firing Reasons			
	Pre-firing Period (Years)		
	−3 to 0	−2 to 0	−1 to 0
All	0.33 (1.27)	−2.11 (1.27)	−0.72 (0.68)
Merger	6.86 (2.74)	5.50 (1.38)	4.17 (1.51)
Regulatory action	−2.98 (5.31)	−1.87 (3.83)	−1.45 (3.19)
Turnover	4.49 (3.11)	−0.62 (4.74)	1.24 (3.52)
Performance	−4.14 (1.26)	−7.01 (1.80)	−3.71 (0.88)
Reorganization	3.22 (1.14)	0.33 (1.29)	−1.37 (0.93)
Reallocation	1.42 (1.75)	0.30 (1.13)	0.79 (1.27)
Not reported	4.00 (2.36)	−0.38 (0.98)	−0.62 (0.70)
Missing	3.27 (2.53)	1.29 (2.45)	2.25 (1.35)

Panel B: Selectivity-Corrected Regressions Using 3-Year Pre-firing Returns for Performance-Based Firings			
Constant	−10.76	−6.15	−13.10
	(11.91)	(17.48)	(19.93)
Headline-sensitive indicator	−5.71	−8.16	−0.52
	(9.18)	(12.61)	(8.50)
Headline-resistant indicator	−4.15	−3.05	–
	(9.22)	(13.25)	
Log (plan sponsor size)	0.68	1.39	2.35
	(0.62)	(0.83)	(1.70)
Consultant indicator	8.41	6.42	−14.73
	(10.25)	(14.80)	(15.73)
Allocation index	–	12.36	–
		(7.96)	
Overfunded indicator	–	–	5.65
			(6.12)
Number of observations	212	159	80

international equity. As before, pre-firing returns are generally statistically indistinguishable from zero. After firing, in the first 2 years, the cumulative excess returns are positive but with large standard errors. In some cases, in the third year, the excess returns are large and statistically significant; for the full sample, the 3-year cumulative excess return is 3.3% with a standard error of 1.4%.

Table IX

Investment Manager Excess Returns before and after Firing

Panel A presents average cumulative excess returns computed by summing quarterly excess returns. Information on benchmarks is provided in Table A1. Heteroskedasticity, serial, and cross-correlation consistent standard errors are calculated using the procedure described in Jegadeesh and Karceski (2004). Panel B shows information ratios calculated as the average excess return scaled by the standard deviation of the excess return. Panel C shows estimates of alphas from calendar-time factor regressions with standard errors in parentheses. For domestic equity mandates, we use the Fama and French (1993) three-factor model with market, size, and book-to-market factors. For fixed income mandates, we employ a three-factor model with the Lehman Brothers Aggregate Bond Index return, a term spread computed as the difference between the long-term government bond return and the T-bill return, and a default spread computed as the difference between the corporate bond return and the long-term government bond return. For international equity mandates, we use an international version of the domestic equity three-factor model. In all pre- and post-return comparisons, we require a balanced sample (i.e., that returns be available in matched pre- and post-firing horizons).

	Pre-firing Period (Years)			Post-firing Period (Years)		
	−3 to 0	−2 to 0	−1 to 0	0 to 1	0 to 2	0 to 3
Panel A: Cumulative Excess Returns						
Full sample	2.27	−2.06	−0.74	0.98	1.47	3.30
	(2.10)	(1.20)	(0.61)	(0.77)	(1.27)	(1.46)
Domestic equity	2.63	−3.28	−1.26	0.83	1.15	3.44
	(3.41)	(1.38)	(0.71)	(1.08)	(1.76)	(2.57)
International equity	9.15	3.72	2.42	1.52	2.66	4.10
	(0.82)	(1.87)	(1.61)	(1.35)	(3.11)	(3.59)
Fixed income	−1.54	−1.47	−0.86	0.91	1.51	2.19
	(0.86)	(1.39)	(0.62)	(0.55)	(1.04)	(1.58)
Panel B: Information Ratios						
Full sample	0.36	−0.37	−0.09	0.76	1.49	2.12
Domestic equity	0.63	−0.31	−0.15	0.30	0.97	1.39
International equity	2.18	0.74	0.67	0.12	0.66	0.62
Fixed income	−1.09	−1.09	−0.28	2.21	3.23	4.35
Panel C: Calendar-Time Alphas from Factor Regressions						
Domestic equity	−0.06	−0.42	−0.57	0.45	0.14	0.10
	(0.22)	(0.19)	(0.21)	(0.55)	(0.36)	(0.32)
International equity	0.42	0.01	−0.63	1.00	0.64	0.57
	(0.25)	(0.26)	(0.68)	(0.52)	(0.30)	(0.27)
Fixed income	0.03	0.15	0.19	0.33	0.30	0.30
	(0.14)	(0.11)	(0.13)	(0.09)	(0.09)	(0.08)

Investment manager termination could be correlated with changes in portfolio risk before and after termination and affect our inferences. For example, Brown, Harlow, and Starks (1996), Chevalier and Ellison (1999), and Busse (2001) show that underperforming mutual fund managers increase portfolio risk in an attempt to generate superior returns. Gallo and Lockwood (1997) show correlated changes in investment style. Such behavior may be prevalent in institutional investment management firms as well. Our calendar-time factor models allow us to test if these pre- and post-event betas are different from each other. Although we do not display the results, we mostly fail to reject the null hypothesis of constant beta. We suspect two reasons for this. First, most investment management firms have a large stable of clients. Losing one or two clients is unlikely to dramatically influence risk-taking incentives. Second, plan sponsor monitoring of tracking error (Del Guercio and Tkac (2002)) is likely to reduce incentives to change risk profiles dramatically.

C. Discussion

As a whole, our data appear to indicate that plan sponsors show limited timing ability in terminating investment managers. In the case of nonperformance terminations, a priori, one should not expect over- or underperformance subsequent to termination. In untabulated results, that is exactly what we find; post-firing excess returns for nonperformance-based firings are essentially zero. In the case of performance-based termination, expectations of post-firing excess returns depend on the perspective of the evaluator. The plan sponsor terminating the investment manager presumably expects post-firing returns to be negative. Counterfactually, we find that the 3-year post-firing excess return for performance-based terminations is 4.20% with a standard error of 1.87%. An independent observer could argue that post-firing excess returns should be zero (under mean reversion) or even positive, either under diseconomies of scale in investment management or if termination disciplines the investment manager. The diseconomies channel is simply that if the manager is capacity constrained, then removal of a mandate might allow the investment manager to improve returns, perhaps through lower trading costs. The disciplinary channel implies that termination improves performance by inducing greater effort. Both channels imply that post-firing returns should be correlated with the size of the lost mandate scaled by assets under management. In unreported regressions with post-firing excess returns as the dependent variable, we find that the coefficient on this scaled mandate is positive and significant (the coefficient is 0.008 with a t-statistic of 1.96), even in the presence of other control variables.

The extent to which such (mis)timing damages the performance of the plan sponsor depends on the performance of the investment managers hired to replace terminated managers. In other words, the appropriate comparison is the returns that the plan sponsor earned (post-hiring) relative to what it would have earned (post-firing). Although it is tempting to simply compare post-hiring returns in Table IV with post-firing returns in Table IX and conduct a cross-sectional analysis, we refrain from doing so because firing and hiring decisions are coordinated using complicated mechanisms. We proceed to an analysis of such “round-trips” below.

V. Round-Trip Termination and Selection of Investment Managers

The best way to illustrate the complexity of a round-trip termination and selection decision is by way of examples.

Example 1

In the first quarter of 2000, the St. Louis Employees Retirement System terminated 1,838 investment advisors for its core long-term fixed income portfolio, reportedly because of poor performance. It then hired Reams Asset Management to handle this $45 million portfolio. Watson Wyatt Investment Consulting assisted in the search.

Example 2

In the first quarter of 2002, the Arapahoe County Employees Retirement System hired Barclays Global Investors to manage $15 million in passive global large-cap equity, Artisan Partners for a $10-million active international all-cap equity mandate, Brazos for $9 million in active domestic micro-cap equity, and Royce for $5 million in active domestic small-cap equity. The Barclays's hiring was funded by reallocating $15 million from a $44-million active domestic large-cap growth equities portfolio managed by Fayez Sarofim. Artisan's allocation came from terminating a $10-million active international all-cap equities portfolio managed by Brinson Partners. Brazos and Royce were funded by terminating a $14-million active domestic mid-cap growth equities portfolio managed by Denver Investment Advisors.

The first example is a straightforward round-trip firing and hiring decision in which the mandate size and type is the same, and the reason for the decision clearly delineated. The second contains two round-trip observations: (1) Denver Investment Advisors is terminated and replaced by Brazos and Royce. The mandates for the hired investment managers are different from the terminated investment manager and the allocation of the $14 million portfolio is not even. (2) Brinson Partners is terminated and replaced by Artisan Partners in the same mandate. Note that the Barclays Global Investors hiring does not create a round-trip observation since it is not the result of a termination but an allocation adjustment for an ongoing investment manager.

A. Sample Construction and Description

Because of the complexity of the process described above, we cannot mechanically associate hiring and firing decisions, and therefore build a sample using manual procedures. We start with our sample of firing decisions. For each firing decision, we match hiring decisions by the same plan sponsor up to one quarter after the firing date.[13] This produces 2,206 candidate firing–hiring decisions, which contain duplications, often because a hiring decision can be associated with more than one firing decision and vice versa. For each candidate observation, we then search for articles detailing the decisions in the following trade journals: *P&I*, *Investment Management Weekly*, *Money Management Letter*, and *Dow Jones Money Management Alert*. We mark each round-trip with an ID that allows us to track these decisions and eliminate duplications. This process

[13] We restrict our search for matching hiring decisions to one quarter after the firing to limit the amount of manual data collection required.

identifies 663 round-trip firing/hiring decisions. We then match these round-trip decisions with our returns database, keeping only decisions for which we have some returns. As before, this eliminates decisions involving investments in hedge funds, venture capital funds, and private equity. Our final sample consists of 412 round-trip firing/hiring decisions between 1996 and 2003.

On average, each round-trip decision is associated with the firing and hiring of 1.1 investment managers, with a maximum of 11 investment managers hired or 7 investment managers fired in a particular decision. The average mandate size for firing is \$116 million while the average mandate size for hiring is \$102 million.

B. Round-Trip Performance

If more than one firm is fired (or hired), we compute the excess return for that round-trip observation as the average across the fired (or hired) firms. In Example 2 described above, pre- and post-firing returns for Denver International Advisors would be compared to the average of the pre- and post-hiring returns of Brazos and Royce. Both hired and fired firms are required to have returns over a particular evaluation horizon.

Panel A of Table X shows average pre- and post-event cumulative excess returns for fired and hired firms for the entire sample. Consistent with earlier results, the pre-firing returns for the overall sample fired firms are statistically indistinguishable from zero because they mix different termination reasons. Post-firing returns are positive, and interestingly, statistically significant at all three horizons. Also mirroring results from earlier tables, pre-hiring excess returns are large and positive. In general, this pattern of returns is reassuring because it suggests that our round-trip sample is similar to the earlier (larger) hiring and firing samples. In addition to hired and fired firm's returns, we also report return differences (hired firm's excess returns minus fired firm's excess returns) with corresponding standard errors. Prior to the firing/hiring decision, the return differences are large, positive, and statistically significant. The 3-year (1-year) cumulative excess return difference prior to the firing/hiring is 9.5% (4.6%) with a standard error of 2.5% (1.00%). After the hiring/firing decision, the performance of the fired firms exceeds that of the newly hired firms over all three horizons but with larger standard errors; the 3-year cumulative excess return difference is –1.03% but with a standard error of 1.1%.

We would like to understand the relation between the opportunity costs described above and plan sponsor attributes. Unfortunately, our cross-sectional analysis is hindered by small sample sizes; we cannot estimate cross-sectional regressions of the form reported in Table V. As a result, we report pre- and post-event return differentials for various categories of the data in Panel B of Table X.[14] The p-values for differences in returns between subcategories are also shown. Not surprisingly, pre-event return differences are significantly higher

[14] We only report results for subcategories with reasonable sample sizes. Also, we report return differentials, rather than separate firing and hiring returns to conserve space.

Table X

Round-Trip Excess Returns for Investment Managers

Returns are cumulated separately for hired and fired firms. In Panel A, we show the separate returns for hired and fired investment managers, as well as the return differential for the entire sample of round-trips. In Panel B, we show only the return differential for various subsamples. Heteroskedasticity and serial correlation consistent standard errors are calculated using the procedure described in Jegadeesh and Karceski (2004) and appear in parentheses. Low and high cutoffs for the allocation index are based on the bottom and top quartiles. Similarly, small and large cutoffs for sponsor size are based on the bottom and top quartiles.

	Pre-event Period			Post-event Period		
	−3 to 0	−2 to 0	−1 to 0	0 to 1	0 to 2	0 to 3
Panel A: Cumulative Excess Returns						
Fired firms	2.03	−1.57	−0.11	1.83	3.14	4.26
	(1.56)	(1.51)	(0.83)	(0.82)	(1.47)	(1.45)
Hired firms	11.55	7.55	4.46	1.34	2.26	3.23
	(3.11)	(1.60)	(1.52)	(0.42)	(0.56)	(0.41)
Return differential (hired–fired)	9.52	9.12	4.56	−0.48	−0.88	−1.03
	(2.47)	(2.30)	(1.00)	(0.78)	(1.33)	(1.14)
Number of round-trips	331	389	412	412	389	331
Panel B: Return Differentials (Hired–Fired Returns) for Subsamples						
Performance	13.13	12.36	6.13	−0.66	−0.56	−0.79
	(2.67)	(2.94)	(1.27)	(1.34)	(1.73)	(1.79)
Nonperformance	7.89	7.58	3.80	−0.40	−1.04	−1.14
	(2.81)	(2.35)	(0.96)	(0.60)	(1.14)	(0.88)
p-value for difference	0.06	0.10	0.02	0.81	0.56	0.73
Headline risk-sensitive	9.55	9.57	4.55	−0.26	−0.76	−0.68
	(2.33)	(2.48)	(0.70)	(0.66)	(1.41)	(1.29)
Headline risk-resistant	9.57	7.62	4.98	−1.46	−1.13	−2.18
	(2.51)	(2.51)	(2.72)	(1.45)	(1.85)	(2.29)
p-value for difference	0.99	0.58	0.87	0.30	0.73	0.38
Small plan sponsors	6.14	5.36	3.32	−0.54	−1.34	−1.39
	(1.91)	(1.62)	(1.02)	(1.14)	(1.37)	(1.33)
Large plan sponsors	13.21	11.68	4.80	−0.30	0.19	0.53
	(2.31)	(1.50)	(0.55)	(0.38)	(0.88)	(0.51)
p-value for difference	0.26	0.22	0.42	0.84	0.25	0.07
Low allocation index	11.59	10.73	4.39	−1.49	−1.79	−2.17
	(2.78)	(2.44)	(1.32)	(1.18)	(2.08)	(1.87)
High allocation index	10.61	9.87	5.33	0.06	−0.77	0.25
	(3.72)	(3.21)	(1.07)	(1.02)	(1.16)	(0.97)
p-value for difference	0.75	0.73	0.29	0.10	0.45	0.14
No consultant	3.24	2.08	2.00	−1.11	−1.04	−1.11
	(1.49)	(1.04)	(1.04)	(1.03)	(0.79)	(0.79)
Consultant	10.69	10.25	4.97	−0.39	−0.86	−1.02
	(2.27)	(2.19)	(1.01)	(0.92)	(1.67)	(1.53)
p-value for difference	0.10	0.05	0.03	0.57	0.85	0.92

for performance-based terminations than non-performance-based firings. Post-event return differentials are negative for both groups, but statistically indistinguishable from each other. Pre-event return differences are also larger for round-trips that use consultants but post-event return differentials are not statistically significant. In fact, for all the categories that we examine (headline risk, sponsor size, allocation index, and consultant use), the post-event return differentials across subcategories are not different from each other.

C. Discussion

How does one interpret the overall evidence from round-trips? The opportunity costs are positive but with high standard errors. If one adds transition costs discussed in the introduction (say, 1.0% to 2.0%) to these opportunity costs, the overall costs of firing and hiring investment managers rise further.[15] Moreover, if the costs associated with hiring and firing investment managers are important, then at the margin they should play a role in retention decisions. Typically, an investment management firm is hired for a given term, but then can be "rehired" for a subsequent term. If replacement costs are relevant, then the pre-rehiring performance that justifies retention should be lower than for brand new hiring. To determine if that is the case, we create a sample of retentions. We examine a random sample of 350 plan sponsors in Nelson's Directory of Plan Sponsors (2005). Nelson's reports the name of investment managers with mandates from each plan sponsor as of 2004, the year that investment manager was originally hired, and the investment mandate. We manually record this information for investment management firms that are in our returns database, where the mandate amount is recorded and where the original hiring year is before 2000. We then assume that a retention decision is made every 3 years. For example, if XYZ Asset Management was originally hired by ABC Plan Sponsor in 1996, we assume a retention decision is made in 1999 and 2002. In total, our sample consists of 1,867 retention decisions. We then compute pre-retention returns in the same manner as before and compare them to pre-hiring returns for the same plan sponsors. We find that the average 1-year (3-year) cumulative excess return for retentions is 2.4% (6.1%), compared with 4.9% (14.7%) for hiring decisions by the same plan sponsors. This suggests that in making retention decisions, plan sponsors incorporate the costs associated with hiring and firing.

VI. Conclusions

To summarize, we find that plan sponsors hire investment managers after superior performance but on average, post-hiring excess returns are zero. Plan sponsors fire investment managers for many reasons, including but not exclusively for underperformance. But, post-firing excess returns are frequently positive and sometimes statistically significant. Our sample of round-trips shows that if plan sponsors had stayed with fired investment managers, their excess

[15] Subtracting a constant from the mean return obviously does not change the standard errors and will "make" the excess returns statistically significant.

returns would be no different from those actually delivered by newly hired managers.

It could be the case that the costs documented and discussed above have compensating benefits that we are unable to measure. From an efficiency perspective, terminating investment managers could be critical to maintaining discipline among incumbents and maintaining a competitive marketplace. It is also possible that the agency relationships described by Lakonishok et al. (1992) create such high barriers to change so as to make it impossible to eliminate the costs. Some of our cross-sectional results are consistent with both of the above possibilities, especially since variation in the efficacy of hiring and firing appears to be related to the economic circumstances of plan sponsors. Although beyond the scope of this paper, there are several other analyses that could enhance our understanding of this form of delegated investment management. For instance, as pointed out by Hart (1992), it is useful to consider whether broad asset class allocations are efficient or reflect nonvalue maximizing behavior. Given the magnitude of assets under the jurisdiction of plan sponsors, correlated shifts in asset allocations could have important implications for asset pricing. We leave this to future research.

Appendix : Standard Error Calculation

The sample comprises N hiring/firing decisions of investment managers by plan sponsors ("events"). We wish to test whether the managers exhibit excess return performance from the event date through an H-quarter holding period. We define the H-quarter cumulative excess return (CER) for investment manager i that starts at the beginning of the event quarter t as the cumulative excess return:

$$CER_i(t, H) = \sum_{s=t}^{t+H-1} (R_{i,s} - R_{b,s}), \tag{A1}$$

where $R_{i,s}$ is the return on the mandate type by the investment manager i in quarter s, and $R_{b,s}$ is the return on the benchmark b in quarter s. We define:

$$\overline{CER}_{\text{sample}}(H) = \frac{1}{N} \sum_{i=1}^{N} CER_i(t, H). \tag{A2}$$

Let N_t equal the number of events in the sample in quarter t, and let N be the total number of events in the sample. Therefore $N = \sum_{t=1}^{T} N_t$. We define the average abnormal return for each event quarter t across all events in that quarter (we refer to this group of events as a quarterly cohort) as

$$\overline{CER}(t, H) = \begin{cases} \dfrac{1}{N_t} \displaystyle\sum_{i=1}^{N_t} CER_i(t, H), & \text{if } N_t > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{A3}$$

Let $\overline{CER}(H)$ be a $T \times 1$ column vector where the t^{th} element equals $\overline{CER}(t, H)$. $\overline{CER}(H)$ is the average long-run excess return of each quarterly cohort. Define w as a $T \times 1$ column vector of weights where the t^{th} element is the ratio of the number of events that occur in quarter t divided by N. Specifically, $w(t) = N_t/N$. Note that the sample average excess return is equal to the quarterly weight vector w times the average excess return of each quarterly cohort:

$$\overline{CER}_{\text{sample}}(H) = w'\overline{CER}(H). \tag{A4}$$

The variance of $\overline{CER}_{\text{sample}}(H)$ is given by

$$\sigma^2\left(\overline{CER}_{\text{sample}}(H)\right) = w'Vw, \tag{A5}$$

where V is the $T \times T$ variance covariance matrix of $\overline{CER}(H)$.

Our estimator for V allows for heteroskedasticity as well as serial correlation and is denoted as HSC. The st^{th} element of HSC is

$$hsc_{st} = \begin{cases} \dfrac{(H-l)}{l}\overline{CER}(s, H)\overline{CER}(t, H), & \text{if } l = |s-t| < H \\ 0 & \text{otherwise} \end{cases}. \tag{A6}$$

This estimator uses the Newey and West (1987) weighting scheme and ensures that HSC is positive definite.

Table A1

Investment Mandates and Indices

Investment Mandate	Description	Index
Domestic Equity		
Largecap	Large-cap equity	S&P 500
Largecapcore	Large-cap—between growth & value	S&P 500
Largecapgrowth	Large-cap—growth	S&P 500/BARRA Growth
Largecapvalue	Large-cap—value	S&P 500/BARRA Value
Midcap	Mid-cap equity	S&P Midcap 400
Midcapcore	Mid-cap—between growth and value	S&P Midcap 400
Midcapgrowth	Mid-cap—growth	S&P/BARRA Mid Cap Growth
Midcapvalue	Mid-cap—value	S&P/BARRA Mid Cap Value
Smallcap	Small-cap equity	S&P Small Cap 600
Smallcapcore	Small-cap—between growth and value	S&P Small Cap 600
Smallcapgrowth	Small-cap—growth	S&P/BARRA Small Cap Growth
Smallcapmicro	Small-cap—value	S&P Small Cap 600
Smallcapvalue	Small-cap equity	S&P/BARRA Small Cap Value

(continued)

Table A1—*Continued*

Investment Mandate	Description	Index
Smid	Small to mid-cap equity	Russell 2500
Smidcapcore	Small to mid-cap—between growth and value	Russell 2500
Smidcapgrowth	Small to mid-cap—growth	Russell 2500 Growth
Smidcapvalue	Small to mid-cap—indexed	Russell 2500 Value
Equitygrowth	All equity–growth	Russell 3000 Growth
Equityvalue	All equity—value	Russell 3000 Value
Equitycombined	All equity	Russell 3000
International equity		
Emergmkteq	Emerging market equity	MSCI Emerging Mkts Free
Europeincuk	Europe incl. U.K.	MSCI Europe 15
Europeincuksm	Europe incl. U.K.—small-cap	MSCI Europe S/C
Globaleq	Global equity (incl. U.S.)	MSCI World Free
Intleq	International equity	MSCI EAFE Free
Intleqsmall	International equity—small-cap	MSCI EAFE S/C
Pacbasinincj	Pacific basin incl. Japan	MSCI AC Pacific Free
Fixed income		
Convertibles	Convertibles	Merrill Lynch Inv Grade Convertible
Fixed1–3yrs	Duration between 1 and 3 years	Merrill Lynch Govt/Corp 1–3 Years
Fixedcore	Inv. and non-inv. grade, duration 3–7 years	Lehman Aggregate
Fixedcoreinvest	Inv. grade, duration 3–7 years	Lehman Aggregate
Fixedcoreopportun	Non-inv. grade, duration 3–7 years	Lehman Aggregate
Fixedhighyield	High yield securities	Lehman High Yield Composite
Shortterm	Duration between 1 and 2.4 years	Citigroup 3-Month T-Bill
Fixedintermed	Duration between 2 and 4.6 years	Lehman Int. Aggregate
Fixedlongdura	Duration greater than 6 years	Lehman Long Govt/Credit
Mortgageb	Mortgage-backed securities	Lehman Mortgages
Fixedcombined	All fixed income	Lehman Aggregate
Emergmktdebt	Emerging market debt	JP Morgan ELMI+
Globalfixhedg	Global fixed income—hedged	Lehman Global Aggregate (Hedged)
Globalfixunhedg	Global fixed income—unhedged	Lehman Global Aggregate (Unhedged)
Intlfixedhedg	International fixed income—hedged	Citigroup Non-US WGBI (Hedged)
Intlfixedunhedg	International fixed income—unhedged	Citigroup Non-US WGBI (Unhedged)
Others		
Realestate	Real estate	NCREIF Property
Realestateselect	Real estate select	NCREIF Property
REITs	REITs	NAREIT
TAA	Tactical asset allocation	Average of S&P 500 and Lehman Aggregate
Balanced	Balanced	Average of S&P 500 and Lehman Aggregate

REFERENCES

Allen, Franklin, 2001, Do financial institutions matter? *Journal of Finance* 56, 1165–1176.

Barber, Brad, and Terrance Odean, 2000, Trading is hazardous to your wealth: The common stock investment performance of individual investors, *Journal of Finance*, 773–806.

Barber, Brad, Terrance Odean, and Lu Zheng, 2005, Out of sight, out of mind: The effects of expenses on mutual fund flows, *Journal of Business* 78, 2095–2120.

Barberis, Nicholas, and Andrei Shleifer, 2003, Style investing, *Journal of Financial Economics* 68, 161–199.

Bollen, Brian, 2004, Lost in transition? *Financial News*, April 19.

Bollen, Nicholas, and Jeffrey Busse, 2005, Short-term persistence in mutual fund performance, *Review of Financial Studies* 18, 569–597.

Brickley, James, Ronald Lease, and Clifford Smith, 1988, Ownership structure and voting on antitakeover amendments, *Journal of Financial Economics* 20, 267–291.

Brown, Keith, W.V. Harlow, and Laura Starks, 1996, Of tournaments and temptations: An analysis of managerial incentives in the mutual fund industry, *Journal of Finance* 51, 85–110.

Brown, Stephen, and William Goetzmann, 1995, Performance persistence, *Journal of Finance*, 50, 679–698.

Brull, Steven, 2006, Rich plan, poor plan, *Institutional Investor* 40, 30–35.

Busse, Jeffrey, 2001, Another look at mutual fund tournaments, *Journal of Financial and Quantitative Analysis* 36, 53–73.

Busse, Jeffrey, Amit Goyal, and Sunil Wahal, 2007, Performance persistence in institutional investment management, Working paper, Arizona State University.

Carhart, Mark, 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57–82.

Carhart, Mark M., Jennifer N. Carpenter, Anthony W. Lynch, and David K. Musto, 2002, Mutual fund survivorship, *Review of Financial Studies* 5, 1439–1463.

Chevalier, Judith, and Glen Ellison, 1999, Career concerns of mutual fund managers, *Quarterly Journal of Economics* 114, 389–432.

Christopherson, Jon A., Wayne Ferson, and Debra Glassman, 1998, Conditioning manager alphas on economic information: Another look at the persistence of performance, *Review of Financial Studies* 11, 111–142.

Coggin, T. Daniel, Frank Fabozzi, and Shafiqur Rahman, 1993, The investment performance of U.S. equity pension fund managers: An empirical investigation, *Journal of Finance* 48, 1039–1055.

Cornell, Bradford, and Richard Roll, 2005, A delegated agent asset-pricing model, *Financial Analysts Journal* 61, 57–69.

Daniel, Kent, Mark Grinblatt, Sheridan Titman, and Russ Wermers, 1997, Measuring mutual fund performance with characteristic-based benchmarks, *Journal of Finance* 52, 1035–1058.

Del Guercio, Diane, and Paula Tkac, 2002, The determinants of the flow of funds of managed portfolios: Mutual funds versus pension funds, *Journal of Financial and Quantitative Analysis* 37, 523–557.

Elton, Edwin, Martin Gruber, and Christopher Blake, 1996, The persistence of risk-adjusted mutual fund performance, *Journal of Business* 69, 133–157.

Elton, Edwin, Martin Gruber and Christopher Blake, 2006, Participant reaction and the performance of funds offered by 401(k) plans, Working paper, New York University.

Elton, Edwin, Martin Gruber, Sanjiv Das, and Mathew Hlavka, 1992, Efficiency with costly information: A reinterpretation of the evidence for managed portfolios, *Review of Financial Studies* 6, 1–22.

Fabozzi, Frank, 1997, *Pension Fund Investment Management* (Frank J. Fabozzi Associates, New Hope, PA).

Fama, Eugene, and Kenneth French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Franzoni, Francesco, and Jose M. Marín, 2006, Pension plan funding and stock market efficiency, *Journal of Finance* 61, 921–956.

Gallo, J.G., and Larry Lockwood, 1997, Benefits of proper style classification of equity portfolio managers, *Journal of Portfolio Management* 23, 47–55.

Goetzmann, William, and Roger Ibbotson, 1994, Do winners repeat? Patterns in mutual fund performance, *Journal of Portfolio Management* 20, 9–18.

Grinblatt, Mark, and Sheridan Titman, 1992, The persistence of mutual fund performance, *Journal of Finance* 47, 1977–1984.

Grinblatt, Mark, Sheridan Titman, and Russ Wermers, 1995, Momentum investing strategies, portfolio performance and herding: A study of mutual fund behavior, *American Economic Review* 85, 1088–1105.

Gruber, Martin, 1996, Another puzzle: The growth in actively managed mutual funds, *Journal of Finance* 51, 783–810.

Hart, Oliver, 1992, Comments on "the structure and performance of the money management industry," *Brookings Papers: Microeconomics* 380–384.

Heisler, Jeffrey, Christopher R. Knittel, John J. Neumann, and Scott Stewart, 2007, Why do institutional plan sponsors fire their investment managers? *Journal of Business and Economic Studies* 13, 88–115.

Hendricks, Darryll, Jayendu Patel, Richard Zeckhauser, 1993, Hot hands in mutual funds: Short-run persistence of performance, 1974–1988, *Journal of Finance* 48, 93–130.

Investment Company Institute, 2004, Trends in mutual fund investing, www.ici.org/stats/index.html.

Ippolito, Richard, 1989, Efficiency with costly information: A study of mutual fund performance, 1965–1984, *Quarterly Journal of Economics* 104, 1–23.

Jegadeesh, Narasimhan, and Jason Karceski, 2004, Long-run performance evaluation: Correlation and heteroskedasticity-consistent tests, Working paper, Emory University.

Jensen, Michael, 1968, The performance of mutual funds in the period 1945–1964, *Journal of Finance* 48, 389–416.

Lakonishok, Josef, Andrei Shleifer, and Robert Vishny, 1992, The structure and performance of the money management industry, *Brookings Papers: Microeconomics*, 339–379.

Logue, Dennis E., and Jack S. Rader, 1998, *Managing Pension Plans: A Comprehensive Guide to Improving Plan Performance* (Harvard Business School Press, Boston, MA).

Madalla, G.S., 1983, *Limited Dependent Variables and Qualitative Variables in Econometrics*, Econometric Society Monographs, Number 3. (Cambridge University Press, Cambridge).

Money Market Directory of Investment Managers and Plan Sponsors, 2004 (Charlottesville, VA).

Nelson's Directory of Investment Managers, 2002 and 2004 (Nelson Publications, Port Chester, NY).

Nelson's Directory of Plan Sponsors, 2005 (Nelson Publications, Port Chester, NY).

Newey, Whitney, and Kenneth West, 1987, A simple, positive definite, heteroskedasticity, and autocorrelation consistent covariance matrix, *Econometrica* 55, 703–708.

Odean, Terrance, 1998, Are investors reluctant to realize their losses? *Journal of Finance* 53, 1775–1798.

Odean, Terrance, 1999, Do investors trade too much? *American Economic Review* 89, 1279–1298.

Proszek, Stan, 2002, Transition management: Simple—but not easy, *Benefits and Pensions Monitor*, October 12.

Sirri, Erik, and Peter Tufano, 1998, Costly search and mutual fund flows, *Journal of Finance* 53, 1589–1622.

Travers, Frank J., 2004, *Investment Manager Analysis: A Comprehensive Guide to Portfolio Selection, Monitoring and Optimization*. (John Wiley & Sons, New York).

Wermers, Russ, 2000, Mutual fund performance: An empirical decomposition into stock-picking talent, style, transactions costs, and expenses, *Journal of Finance* 55, 1655–1695.

Werner, Bob, 2001, The true costs and benefits of portfolio transition management, www.russell.com/AU/press_room/Press_Releases/PR20011004_AU_p.asp.

Zheng, Lu, 1999, Is money smart? A study of mutual fund investors' fund selection ability, *Journal of Finance* 54, 901–933.

First draft: October 2007
This draft: December 2009
Not for quotation: Comments welcome

Luck versus Skill in the Cross Section of Mutual Fund Returns

Eugene F. Fama and Kenneth R. French*

Forthcoming in the Journal of Finance

Abstract

The aggregate portfolio of actively managed U.S. equity mutual funds is close to the market portfolio, but the high costs of active management show up intact as lower returns to investors. Bootstrap simulations suggest that few funds produce benchmark adjusted expected returns sufficient to cover their costs. If we add back the costs in fund expense ratios, there is evidence of inferior and superior performance (non-zero true α) in the extreme tails of the cross section of mutual fund α estimates.

* Booth School of Business, University of Chicago (Fama) and Amos Tuck School of Business, Dartmouth College (French). We are grateful for the comments of Juhani Linnainmaa, Sunil Wahal, Jerry Zimmerman, and seminar participants at the University of Chicago, the California Institute of Technology, UCLA, and the Meckling Symposium at the University of Rochester. Special thanks to John Cochrane, and the journal editor, associate editor, and referees.

There is a constraint on the returns to active investing that we call equilibrium accounting. In short (details later), suppose that when returns are measured before costs (fees and other expenses), passive investors get passive returns; that is, they have zero α (abnormal expected return) relative to passive benchmarks. This means active investment must also be a zero sum game – aggregate α is zero before costs. If some active investors have positive α before costs, it is dollar for dollar at the expense of other active investors. After costs, that is, in terms of net returns to investors, active investment must be a negative sum game. (Sharpe (1991) calls this the arithmetic of active management.)

We examine mutual fund performance from the perspective of equilibrium accounting. For example, at the aggregate level, if the value-weight (VW) portfolio of active funds has a positive α before costs, we can infer that the VW portfolio of active investments outside mutual funds has a negative α. In other words, active mutual funds win at the expense of active investments outside mutual funds. In fact, we find that the VW portfolio of active funds that invest primarily in U.S. equities is close to the market portfolio, and estimated before expenses, its α relative to common benchmarks is close to zero. Since the VW portfolio of active funds produces α close to zero in gross (pre-expense) returns, α estimated on the net (post-expense) returns realized by investors is negative by about the amount of fund expenses.

The aggregate results imply that if there are active mutual funds with positive true α, they are balanced by active funds with negative α. We test for the existence of such funds. The challenge is to distinguish skill from luck. Given the multitude of funds, many have extreme returns by chance. A common approach to this problem is to test for persistence in fund returns, that is, whether past winners continue to produce high returns and losers continue to underperform (for example, Grinblatt and Titman (1992), Carhart (1997)). Persistence tests have an important weakness. They rank funds on short-term past performance, so there may be little evidence of persistence because the allocation of funds to winner and loser portfolios is largely based on noise.

We take a different tack. We use long histories of individual fund returns and bootstrap simulations of return histories to infer the existence of superior and inferior funds. We compare the actual cross-section of fund α estimates to the results from 10,000 bootstrap simulations of the cross-section. The returns of the funds in a simulation run have the properties of actual fund returns, except we set true α to zero in the return population from which simulation samples are drawn. The simulations thus describe the distribution of α

estimates when there is no abnormal performance in fund returns. Comparing the distribution of α estimates from the simulations to the cross-section of α estimates for actual fund returns allows us to draw inferences about the existence of skilled managers.

For fund investors the simulation results are disheartening. When α is estimated on net returns to investors, the cross-section of precision-adjusted α estimates, $t(\alpha)$, suggests that few active funds produce benchmark adjusted expected returns that cover their costs. Thus, if many managers have sufficient skill to cover costs, they are hidden by the mass of managers with insufficient skill. On a practical level, our results on long-term performance say that true α in net returns to investors is negative for most if not all active funds, including funds with strongly positive α estimates for their entire histories.

Mutual funds look better when returns are measured gross, that is, before the costs included in expense ratios. Comparing the cross-section of $t(\alpha)$ estimates from gross fund returns to the average cross-section from the simulations suggests that there are inferior managers whose actions reduce expected returns and there are superior managers who enhance expected returns. If we assume that the cross section of true α has a normal distribution with mean zero and standard deviation σ, then σ around 1.25% per year seems to capture the tails of the cross section of α estimates for our full sample of actively managed funds.

The estimate of the standard deviation of true α, 1.25% per year, does not imply much skill. It suggests, for example, that fewer than 16% of funds have α greater than 1.25% per year (about 0.10% per month), and only about 2.3% have α greater than 2.50% per year (about 0.21% per month) – before expenses.

The simulation tests have power. If the cross section of true α for gross fund returns is normal with mean zero, the simulations strongly suggest that the standard deviation of true α is between 0.75% and 1.75% per year. Thus, the simulations rule out values of σ rather close to our estimate, 1.25%. The power traces to the fact that a large cross section of funds produces precise estimates of the percentiles of $t(\alpha)$ under different assumptions about σ, the standard deviation of true α. This precision allows us to put σ in a rather narrow range.

Readers suggest that our results are consistent with the predictions of Berk and Green (2004). We outline their model in Section II, after the tests on mutual fund aggregates (Section I), and before the bootstrap simulations (Sections III and IV). Our results reject most of their predictions about mutual fund returns.

Given the prominence of their model, our contrary evidence seems an important contribution. The paper closest to ours is Kosowski et al. (2006). They do bootstrap simulations that seem to produce stronger evidence of manager skill. We contrast their tests and ours in Section V, after presenting our results. Section VI concludes.

I. The Performance of EW and VW Portfolios of U.S. Equity Mutual Funds

Our mutual fund sample is from the CRSP (Center for Research in Security Prices) database. We include only funds that invest primarily in U.S. common stocks, and we combine, with value weights, different classes of the same fund into a single fund. (See French (2008).) To focus better on the performance of active managers, we exclude index funds from all our tests. The CRSP data start in 1962, but we concentrate on the period after 1983. During the period 1962 to 1983 about 15% of the funds on CRSP report only annual returns, and the average annual EW return for these funds is 5.29% lower than for funds that report monthly returns. As a result, the EW average return on all funds is a nontrivial 0.65% per year lower than the EW return of funds that report monthly returns. Thus, during 1962 to 1983 there is selection bias in tests like ours that use only funds that report monthly returns. After 1983 almost all funds report monthly returns. (Elton, Gruber, and Blake (2001) discuss CRSP data problems for the period before 1984.)

A. The Regression Framework

Our main benchmark for evaluating fund performance is the three-factor model of Fama and French (1993), but we also show results for Carhart's (1997) four-factor model. To measure performance, these models use two variants of the time-series regression,

$$R_{it} - R_{ft} = a_i + b_i(R_{Mt} - R_{ft}) + s_i SMB_t + h_i HML_t + m_i MOM_t + e_{it}. \qquad (1)$$

In this regression, R_{it} is the return on fund i for month t, R_{ft} is the riskfree rate (the one-month U.S. Treasury bill rate), R_{Mt} is the market return (the return on a value-weight portfolio of NYSE, Amex, and NASDAQ stocks), SMB_t and HML_t are the size and value-growth returns of Fama and French (1993), MOM_t is our version of Carhart's (1997) momentum return, a_i is the average return left unexplained by the benchmark model (the estimate of α_i), and e_{it} is the regression residual. The full version of (1) is Carhart's four-factor model, and the regression without MOM_t is the Fama-French three-factor model. The construction of SMB_t

and HML_t follows Fama and French (1993). The momentum return, MOM_t, is defined like HML_t, except that we sort on prior return rather than the book-to-market equity ratio. (See Table I.)

Regression (1) allows a more precise statement of the constraints of equilibrium accounting. The VW aggregate of the U.S. equity portfolios of all investors is the market portfolio. It has a market slope equal to 1.0 in (1), zero slopes on the other explanatory returns, and a zero intercept – before investment costs. This means that if the VW aggregate portfolio of passive investors also has a zero intercept before costs, the VW aggregate portfolio of active investors must have a zero intercept. Thus, positive and negative intercepts among active investors must balance out – before costs.

There is controversy about whether the average SMB_t, HML_t, and MOM_t returns are rewards for risk or the result of mispricing. For our purposes, there is no need to take a stance on this issue. We can simply interpret SMB_t, HML_t, and MOM_t as diversified passive benchmark returns that capture patterns in average returns during our sample period, whatever the source of the average returns. Abstracting from the variation in returns associated with $R_{Mt} - R_{ft}$, SMB_t, HML_t, and MOM_t then allows us to focus better on the effects of active management (stock picking), which should show up in the three-factor and four-factor intercepts.

From an investment perspective, the slopes on the explanatory returns in (1) describe a diversified portfolio of passive benchmarks (including the riskfree security) that replicates the exposures of the fund on the left to common factors in returns. The regression intercept then measures the average return provided by a fund in excess of the return on a comparable passive portfolio. We interpret a positive expected intercept (true α) as good performance, and a negative expected intercept signals bad performance.[1]

Table I shows summary statistics for the explanatory returns in (1) for January 1984 through September 2006 (henceforth 1984 to 2006), the period used in our tests. The momentum factor (MOM_t) has the highest average return, 0.79% per month (t = 3.01), but the average values of the monthly market premium (R_{Mt} - R_{ft}) and the value-growth return (HML_t) are also large, 0.64% (t = 2.42) and 0.40% (t = 2.10), respectively. The size return, SMB_t, has the smallest average value, 0.03% per month (t = 0.13).

Table I about here.

B. Regression Results for EW and VW Portfolios of Active Funds

Table II shows estimates of regression (1) for the monthly returns of 1984 to 2006 on equal-weight (EW) and value-weight (VW) portfolios of the funds in our sample. In the VW portfolio, funds are weighted by assets under management (AUM) at the beginning of each month. The EW portfolio weights funds equally each month. The intercepts in (1) for EW fund returns tell us whether funds on average produce returns different from those implied by their exposures to common factors in returns, whereas VW returns tell us about the fate of aggregate wealth invested in funds. Table II shows estimates of (1) for fund returns measured gross and net of fund expenses. Net returns are those received by investors. Monthly gross returns are net returns plus 1/12th of a fund's expense ratio for the year.

Table II about here.

The market slopes in Table II are close to 1.0, which is not surprising since our sample is funds that invest primarily in U.S. stocks. The HML_t and MOM_t slopes are close to zero. Thus, in aggregate active funds show little exposure to the value/growth and momentum factors. The EW portfolio of funds produces a larger SMB_t slope (0.18) than the VW portfolio (0.07). We infer that smaller funds are more likely to invest in small stocks, but total dollars invested in active funds (captured by VW returns) show little tilt toward small stocks.

The intercepts in the estimates of (1) summarize the average performance of funds (EW returns) and the performance of aggregate wealth invested in funds (VW returns) relative to passive benchmarks. In terms of net returns to investors, performance is poor. The three-factor and four-factor (annualized) intercepts for EW and VW net returns are negative, ranging from -0.81% to -1.00% per year, with *t*-statistics from -2.05 to -3.02. These results are in line with previous work (e.g., Jensen (1968), Malkiel (1995), Gruber (1996)).

The intercepts in (1) for EW and VW net fund returns tell us whether on average active managers have sufficient skill to generate returns that cover the costs funds impose on investors. Gross returns come closer to testing whether managers have any skill. For EW gross fund returns, the three-factor and four-factor intercepts for 1984 to 2006 are positive, 0.36% and 0.39% per year, but they are only 0.85 and 0.90 standard errors from zero. The intercepts in (1) for VW gross returns are quite close to zero, 0.13% per year (t = 0.40) for the three-factor version of (1) and -0.05% per year (t = -0.15) for the four-factor model.

Table II also shows estimates of the CAPM version of (1), in which $R_{Mt} - R_{ft}$ is the only explanatory return. The annualized CAPM intercept for VW gross fund returns for 1984 to 2006, -0.18% per year (t = -0.49) is again close to zero and similar to the estimates for the three-factor and four-factor models. It is not surprising that the intercepts for the three models are so similar (-0.18%, 0.13%, and -0.05% per year) since VW fund returns produce slopes close to zero for the non-market explanatory returns in (1).

We can offer an equilibrium accounting perspective on the results in Table II. When we add back the costs in expense ratios, α estimates for VW gross fund returns are close to zero. Thus, before expenses, there is no evidence that total wealth invested in active funds gets any benefits or suffers any losses from active management. VW fund returns also show little exposure to the size, value, and momentum returns, and the market return alone explains 99% of the variance of the monthly VW fund return. Together these facts say that during 1984 to 2006, active mutual funds in aggregate hold a portfolio that, before expenses, mimics market portfolio returns. The return to investors is, however, reduced by the high expense ratios of active funds. These results echo equilibrium accounting, but for a subset of investment managers where the implications of equilibrium accounting for aggregate investor returns need not hold.

C. Measurement Issues in the Tests on Gross Returns

The benchmark explanatory returns in (1) are before all costs. This is appropriate in tests on net fund returns where the issue addressed is whether managers have sufficient skill to produce expected returns that cover their costs. Gross returns pose more difficult measurement issues.

The issue in the tests on gross fund returns is whether managers have skill that causes expected returns to differ from those of comparable passive benchmarks. For this purpose, one would like fund returns measured before all costs and non-return revenues. This would put funds on the same pure return basis as the benchmark explanatory returns, so the regressions could focus on manager skill. Our gross fund returns are before the costs in expense ratios (including management fees), but they are net of other costs, primarily trading costs, and they include the typically small revenues from securities lending.

We could attempt to add trading costs to our estimates of gross fund returns. Funds do not report trading costs, however, and estimates are subject to large error. For example, trading costs are likely to vary across funds because of differences in style tilts, trading skill, and the extent to which a fund demands immediacy in trade execution. Trading costs also vary through time. Our view is that estimates of trading costs for individual funds, especially actively managed funds, are fraught with error and potential bias, and are likely to be misleading. We prefer to stay with our simple definition of gross returns (net returns plus the costs in expense ratios), with periodic qualifications to our inferences.

An alternative approach (suggested by a referee) is to put the passive benchmarks produced by combining the explanatory returns in (1) in the same units as the gross fund returns on the left of (1). This involves taking account of the costs not covered in expense ratios that would be borne by an efficiently managed passive benchmark with the same style tilts as the fund whose gross returns are to be explained. An appendix discusses this approach in detail. The bottom line is that for efficiently managed passive funds, the costs missed in expense ratios are close to zero. Thus, adjusting the benchmarks produced by (1) for estimates of these costs is unnecessary.

This does not mean our tests on gross fund returns capture the pure effects of skill. Though it appears that all substantial costs incurred by efficiently managed passive funds are in their expense ratios, this is less likely to be true for actively managed funds. The typical active fund trades more than the typical passive fund, and active funds are likely to demand immediacy in trading that pushes up costs. Our tests on gross returns thus produce α estimates that capture skill, less whatever are the net costs (costs minus non-return revenues) missed by expense ratios. Equivalently, the tests say that a fund's management has skill only if it is sufficient to cover the missing costs (primarily trading costs). This seems like a reasonable definition of skill since an efficiently managed passive fund can apparently avoid these costs. More important, this is the definition of skill we can accurately test, given the unavoidable absence of accurate trading cost estimates for active funds.

The fact that our gross fund returns are net of the costs missed in expense ratios does, however, affect the inferences about equilibrium accounting we can draw from the aggregate results in Table II. Since the α estimates for VW gross fund returns in Table II are close to zero, they suggest that funds in aggregate show sufficient skill to produce expected returns that cover some or all the costs missed in expense ratios. If this is the correct inference (precision is an issue), equilibrium accounting then says that the costs recovered by funds are matched by equivalent losses on investments outside mutual funds.

II. Berk and Green (2004)

Readers contend that our results (Table II and below) are consistent with Berk and Green (2004). Their model is attractive theory, but our results reject most of its predictions about mutual fund returns.

In their world, a fund is endowed with a permanent α, before costs, but it faces costs that are an increasing convex function of assets under management (AUM). Investors use returns to update estimates of α. A fund with a positive expected α before costs attracts inflows until AUM reaches the point where expected α, net of costs, is zero. Outflows drive out funds with negative expected α. In equilibrium, all active funds (and thus funds in aggregate) have positive expected α before costs and zero expected α net of costs.

Our evidence that the aggregate portfolio of mutual funds has negative α net of costs contradicts the predictions of Berk and Green (2004). The results below on the net returns of individual funds also reject their prediction that all active managers have zero α net of costs. In fact, our results say that for most if not all funds, true α in net returns is negative.

Finally, equilibrium accounting poses a theoretical problem for Berk and Green (2004). Their model focuses on rational investors who optimally choose among passive and active alternatives. In aggregate, their investors have positive α before costs and zero α after costs. Equilibrium accounting, however, says that in aggregate investors have zero α before costs and negative α after costs.

III. Bootstrap Simulations

Table II says that on average active mutual funds do not produce gross returns above (or below) those of passive benchmarks. This may just mean that managers with skill that allows them to outperform the benchmarks are balanced by inferior managers who underperform. We turn now to simulations that use individual fund returns to infer the existence of superior and inferior managers.

A. Setup

To lessen the effects of "incubation bias" (see below), we limit the tests to funds that reach five million 2006 dollars in assets under management (AUM). Since the AUM minimum is in 2006 dollars, we include a fund in 1984, for example, if it has more than about \$2.5 million in AUM in 1984. Once a fund passes the AUM minimum, it is included in all subsequent tests, so this requirement does not create selection bias. We also show results for funds after they pass \$250 million and \$1 billion. Since we estimate benchmark regressions for each fund, we limit the tests to funds that have at least eight months of returns after they pass an AUM bound, so there is a bit of survival bias. To avoid having lots of new funds with short return histories, we only use funds that appear on CRSP at least five years before the end of our sample period.

Fund management companies commonly provide seed money to new funds to develop a return history. Incubation bias arises because funds typically open to the public – and their pre-release returns are included in mutual fund databases – only if the returns turn out to be attractive. The \$5 million AUM bound for admission to the tests alleviates this bias since AUM is likely to be low during the pre-release period.

Evans (2007) suggests that incubation bias can be minimized by using returns only after funds receive a ticker symbol from NASDAQ, which typically means they are available to the public. Systematic data on ticker symbol start dates are available only after 1998. We have replicated our tests for 1999 to 2006 using CRSP start dates for new funds (as in our reported results) and then using NASDAQ ticker dates (from Evans). Switching to ticker dates has almost no effect on aggregate fund returns (as in Table II), and trivial effects on the cross-section of $t(\alpha)$ estimates for funds (as in Table III below). We conclude that incubation bias is probably unimportant in our results for 1984 to 2006.

Our goal is to draw inferences about the cross section of true α for active funds, specifically, whether the cross section of α estimates suggests a world where true α is zero for all funds or whether there is non-zero true α, especially in the tails of the cross section of α estimates. We are interested in answering this question for 12 different cross sections of α estimates – gross and net returns, for the three-factor and four-factor benchmarks, and the three AUM samples. Thus, we use regression (1) to estimate each fund's three-factor or four-factor α for gross or net returns for the part of 1984 to 2006 after the fund passes each AUM bound.

The tests for non-zero true α in actual fund returns use bootstrap simulations on returns that have the properties of fund returns, except that true α is set to zero for every fund. To set α to zero, we subtract a fund's α estimate from its monthly returns. For example, to compute three-factor benchmark-adjusted gross returns for a fund in the $5 million group, we subtract its three-factor α estimated from monthly gross returns for the part of 1984 to 2006 that the fund is in the $5 million group from the fund's monthly gross returns for that period. We calculate benchmark-adjusted returns for the three-factor and four-factor models, for gross and net returns, and for the three AUM bounds. The result is 12 populations of benchmark-adjusted (zero-α) returns. (CAPM simulation results are in the Appendix.)

A simulation run is a random sample, with replacement, of the calendar months of 1984 to 2006. Each simulation run has 273 months, like January 1984 to September 2006. For each of the 12 sets of benchmark-adjusted returns, we estimate, fund by fund, the relevant benchmark model on the simulation draw of months of adjusted returns, dropping funds that are in the simulation run for less than eight months. Each run thus produces 12 cross-sections of α estimates using the same random sample of months from 12 populations of adjusted (zero-α) fund returns.

We do 10,000 simulation runs to produce 12 distributions of t-statistics, $t(\alpha)$, for a world in which true α is zero. We focus on $t(\alpha)$, rather than estimates of α, to control for differences in precision due to differences in residual variance and in the number of months funds are in a simulation run.

Note that setting true α equal to zero builds different assumptions about skill into the tests on gross and net fund returns. For net returns, setting true α to zero is a world where every manager has sufficient skill to generate expected returns that cover all costs. In contrast, setting true α to zero in gross returns is a world where every fund manager has just enough skill to produce expected returns that cover the costs missed in expense ratios.

Our simulation approach has an important advantage. Because a simulation run is the same random sample of months for all funds, the simulations capture the cross-correlation of fund returns and its effects on the distribution of $t(\alpha)$ estimates. Since we jointly sample fund and explanatory returns, we also capture any correlated heteroscedasticity of the explanatory returns and disturbances of a benchmark model. We shall see that these details of our approach are important for inferences about true α in actual fund returns.

Defining a simulation run as the same random sample of months for all funds also has a cost. If a fund is not in the tests for the entire 1984 to 2006 period, it is likely to show up in a simulation run for more or less than the number of months it is in our sample. This is not serious. We focus on $t(\alpha)$, and the distribution of $t(\alpha)$ estimates depends on the number of months funds are in a simulation run through a degrees of freedom effect. The distributions of $t(\alpha)$ estimates for funds that are oversampled in a simulation run have more degrees of freedom (and thinner extreme tails) than the distributions of $t(\alpha)$ for the actual returns of the funds. Within a simulation run, however, oversampling of some funds should roughly offset under-sampling of others, so a simulation run should produce a representative sample of $t(\alpha)$ estimates for simulated returns that have the properties of actual fund returns, except that true α is zero for every fund. Oversampling and under-sampling of fund returns in a simulation run should also about balance out in the 10,000 runs used in our inferences.

A qualification of this conclusion is in order. In a simulation run, as in the tests on actual returns, we discard funds that have less than eight months of returns. This means we end up with a bit more oversampling of fund returns. As a result, the distributions of $t(\alpha)$ estimates in the simulations tend to have more degrees of freedom (and thinner tails) than the estimates for actual fund returns. This means our tests are a bit biased toward finding false evidence of performance in the tails of $t(\alpha)$ estimates for actual fund returns.

There are two additional caveats. (i) Random sampling of months in a simulation run preserves the cross-correlation of fund returns, but we lose any effects of autocorrelation. The literature on autocorrelation of stock returns (for example, Fama (1965)) suggests that this is a minor problem. (ii) Because we randomly sample months, we also lose any effects of variation through time in the regression slopes in (1). (The issues posed by time-varying slopes are discussed by Ferson and Schadt (1996).) Capturing time variation in the regression slopes poses thorny problems, and we leave this potentially important issue for future research.

To develop perspective on the simulations, we first compare, in qualitative terms, the percentiles of the cross-section of $t(\alpha)$ estimates from actual fund returns and the average values of the percentiles from the simulations. We then turn to likelihood statements about whether the cross-section of $t(\alpha)$ estimates for actual fund returns points to the existence of skill.

B. First Impressions

When we estimate a benchmark model on the returns of each fund in an AUM group, we get a cross-section of $t(\alpha)$ estimates that can be ordered into a cumulative distribution function (CDF) of $t(\alpha)$ estimates for actual fund returns. A simulation run for the same combination of benchmark model and AUM group also produces a cross-section of $t(\alpha)$ estimates and its CDF for a world in which true α is zero. In our initial examination of the simulations we compare (i) the values of $t(\alpha)$ at selected percentiles of the CDF of the $t(\alpha)$ estimates from actual fund returns and (ii) the averages across the 10,000 simulation runs of the $t(\alpha)$ estimates at the same percentiles. For example, the first percentile of three-factor $t(\alpha)$ estimates for the net returns of funds in the \$5 million AUM group is -3.87, versus an average first percentile of -2.50 from the 10,000 three-factor simulation runs for the net returns of funds in this group (Table III).

Table III about here.

For each combination of gross or net returns, AUM group, and benchmark model, Table III shows the CDF of $t(\alpha)$ estimates for actual returns and the average of the 10,000 simulation CDFs. The average simulation CDFs are similar for gross and net returns and for the two benchmark models. This is not surprising since true α is always zero in the simulations. The dispersion of the average simulation CDFs decreases from lower to higher AUM groups. This is at least in part a degrees of freedom effect; on average funds in lower AUM groups have shorter sample periods.

B.1. Net Returns

The Berk and Green (2004) prediction that most fund managers have sufficient skill to cover their costs fares poorly in Table III. The left tail percentiles of the $t(\alpha)$ estimates from actual net fund returns are far below the corresponding average values from the simulations. For example, the 10th percentiles of the actual $t(\alpha)$ estimates, -2.34, -2.37, and -2.53 for the \$5 million, \$250 million, and \$1 billion groups, are much more extreme than the average estimates from the simulation, -1.32, -1.31, and -1.30. The right tails of the $t(\alpha)$ estimates also do not suggest widespread skill sufficient to cover costs. In the tests that use the three-factor model, the $t(\alpha)$ estimates from the actual net returns of funds in the \$5 million group are below the average values from the simulations for all percentiles below the 98th. For the \$1 billion group, only the 99th percentile of three-factor $t(\alpha)$ for actual net fund returns is above the average simulation 99th percentile, and then only slightly. For the \$250 million group, the percentiles of three-factor $t(\alpha)$ for actual net fund returns are all below the averages from the simulations. Figure 1 is a picture of the actual and average simulated CDFs for the \$5 million AUM group.

Figure 1 about here.

The evidence of skill sufficient to cover costs is even weaker with an adjustment for momentum exposure. In the tests that use the four-factor model, the percentiles of the $t(\alpha)$ estimates for actual net fund returns are always below the average values from the simulations. In other words, the averages of the percentile values of four-factor $t(\alpha)$ from the simulations of net returns (where by construction skill suffices to cover costs) always beat the corresponding percentiles of $t(\alpha)$ for actual net fund returns.

There is a glimmer of hope for investors in the tests on net returns. Even in the four-factor tests, the 99^{th} and, for the \$5 million group, the 98^{th} percentiles of the $t(\alpha)$ estimates for actual fund returns are close to the average values from the simulations. This suggests that some fund managers have enough skill to produce expected benchmark adjusted net returns that cover costs. This is, however, a far cry from the prediction of Berk and Green (2004) that most if not all fund managers can cover their costs.

B.2. Gross Returns

It is possible that the fruits of skill do not show up more generally in net fund returns because they are absorbed by expenses. The tests on gross returns in Table III show that adding back the costs in expense ratios pushes up $t(\alpha)$ for actual fund returns. For all AUM groups, however, the left tail of three-factor $t(\alpha)$ estimates for actual gross fund returns is still to the left of the average from the simulations. For example, in the simulations the average value of the fifth percentile of $t(\alpha)$ for gross returns for the \$5 million group is -1.71, but the actual fifth percentile from actual fund returns is much lower, -2.19. Thus, the left tails of the CDFs of three-factor $t(\alpha)$ suggest that when returns are measured before expenses, there are inferior fund managers whose actions result in negative true α relative to passive benchmarks.

Conversely, the right tails of three-factor $t(\alpha)$ suggest that there are superior managers who enhance expected returns relative to passive benchmarks. For the \$5 million AUM group, the CDF of $t(\alpha)$ estimates for actual gross fund returns moves to the right of the average from the simulations at about the 60^{th} percentile. For example, the 95^{th} percentile of $t(\alpha)$ for funds in the \$5 million group averages 1.68 in the simulations, but the actual 95th percentile is higher, 2.04. For the two larger AUM groups the crossovers occur at higher percentiles, around the 80th for the \$250 million group and the 90th for the \$1 billion group. Figure 2 graphs the results for the three-factor benchmark and the \$5 million AUM group.

Figure 2 about here.

The four-factor results for gross returns in Table III are similar to the three-factor results, with a minor nuance. Adding a momentum control tends to shrink slightly the left and right tails of the cross-sections of $t(\alpha)$ estimates for actual fund returns. This suggests that funds with negative three-factor α estimates tend to have slight negative MOM_t exposure and funds with positive three-factor α tend to have slight positive exposure. Controlling for momentum pulls the α estimates toward zero, but only a bit.

Finally, the average simulation distribution of $t(\alpha)$ for the \$5 million fund group is like a t distribution with about 24 degrees of freedom. The average sample life of these funds is 112 months, so we can probably conclude that the simulation distributions of $t(\alpha)$ are more fat-tailed than can be explained by degrees of freedom. This may in part be due to fat-tailed distributions of stock returns (Fama (1965)). A referee suggests that active trading may also fatten the tails of fund returns. And properties of the joint distribution of fund returns may have important effects on the cross-section of $t(\alpha)$ estimates – a comment of some import in our later discussion of Kosowski et al. (2006).

C. Likelihoods

Comparing the percentiles of $t(\alpha)$ estimates for actual fund returns with the simulation averages gives hints about whether manager skill affects expected returns. Table III also provides likelihoods, in particular, the fractions of the 10,000 simulation runs that produce lower values of $t(\alpha)$ at selected percentiles than actual fund returns. These likelihoods allow us to judge more formally whether the tails of the cross-section of $t(\alpha)$ estimates for actual fund returns are extreme relative to what we observe when true α is zero.

Specifically, we infer that some managers lack skill sufficient to cover costs if low fractions of the simulation runs produce left tail percentiles of $t(\alpha)$ below those from actual net fund returns, or equivalently, if large fractions of the simulation runs beat the left tail $t(\alpha)$ estimates from actual net fund returns. Likewise, we infer that some managers produce benchmark-adjusted expected returns that more than cover costs if large fractions of the simulation runs produce right tail percentiles of $t(\alpha)$ below those from actual net fund returns.

The logic is similar for gross returns, but the question is whether there are managers with skill sufficient to cover the costs (primarily trading costs) missing from expense ratios.

There are two problems in drawing inferences from the likelihoods in Table III. (i) Results are shown for many percentiles so there is a multiple comparisons issue. (ii) The likelihoods for different percentiles are correlated. One approach to these problems is to focus on a given percentile of each tail of $t(\alpha)$, for example, the 5th and the 95th, and draw inferences entirely from them. This discards lots of information. We prefer to examine all the likelihoods, with emphasis on the extreme tails, where performance is most likely to be identified. As a result, our inferences from the formal likelihoods are somewhat informal.

C.1. Net Returns

The likelihoods in Table III confirm that skill sufficient to cover costs is rare. Below the 80th percentile, the three-factor $t(\alpha)$ estimates for actual net fund returns beat those from the simulations in less than 1.0% of the net return simulation runs. For example, the 70th percentile of the cross-section of three-factor $t(\alpha)$ estimates from the net returns of \$5 million funds (our full sample) is 0.08, and only 0.51% (about half of one percent) of the 10,000 simulation runs for this group produce 70th percentile $t(\alpha)$ estimates below 0.08. It seems safe to conclude that most fund managers do not have enough skill to produce benchmark adjusted net returns that cover costs. This again is bad news for Berk and Green (2004) since their model predicts that skill sufficient to cover costs is the general rule.

The likelihoods for the most extreme right tail percentiles of the three-factor $t(\alpha)$ estimates in Table III also confirm our earlier conclusion that some managers do have sufficient skill to cover costs. For the \$5 million group, the 97th, 98th, and 99th percentiles of the cross section of three-factor $t(\alpha)$ estimates from actual net fund returns are close to the average values from the simulations, and 49.35% to 58.70% of the $t(\alpha)$ estimates from the 10,000 simulation runs are below those from actual net returns. The likelihoods that the highest percentiles of the $t(\alpha)$ estimates from the net returns of funds in the \$5 million group beat those from the simulations drop below 40% when we use the four-factor model to measure α, but the likelihoods nevertheless suggest that some fund managers have enough skill to cover costs.

Some perspective is helpful. For the \$5 million group, about 30% of funds produce positive net return α estimates. The likelihoods in Table III tell us, however, that most of these funds are just lucky; their managers are not able to produce benchmark adjusted expected returns that cover costs. For example, the 90th percentile of the $t(\alpha)$ estimates for actual net fund returns is near 1.00. The average standard error of the α estimates is 0.28% (monthly), which suggests that funds around the 90th percentile of $t(\alpha)$ beat our benchmarks by more than 3.3% per year for the entire period they are in the sample. These managers are sure to be anointed as highly skilled active investors. But about 90% of the net return simulation runs produce 90th percentiles of $t(\alpha)$ that beat those from actual fund returns. It thus seems that, like funds below the 90th percentile, most funds around the 90th percentile do not have managers with sufficient skill to cover costs; that is, true net return α is negative.

The odds that managers have enough skill to cover costs are better for funds at or above the 97th percentile of the $t(\alpha)$ estimates. In the \$5 million group, funds at the 97th, 98th, and 99th percentiles of three-factor $t(\alpha)$ estimates do about as well as would be expected if all fund managers were able to produce benchmark-adjusted expected returns that cover costs. But this just means that our estimate of true net return three-factor α for these funds is close to zero. If we switch to the four-factor model, the estimate of true α is negative for all percentiles of the $t(\alpha)$ estimates since the percentiles from actual net fund returns beat those from the simulations in less than 40% of the simulation runs.

What mix of active funds might generate the net return results in Table III? Suppose there are two groups of funds. Managers of good funds have just enough skill to produce zero α in net returns; bad funds have negative α. When the two groups are mixed, the expected cross section of $t(\alpha)$ estimates is entirely to the left of the average of the cross sections from the net return simulation runs (in which all managers have sufficient skill to cover costs). Even the extreme right tail of the $t(\alpha)$ estimates for actual net fund returns will be weighed down by bad managers who are extremely lucky but have smaller $t(\alpha)$ estimates than if they were extremely lucky good managers. In our tests, most of the cross section of $t(\alpha)$ estimates for actual net fund returns is way left of what we expect if all managers have zero true α. Thus most funds are probably in the negative true α group. At least for the \$5 million AUM sample, the 97th, 98th, and 99th percentiles of the three-factor $t(\alpha)$ estimates for actual net fund returns are similar to the simulation averages. This suggests that buried in the results are fund managers with more than enough skill to cover costs, and the lucky among them pull up the extreme right tail of the net return $t(\alpha)$ estimates. Unfortunately, these good funds are indistinguishable from the lucky bad funds that land in the top percentiles of the $t(\alpha)$ estimates but have negative true α. As a result, our estimate of the three-factor net return α for a portfolio of the top three percentiles of the $5 million group is near zero; the positive α of the lucky (but hidden) good funds is offset by the negative α of the lucky bad funds. And when we switch to the four-factor model, our estimate of true α turns negative even for the top three percentiles of the $t(\alpha)$ estimates.

Finally, our tests exclude index funds, but we can report that for 1984 to 2006 the net return three-factor α estimate for the VW portfolio of index funds (in which large low cost funds get heavy weight) is -0.16% per year (-0.01% per month, t = -0.61), and four-factor α is 0.01% per year (t = 0.02). Since large low cost index funds are not subject to the vagaries of active management, it seems reasonable to infer that the net return true α for a portfolio of these funds is close to zero. In other words, going forward we expect that a portfolio of low cost index funds will perform about as well as a portfolio of the top three percentiles of past active winners, and better than the rest of the active fund universe.

C.2. Gross Returns

The simulation tests for net returns ask whether active managers have sufficient skill to cover all their costs. In the tests on gross returns, the bar is lower. Specifically, the issue is whether managers have enough skill to at least cover the costs (primarily trading costs) missing from expense ratios.

The three-factor gross return simulations for the $5 million AUM group suggest that most funds in the left tail of three-factor $t(\alpha)$ estimates do not have enough skill to produce benchmark adjusted expected returns that cover trading costs, but many managers in the right tail have such skill. For the 40th and lower percentiles, the three-factor $t(\alpha)$ estimates for the actual gross returns of funds in the $5 million group beat those from the simulations in less than 30% of the simulation runs, falling to less than 6% for the 10th and lower percentiles. Conversely, above the 60th percentile, the three-factor $t(\alpha)$ estimates for actual gross fund returns beat those from the simulations in at least 56% of the simulation runs, rising to more than 90% for the 96th and higher percentiles. As usual, the results are weaker when we switch from three-factor to four-factor benchmarks, but the general conclusions are the same.

For many readers, the important insight of Berk and Green (2004) is their assumption that there are diseconomies of scale in active management, not their detailed predictions about net fund returns (which are rejected in our tests). The right tails of the $t(\alpha)$ estimates for gross returns suggest diseconomies. The extreme right tail percentiles of $t(\alpha)$ are typically lower for the \$250 million and \$1 billion groups than for the \$5 million group, and more of the simulation runs beat the extreme right tail percentiles of the $t(\alpha)$ estimates for the larger AUM funds. In the world of Berk and Green (2004), however, the weeding out of unskilled managers should also lead to left tails for $t(\alpha)$ estimates that are less extreme for larger funds. This prediction is not confirmed in our results. The left tails of the $t(\alpha)$ estimates for the \$250 million and \$1 billion groups are at least as extreme as the left tail for the \$5 million group. This contradiction in the left tails of the $t(\alpha)$ estimates makes us reluctant to interpret the right tails as evidence of diseconomies of scale.

The tests on gross returns point to the presence of skill (positive and negative). We next estimate the size of the skill effects. A side benefit is evidence on the power of the simulation tests.

IV. Estimating the Distribution of True α in Gross Fund Returns

To examine the likely size of the skill effects in gross fund returns we repeat the simulations but with α injected into fund returns. We then examine (i) how much α is necessary to reproduce the cross section of $t(\alpha)$ estimates for actual gross fund returns, and (ii) levels of α too extreme to be consistent with the $t(\alpha)$ estimates for actual fund returns.

Given the evidence that, at least for the \$5 million group (our full sample), the distribution of $t(\alpha)$ estimates in gross fund returns is roughly symmetric about zero (Table III), it is reasonable to assume that true α is distributed around zero. It is also reasonable to assume that extreme levels of skill (good or bad) are rare. Concretely, we assume that each fund is endowed with a gross return α drawn from a normal distribution with a mean of zero and a standard deviation of σ per year.

The new simulations are much like the old. The first step again is to adjust the gross returns of each fund, setting α to zero for the three-factor and four-factor benchmarks and each of the three AUM groups. But now, before drawing the random sample of months for a simulation run, we draw a true α from a normal distribution with mean zero and standard deviation σ per year – the same α for every combination of benchmark model and AUM group for a given fund, but an independent drawing of α for each fund.

It seems reasonable that more diversified funds have less leeway to generate true α. To capture this idea, we scale the α drawn for a fund by the ratio of the fund's (three-factor or four-factor) residual standard error to the average standard error for all funds. We add the scaled α to the fund's benchmark-adjusted returns. We then draw a random sample (with replacement) of 273 months, and for each fund we estimate three-factor and four-factor regressions on the adjusted gross returns of the fund's three AUM samples. The simulations thus use returns that have the properties of actual fund returns, except we know true α has a normal distribution with mean zero and (for the "average" fund) standard deviation σ per year. We do 10,000 simulation runs, and a fund gets a new drawing of α in each run. To examine power, we vary σ, the standard deviation of true α, from 0.0% to 2.0% per year, in steps of 0.25%.

Table IV shows percentiles of the cross section of $t(\alpha)$ estimates for actual gross fund returns (from Table III) and the average $t(\alpha)$ estimates at the same percentiles from the 10,000 simulation runs, for each value of σ. These are useful for judging how much dispersion in true α is consistent with the actual cross section of $t(\alpha)$ estimates. For each σ, the table also shows the fraction of the simulation runs that produce percentiles of $t(\alpha)$ estimates below those from actual fund returns. These are our evidence for inferences about the amount of dispersion in true α we might rule out as too extreme.

Table IV about here.

A. Likely Levels of Performance

If true α comes from a normal distribution with mean zero and standard deviation σ, Table 4 provides two slightly different ways to infer the value of σ. We can look for the value of σ that produces average simulation percentile values of $t(\alpha)$ most like those from actual fund returns. Or we can look for the σ that produces simulation $t(\alpha)$ estimates below those for actual returns in about 50% of the simulation runs. If α has a normal distribution with mean zero and standard deviation σ, we expect the effects of the level of σ to become stronger as we look further into the tails of the cross section of $t(\alpha)$. Thus, we are most interested in values of σ that match the extreme tails of the $t(\alpha)$ estimates for actual gross fund returns.

The normality assumption for true α is an approximation. We do not expect that a single value of σ (the standard deviation of true α) completely captures the tails of the $t(\alpha)$ estimates for actual fund returns, even if we allow a different σ for each tail. With this caveat, the three-factor and four-factor simulations for the \$5 million group suggest that σ around 1.25% to 1.50% per year captures the extreme left tail of the $t(\alpha)$ estimates for actual gross fund returns, and 1.25% works for the right tail. For the \$250 million and \$1 billion groups, the three-factor simulations again suggest σ around 1.25% to 1.50% per year for the left tail of the $t(\alpha)$ estimates for gross fund returns, but for the right tail, σ is lower, 0.75% to 1.00% per year. In the four-factor simulations for the \$250 and \$1 billion groups σ = 1.25% per year seems to capture the extreme left tail of the $t(\alpha)$ estimates for gross fund returns, but the estimate of σ for the right tail is again lower 0.75% per year. (To save space, Table IV shows results only for the \$5 million and \$1 billion AUM groups.)

The estimates do not suggest much performance, especially for larger funds. Thus, σ = 1.25% says that about one sixth of funds have true gross return α greater than 1.25% per year (about 0.10% per month) and only about 2.4% have true α greater than 2.50% per year (0.21% per month). For perspective, the average of the OLS standard errors of individual fund α estimates – the average imprecision of α estimates – is 0.28% per month (3.4% per year). Moreover, much lower right tail σ estimates for the \$250 million and \$1 billion funds say that lots of the right tail performance observed in the full (\$5 million) sample is due to tiny funds.

Our gross fund returns are net of trading costs. Returning trading costs to funds (if that is deemed appropriate) would increase the $t(\alpha)$ estimates in both the left and the right tails, which, depending on the (unknown) magnitudes, may move them toward more similar estimates of σ.

B. Unlikely Levels of Performance

What levels of σ can we reject? The answer depends on how confident we wish to be about our inferences. Suppose we are willing to accept a 20% chance of setting a lower bound for σ that is too high and a 20% chance of setting an upper bound that is too low. These bounds imply a narrower range than we would have with standard significance levels, but they are reasonable if our goal is to provide perspective on likely values of σ.

Under the 20% rule, the lower bound for the left tail estimate of σ is the value that produces left tail percentile $t(\alpha)$ estimates below those from actual fund returns in about 20% of the simulation runs. The upper bound for the left tail σ is the value that produces left tail percentiles of $t(\alpha)$ below those from actual fund returns in about 80% of the simulation runs. Conversely, under the 20% rule, the lower bound for the right tail

σ estimate produces right tail percentile $t(\alpha)$ estimates below those from actual fund returns in about 80% of the simulation runs. And the upper bound for the right tail σ produces right tail percentiles of $t(\alpha)$ below those from actual fund returns in about 20% of the simulation runs.

In brief, applying the 20% rule leads to intervals for σ that are equal to the point estimates of the preceding section plus and minus 0.5%. For example, 1.25% per year works fairly well as the left tail estimate of σ for all AUM groups and for the three-factor and four-factor models, and the interval for the left tail σ estimates is 0.75% to 1.75%. For the \$5 million group, σ = 1.25% also works for the right tail, and the interval is again 0.75% to 1.75%. For the \$250 million and \$1 billion groups, the right tail estimate of σ drops to about 0.75% per year, and the 20% rule leads to an interval for σ from 0.25% to 1.25% per year.

What do these results say about the power of the simulation approach? The upper bound on σ for the \$5 million group, 1.75% per year, translates to a monthly σ for the cross section of true α of about 0.146%. Suppose the standard error of each fund's α estimate is 0.28% per month (the sample average). With a monthly σ of 0.146%, the standard deviation of the cross section of α estimates – caused by measurement error and dispersion in true α – is $(0.146^2 + 0.28^2)^{1/2} = 0.316\%$. This is only a bit bigger than 0.299%, the standard deviation implied by our estimate of σ for the \$5 million group, 1.25% per year. The fact that the simulations assign a relatively low probability to $\sigma \geq 1.75\%$ despite the small difference between the implied standard deviations of the α estimates for σ = 1.25% (the point estimate) and σ = 1.75% suggests that the simulations have power. The source of the power is our large sample of funds (3156 in the \$5 million group). With so many funds, the percentiles of $t(\alpha)$ are estimated precisely, which produces power to draw inferences about σ. (We thank a referee for this insight.)

V. Kosowski et al. (2006)

The paper closest to ours is Kosowski et al. (2006). They use bootstrap simulations to draw inferences about performance in the cross-section of four-factor $t(\alpha)$ estimates for net fund returns. Their main inference is more positive than ours. They find that the 95th and higher percentiles of four-factor $t(\alpha)$ estimates for net fund returns are above the same simulation percentiles in more than 99% of simulation runs. This seems like strong evidence that among the best funds, many have more than sufficient skill to cover costs. Our simulations on net returns uncover much less evidence of skill. Two things account for their stronger results, simulation approach and time period.

We jointly sample fund (and explanatory) returns, whereas Kosowski et al. (2006) do independent simulations for each fund. The benefit of their approach is that the number of months a fund is in a simulation run always matches the fund's actual number of months of returns. The cost is that their simulations take no account of the correlation of α estimates for different funds that arises because a benchmark model does not capture all common variation in fund returns. They summarize but do not show simulations that jointly sample the four-factor residuals of funds. But they never jointly sample fund returns and explanatory returns, which means (for example) they miss any effects of correlated movement in the volatilities of four-factor explanatory returns and residuals. In fact, in the results they show, the explanatory returns do not vary across simulation runs; the historical sequence of explanatory returns is used in every run.

Their rules for including funds in the simulation tests are also different. They include the complete return histories of all funds that survive more than 60 months (so there is survival bias). We include funds after they pass \$5 million in AUM if they have at least eight months of returns thereafter (less survival bias).

Table V shows simulation results for their 1975 to 2002 period using (i) their rules for including funds and (ii) our rules. Note that both sets of simulations use our approach to drawing simulation samples; that is, a simulation run uses the same random sample of months for all funds, which allows for all effects implied by the joint distribution of fund returns, and of fund and explanatory returns.

Table V about here.

The rules used to include funds affect the cross section of $t(\alpha)$ estimates for actual fund returns. Specifically, the right tail $t(\alpha)$ estimates for actual fund returns are less extreme for our sample. This suggests that their rule that a fund must have at least 60 months of returns produces more survival bias than our eight month rule. Another possibility is that some funds have high returns when they are tiny but do not do as well after they pass \$5 million. And this may in part be due to an incubation bias in the fund sample of Kosowski et al. (2006), since they include a fund's entire return history if the fund survives for 60 months.

For either sample of funds, joint sampling of fund returns (our approach) affects the simulation results. Kosowski et al. (2006) report that more than 99% of their simulation runs produce 95th percentile four-factor

$t(\alpha)$ estimates below the 95th percentile from actual net fund returns. In Table V, the number drops to 82.42% for the fund sample selected using their rules and 68.32% using our rules. Skipping the details, we can report that the stronger performance results from the fund sample chosen using their rules is due to the 60 month survival rule. If the survival rule is reduced to eight months, their rules for including funds produce simulation results close to ours. The important point, however, is that whatever inclusion rules are used, failure to account for the joint distribution of fund returns, and of fund and explanatory returns, biases the inferences of Kosowski et al. (2006) toward positive performance. (Cuthbertson et al. (2008) apply the simulation approach of Kosowski et al. to UK mutual funds, with similar results and, we guess, similar problems.)

Time period is also an important source of differences in results. Our simulations for 1984 to 2006 produce much less evidence of funds with sufficient skill to cover costs. In Table III, the CDFs of four-factor $t(\alpha)$ estimates for the net fund returns of 1984 to 2006 are always to the left of the average CDFs from the net return simulations (in which funds have sufficient skill to cover costs). Even in the extreme right tail of four-factor $t(\alpha)$ for net returns, more than 60% of the simulation runs beat the $t(\alpha)$ estimates for actual fund returns. But when our approach is applied to the 1975 to 2002 period of Kosowski et al. (2006), the 90th and higher percentiles of $t(\alpha)$ for net fund returns are above the average values from the simulations (Table V). And for the 97th and higher percentiles, less than 20% of the simulation runs beat the $t(\alpha)$ estimates for actual fund returns.

What do we make of the stronger results for 1975 to 2002 versus 1984 to 2006? One story is that in olden times there were fewer funds and a larger percentage of managers with skill sufficient skill to cover costs. Over time the skilled managers lost their edge or went on to more lucrative pursuits (for example, hedge funds). Or perhaps, the entry of hordes of mediocre managers posing as skilled (Cremers and Petajisto (2008)) buries the tracks of true skill. Stronger results for 1975 to 2002 may also be due to biases in the CRSP data that are more prevalent in earlier years (Elton, Gruber, and Blake (2001)). Whatever the explanation, the stronger evidence for performance during 1975 to 2002 is interesting, but irrelevant for today's investors.

IV. Conclusions

For 1984 to 2006, when the CRSP database is relatively free of biases, mutual fund investors in aggregate get net returns that underperform CAPM, three-factor, and four-factor benchmarks by about the costs in expense ratios. Thus, if there are fund managers with enough skill to produce benchmark adjusted expected returns that cover costs, their tracks are hidden in the aggregate results by the performance of managers with insufficient skill.

When we turn to individual funds, the challenge is to distinguish skill from luck. With 3156 funds in our full ($5 million AUM) sample, some do extraordinarily well and some do extraordinarily poorly just by chance. To distinguish between luck and skill, we compare the distribution of $t(\alpha)$ estimates from actual fund returns with the distribution from bootstrap simulations in which all funds have zero true α. The tests on net returns say that few funds have enough skill to cover costs. The distribution of three-factor $t(\alpha)$ estimates from net fund returns is almost always to the left of the zero α distribution. The extreme right tail of the three-factor $t(\alpha)$ estimates for net fund returns, however, is roughly in line with the simulated distribution. This suggests that some managers do have sufficient skill to cover costs. But the estimate of net return three-factor true α is about zero even for the portfolio of funds in the top percentiles of historical three-factor $t(\alpha)$ estimates, and the estimate of four-factor true α is negative. Moreover, the estimate of true α for funds in the top percentiles is no better than the estimated α (also near zero) for large efficiently managed passive funds.

The simulation results for gross fund returns say that when returns are measured before the costs in expense ratios, there is stronger evidence of manager skill, negative as well as positive. For our $5 million AUM sample, true three-factor or four-factor gross return α seems to be symmetric about zero with a cross section standard deviation of about 1.25% per year (about ten basis points per month). For larger ($250 million and $1 billion AUM) funds, the standard deviation for the left tail is again about 1.25% per year, but the right tail standard deviation of true α falls to about 0.75%.

Appendix A – CAPM Bootstrap Simulations

Table AI replicates the bootstrap simulations in Table III for a CAPM benchmark, that is, regression (1) with the excess market return as the only explanatory variable. The CAPM results are different. The CAPM tests on net returns produce what seems like strong evidence that some fund managers have sufficient skill to cover costs. Thus, for percentiles above the 90^{th}, the CAPM $t(\alpha)$ estimates for actual net fund returns are always above the averages from the net return simulations (in which all managers have sufficient skill to cover costs), and the $t(\alpha)$ estimates for actual fund returns typically beat those from the simulations in more than 80% of simulation runs. Relative to the three-factor and four-factor tests in Table III, the CAPM tests on gross returns in Table AI also produce what seems like stronger evidence that some managers have skill that leads to positive true α, while others have negative true α.

Table AI about here.

In fact, the CAPM results just illustrate well-known patterns in average returns that cause problems for the CAPM during our sample period. Actual mutual fund returns contain the effects of size, value/growth, and momentum tilts in fund portfolios that are missed by the CAPM. Thus, even passive funds that tilt toward small stocks, or value stocks, or positive momentum stocks are likely to produce positive α estimates in CAPM tests, despite the fact that their managers make no effort to pick individual stocks. The CAPM simulations allow for the relation between average return and market exposure, but they wash out all other patterns in average returns when they subtract each fund's CAPM α estimate from its returns. As a result, the CAPM simulations say that actual fund returns have non-zero true α.

Which patterns in average returns left unexplained by the CAPM are most responsible for the differences between the CAPM simulation results and the results for the three-factor and four-factor models? Table 3 says that adding the momentum factor to the three-factor model has minor effects on estimates of $t(\alpha)$. Since the momentum return MOM_t has the highest average premium during our sample period, we infer that long-term exposure to momentum is probably rare among mutual funds. The average size (SMB_t) premium is trivial during our 1984 to 2006 sample period (0.03% per month, Table I), so size tilts probably are not driving the different results for the CAPM. That leaves the value (HML_t) premium as the focus of the story. Funds in the right tail of the CAPM $t(\alpha)$ estimates are more likely to have positive HML_t exposure that makes them look good in CAPM tests, and funds in the left tail are likely to have negative HML_t exposure.

In short, the CAPM tests are a lesson about how failure to account for common patterns in returns and average returns can affect inferences about the skill of fund managers.

Appendix B – Measurement Issues in Gross Returns

The question in the tests on gross fund returns is whether managers have skill that causes expected returns to differ from those of comparable passive benchmarks. For this purpose, we would like to have fund returns measured before all costs but net of non-return income like revenues from securities lending. This would put funds on the same pure return basis as the benchmark explanatory returns, so the tests could focus on the effects of skill. Our gross fund returns are before the costs in expense ratios, but they are net of other costs, primarily trading costs, and they include income from securities lending.

We could attempt to add trading costs to our estimates of gross fund returns. Funds do not report trading costs, however, and even when turnover is available, estimates of trading costs are subject to large error (Carhart (1997)). For example, trading costs are likely to vary across funds because of differences in style tilts, trading skill, and the extent to which a fund is actively managed and demands immediacy in trade execution. Trading costs can also vary through time because of changes in a fund's management and general changes in the costs of trading. All this leads us to conclude that estimates of trading costs for individual funds, especially actively managed funds, are fraught with error and potential bias, and so can be misleading. As a result, we do not take that route in our tests on gross returns.

An alternative approach (suggested by a referee) is to put the passive benchmarks produced by combining the explanatory returns in (1) in the same units as the gross fund returns on the left of (1). This involves taking account of the costs (primarily trading costs) not covered in expense ratios that would be borne by an efficiently managed passive benchmark with the same style tilts as the fund whose gross returns are to be explained.

Vanguard's index funds are good candidates for this exercise since, except for momentum, Vanguard provides index funds (Total Stock Market Index Fund, Growth Index Fund, Value Index Fund, Small-Cap Index Fund, Small-Cap Growth Index Fund, and Small-Cap Value Index Fund), that track well-defined target passive portfolios much like the market portfolio and the components of SMB_t and HML_t in (1). (We thank an Associate Editor for this insight.) Because the Vanguard index funds closely track their targets and stock picking skill is not an issue, we can estimate the average annual costs not included in a fund's expense ratio. Specifically, we add a fund's expense ratio to its reported average annual return for the ten years through 2008 and then subtract the result from the average annual return of the fund's target for the same period. (The same calculation for an actively managed fund would include the effects of skill, as well as the costs not in expense ratios.) For every Vanguard index fund, this estimate of the costs missed in expense ratios is negative; that is, the fund's target return, which is before all costs, beats the fund's actual net return by less than the fund's expense ratio. If anything, Vanguard's small cap index funds do better on this score than its large cap funds – a clear warning that presumptions about trading costs can be misleading.

The Vanguard results are probably not unusual. We can report that the CAPM, three-factor, and four-factor α estimates for 1984 to 2006 for the net returns on a VW portfolio of index funds (which is dominated by large funds with low expense ratios) are close to zero, 0.08%, -0.16%, and 0.01% per year (t = 0.18, -0.61, and 0.02). In other words, in aggregate, wealth invested in index funds seems to earn average returns that cover costs, including trading costs.

Passive mutual funds that focus on momentum do not as yet exist, so we do not have estimates of trading costs for such funds. Existing work (Grundy and Martin (2001), Korajczyk and Sadka (2004)) suggests that the costs are significant. In our tests, however, the cross-sections of four-factor α estimates for funds are similar to the cross-sections of three-factor estimates, and the three-factor and four-factor tests produce much the same inferences. Given the large average MOM_t return, these results suggest that non-trivial long-term exposure to MOM_t is rare, so ignoring MOM_t trading costs is inconsequential. Moreover, the discussion of results in the text centers primarily on the three-factor model. The four-factor results are primarily a robustness check.

The Vanguard evidence and the results for a VW portfolio of index funds suggest that for the market and the components of SMB_t and HML_t, comparable efficiently managed passive mutual funds can enhance returns through trading, securities lending, and perhaps in other ways, so that their total costs are close to their expense ratios. Thus, our three-factor α estimates for the gross returns of funds would hardly change if we adjusted their passive benchmarks for the costs missed in expense ratios.

This does not mean our tests on gross returns capture the pure effects of skill. Though expense ratios seem to capture the total costs of efficiently managed passive funds, this is less likely to be true for actively managed funds. The typical active fund trades more than the typical passive fund, and active funds are likely to demand immediacy in trading that produces positive costs. Because of their high turnover, active funds also have fewer opportunities to generate revenues via securities lending (which are also trivial for the Vanguard funds). In short, it seems more likely that for active funds the costs not included in expense ratios are positive. Thus, our tests on the gross returns of funds produce α estimates that capture the effects of skill, less any costs missed by the expense ratios of the funds.

Equivalently, our tests on gross returns say that a fund's management has skill only if it is sufficient to cover the costs (primarily trading costs) not included in its expense ratio. This is a reasonable definition of skill since a comparable efficiently managed passive fund would apparently avoid these costs. More important, this definition of skill is the only one we can accurately test in the absence of accurate estimates of the trading costs of active funds (impossible with available data).

It is fortuitous that efficiently managed passive benchmarks do not seem to have substantial costs missed in their expense ratios since accurate adjustment for such costs is non-trivial, perhaps impossible. For example, consider an actively managed small value fund. The passive benchmark for the fund produced by the three-factor version of (1) is likely to imply positive weights on the market, *SMB*, and *HML*, which implies positive weights on the market (*M*), small stocks (*S*), and value stocks (*H*) and negative weights on big stocks (*B*) and growth stocks (*L*). Suppose that (contrary to our estimates) efficiently managed passive funds have non-trivial trading costs. We might then increase the three-factor gross return α estimate for an active fund for the trading costs of the long positions in *M*, *S*, and *H* and the short positions in *B* and *L* that passively replicate the small value style of the active fund. But this is overkill. The three-factor model produces a passive clone for an actively managed fund by inefficiently combining five passive portfolios. A small value fund simply buys a diversified portfolio of small value stocks and only bears the trading costs of these stocks. As a result, even a passive small value fund evaluated with the three-factor model is likely to produce a positive α estimate if we enhance the estimate with positive trading costs for the five components of its three-factor clone.

If we wish to adjust the tests on gross returns for the trading costs of an efficiently managed passive fund with the same style tilts as the active fund to be evaluated, the correct procedure is to add an estimate of the trading costs of a comparable efficiently managed passive fund to the active fund's gross return α estimate. For example, a small value active fund would be reimbursed for the trading costs (more precisely, for all the costs missed in the expense ratio) of an efficiently managed passive fund with the same style tilts. This is non-trivial since a style group includes active funds with widely different style tilts, and we need an efficiently managed passive clone for every active fund. Fortunately, the costs missed in expense ratios are apparently close to zero for efficiently managed passive funds, and ignoring them (as we do in our tests) is inconsequential for inferences.

References

Berk, Jonathan B., and Richard C. Green, 2004, Mutual fund flows in rational markets, *Journal of Political Economy* 1269-1295.

Carhart, Mark M., 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57-82.

Cremers, Martijn, and Antti Petajisto, 2008, How active is your fund manager? A new measure that predicts performance," *Review of Financial Studies*, forthcoming.

Cuthbertson, Keith, Dirk Nitzche, and Niall O'Sullivan, 2008, UK Mutual fund performance: Skill or luck? *Journal of Empirical Finance*, forthcoming.

Dybvig, Philip H. and Stephen A. Ross, 1985, The Analytics of performance measurement using a security market line, *Journal of Finance* 40, 401-416.

Elton, Edwin J., Martin J. Gruber, and Christopher R. Blake, 2001, A first look at the accuracy of the CRSP mutual fund database and a comparison of the CRSP and Morningstar mutual fund databases, *Journal of Finance* 56, 2415-2430.

Evans, Richard, 2007, The incubation bias, muscript, Darden Graduate School of Business, University of Virginia, July 30.

Fama, Eugene F., 1965, The behavior of stock market prices, *Journal of Business* 38, 34-105.

Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3-56.

Ferson, Wayne E., and Rudi W. Schadt, 1996, Measuring fund strategy and performance in changing economic conditions, *Journal of Finance* 51, 425-462.

French, Kenneth R., 2008, The cost of active investing, *Journal of Finance*, 63, forthcoming.

Grinblatt, Mark, and Sheridan Titman, 1992, Performance persistence in mutual funds, Journal of Finance 47, 1977-1984.

Gruber, Martin J. 1996, Another puzzle: The growth of actively managed mutual funds, *Journal of Finance* 51, 783-810.

Grundy, Bruce D. and J. Spencer Martin, 2001, Understanding the nature of the risks and the sources of the rewards to momentum investing, *Journal of Financial Studies* 14, 29-78.

Jensen, Michael C., 1968, The performance of mutual funds in the period 1945-1964, *Journal of Finance* 23, 2033-2058.

Korajczyk, Robert A. and Ronnie Sadka, 2004, Are momentum profits robust to trading costs?, *Journal of Finance* 59, 1039-1082.

Kosowski, Robert, Allan Timmermann, Russ Wermers, and Hal White, 2006, Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis, *Journal of Finance* 61, 2551-2595.

Malkiel, Burton G. 1995, Returns from investing in equity mutual funds: 1971-1991, *Journal of Finance* 50, 549-572.

Sharpe, William F., 1991, The arithmetic of active management, *Financial Analysts Journal* 47, January/February, 7–9.

Footnotes

[1] Formal justification for this definition of good and bad performance is provided by Dybvig and Ross (1985). Given a riskfree security, their Theorem 5 implies that if the intercept in (1) is positive, there is a portfolio with positive weight on fund *i* and the portfolio of the explanatory portfolios on the right of (1) that has a higher Sharpe ratio than the portfolio of the explanatory portfolios. Similarly, if the intercept is negative, there is a portfolio with negative weight on fund *i* that has a higher Sharpe ratio than the portfolio of the explanatory portfolios.

Table I — Summary statistics for monthly explanatory returns for the three-factor and four-factor models

R_M is the return on a value-weight market portfolio of NYSE, Amex, and NASDAQ stocks, and R_f, is the one-month Treasury bill rate. The construction of SMB_t and HML_t follows Fama and French (1993). At the end of June of each year *k*, we sort stocks into two size groups. Small includes NYSE, Amex, and NASDAQ stocks with June market capitalization below the NYSE median and Big includes stocks with market cap above the NYSE median. We also sorts stocks into three book-to-market equity (*B*/*M*) groups, Growth (NYSE, Amex, and NASDAQ stocks in the bottom 30% of NYSE *B*/*M*), Neutral (middle 40% of NYSE *B*/*M*), and Value (top 30% of NYSE *B*/*M*). Book equity is for the fiscal year ending in calendar year *k*-1, and the market cap in *B*/*M* is for the end of December of *k*-1. The intersection of the (independent) size and *B*/*M* sorts produces six value-weight portfolios, refreshed at the end of June each year. The size return, SMB_t, is the simple average of the month *t* returns on the three Small stock portfolios minus the average of the returns on the three Big stock portfolios. The value-growth return, HML_t, is the simple average of the returns on the two Value portfolios minus the average of the returns on the two Growth portfolios. The momentum return, MOM_t, is defined like HML_t, except that we sort on prior return rather than *B*/*M* and the momentum sort is refreshed monthly rather than annually. At the end of each month *t*-1 we sort NYSE stocks on the average of the eleven months of returns to the end of month *t*-2. (Dropping the return for month *t*-1 is common in the momentum literature.) We use the 30th and 70th NYSE percentiles to assign NYSE, Amex, and NASDAQ stocks to Low, Medium, and High momentum groups. The intersection of the size sort for the most recent June and the independent momentum sort produces six value-weight portfolios, refreshed monthly. The momentum return, MOM_t, is the simple average of the month *t* returns on the two High momentum portfolios minus the average of the returns on the two Low momentum portfolios. The table shows average monthly return, the standard deviation of monthly returns, and the *t*-statistic for the average monthly return. The period is January 1984 through September 2006.

	Average Return				Standard Deviation				*t*-statistic			
	R_M -R_f	*SMB*	*HML*	*MOM*	R_M-R_f	*SMB*	*HML*	*MOM*	R_M-R_f	*SMB*	*HML*	*MOM*
1984-2006	0.64	0.03	0.40	0.79	4.36	3.38	3.17	4.35	2.42	0.13	2.10	3.01

Table II – Intercepts and slopes in variants of regression (1) for equal-weight (EW) and value-weight (VW) portfolios of actively managed mutual funds

The table shows the annualized intercepts (12**a*) and *t*-statistics for the intercepts (*t*(*Coef*)) for the CAPM, three-factor, and four-factor versions of regression (1) estimated on equal-weight (EW) and value-weight (VW) net and gross returns on the portfolios of actively managed mutual funds in our sample. The table also shows the regression slopes (*b*, *s*, *h*, and *m*, for R_M-R_f, *SMB*, *HML* and *MOM*, respectively), *t*-statistics for the slopes, and the regression R^2, all of which are the same to two decimals for gross and net returns. For the market slope, *t*(*Coef*) tests whether *b* is different from 1.0. Net returns are those received by investors. Gross returns are net returns plus 1/12th of a fund's expense ratio for the year. When a fund's expense ratio for a year is missing, we assume it is the same as other actively managed funds with similar assets under management (AUM). The period is January 1984 through September 2006. On average there are 1308 funds and their average AUM is $648.0 million.

	12**a*						
	Net	Gross	*b*	*s*	*h*	*m*	R^2
EW Returns							
Coef	-1.11	0.18	1.01				0.96
t(*Coef*)	-1.80	0.31	1.12				
Coef	-0.93	0.36	0.98	0.18	-0.00		0.98
t(*Coef*)	-2.13	0.85	-1.78	16.09	-0.24		
Coef	-0.92	0.39	0.98	0.18	-0.00	-0.00	0.98
t(*Coef*)	-2.05	0.90	-1.78	16.01	-0.25	-0.14	
VW Returns							
Coef	-1.13	-0.18	0.99				0.99
t(*Coef*)	-3.03	-0.49	-2.10				
Coef	-0.81	0.13	0.96	0.07	-0.03		0.99
t(*Coef*)	-2.50	0.40	-5.42	7.96	-3.22		
Coef	-1.00	-0.05	0.97	0.07	-0.03	0.02	0.99
t(*Coef*)	-3.02	-0.15	-5.03	7.78	-3.03	2.60	

Table III - Percentiles of $t(\alpha)$ estimates for actual and simulated fund returns: January 1984 to September 2006

The table shows values of $t(\alpha)$ at selected percentiles (Pct) of the distribution of $t(\alpha)$ estimates for actual (Act) net and gross fund returns. The table also shows the percent of the 10,000 simulation runs that produce lower values of $t(\alpha)$ at the selected percentiles than those observed for actual fund returns (% < Act). Sim is the average value of $t(\alpha)$ at the selected percentiles from the simulations. The period is January 1984 to September 2006 and results are shown for the three- and four-factor models for the $5 million, $250 million, and $1 billion AUM fund groups. There are 3156 funds in the $5 million group, 1422 in the $250 million group, and 660 in the $1 billion group.

	5 Million			250 Million			1 Billion		
Pct	Sim	Act	%<Act	Sim	Act	%<Act	Sim	Act	%<Act
3-Factor Net Returns									
1	-2.50	-3.87	0.08	-2.45	-3.87	0.10	-2.39	-4.39	0.01
2	-2.17	-3.42	0.06	-2.13	-3.38	0.13	-2.09	-3.55	0.09
3	-1.97	-3.15	0.07	-1.94	-3.15	0.12	-1.91	-3.36	0.07
4	-1.83	-2.99	0.06	-1.80	-3.04	0.10	-1.78	-3.16	0.07
5	-1.71	-2.84	0.08	-1.69	-2.91	0.10	-1.67	-2.99	0.10
10	-1.32	-2.34	0.05	-1.31	-2.37	0.10	-1.30	-2.53	0.08
20	-0.87	-1.74	0.03	-0.86	-1.87	0.04	-0.86	-1.98	0.03
30	-0.54	-1.27	0.06	-0.54	-1.41	0.06	-0.54	-1.59	0.02
40	-0.26	-0.92	0.05	-0.27	-1.03	0.07	-0.27	-1.19	0.02
50	-0.01	-0.62	0.04	-0.01	-0.71	0.06	-0.01	-0.82	0.03
60	0.25	-0.29	0.11	0.25	-0.39	0.19	0.24	-0.51	0.05
70	0.52	0.08	0.51	0.52	-0.08	0.25	0.52	-0.20	0.08
80	0.85	0.50	3.20	0.84	0.37	1.68	0.84	0.25	0.85
90	1.30	1.01	8.17	1.29	0.89	5.19	1.28	0.82	4.81
95	1.68	1.54	30.55	1.66	1.36	14.17	1.64	1.34	17.73
96	1.80	1.71	40.06	1.76	1.49	17.24	1.74	1.52	26.33
97	1.94	1.91	49.35	1.90	1.69	25.92	1.87	1.79	42.86
98	2.13	2.17	58.70	2.08	1.90	30.43	2.04	2.02	50.07
99	2.45	2.47	57.42	2.36	2.29	43.92	2.31	2.40	63.11
4-Factor Net Returns									
1	-2.55	-3.94	0.04	-2.47	-3.94	0.08	-2.40	-4.22	0.01
2	-2.20	-3.43	0.04	-2.14	-3.43	0.09	-2.09	-3.48	0.08
3	-2.00	-3.08	0.13	-1.95	-3.07	0.25	-1.91	-3.11	0.23
4	-1.85	-2.88	0.13	-1.80	-2.88	0.22	-1.77	-2.95	0.21
5	-1.73	-2.74	0.12	-1.69	-2.78	0.18	-1.66	-2.86	0.14
10	-1.33	-2.23	0.14	-1.30	-2.34	0.14	-1.29	-2.48	0.07
20	-0.86	-1.67	0.10	-0.85	-1.80	0.11	-0.84	-1.96	0.05
30	-0.53	-1.25	0.12	-0.52	-1.39	0.10	-0.52	-1.54	0.04
40	-0.25	-0.88	0.21	-0.25	-1.04	0.14	-0.25	-1.23	0.05
50	0.01	-0.60	0.18	0.01	-0.76	0.11	0.01	-0.87	0.07
60	0.26	-0.29	0.25	0.27	-0.42	0.29	0.26	-0.49	0.19
70	0.54	0.02	0.37	0.54	-0.13	0.24	0.54	-0.18	0.24
80	0.87	0.44	1.76	0.86	0.27	0.72	0.86	0.17	0.45
90	1.33	1.04	10.62	1.31	0.86	4.40	1.30	0.86	7.07
95	1.72	1.53	23.82	1.69	1.37	14.35	1.67	1.31	14.13
96	1.84	1.67	28.21	1.80	1.51	18.23	1.78	1.45	17.16
97	1.99	1.84	31.30	1.94	1.65	18.62	1.91	1.57	17.05
98	2.19	2.09	39.12	2.12	1.79	15.57	2.08	1.76	18.86
99	2.52	2.40	36.96	2.42	2.22	29.88	2.36	2.26	42.00

Table III (continued)

Pct	5 Million Sim	5 Million Act	5 Million %<Act	250 Million Sim	250 Million Act	250 Million %<Act	1 Billion Sim	1 Billion Act	1 Billion %<Act
3-Factor Gross Returns									
1	-2.49	-3.07	4.11	-2.45	-3.16	3.16	-2.39	-3.29	1.88
2	-2.17	-2.68	4.79	-2.13	-2.67	6.01	-2.09	-2.70	5.64
3	-1.97	-2.48	4.20	-1.94	-2.51	4.47	-1.91	-2.51	5.12
4	-1.83	-2.31	4.41	-1.80	-2.35	4.68	-1.78	-2.33	5.77
5	-1.71	-2.19	4.15	-1.69	-2.18	5.99	-1.67	-2.18	6.52
10	-1.32	-1.72	5.75	-1.31	-1.77	5.94	-1.30	-1.86	4.15
20	-0.87	-1.10	13.61	-0.86	-1.24	7.18	-0.86	-1.43	2.52
30	-0.54	-0.71	20.03	-0.54	-0.79	15.10	-0.54	-1.00	4.28
40	-0.26	-0.36	29.74	-0.27	-0.43	23.84	-0.27	-0.59	10.25
50	-0.01	-0.06	38.87	-0.01	-0.15	26.28	-0.01	-0.28	13.48
60	0.25	0.28	56.05	0.25	0.14	31.47	0.24	0.05	21.21
70	0.52	0.63	71.81	0.52	0.48	43.62	0.52	0.35	26.70
80	0.85	1.06	85.21	0.84	0.88	58.14	0.84	0.79	44.31
90	1.30	1.59	90.01	1.29	1.41	69.39	1.28	1.34	60.63
95	1.68	2.04	92.10	1.66	1.81	72.89	1.64	1.78	70.37
96	1.80	2.20	93.73	1.76	1.93	73.44	1.74	1.96	77.00
97	1.94	2.44	95.97	1.90	2.19	84.36	1.87	2.22	85.47
98	2.13	2.72	97.29	2.08	2.47	89.30	2.04	2.37	83.72
99	2.45	3.03	96.66	2.36	2.83	90.95	2.31	2.97	94.63
4-Factor Gross Returns									
1	-2.55	-3.06	5.49	-2.47	-3.02	6.72	-2.40	-3.34	1.67
2	-2.20	-2.71	4.99	-2.14	-2.63	7.84	-2.09	-2.48	14.14
3	-2.00	-2.46	5.46	-1.95	-2.43	7.33	-1.91	-2.40	8.43
4	-1.85	-2.27	6.39	-1.80	-2.33	5.73	-1.77	-2.25	8.66
5	-1.73	-2.11	7.71	-1.69	-2.12	8.62	-1.66	-2.11	9.52
10	-1.33	-1.62	12.27	-1.30	-1.71	8.63	-1.29	-1.85	4.69
20	-0.86	-1.09	16.23	-0.85	-1.19	11.13	-0.84	-1.34	5.29
30	-0.53	-0.65	28.46	-0.52	-0.75	19.76	-0.52	-0.92	8.75
40	-0.25	-0.33	35.43	-0.25	-0.45	22.31	-0.25	-0.57	12.54
50	0.01	-0.02	44.53	0.01	-0.16	26.29	0.01	-0.29	14.40
60	0.26	0.28	53.17	0.27	0.09	25.86	0.26	0.05	22.48
70	0.54	0.62	64.90	0.54	0.48	43.11	0.54	0.36	27.78
80	0.87	0.98	70.19	0.86	0.85	50.07	0.86	0.82	47.07
90	1.33	1.58	84.76	1.31	1.36	58.66	1.30	1.41	65.72
95	1.72	2.05	88.77	1.69	1.87	73.81	1.67	1.83	70.55
96	1.84	2.21	91.03	1.80	2.01	76.27	1.78	1.95	70.91
97	1.99	2.39	92.01	1.94	2.21	81.22	1.91	2.04	66.61
98	2.19	2.58	91.20	2.12	2.43	83.35	2.08	2.30	74.26
99	2.52	3.01	93.44	2.42	2.72	81.41	2.36	2.57	71.98

Table IV - Percentiles of $t(\alpha)$ estimates for actual and for simulated gross fund returns with injected α

The table shows values of $t(\alpha)$ at selected percentiles (Pct) of the distribution of $t(\alpha)$ estimates for Actual gross fund returns (repeated from Table 3). The table also shows the average values of the $t(\alpha)$ estimates at the same percentiles from the 10,000 simulations, for seven values of σ (the annual standard deviation of injected α). The final seven columns of the table show, for each value of σ, the percents of the 10,000 simulation runs that produce lower $t(\alpha)$ estimates at the selected percentiles than actual fund returns. The period is January 1984 to September 2006 and results are shown for the three- and four-factor models for the \$5 million and \$1 billion AUM fund groups.

	Actual	Average $t(\alpha)$ from Simulations							Percent of Simulations below Actual						
Pct	$t(\alpha)$	0.50	0.75	1.0	1.25	1.50	1.75	2.00	0.50	0.75	1.0	1.25	1.50	1.75	2.00
3-Factor α, AUM > 5 Million															
1	-3.07	-2.63	-2.78	-2.99	-3.24	-3.54	-3.87	-4.23	7.46	15.74	36.37	69.29	92.13	99.00	99.94
2	-2.68	-2.27	-2.38	-2.52	-2.69	-2.89	-3.10	-3.34	8.03	13.67	26.30	49.41	76.25	93.77	99.04
3	-2.48	-2.06	-2.15	-2.27	-2.40	-2.55	-2.72	-2.91	6.55	10.94	19.32	35.16	57.73	80.85	94.54
4	-2.31	-1.91	-1.99	-2.08	-2.20	-2.33	-2.47	-2.62	6.85	10.63	17.71	30.94	49.77	71.63	88.96
5	-2.19	-1.78	-1.85	-1.94	-2.04	-2.16	-2.28	-2.41	6.36	9.68	15.54	26.25	41.95	61.44	80.67
10	-1.72	-1.37	-1.42	-1.48	-1.55	-1.62	-1.70	-1.78	7.86	10.82	15.39	22.37	32.27	44.75	59.63
90	1.59	1.35	1.40	1.46	1.53	1.60	1.68	1.76	86.35	81.64	74.23	64.06	51.23	36.75	22.61
95	2.04	1.75	1.83	1.92	2.02	2.13	2.26	2.39	88.27	82.46	72.10	56.14	36.57	18.02	5.63
96	2.20	1.87	1.96	2.06	2.17	2.30	2.45	2.60	90.76	85.40	74.75	57.87	36.04	16.06	4.08
97	2.44	2.03	2.12	2.23	2.37	2.53	2.70	2.88	93.73	89.76	80.72	63.35	38.59	15.13	3.39
98	2.72	2.23	2.35	2.49	2.66	2.85	3.07	3.31	95.29	91.75	82.56	61.96	32.18	8.75	1.24
99	3.03	2.58	2.74	2.95	3.20	3.49	3.82	4.18	93.48	85.84	63.90	29.57	5.82	0.41	0.02

	Actual	Average $t(\alpha)$ from Simulations							Percent of Simulations below Actual						
Pct	$t(\alpha)$	0.25	0.50	0.75	1.0	1.25	1.50	1.75	0.25	0.50	0.75	1.0	1.25	1.50	1.75
3-Factor α, AUM > 1 Billion															
1	-3.29	-2.42	-2.54	-2.73	-2.99	-3.31	-3.68	-4.09	2.20	3.63	8.89	22.71	48.69	76.60	92.24
2	-2.70	-2.12	-2.21	-2.34	-2.52	-2.73	-2.98	-3.25	6.27	9.11	15.83	28.78	51.16	75.26	90.50
3	-2.51	-1.94	-2.01	-2.12	-2.27	-2.44	-2.63	-2.84	5.82	8.10	13.21	22.97	39.66	61.21	80.93
4	-2.33	-1.80	-1.87	-1.97	-2.09	-2.23	-2.40	-2.57	6.43	8.75	13.73	22.41	36.16	55.28	74.51
5	-2.18	-1.69	-1.75	-1.84	-1.95	-2.08	-2.22	-2.37	7.42	9.81	14.45	22.65	35.12	51.86	70.08
10	-1.86	-1.32	-1.36	-1.42	-1.49	-1.58	-1.67	-1.77	4.46	5.54	7.88	11.48	17.22	25.15	36.41
90	1.34	1.29	1.34	1.40	1.47	1.55	1.64	1.74	58.48	52.67	43.86	34.00	23.16	13.78	6.93
95	1.78	1.66	1.72	1.81	1.92	2.04	2.18	2.33	67.79	60.84	49.40	35.18	20.70	9.71	3.14
96	1.96	1.77	1.83	1.93	2.05	2.19	2.35	2.53	74.69	68.02	56.45	40.71	24.10	11.07	3.42
97	2.22	1.90	1.97	2.08	2.23	2.39	2.58	2.78	84.12	79.12	68.67	52.24	32.34	14.98	4.56
98	2.37	2.07	2.16	2.29	2.46	2.67	2.90	3.16	82.02	75.12	61.59	41.62	21.00	6.82	1.56
99	2.97	2.35	2.46	2.64	2.88	3.18	3.52	3.90	93.92	90.90	81.49	61.18	32.94	11.99	2.84

Table IV (Continued)

Pct	Actual $t(\alpha)$	Average $t(\alpha)$ from Simulations 0.50	0.75	1.0	1.25	1.50	1.75	2.00	Percent of Simulations below Actual 0.50	0.75	1.0	1.25	1.50	1.75	2.00
4-Factor α, AUM > 5 Million															
1	-3.06	-2.69	-2.85	-3.06	-3.33	-3.63	-3.97	-4.34	10.99	22.26	47.07	78.71	95.82	99.64	99.97
2	-2.71	-2.31	-2.42	-2.57	-2.74	-2.94	-3.16	-3.41	8.36	14.61	28.65	52.04	78.54	94.44	99.25
3	-2.46	-2.09	-2.18	-2.30	-2.44	-2.60	-2.77	-2.96	8.82	14.08	25.32	42.58	66.01	86.35	96.53
4	-2.27	-1.93	-2.01	-2.11	-2.23	-2.36	-2.51	-2.66	9.85	15.06	24.90	39.73	60.10	80.07	93.38
5	-2.11	-1.80	-1.87	-1.96	-2.07	-2.18	-2.31	-2.45	11.46	16.87	26.22	39.83	57.77	76.75	90.48
10	-1.62	-1.38	-1.43	-1.49	-1.56	-1.64	-1.72	-1.80	16.05	21.02	27.97	37.70	49.46	63.12	76.66
90	1.58	1.38	1.43	1.50	1.56	1.64	1.72	1.80	79.81	74.21	66.46	55.96	43.49	30.25	18.17
95	2.05	1.80	1.87	1.96	2.07	2.18	2.31	2.44	83.71	76.60	66.13	50.01	31.99	15.33	4.86
96	2.21	1.92	2.00	2.11	2.22	2.36	2.50	2.66	86.46	79.25	68.28	50.91	30.64	12.88	3.36
97	2.39	2.08	2.17	2.29	2.43	2.59	2.76	2.95	87.10	79.52	66.70	46.44	24.15	7.38	1.23
98	2.58	2.29	2.41	2.55	2.72	2.92	3.14	3.38	85.21	75.29	57.57	32.34	10.71	1.74	0.15
99	3.01	2.66	2.81	3.02	3.28	3.58	3.91	4.27	88.10	75.85	49.98	19.30	3.25	0.19	0.01

Pct	Actual $t(\alpha)$	Average $t(\alpha)$ from Simulations 0.25	0.50	0.75	1.0	1.25	1.50	1.75	Percent of Simulations below Actual 0.25	0.50	0.75	1.0	1.25	1.50	1.75
4-Factor α, AUM > 1 Billion															
1	-3.34	-2.44	-2.56	-2.76	-3.03	-3.36	-3.74	-4.16	2.00	3.35	8.55	22.91	48.31	75.80	92.22
2	-2.48	-2.12	-2.22	-2.36	-2.54	-2.77	-3.02	-3.30	16.25	22.21	35.17	54.99	75.98	91.84	97.93
3	-2.40	-1.93	-2.01	-2.13	-2.28	-2.46	-2.66	-2.88	9.63	13.33	20.63	34.25	54.13	74.16	89.62
4	-2.25	-1.80	-1.87	-1.97	-2.10	-2.25	-2.42	-2.60	9.86	13.41	19.68	30.72	47.91	66.77	83.79
5	-2.11	-1.68	-1.75	-1.84	-1.96	-2.09	-2.24	-2.40	10.76	13.64	19.81	29.76	44.52	62.13	78.56
10	-1.85	-1.30	-1.35	-1.41	-1.49	-1.58	-1.67	-1.78	5.10	6.48	8.67	12.86	19.05	27.52	38.83
90	1.41	1.32	1.36	1.43	1.50	1.59	1.69	1.79	63.74	58.42	50.10	40.62	29.70	19.17	10.28
95	1.83	1.69	1.76	1.85	1.96	2.09	2.24	2.40	68.10	61.80	50.88	37.12	22.46	10.67	3.71
96	1.95	1.80	1.87	1.97	2.10	2.25	2.41	2.59	68.50	61.33	49.38	34.72	19.04	7.98	2.30
97	2.04	1.94	2.02	2.13	2.28	2.45	2.64	2.86	64.06	55.68	41.85	26.04	11.89	3.83	0.74
98	2.30	2.12	2.21	2.35	2.53	2.74	2.98	3.25	71.76	62.55	47.39	28.62	11.89	3.45	0.70
99	2.57	2.40	2.52	2.71	2.96	3.26	3.61	4.00	68.67	57.31	38.54	18.10	5.17	0.98	0.11

Table V – Percentiles of four-factor $t(\alpha)$ for actual and simulated fund returns: 1975 to 2002

The table shows values of four-factor $t(\alpha)$ at selected percentiles (Pct) of the distribution of $t(\alpha)$ for actual (Act) net and gross fund returns for funds selected using the exclusion rules of Kosowski et al. (2006) and for funds in our $5 million AUM group selected using our exclusion rules. The period is 1975 to 2002 (as in Kosowski et al. (2006)). The table also shows the fractions (%<Act) of the 10,000 simulation runs that produce lower values of $t(\alpha)$ at the selected percentiles than those observed for actual fund returns. Sim is the average value of $t(\alpha)$ at the selected percentiles from the simulations.

	Kosowski et al Exclusion Rules			Our Exclusion Rules		
Pct	Sim	Act	%<Act	Sim	Act	%<Act
1	-2.48	-3.69	0.18	-2.46	-3.70	0.16
2	-2.16	-3.25	0.19	-2.14	-3.17	0.30
3	-1.96	-2.87	0.53	-1.95	-2.80	0.70
4	-1.82	-2.55	1.34	-1.80	-2.63	0.69
5	-1.70	-2.36	1.90	-1.69	-2.41	1.36
10	-1.31	-1.92	2.17	-1.30	-1.95	1.66
20	-0.85	-1.41	2.15	-0.85	-1.41	2.17
30	-0.52	-1.01	3.18	-0.52	-1.00	3.54
40	-0.25	-0.65	5.75	-0.24	-0.66	5.35
50	0.01	-0.33	9.19	0.01	-0.34	8.50
60	0.27	-0.02	12.20	0.27	-0.03	11.92
70	0.55	0.29	16.51	0.55	0.27	14.86
80	0.87	0.73	32.80	0.87	0.69	28.11
90	1.32	1.44	68.19	1.32	1.34	56.29
95	1.69	1.97	82.42	1.69	1.81	68.32
96	1.80	2.18	88.38	1.80	2.00	75.70
97	1.94	2.38	90.73	1.94	2.25	83.74
98	2.12	2.59	91.38	2.12	2.51	87.57
99	2.40	3.07	95.79	2.42	2.83	88.37

Table AI - Percentiles of CAPM $t(\alpha)$ estimates for actual and simulated fund returns

The table shows values of $t(\alpha)$ at selected percentiles (Pct) of the distribution of CAPM $t(\alpha)$ estimates for actual (Act) net and gross fund returns. The table also shows the percent of the 10,000 simulation runs that produce lower values of $t(\alpha)$ at the selected percentiles than those observed for actual fund returns (%<Act). Sim is the average value of $t(\alpha)$ at the selected percentiles from the simulations. The period is January 1984 to September 2006 and results are shown for the \$5 million, \$250 million, and \$1 billion AUM fund groups.

	5 Million			250 Million			1 Billion		
Pct	Sim	Act	%<Act	Sim	Act	%<Act	Sim	Act	%<Act
				Net Returns					
1	-2.36	-3.72	0.25	-2.30	-3.70	0.40	-2.27	-4.10	0.08
2	-2.06	-3.28	0.45	-2.02	-3.29	0.58	-2.00	-3.50	0.24
3	-1.88	-3.00	0.64	-1.85	-3.02	0.79	-1.84	-3.29	0.23
4	-1.75	-2.84	0.62	-1.72	-2.92	0.65	-1.71	-3.18	0.19
5	-1.65	-2.69	0.74	-1.62	-2.76	0.77	-1.62	-3.00	0.27
10	-1.29	-2.16	1.08	-1.28	-2.18	1.64	-1.28	-2.47	0.46
20	-0.86	-1.48	1.93	-0.86	-1.58	1.98	-0.87	-1.79	0.70
30	-0.54	-1.05	2.09	-0.55	-1.11	2.30	-0.56	-1.35	0.44
40	-0.26	-0.65	3.84	-0.27	-0.75	2.50	-0.28	-0.88	0.48
50	0.00	-0.29	8.05	0.00	-0.36	5.23	-0.01	-0.46	1.29
60	0.26	0.08	20.79	0.26	0.06	19.86	0.26	-0.10	4.02
70	0.53	0.49	46.40	0.53	0.47	43.16	0.54	0.31	18.52
80	0.84	0.95	71.01	0.84	0.89	61.89	0.84	0.72	36.21
90	1.26	1.66	91.09	1.25	1.49	79.61	1.24	1.42	73.88
95	1.61	2.31	97.29	1.58	2.09	92.39	1.56	1.91	84.74
96	1.71	2.45	97.55	1.67	2.23	93.43	1.66	2.03	85.94
97	1.84	2.68	98.46	1.79	2.43	95.05	1.77	2.22	89.01
98	2.01	2.89	98.69	1.95	2.60	95.07	1.92	2.47	92.06
99	2.29	3.21	98.88	2.21	2.96	96.51	2.16	2.76	92.96
				Gross Returns					
1	-2.36	-3.04	4.09	-2.30	-3.01	5.35	-2.27	-3.29	2.00
2	-2.06	-2.66	5.29	-2.02	-2.67	6.32	-2.00	-2.93	2.57
3	-1.88	-2.45	5.88	-1.85	-2.45	7.17	-1.84	-2.76	2.37
4	-1.75	-2.26	7.41	-1.72	-2.31	7.54	-1.71	-2.49	3.99
5	-1.65	-2.13	7.82	-1.62	-2.16	8.80	-1.62	-2.34	4.91
10	-1.29	-1.65	11.87	-1.28	-1.66	13.56	-1.28	-1.95	4.93
20	-0.86	-0.95	33.12	-0.86	-1.04	25.14	-0.87	-1.35	7.59
30	-0.54	-0.55	44.63	-0.55	-0.63	35.49	-0.56	-0.88	12.00
40	-0.26	-0.19	62.18	-0.27	-0.26	50.41	-0.28	-0.43	24.27
50	0.00	0.16	77.76	0.00	0.10	67.74	-0.01	-0.05	41.74
60	0.26	0.53	89.27	0.26	0.46	81.45	0.26	0.36	67.57
70	0.53	0.98	96.44	0.53	0.91	91.87	0.54	0.77	82.64
80	0.84	1.44	97.60	0.84	1.37	95.08	0.84	1.18	87.03
90	1.26	2.12	98.96	1.25	1.98	97.29	1.24	1.82	94.23
95	1.61	2.76	99.65	1.58	2.47	98.14	1.56	2.33	96.87
96	1.71	2.89	99.69	1.67	2.72	98.98	1.66	2.46	97.14
97	1.84	3.12	99.77	1.79	2.85	99.01	1.77	2.59	97.16
98	2.01	3.35	99.84	1.95	3.05	99.18	1.92	2.84	98.03
99	2.29	3.72	99.89	2.21	3.37	99.35	2.16	3.34	99.14

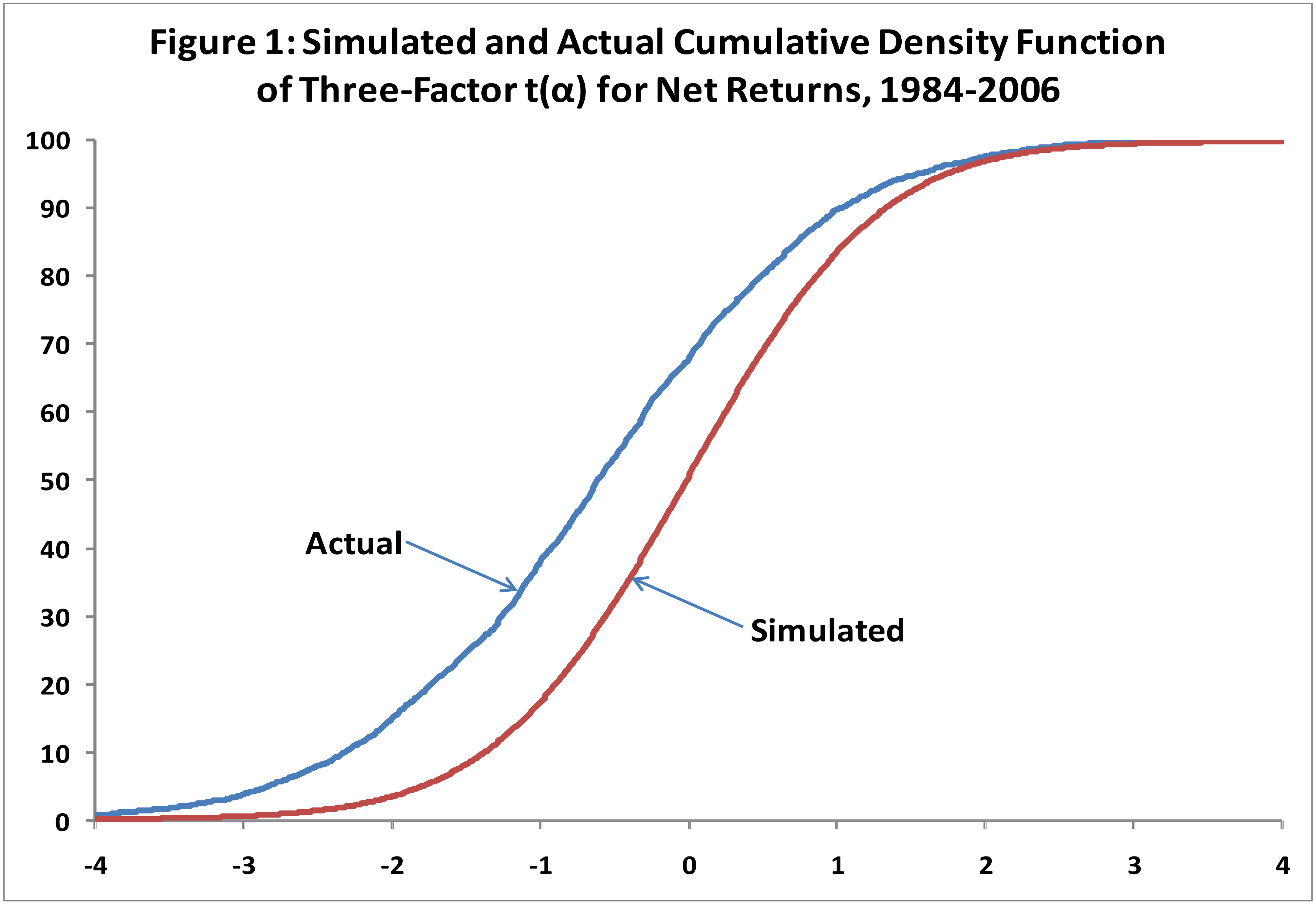


Figure 1: Simulated and Actual Cumulative Density Function
of Three-Factor t(α) for Net Returns, 1984-2006
100
90
80
70
60
50
40
30
20
10
0
-4
-3
-2
-1
0
1
2
3
4
Actual
Simulated

Figure 2: Simulated and Actual Cumulative Density Function of Three-Factor t(α) for Gross Returns, 1984-2006