MUTUAL FUND PERFORMANCE: MEASUREMENT AND EVIDENCE

Keith Cuthbertson*, Dirk Nitzsche* and Niall O’Sullivan**

This version February 2010

Abstract:

The paper provides a critical review of empirical findings on the performance of mutual funds, mainly for the US and UK. Ex-post, there are around 0-5% of top performing UK and US equity mutual funds with truly positive-alpha performance (after fees) and around 20% of funds that have truly poor alpha performance, with about 75% of active funds which are effectively zero-alpha funds. Key drivers of relative performance are, load fees, expenses and turnover. There is little evidence of successful market timing. Evidence suggests past winner funds persist, when rebalancing is frequent (i.e. less than one year) and when using sophisticated sorting rules (e.g. Bayesian approaches) - but transactions costs (load and advisory fees) imply that economic gains to investors from winner funds may be marginal. The US evidence clearly supports the view that past loser funds remain losers. Broadly speaking results for bond mutual funds are similar to those for equity funds. Sensible advice for most investors would be to hold low cost index funds and avoid holding past ‘active’ loser funds. Only sophisticated investors should pursue an active ex-ante investment strategy of trying to pick winners - and then with much caution.

Keywords: Mutual fund performance, persistence, smart money.
JEL Classification: C15, G11, C14.

* Cass Business School, City University, London, UK
** Department of Economics, University College Cork, Ireland

Corresponding Author:
Professor Keith Cuthbertson, Cass Business School, 106 Bunhill Row, London, EC1Y 8TZ. Tel.: +44-(0)-20-7040-5070, Fax: +44-(0)-20-7040-8881, E-mail: k.cuthbertson@city.ac.uk.

We thank David Barr, Don Bredin, Ales Cerny, Liam Gallagher, Anil Keswani, Ian Marsh, Joelle Mifre, Michael Moore, Mark Taylor, Lorenzo Trapani, Giovanni Urga and seminar participants at University of Bari, University of Newcastle, Trinity College Dublin (INFINITI, conference), University of Venice (Ca Foscarini), Barclays Global Investors (London), Investment Managers Association Conference: “Challenges for Fund Management” (London), European Financial Management Association (Madrid), AFFI conference (Paris), QASS conference (Samos), FMA/EDHEC conference on fund performance (Nice), Fund Management Research conference (London) for discussions and comments.

1 An earlier version of this paper appeared with the title “Mutual Fund Performance” and focussed on empirical results. This updated version incorporates an analysis of theoretical as well as empirical issues.
MUTUAL FUND PERFORMANCE: MEASUREMENT AND EVIDENCE

CONTENTS

1. INTRODUCTION

2. PERFORMANCE
   2.1 Industry Models
   2.2 Performance Measures
       A. Factor Models
       B. Characteristic Based Measures

3. EX POST PERFORMANCE: EVIDENCE
   3.1 Measuring Returns
   3.2 Size, Power and False Discoveries
   3.3 Data Issues
   3.4 US and UK Results
   3.5 Luck and Performance: A Tale of Two Tails
   3.6 Market Timing

4. PICKING WINNERS: PERSISTENCE

5. PERSISTENCE: EVIDENCE
   5.1 Statistical Measures
   5.2 Recursive Portfolio Method
   5.3 Stock Holdings And Trade Data
   5.4 Other Sources of Persistence

6. FUND FLOWS AND PERFORMANCE
   6.1 Performance and Fund Characteristics
   6.2 Investment Flows and Performance: Is Money Smart?
   6.3 Mutual Fund Managers

7. CONCLUSIONS
1. INTRODUCTION

The mutual fund industry in the USA and UK has increased dramatically over the last 30 years and now accounts for a substantial amount of private sector saving and substantial new net inflows of saving into risky financial assets. For example, at end-2005 US mutual funds held $8.9 trillion in total fund assets (about half of the world’s fund assets) across a total of around 8,500 funds - with nearly half of US households holding funds comprising around 20% of their total financial assets (Investment Company Institute 2006).

Mutual funds are pooled investments which provide liquidity and enable investors to enjoy economies of scale in gaining access to well diversified portfolios of securities which are often differentiated by funds styles such as aggressive growth, growth and income, growth, equity-income and small companies\(^2\). Most funds are ‘active’ in that they either try to pick ‘winner stocks’ or they engage in market timing (i.e. predicting relative returns of broad asset classes) and these managed funds generally charge higher fees than ‘index’ or ‘tracker’ funds (which mimic movements in broad market indexes)\(^3\). In the US and UK about 70% of institutional funds are actively managed and this rises to over 90% for retail funds. Not surprisingly, the mutual fund industry is larger in countries with strong rules, laws and regulation, where the population is more educated and wealthier and where defined contribution pension plans are more prevalent (Khorana, Servaes and Tufano 2005).

The rationale for managed funds is that they “add value” by using private information and manager skill to produce “abnormal performance”. Two aspects of fund performance are of key importance. First, whether managed funds have an (ex-post) abnormal performance which is positive and whether this outperformance accrues to the investor\(^4\). Here the underlying issues are the popular notion of ranking funds as ‘star funds’ or the common practice of counting funds

---

\(^2\) In the UK mutual funds are often referred to as Unit Trusts although their correct designation is Open Ended Investment Companies, OEIC. In the US ‘unit trusts’ purchase assets but do not subsequently trade them. In the US ‘self declared’ fund styles are overseen by the SEC but it is not always the case that the style name accurately represents the underlying assets in the portfolio. The SEC rules mandate that a fund name must imply that it has at least 80% of their assets in securities of this type/name but there is much leeway in interpretation of the rule. Morningstar, the Thompson/CDA-Spectrum files and the CRSP mutual fund files have somewhat different investment categories from each other – so allocation to a particular category requires some judgment – Wermers (2003a). We use the terms ‘category’ and ‘style’ interchangeably.

\(^3\) With the recent appearance of Exchange Traded Funds ETFs, investors may also ‘track’ a diversified position in a given style category (e.g. small stocks, telecom stocks). ETFs are also redeemable at market value at any time of the trading day (and for example, not just at 4pm New York time as for US mutual funds) and ETFs often have special tax privileges.

\(^4\) Taxes on capital gains and dividend disbursements also influence the return to investors, although lack of data on individuals’ tax liabilities makes any adjustments difficult - so most studies use pre-tax returns.
which are statistically significant to give some indication of the performance of the managed fund industry as a whole. In fact, the mutual fund literature provides excellent examples of the application of recent statistical advances such as separating skill from luck when funds are ranked by performance (i.e. the theory of order statistics), the possibility of “false discoveries” and the more familiar problems of size, power and other data issues in applied work – this should be of interest to economists generally and not just specialists in the mutual fund area. The second key issue is whether abnormal fund performance can be identified ex-ante and for how long it persists in the future. Note that a fund can be beneficial to long-term investors even if it does not have persistent abnormal performance relative to other funds - it may simply implement successful stock picking strategies at infrequent intervals (thus also saving on transactions costs). However, if fund returns are predictable then it may be possible for investors to re-allocate their savings towards ‘winner funds’ and enhance their abnormal returns (relative to a passive index strategy) – in short, “money may be smart”. Any abnormal returns to investors must outweigh any switching costs between funds – these include search costs, load fees and any advisory fees. As the number of possible rules for predicting fund returns is very large then issues of data mining and data snooping come to the fore. We also want to ascertain whether the smart money effect yields predictions consistent with theoretical models of industry behavior.

Given statistical problems in assessing performance, it may be the case that “other information” outside that of the particular fund’s returns and asset holdings might be useful assessing the performance – this has recently been explored using various Bayesian approaches. Finally we need to determine whether returns to active management accrue predominately to investors or to the fund (managers) – how far is the difference between gross and net returns due to fund characteristics such as turnover, advertising costs, management expenses, load fees, the payment of incentive fees or the extent of managerial ownership of the funds?

Measures of ex-post performance and predictability in performance of mutual funds can be interpreted as direct tests of the Efficient Markets Hypothesis (EMH) in a market where entry barriers are relatively low, there are many professional traders who operate in a competitive environment and information is available at relatively low cost – precisely the conditions under which the EMH is expected to be valid. Hence examining mutual funds provides a way of testing the behavior of investors against the classic paradigm of finance theory where individuals are assumed to make rational decisions in relatively frictionless and low information cost markets, which leads to the elimination of inferior financial products and the growth in successful ones.
Evidence on mutual fund performance also throws light on aspects of the behavioral finance literature versus 'standard' asset pricing models. The behavioral finance literature (see Barberis and Thaler 2003 for a survey) has provided theoretical models and empirical evidence which suggests that active stock picking 'styles' such as value-growth (LaPorta et al 1997, Chan et al 1996, Chan and Lakonishok 2004) and momentum (Jegadeesh and Titman 1993, 2001, Chan et al 2000, Hon and Tonks 2003), as well as market timing strategies (Pesaran and Timmermann 1994, 1995, 2000, Ang and Bekaert 2007) can earn abnormal returns after correcting for risk and transactions costs. Large sections of the managed fund industry follow active strategies and there is the possibility that managed funds may be able to take advantage of less informed noise traders. Studies of predictability/persistence in fund performance therefore provide additional evidence on whether anomalies are pervasive enough to influence the performance of a substantial part of the market.

The main aim of this paper is to provide a critical review of empirical findings on the performance of mutual funds, concentrating on US and UK studies and on the literature published over the last 15 years where innovation and data advances have been most marked. While this paper does not aim to provide a detailed critique of the theory behind asset pricing models or of econometric techniques in performance measurement, we do address these issues where they affect the interpretation of empirical results. In short, we critically appraise rival theoretical, methodological and practical issues in evaluating mutual fund performance, place particular emphasis on providing quantitative (rather than just qualitative) results and on standardizing results across studies (where possible) - the reader can then more easily compare results across somewhat diverse approaches.

The rest of this article is organized as follows. In section 2.1 we discuss models of the industry as a whole. Issues in interpreting performance measures form the centerpiece of section 2.2, where we critically appraise the theoretical basis for commonly used measures of abnormal performance such as alpha, the Sharpe ratio and various non-parametric measures. Abnormal performance is often discussed in terms of "security selection" and "market timing" and we discuss the theoretical requirements for separately identifying these two aspects of performance – that is, performance attribution. Having carefully set out the theoretical basis for performance measures, we proceed in section 3 to discuss empirical issues which are examined with reference to US and UK studies. These include biases in performance statistics due to data

Lehmann and Timmermann (2007) provide a comprehensive survey of theoretical issues in performance measurement, which we draw on for this paper. Space constraints imply we cannot survey all the empirical literature on other countries – indeed many countries have little reliable mutual fund (MF) data or the MF industry is relatively undeveloped. While there are some useful academic surveys on the performance of the mutual fund industry these are mainly in books and now require updating – see for example, Friend, Blume and Crockett (1970), Sirri and Tufano (1993), Grinblatt and Titman (1995), Pozen (1998) and Bogle (1999).
deficiencies, issues of the size and power of tests, the problems of non-normal specific risks which may be correlated across funds (e.g. due to herding) and the possibility of “false discoveries” amongst winner and loser funds. Section 4 analyses different approaches to measuring persistence while in section 5 we examine the empirical evidence. In section 6 we analyze the relationship between fund characteristics, fund flows and performance and the question of whether ‘money is smart’. In section 7 we present our conclusions.

2. PERFORMANCE

In the first part of this section we discuss theoretical models of the mutual fund industry and their implications for performance and performance persistence. Then we examine the theoretical basis of risk adjusted performance measures.

2.1 INDUSTRY MODELS

Is it reasonable to expect some funds to under or over-perform their benchmarks over very long horizons or that investors can use ex-ante trading rules which result in positive and persistent abnormal returns? In short, are there models which provide a rationale for ‘smart money’ behavior in which investors can ‘beat the market’. In a seminal article Grossman and Stiglitz (1980) argue that in equilibrium, expected abnormal returns should not be zero, otherwise there would be no incentive to gather and process costly information. Taking up the idea that information processing is costly, Berk and Green (2004) use a general equilibrium competitive model (with no moral hazard or asymmetric information) to analyze fund flows, ex-post returns and performance persistence. The model is very similar to the standard perfectly competitive model of the firm where lower costs are followed by decreasing returns to scale.

In the Berk and Green (2004) model, low barriers to entry ensure that any short-term abnormal profits (e.g. due to lower costs or manager skill) are quickly competed away. Hence fund managers are skillful, but in equilibrium firms earn zero abnormal profits and past performance cannot be used either to predict future performance or to infer the average skill level.

of managers. Fund managers do have differential skill (e.g. stock picking ability) at a gross return level, which (Bayesian) investors learn about based on past returns. Investors chase recent past winner funds who then attract disproportionate cash inflows. The performance-flow relationship is endogenous in the model and is nonlinear (convex) because the cost function is non-linear – this is consistent with empirical evidence presented in section 6.2 and with evidence that new funds are more likely to be started after high inflows either into particular fund families or into all funds as a whole (Khorana 1996). In the model net cash inflows are subject to diminishing returns, so successful funds find it ever more difficult to earn abnormal net returns - any profits are therefore short-lived and in equilibrium funds earn zero abnormal returns. Evidence supporting economies of scale is found by Khorana et al (2008) who show that larger (net asset value) funds and funds which belong to large fund complexes, have lower fees. But Pollet and Wilson (2008) provide evidence that diminishing returns to scale may arise from liquidity constraints (and possible adverse price impact costs), since fund managers overwhelmingly respond to asset growth by increasing their “ownership shares” rather than adding shares of companies not yet in their portfolios. Further evidence on diseconomies is provided by Edelen, Evans and Kadlec (2007) who directly estimate trade costs (commissions, spreads and price impact), which adversely affect performance.

Diseconomies of scale might also eventually ensue if high fund inflows lead to relatively large “liquidity purchases” by managers rather than discretionary stock purchases based on private information or, if pricing anomalies gradually disappear as managers attempt to exploit them. We discuss the former in section 5.2 and evidence on the latter can be found in the anomalies literature (Barberis and Thaler 2003).

In Berk and Green (2004), managers aim at all times to maximize fees but given the performance-flow relationship this involves maximizing expected returns\(^7\). The manager faces increasing marginal costs on actively managed funds which establishes an optimal size for managed funds with any excess invested in an index fund (at zero cost). As fees are also paid on index funds, this is how the active managers extract their rents. Indeed, it is the response of fund flows to past returns which facilitate the competitive process, but as it is the fund managers who are skilled, then they (and not the investors) expropriate the economic rents. However, in equilibrium neither the size of the fund nor the manager’s ability help predict investor returns (which are net of management fees). When investors do not know who the skilled managers are, they infer this from past returns (in a Bayesian framework), so fund inflows are generated by

\(^7\) This rules out dynamic strategies such as managers with good (bad) performance early in the year, reducing (increasing) risk later in the year (Brown, Harlow and Starks 1996, Chevalier and Ellison 1999a). However, gaming of performance measures could result in such an outcome (Goetzmann et al 2007).
rational investors but it is this (plus diminishing returns), which makes future returns to investors unpredictable.

A key prediction of the model is that extremely good performance should result in large fund inflows, while poorly performing funds experience relatively smaller outflows (i.e. the performance-flow relationship is non-linear). However, funds with very large losses should close down. It is these competitive forces that ensure that both 'good' and 'bad' performance does not persist in equilibrium.

Research which looks at the determinants of fund fees, funds starts and the size of funds suggests that we would not expect to see substantial differences in the degree of competition or barriers to entry in the US and UK fund markets (Khorana, Servaes and Tufano 2006, Otten and Schweitzer 2002). The main differences seem to be somewhat stronger investor protection in the US (i.e. class action suits in the US versus caveat emptor in the UK) and differential tax treatment of realized capital gains by the fund.

The model of Lynch and Musto (2003) shares some common assumptions with Berk and Green (2004), since investors learn about managers’ abilities based on past returns. But in the Lynch and Musto model exogenous differences in ability lead to differences in performance persistence, because there are no diminishing returns to funds which experience cash inflows. The model does not seek to explain persistence per se but it is the effect of differential persistence between past winner and loser funds which leads to a convex performance-flow relationship. The decision about changing strategy is the key feature of the model and is assumed to be more likely for past losers than past winners. If past winners persist (because their successful strategy does not change), then we would expect past performance to lead to large inflows into winner funds. Similarly if past losers persist then we expect to see cash outflows. But why do severe past losers barely experience lower future outflows than moderate past losers? It is the latter that the model focuses on.

Lynch and Musto use a two period model and when time-1 fund return is low the manager is likely to switch strategies and this decreases the predictive power of past returns for future returns - hence the model predicts past poor performing funds exhibit weak persistence – although one would expect that with no change in strategy and persistent poor performance, such funds would eventually experience substantial outflows.

These two key models of the flow-performance relationship give clear but different predictions. In both models managers have differential skill, which investors infer from past
performance. But in the Berk and Green model, the average ex-post abnormal fund performance (net of costs) is zero and performance persistence (of both past winners and losers) is weak or non existent in equilibrium. In the Lynch and Musto model the performance of past winners persists (because they do not change their successful strategy) but past losers experience less persistence. For both models two key questions remain. First, how long does any disequilibrium and hence (good or bad) abnormal performance last? Second, is abnormal performance economically significant and exploitable by either fund managers or investors - these are empirical questions to which much of this survey is devoted.

The equilibrium model of Nanda, Narayanan and Warther (2000) is somewhat different to the above models since it is more concerned with explaining the co-existence of load and no-load funds. It is based on the idea that some (but not all) investors create systematic liquidity costs for funds – the latter can be reduced by funds charging loads. Managers whose ability is revealed to be high (i.e. those with skill) form load funds - but in order to attract investors with low liquidity needs they share some of the rents by charging lower management fees, so that investors in funds with load fees earn positive expected returns in equilibrium. However, managers whose ability is revealed to be low, form no-load funds that attract investors with high liquidity needs. The model therefore has two key predictions for the performance-cost relationship. First, investor returns in load funds exceed those in no-load funds and second, investors in funds with high-load fees will earn a higher return than those in low-load funds – we investigate the empirical evidence for these predictions in section 6.1.

The above theoretical models seek to explain some of the stylized facts of the performance-flow relationship and the strength of short-run and long-run abnormal performance of past winner and loser funds within the fund industry as a whole. We now turn to the rationale for various commonly used measures of abnormal performance.

### 2.2 PERFORMANCE MEASURES

There are strong links between the asset pricing and performance evaluation literature - abnormal performance in the performance evaluation literature can be interpreted as mispricing in the asset pricing literature. In addition, equilibrium portfolio holdings in the asset pricing literature can provide a benchmark portfolio against which abnormal fund performance is measured. For

---

8 The partial equilibrium model of Huang, Wei and Yan (2007) also has investors learning about fund performance in a Bayesian framework as in Berk and Green (2004) and this results in differential fund inflows based on past performance. But Huang et al (2007) also include “participation costs” (i.e. costs of search, learning and fees) in the model which apply to “new” or “switching” investors. The slope of the performance-flow relationship then depends on the level of participation costs. But the model does not seek to explain fund performance and persistence, only non-linearities in the flow-performance relationship – see section 6.2.
example, in the static mean-variance portfolio model the “optimal” market (tangency) portfolio provides the benchmark return in the CAPM (with a zero unconditional Jensen’s alpha). If investors are sceptical about the performance of managed funds, then the market portfolio is a natural benchmark portfolio since all mean-variance investors (on the margin) who have homogenous views about expected returns and the variance-covariance matrix, will hold the market portfolio.

A reasonable way of thinking about the usefulness of managed funds, which fits with the stylised view of the investment process, is that the representative investor starts with a benchmark portfolio based on the view that abnormal returns on assets are zero, given the investor’s (publicly) available information, $I_t$. The question then arises as to whether managed funds can “add-value” by using superior information and skill. If managed funds earn “abnormal returns” (e.g. non-zero alphas), then by combining these funds “optimally” with the market portfolio of assets, a higher Sharpe ratio can be achieved.\(^9\)

Note that this interpretation assumes that in practice most (but not all) marginal investors assume asset prices are determined in an efficient market, but the market for managed funds has an element of inefficiency, since some managed funds are capable of producing abnormal returns (Grossman and Stiglitz 1980). Some may find the distinction between “marginal investors” who use public information and managed funds with private information a little “forced” but given costs of acquiring and processing information, it does not appear an unreasonable working assumption when looking for a rationale for managed funds.

---

\(^9\) This can be achieved by maximising the Sharpe ratio which now depends on the managed funds and the market portfolio and if there is zero contemporaneous correlation across residuals of each of these portfolios, then the weight of the managed funds in the overall portfolio is proportional to the appraisal ratio, $AR_t = \alpha_t / \sigma^2(\varepsilon_t)$ (Elton, Gruber and Blake 1996b). Note that ratio $\alpha / \sigma(\varepsilon)$ is in our view more usually (but not always) referred to as the “Information Ratio” both in the academic literature and especially amongst practitioners (see for example, Waring and Siegel 2003, Grinold and Khan 2000). In fact, both the Appraisal Ratio and the Information Ratio are often defined in different ways by practitioners and academics (and are also sometimes treated as equivalent). We want to distinguish between the two ratios since they have two different functions in the optimal portfolio literature: $AR = \alpha / \sigma^2(\varepsilon)$ determines optimal weights and $IR = \alpha / \sigma(\varepsilon)$ (squared) determines the additional contribution to the (squared) Sharpe ratio when combining the “passive” with the “active portfolio” (Treynor and Black 1973).
The above analysis can be generalised to an intertemporal framework using the stochastic discount factor (SDF) approach\(^{10}\). Assume the investor has an indirect utility function \(V(W_t, x_t)\) which depends on wealth \(W_t\) and a state vector of “other variables” \(x_t\) which may affect the investor’s asset allocation decisions (so that utility may be state dependent and non-separable). The Euler equation (first order condition) for any asset-i is:

\[
E[m_{t+1} | I_t] = 1
\]

where \(m_{t+1} = V'(W_{t+1}, x_{t+1})/V'(W_t, x_t)\) is the intertemporal marginal rate of substitution, \(MRS\).

If there is a riskless asset with return \(R_{f,t+1}\) (known at time-\(t\)) then:

\[
\begin{align*}
E[m_{t+1} | I_t] & = 1/R_{f,t+1} \\
E[m_{t+1}(R_{t+1} - R_{f,t+1}) | I_t] & = 0
\end{align*}
\]

It is also important to note that the Euler equation can also be derived from standard no-arbitrage arguments (see Chen and Knez 1996, Cochrane 2001, Lehmann and Timmermann 2007, Wickens 2008) - hence there are two complementary approaches which provide a rationale for performance measurement of managed funds.

If we now take the conditional population projection of \(m_{t+1}\) on a set of \(N\) risky asset returns, \(R_{t+1}\) (which are not perfectly correlated):

\[
m_{t+1} = \delta_{0t} + \delta_t R_{t+1} + \epsilon_{m,t+1}
\]

then \(\delta_t\) may be interpreted as determining optimal portfolio weights \(w_{it} = \delta_t / \delta_t' i\) (where \(i\) is an \(N\times1\) vector of ones) while benchmark portfolio returns are \(R_{b,t+1} = w_{it}' R_{t+1}\). Here \(\delta_t\) is the optimum portfolio of \(N\) risky assets for hedging fluctuations in the intertemporal \(MRS\) for any marginal investor. The \(N\)-asset excess returns \(r_{t+1} = R_{t+1} - i R_{f,t+1}\) should obey:

\[
E[r_{t+1} | I_t] = \beta_{0t} E[r_{St+1} | I_t]
\]

\(^{10}\) This section draws heavily on Lehmann and Timmermann (2007).
(with zero intercept), since portfolio \( \delta_t \) with excess return \( r_{\delta, t+1} = R_{\delta, t+1} - i R_{f, t+1} \) is a conditionally mean-variance efficient ("passive") portfolio. Portfolio \( \delta_t \) (which does not include managed funds) is a natural benchmark portfolio for an investor who is sceptical about fund performance and if the marginal investor uses only publicly available information then she cannot improve on portfolio \( \delta_t \). However, suppose a managed fund- \( p \) embodies private information which is only available to the investor via the fund. The projection of the managed fund (excess) return \( r_{pt, t+1} \) on the benchmark (excess) return \( r_{\delta, t+1} \) is:

\[
r_{pt, t+1} = \alpha_{pt} + \beta_{pt} r_{\delta, t+1} + \varepsilon_{pt, t+1}
\]

where it is important to note that alpha and beta are conditional on public information available to the investor \( I_t \) (and not any superior private information available to the managed fund). The Euler equation for the marginal investor with respect to the managed fund is

\[
E[m_{t+1} r_{pt+1} | I_t] = 0
\]

where the conditional appraisal ratio is \( AR = \alpha_{pt} / \text{Var}(e_{pt+1} | I_t) \). First note that \( \delta_{pt} \) is non-zero only if the conditional Jensen’s alpha is non-zero. Second, the “size” of \( \delta_{pt} \), which

\[
\delta_{pt} = \frac{E[e_{m, t+1} e_{pt+1} | I_t]}{\text{Var}(e_{pt+1} | I_t)} = \frac{E[m_{t+1} e_{pt+1} | I_t]}{\text{Var}(e_{pt+1} | I_t)} = -\frac{\alpha_{pt}}{R_{f, t+1} \text{Var}(e_{pt+1} | I_t)}
\]

where \( \alpha_{pt} \) is given by

\[
\alpha_{pt} = -R_{f, t+1} E[m_{t+1} e_{pt+1} | I_t]
\]

In the projection of \( m_{t+1} \) on the N-vector \( R_{t+1} \) and (the scalar) \( R_{pt+1} \) the coefficient \( \delta_{pt} \) on \( R_{pt+1} \) is

\[
0 = E[m_{t+1} r_{pt+1} | I_t] = E[m_{t+1} (\alpha_{pt} + \beta_{pt} r_{\delta, t+1} + \varepsilon_{pt+1}) | I_t] = R_{\delta, t+1} \alpha_{pt} + E[m_{t+1} e_{pt+1} | I_t] \quad \text{where} \quad E[m_{t+1} | I_t] = R_{\delta, t+1}^{\top}
\]

and

\[
\beta_{pt} E[m_{t+1} r_{\delta, t+1} | I_t] = 0 \quad \text{from the first order conditions for the investor with respect to the benchmark portfolio.}
\]

\[1\]

\[2\]

\[3\]

Equation \([3]\), \( m_{t+1} = \delta_0 + \delta_{f, t+1} + \varepsilon_{m, t+1} \) and the Euler equation \([2a]\) imply \( E[m_{t+1} | I_t] = R_{\delta, t+1}^{\top} \)

\[
= \delta_0 + \delta E[R_{\delta, t+1} | I_t] \quad \text{hence} \quad \varepsilon_{m, t+1} = m_{t+1} - R_{f, t+1}^{\top} - \delta (R_{\delta, t+1} - E[R_{\delta, t+1} | I_t]) \quad \text{Since the benchmark return is} \quad R_{\delta, t+1} = w_{\delta} R_{f, t+1} \quad \text{then} \quad R_{\delta, t+1} - E[R_{\delta, t+1} | I_t] \quad \text{is orthogonal to} \quad e_{pt+1}, \quad \text{and} \quad E[e_{m, t+1} e_{pt+1} | I_t] = E[m_{t+1} e_{pt+1} | I_t] = -\alpha_{pt} / R_{f, t+1} \quad \text{using} \quad [6].
determines the weight of the managed fund in the portfolio, depends directly on the (conditional) appraisal ratio $AR$ of the fund (Treynor-Black 1973). Also, large values of alpha imply large values of $E[m_{t+1},\epsilon_{p,t+1} \mid I_t]$. In the intertemporal model, a non-zero alpha implies that combining the managed fund with the benchmark portfolio (and the risk-free asset) can yield an improvement in hedging the intertemporal $MRS$ – and this provides a theoretical basis for alpha in an intertemporal model of portfolio choice. Alpha also has a simple intuitive interpretation – it is the return on a zero net investment (and zero-beta) strategy of investing $1$ in the managed portfolio, $1 - \beta_{pt}$ dollars in the risk-free asset and short selling $\beta_{pt}$ dollars of the benchmark portfolio $\delta_t$. It therefore allows investors to attain the mean-variance efficient frontier by augmenting the N-asset portfolio (and the risk-free asset) with the managed fund. (This approach is easily extended to M-managed funds.)

The SDF (and the no-arbitrage) approach makes clear that it is the conditional alpha and beta that we are trying to measure and the benchmark portfolio in this approach is not the market portfolio of all tradeable assets (as in the CAPM) but is the mean-variance efficient portfolio of investors who use only public information.

**BENCHMARK PORTFOLIO**

If the benchmark portfolio chosen in practice really is the conditional mean-variance efficient portfolio $\delta_t$, then we could reasonably describe (the conditional) alpha as the “abnormal performance” of the managed fund. The difficulty is that in practice we cannot be sure we have the mean-variance efficient portfolio. First, because it is virtually impossible to implement the mean-variance problem with more than about 20 assets and second because of the large degree of uncertainty in estimates of the portfolio weights $\delta_t$ (Britten-Jones 1999). Without precise knowledge of the benchmark portfolio our estimated conditional Jensen’s alpha could reflect benchmark error and hence a non-zero alpha might not reflect superior information and skill of the managed fund (Roll 1977). However, we can still fall back on the weaker proposition that a non-zero alpha implies the mean-variance trade-off based on portfolio $\delta_t$ (and the risk-free asset) can be improved by adding the managed fund p.

**A. FACTOR MODELS**

The link between the SDF approach and linear factor models is the assumption that $m_{t+1}$ is an exact linear combination of observable factors:
For some equilibrium models the weights $\psi_{kt}$ are given by the theoretical model (e.g. the CAPM) but for multifactor models these are usually determined empirically. In the Harvey-Siddique (2000) model, the above equation is augmented by a non-linear term in the square of the market return and this leads to a model in which coskewness is priced – assets which make portfolio returns more left-skewed are less desirable and should result in higher expected returns.

If $f_{t+1} = \{f_{t+1}, f_{2t+1}, ..., f_{kt+1}\}'$ are the K-factors and $\beta_{j} = \{\beta_{j1}, \beta_{j2}, ..., \beta_{jk}\}$ are the k-factor loadings then a mutual fund-p at time t, with asset proportions $w_{jt}$ ($j = 1, 2, ..., p$) has a required return (to compensate for factor risk) equal to $\beta_{j}E(f_{t+1} | I_{t})$ and the fund abnormal performance is given by $\alpha_{pt}$ in the regression (Jensen 1968, Lehmann and Modest 1987):

$$r_{pt+1} = \sum_{j=1}^{p}(w_{j}\alpha_{ji} + w_{j}\beta_{j})f_{t+1} + \sum_{j=1}^{p}w_{j}\varepsilon_{jt+1} = \alpha_{pt} + \beta_{pt}f_{t+1} + \varepsilon_{pt+1}$$

where $\varepsilon_{pt} = \sum_{j=1}^{p}w_{j}\varepsilon_{ji}$, $\alpha_{pt} = \sum_{j=1}^{p}(w_{j}\alpha_{ji})$, $\beta_{pt} = \sum_{j=1}^{p}(w_{j}\beta_{j})$. Note that the fund’s parameters will be time varying if either the factor betas or the fund weights are time varying. As we shall see, time variation in factor loadings in the empirical literature has been examined using rolling regressions, loadings depending on observable (macro) variables, switching models and time-varying parameter models based on unobservable factors (Kalman filter).

As in the asset pricing literature, it is impossible in the performance measurement literature to completely rule out mismeasured benchmarks (unless the fund declares an explicit set of observable benchmarks and their changing weights over time). In practice, benchmarks used in the performance measurement literature often depend on empirical results from the asset pricing literature and use factors which seem to price a wide variety of different portfolios.
Fama-French (1992, 1993) find that the market return, ‘size’ (SMB) and ‘book-to-market’ (HML) risk factors\(^{13}\) explain average returns on a wide range of different portfolios of US stocks and they interpret these factors in terms of systematic economic risk. However, the cross-section of average returns on stocks sorted into portfolios on the basis of their prior one-year raw returns are not well explained by the Fama-French 3F model but a one-year ‘momentum’ factor (MOM) captures this ‘anomaly’ (Jegadeesh and Titman 1993, 2001, Rouwenhorst 1998)\(^{14}\). The interpretation of MOM as a risk factor is more tenuous than the other factors mentioned above. However, Harvey and Siddique (2000) show that MOM may be a proxy for conditional skewness as momentum winner portfolios have more negative skew than loser portfolios, while Pastor and Stambaugh (2003) note that a factor mimicking portfolio based on a liquidity measure substantially affects the importance of the momentum variable in stock return regressions. The factors can be interpreted as systematic risk in the CAPM or Arbitrage Pricing Theory APT or as variables which hedge changes in the intertemporal MRS in the SDF approach. However, although the above interpretations are theoretically defensible they might not be foremost in the minds of many investment practitioners.

**STYLE FACTORS**

Practitioners using factor models probably interpret the alpha of the 4F model primarily in terms of “outperformance” relative to mechanical “style factors” that could easily be implemented by individual investors (Sharpe 1992). Investors could form zero investment portfolios based on the excess market return, the HML, SMB and MOM style factors. These passive portfolios have specific exposures to each factor – but the factors need not represent systematic economic risk. A positive 4F-alpha, (net of management fees) for a managed fund then implies superior performance, relative to that attainable by an equal exposure to the factors undertaken by individual investors, using only publicly available information.

**UNCONDITIONAL MODELS**

Empirical studies often assume factor loadings are time invariant and estimate unconditional factor models, one of the most popular being Carhart’s (1997) four factor (4F) model:

\[
 r_{pt+1} = \alpha_p + \beta_{1p} r_{mt+1} + \beta_{2p} \text{SMB}_{t+1} + \beta_{3p} \text{HML}_{t+1} + \beta_{4p} \text{MOM}_{t+1} + \epsilon_{pt+1}
\]

\(^{13}\) The Fama-French size factor SMB (‘small minus big’) is the difference between the returns on a portfolio of small stocks (e.g. smallest third) and large stocks. The book-to-market value factor, HML (‘high minus low’), is a measure of the difference between the returns on high versus low BMV stocks. To construct these series the portfolios are usually rebalanced every year.

\(^{14}\) The MOM variable is the difference in returns between a portfolio of previously high return stocks (e.g. top 1/3\(^{rd}\)) and previously poor return stocks (e.g. bottom 1/3\(^{rd}\)), usually rebalanced annually.
where \( r_m \) is the excess return on the market portfolio, \( SMB \), \( HML \) and \( MOM \) are zero investment factor mimicking portfolios for size, book-to-market value and momentum effects, respectively. If \( \beta_{mp} = 0 \) the model is the Fama-French (1992, 1993) 3F model while Jensen’s (1968) alpha is the intercept from the CAPM one-factor (or market) model.

**MARKET TIMING**

In addition to security selection skills, models of portfolio performance also attempt to identify whether fund managers have the ability to market-time. Can fund managers successfully forecast the future direction of the market (or other factors) and alter the market (factor) beta accordingly (see Admati and Ross 1985, Admati et al 1986)? It is natural to attempt to decompose the managed portfolio return into that due to the benchmark return \( \beta_{pt} r_{\delta t+1} \), known as the return to *market timing* and the remainder \( \alpha_{pt} + \epsilon_{pt+1} \) which is referred to as the return to *security selection*. Separating out these two distinct sources of performance (i.e. *performance attribution*) is extremely difficult. The unconditional Jensen regression is:

\[
[11] 
\begin{align*}
 r_{pt+1} &= \alpha_p + \beta_p r_{\delta t+1} + v_{pt+1} 
\end{align*}
\]

Using [11], substituting from [5] for \( r_{pt+1} \) and taking unconditional expectations:

\[
[12] 
\begin{align*}
 \alpha_p &= E[\alpha_{pt}] + E[(\beta_{pt} - \beta_p) r_{\delta t+1}] = E[\alpha_{pt}] + \text{cov}[\beta_{pt}, r_{\delta t+1}] 
\end{align*}
\]

The unconditional covariance term can then be decomposed into its conditional components so that the unconditional alpha is:

\[
[13] 
\begin{align*}
 \alpha_p &= E[\alpha_{pt}] + \text{cov}[\beta_{pt}, r_{\delta t+1} \mid I_t] + \text{cov}[\beta_{pt}, E r_{\delta t+1} \mid I_t] 
\end{align*}
\]

Successful market timing implies \( \text{cov}[\beta_{pt}, r_{\delta t+1} \mid I_t] \neq 0 \), since *private* information of the managed fund is used to alter beta. But the second term may also be non-zero if the managed fund alters portfolio betas in response to changes in the *expected* return on the benchmark. Hence the unconditional alpha (as well as beta) is unbiased only under rather stringent conditions namely, if benchmark returns are *iid* (unpredictable) and market timing based on private information is unsuccessful (or not attempted). Otherwise, even if benchmark returns are *iid* but market timing is successful then the unconditional alpha reflects both security selection and
market timing – so performance measurement is possible but performance attribution is not (without additional structure in the model). Indeed even if benchmark returns are iid, the first covariance term in [13] can result in a negative unconditional alpha if $\text{cov}[^{\theta}p_{t}, r_{\delta+1} | I_{t}] < 0^{15}$, so market timing may be interpreted as poor security selection (Admati and Ross 1985, Dybvig and Ross 1985).

Modelling market timing (i.e. the $\text{cov}[^{\theta}p_{t}, r_{\delta+1} | I_{t}]$ term), has been attempted both parametrically and non-parametrically, while modelling the $\text{cov}[^{\theta}p_{t}, E[r_{\delta+1} | I_{t}]]$ term often draws on the stock market predictability literature and assumes beta depends on a subset of publicly available information - although “unobserved factor models” (estimated using the Kalman filter) have also recently been used. First we deal with market timing.

Treynor and Mazuy (1966) TM assume a successful market timer adjusts the market factor loading using $^{\theta}p_{t} = \theta_{p} + \gamma_{pm} [r_{m,t+1}]$ while Henriksson and Merton (1981) HM, view market timing as the payoff to a call option on the excess market return $^{\theta}p_{t} = \theta_{p} + \gamma_{pm} [r_{m,t+1}^{+}]$ where $[r_{m,t+1}]^{+} = \max\{0, r_{m,t+1}\}$. Hence, the regression equation for the market model is:

$$r_{p,t+1} = \alpha_{p} + \theta_{p} r_{m,t+1} + \gamma_{pm} f[r_{m,t+1}] + \epsilon_{p,t+1}$$

where $f[r_{m,t+1}] = r_{m,t+1}^{2}$ for the TM model and $f[r_{m,t+1}] = [r_{m,t}]^{+}$ for the HM model and $\gamma_{pm} > 0$ indicates successful market timing.

Separating out the two distinct sources of performance attribution has been attempted by explicit modelling of the signals used by portfolio managers (Admati et al 1986) but it can be shown that the TM regression does not resolve the performance attribution problem unless additional “strong restrictions” are placed on the model (Lehmann and Timmermann 2007).^{16}

### Conditional Models

^{15} This may result from portfolio managers minimising portfolio variance for given unconditional expected returns (Bollen and Busse 2001).

^{16} Admati et al (1986) assume normality of the time-varying beta due to market timing, normality of the residual in the regression of the benchmark return on the time varying beta and use an auxiliary regression of the squared-residual from the T-M regression on the benchmark return and benchmark return squared, in order to separately identify security selection and market timing.
A further attempt to resolve the problem of separating security selection from market timing and to address the issue of time-varying parameters, is to assume that the time varying beta depends partly on a subset of publicly available information $Z_t \subseteq I_t$ so that $\beta_{p,t} \text{ is a linear function of } z_t = Z_t - EZ_t$ and the TM equation becomes (Ferson and Schadt 1996):

$$r_{p,t+1} = \alpha_p + b_{0p} \left( r_{m,t+1} \right) + b_{1p} \left( z_t \cdot r_{m,t+1} \right) + \gamma_{pm} r^2_{m,t+1} + \varepsilon_{p,t+1}$$

Public information $z_t$ as well as the manager’s market timing skill (based on private information) influence beta. Christopherson, Ferson and Glassman (1998) assume that alpha (as well as beta) may depend linearly on $z_t$ and the TM model becomes:

$$r_{p,t+1} = \alpha_{0p} + a_{p} \cdot z_t + b_{0p} \cdot r_{m,t+1} + b_{1p} \left( z_t \cdot r_{m,t+1} \right) + \gamma_{pm} r^2_{m,t+1} + \varepsilon_{p,t+1}$$

The above conditional models are easily generalized to a multifactor framework. Most studies follow Ferson and Schadt (1996) and Christopherson, Ferson and Glassman (1998) where the $z_t$ variables are often taken to include the one-month T-Bill yield, the dividend yield of the market factor and the term spread.

Under the null of no market timing skill, alpha is estimated consistently in these conditional TM equations but unfortunately with successful market timing, alpha is inconsistent. Thus while it appears that conditional factor models are useful in modelling time variation in beta due to public information, nevertheless in the presence of successful market timing it is virtually impossible to separate out timing ability from security selection and alpha will reflect both elements. One method used to sidestep this problem is to measure “overall performance” using

$$r_p = \left( 1/T \right) \sum_{t=1}^{T} (\alpha_{0p} + \gamma_{pm} r_{m,t+1}) \text{ where the summation is over all trading “days” over which the fund parameters have been estimated (Bollen and Busse 2004).}$$

In recent work, Mamaysky et al (2008) adopt a Kalman filter approach to model dynamic factor loadings. Fund portfolio weights and the alphas of individual stocks are assumed to depend on an unobserved factor $F$ – broadly speaking the former is “factor timing” and the latter

---

17 In an attempt to model skill in the presence of market timing Grinblatt and Titman (1989, 1994) use a novel method to estimate alpha without using portfolio weights or a benchmark return - which they called a class period weighting method. In this approach alpha is a weighted linear function of fund returns and the weights correspond to the marginal utility of an investor with a power utility function. But the method requires a measure of the risk aversion parameter and the distinction between market timing and security selection is also not resolved in this approach.
"stock selection". The unobserved factor $F$ is autocorrelated $F_t = vF_{t-1} + \eta_t$. The model can also incorporate portfolio betas depending on observable factors and then if $v = 0$ or $\sigma_\eta = 0$ the model collapses to the above conditional models.

Another conceptual problem with the TM and HM approaches is the inability to distinguish the quality of the manager’s timing information from the aggressiveness of her response – the TM and HM regressions confound these two effects in the $\gamma_{pm}$ parameter. The investor is more concerned with the quality of information as she can control “aggressiveness” by choosing the proportion of her wealth to invest in the fund. Using a non-parametric approach Jiang (2003) provides a test of market timing which can identify the quality of the manager’s market timing ability\footnote{The non-parametric test has good size properties and is robust (to outliers, non-normality, heteroscedasticity, and differences between timing frequency and data frequency) – although it does require serially uncorrelated returns.}. The intuition behind the approach is that the manager should increase beta if she thinks the market will rise in the future. For any two non-overlapping time periods $t_1$ and $t_2$ the statistic $\theta = 2 x \text{prob}(\beta_{t_1} > \beta_{t_2} \mid r_{m,t_{1+1}} > r_{m,t_{2+1}}) - 1$ is greater than zero for successful market timing. With no market timing ability $\beta$ has no correlation with the market return and hence $\text{prob}(\cdot) = 1/2$ and $\theta = 0$ ($\theta < 0$ implies adverse market timing). Hence $\theta$ depends only on how often a manager correctly ranks and appropriately reacts to a market signal, but not on how aggressively she reacts.

Now consider the triplet sampled from any three periods for a fund’s excess returns, $\{r_{p,t_1}, r_{p,t_2}, r_{p,t_3}\}$ where $r_{m,t_3} > r_{m,t_2} > r_{m,t_1}$. A fund manager who has superior market timing ability (regardless of her degree of aggressiveness) should have a higher average beta in the $t_2$ to $t_3$ period, than in the $t_1$ to $t_2$ period. The measured value of beta in these two periods is $\beta_{12} = (r_{p,t_2} - r_{p,t_1})/(r_{m,t_2} - r_{m,t_1})$ and $\beta_{23} = (r_{p,t_3} - r_{p,t_2})/(r_{m,t_3} - r_{m,t_2})$. The sample analogue to $\theta$ is therefore:

$$\hat{\theta}_n = \left(\frac{n}{3}\right)^{-1} \sum_w \text{sign}(\beta_{23} > \beta_{12})$$
where \( w \) represents the triplets in the data where \( w \equiv \{ R_{m,t_1}, R_{m,t_2}, R_{m,t_3} \} \) and \( \text{sign}(.) \) assumes a value of \{1, -1, 0\} if the argument is \{positive, negative, zero\}. Under certain relatively weak assumptions \( \hat{\theta}_n \) is asymptotically \( N(0, \sigma^2_{\theta_n}) \).^{19}

Given the difficulty of performance attribution in parametric models, simulations (based on either TM or HM being true) can throw some light on the relative merits of \( \gamma_{pm} \) and \( \theta \) as measures of market timing (Jiang 2003). For example, consider an information signal \( y_t = r_{it+1} + \eta_{t+1}, \) where \( \sigma_y / \sigma_{\eta} \) represents the quality of the signal and \( \beta_t = \beta_0 + \lambda(r_{m,t+1} - E r_m) \) where the parameter \( \lambda \) measures the strength of the manager’s response and hence the proportion held in the risky portfolio. Simulations show that \( \theta \) is very responsive to the quality of the information signal and relatively invariant with respect to \( \lambda \), whereas \( \gamma_{pm} \) mainly reflects the latter.

A further difficulty in assessing timing ability arises if the frequency of the researcher’s observed data differs from the frequency of the manager’s timing strategy (where the latter may not be uniform or even known) and this introduces bias into \( \gamma_{pm} \) (Goetzmann et al 2000). Ferson and Khang (2002) also show that bias results when expected returns are time varying and managers trade between return observation dates (“interim trading bias”). Simulations confirm these results as \( \gamma_{pm} \) estimated using monthly data is reduced by around two-thirds when returns are generated assuming daily timing in the HM or TM models (Bollen and Busse 2001). However, because the nonparametric measure \( \theta \) does not rely on market returns measured at the “right” frequency, it is quite robust to this “data timing” problem.

Both the parametric and non-parametric approaches rely on selectivity being independent of timing and that the managed fund does not contain derivatives. Jagannathan and Korajczyk (1986) show that buying options or undertaking dynamic trading or if stocks are held passively but they have options like features, can induce spurious timing ability – the latter is known as “passive timing” and all the above features are often referred to as “artificial timing”. However, evidence suggests that mutual funds have not used options strategies extensively (Kosik and Pontiff 1999) although this may be changing in the US and UK in recent years. Simulations show that when selectivity and timing are correlated \( \theta \) provides a more robust measure of the quality

---

of timing information and is influenced less by the aggressiveness of the manager’s response, than are the TM and HM measures\textsuperscript{20}.

**SHARPE RATIO**

The Sharpe ratio is widely used, particularly by investment practitioners and in its conditional form is:

\[
SR_{p_t} = \frac{E[r_{p,t+1} | I_t]}{\sqrt{\text{Var}(R_{p,t+1} | I_t)}}
\]

\(SR_{p_t}\) is the reward-to-variability ratio for a portfolio formed by borrowing at the risk-free rate and investing in the managed portfolio-\(p\). Initially, the Sharpe (1966) ratio was interpreted as the distance of portfolio-\(p\) from the efficient frontier but (as noted above) later work (Jobson and Korkie 1982, Gibbons, Ross and Shanken 1989) showed that this “distance” also depends on portfolio alphas and residual variances and covariances. Hence, we cannot say that portfolio-A is better than portfolio-B if its Sharpe ratio is higher, because the distance from the frontier does not depend solely on the mean and variance of the two competing portfolios (Lehmann and Timmermann 2007). The Sharpe ratio also deals with a mutually exclusive choice between two or more managed portfolios whereas in practice investors are often looking to add a (small) set of managed funds to their existing portfolios and here alpha is more informative. Hence in the academic literature, there is greater emphasis on alpha from multifactor regressions rather than on the Sharpe ratio, although practitioners probably give about equal consideration to both measures\textsuperscript{21}.

Successful market timing also causes difficulties when using the unconditional Sharpe ratio as a measure of performance. First, a successful market timer will have a higher conditional Sharpe ratio (than an uninformed investor) but because a market timer introduces greater volatility in portfolio returns, she may have a lower *unconditional* Sharpe ratio. Second, successful market timing results in managed portfolio returns being non-normal (even if benchmark returns are normal) and therefore the Sharpe ratio may be misleading since it ignores

\[\text{In simulations, selectivity is introduced by randomly dropping 8 (out of a 100 randomly selected) stocks each month from the fund portfolio (i.e. 96% annual turnover rate) which are in the bottom quartile of next period’s returns and adding 8 stocks from the top quartile of next periods returns. For market timing, the weight in the risky portfolio is determined as before by the quality of the signal and the aggressiveness parameter } \lambda \text{ of the fund manager (Jiang 2003).}\]

\[\text{The Sortino ratio (Sortino and van der Meer 1991) is a variant on the Sharpe ratio which just uses the “downside” returns in measuring portfolio variance while the “upside-potential” Sharpe ratio (van der Meer, Plantinga and Forsey 1999) only counts “good” returns in the numerator of the Sharpe ratio - these are both variants on “partial moments”. These measures exhibit the same limitations as the Sharpe ratio, noted above.}\]
skewness and kurtosis - statistical tests are also problematic when returns are non-normal or serially correlated (Lo 2001, 2002).

**DYNAMIC TRADING STRATEGIES**

An additional problem in performance evaluation is the possible use of dynamic trading strategies by managers who manipulate standard performance measures. If portfolio returns are monitored frequently then static manipulation is detectable (Goetzmann, Ingersoll and Ivkovich 2000), but this does not carry over to dynamic portfolio strategies\(^{22}\). For example, a manager who experiences "good luck" in the first half of the performance horizon, with a measured SR that is higher than expected, could reduce leverage so that in the second period she has a lower expected return and standard deviation – this implies she allows past good luck to weight more heavily in the Sharpe ratio measured over the whole evaluation period. (The reverse case also applies - if "bad luck" occurs then she increases leverage in the subsequent part of the evaluation period.) There is no superior private information or "skill" involved here – managers use simple (mechanical) leverage strategies to game the estimated performance measure. Since manipulation of a performance measure relies on leverage, then it can also be accomplished using options\(^{23}\).

Of crucial practical importance is whether such dynamic manipulation survives the transactions costs involved. Goetzmann et al (2007) using simulation, demonstrates that performance manipulation using simple (uninformed) leverage strategies can be gamed to produce statistically and economically significant performance measures even in the presence of high trading costs. For example, with leverage "manipulated" every month (after the first year) but constrained to be between 0.5 and 1.5, the average (unconditional) Jensen’s alpha is 2% p.a. with zero transaction costs and 1.6% p.a. with 20% transactions costs (per round trip) and in the latter case alpha is positive in over 85% of the simulations. Gains from manipulation are a little lower for the TM and HM regressions and for multifactor models. However, such ex-post gains

---

\(^{22}\) Static manipulation is tilting the portfolio away from the (levered) benchmark even when there is no new information. Dynamic manipulation is altering the portfolio over time on the basis of past performance but no new information.

\(^{23}\) Restrictions or covenants to mutual fund activity vary across funds. But in 2000 about 90% of US funds could not buy on margin, around 65% could not short securities, about 25-30% could not trade either index futures or individual options. In addition around 20% were not allowed to borrow money and about 20% of funds could not hold restricted securities (e.g. those obtained by private sale). But of those funds that are not restricted in the above fashion a maximum of only about 12% of funds actually used any one of the above techniques over the 1996-2000 period – and there are no appreciable trends in this figure over this period (Almazan et al 2004). However, with the advent of hedge funds (Stultz 2007), the aggressive marketing of 120/20 mutual funds and the European UCITSIII regulations, it is likely that mutual funds will be less restrictive in their trades in the future.
are fraudulent as they are generated without private information – the quality of which is the basis for much of the fees charged to investors in managed funds.\footnote{Performance measures investigated include the Sharpe, Sortino and “upside-potential” Sharpe ratios, unconditional alpha, and alpha from the TM and HM regressions. Goetzmann et al (2007) also derive a manipulation-proof performance measure – which has an intuitive interpretation in terms of an additively separable power utility function of monthly returns over the performance horizon.}

**B. CHARACTERISTIC BASED MEASURES**

It was noted above that in measuring abnormal fund performance we may require a parametric model of time variation in portfolio weights. However, data on a fund’s asset holdings or trades (i.e. buys/sells) allows one to construct a series for fund returns which accurately reflects the changing weights on the characteristics of the stocks held by the fund (e.g. small/large cap stocks, high/low BMV stocks, momentum stocks). Below we outline 'characteristic based' performance measures which use stock holdings and trade data of funds and then discuss some strengths and weaknesses of the approach.

Grinblatt and Titman (1989) pioneered this approach which was then extended by Daniel, Grinblatt, Titman and Wermers DGTW (1997), whereby each stock held (or traded) by the mutual fund is matched with ‘benchmark stocks’ that have similar characteristics in terms of size, BMV and momentum. The variable CS (“Characteristic Selectivity”) measures a fund’s stock selection ability (“stock picking skills”):

\[
CS_t = \sum_{j=1}^{N} w_{j,t-k} [R_{j,t} - BR_{j,t-k}]
\]

where \( R_{j,t} \) is the return on stock-\( j \) in quarter-\( t \) and \( BR_{j,t-k} \) is the return on the benchmark portfolio in quarter \( t \) to which stock-\( j \) was allocated during period \( t-k \) according to its size, BMV and momentum characteristics. The weight \( w_{j,t-k} \) is the fund’s portfolio weight in stock-\( j \) at the end of quarter \( t-k \).\footnote{When we use a trade data measure of CS, then \( w_{j,t-k} \) is the dollar purchases/sales of stock-\( j \) as a proportion of total purchases/sales of all stocks. Note that when using stock holdings, the characteristics based measures assume these stocks are not sold within the quarter – an issue we take up in section 3.4.} If CS is positive then the fund held stocks at the end of quarter \( t-1 \), which have on average outperformed their characteristic benchmark returns in period \( t \). The CS measure can also be applied to stocks sold/purchased during period \( t \), which might be more informative of genuine risk, than stocks passively held. If the ‘characteristics’ are thought to mimic true underlying risks then CS is a genuine risk adjusted abnormal return, otherwise CS is a “style measure” and shows whether the fund has outperformed stocks with similar characteristics.
The variable CT ("characteristic timing") measures whether the fund generates additional return by exploiting the predictability in style returns of the (size, BMV, momentum) benchmark portfolios:

\[ CT_t = \sum_{j=1}^{5} \left[ w_{j,t-1} BR_j (j, t-1) - w_{j,t-5} BR_j (j, t-5) \right] \]

A positive value for CT implies the fund increased its holding in stocks whose benchmark returns have risen (on average) over the past year. CT provides an alternative measure of 'market timing' to the parametric HM and TM measures and the non-parametric approach of Jiang (2003). Finally, a positive value for "average style":

\[ AS_t = \sum_{j=1}^{5} \left[ w_{j,t-5} BR_j (j, t-5) \right] \]

indicates that the fund was holding stocks 5 quarters previously with characteristics that are currently experiencing high returns. This indicates skill if managers hold stocks that are persistently mispriced (e.g. small stocks).

Stock trades (buys/sells) are ‘active’ decisions and are only likely to take place when the trader has superior information which outweighs any additional trading costs - whereas continuing to hold a stock may be a relatively passive decision reflecting no strongly held views. Thus with the characteristics approach it is argued that using stock trades (i.e. buys/sells) may be more successful in uncovering good (or bad) stock picking or timing skills, than using holdings data (or factor models). However note that the above argument may carry less force for funds which experience very large cash inflows and undertake “liquidity” rather than informed discretionary trades – an issue we investigate in section 5.4.

It is argued that the CS index gives a better measure of abnormal return than the alpha from conditional factor models, where time-varying factor loadings have to be modeled parametrically whereas the CS index directly measures any changes in stock holdings. Also, in the presence of non-linear factor return premia the characteristics approach may provide better benchmarks than linear factor models. Note however, that the CS-holdings measure ignores trades that take place between quarterly reporting dates and compared to monthly holdings data we may miss upwards of 20% of a typical fund’s trades (Elton, Gruber, Krasny and Ozelge 2006) which may lower the power of the test – an issue we address in section 3.2. Also, because of a lack of comprehensive data on all asset holdings (and trades), the characteristics based
approach has only been applied to stock holdings and the sales/purchases of stocks by US equity mutual funds. A limitation of the CS index is that it is usually measured gross of (stock) transactions costs and management fees so ‘returns’ to investors are not directly measured. As equity funds also hold other assets (e.g. cash, bills, bonds) then a positive value for any of the above characteristic based measures need not translate into a positive (gross or net) return on the fund as a whole.

Models of abnormal return based on CS or alpha can be applied to individual funds or portfolios of funds (e.g. an average across all fund styles or just all growth funds). However, averages across a large portfolio of funds may be of little practical use if individual investors feel constrained in the number of funds they are willing to hold (perhaps due to load fees and information costs)\(^{26}\) – although this would be mitigated if the chosen portfolio of funds comprised a limited number of fund families or in the case of institutional investors if they are subject to low switching costs. So, performance measures applied to individual funds or ‘small’ portfolios may be of more practical use. Also, it is clearly of interest to see if performance measures (for either individual funds or portfolios of funds) are stable over sub-periods – if they are, then this suggests persistence and the possibility of profitable ex-ante trading rules – an issue we take up in section 4.

3. EX POST PERFORMANCE: EVIDENCE

We begin this section by outlining some data issues in measuring fund returns, then in section 3.2 we discuss size and power of performance statistics and in section 3.3 possible biases due to data deficiencies. In section 3.4 we examine the ex-post average risk adjusted performance of US and UK equity mutual funds while in section 3.5 we discuss the performance of individual funds, particularly those in the tails of the performance distribution. Finally section 3.6 examines the evidence for successful market timing by funds.

3.1 MEASURING RETURNS

The SEC requires that all US mutual funds report an annual total expense ratio, TER (as a percentage of total net assets) which comprises operating expenses (i.e. management, legal, accounting and custodial fees) plus the 12b-1 fee. Part of the 12b-1 fees (if there are no load fees) are used to pay direct marketing costs including compensation to selling brokers – these

\(^{26}\) Rebalancing costs are incurred if the portfolio is equally or value weighted. Rebalancing costs for equally weighted portfolios occurs when funds die or are merged and also to accommodate the long term rise in the number of funds. In addition a value weighted fund requires rebalancing as market prices fluctuate.
are limited to 1% p.a. of the fund’s assets by the National Association of Securities Dealers, NASD). Trading costs include brokers’ commissions, bid-ask spreads and price impact effects but these are not included in the management fee and the total expense ratio, TER.27

Fund returns can be gross or net of various charges depending on whether we wish to measure returns to the fund (or fund managers) or returns to the investor/customer. Net returns are returns to investors (before deduction of any load fees or payment of personal taxes)28 and are calculated as:

\[
1 + R_t^{net} = \left( \frac{NAV_t}{NAV_{t-1}} \right) \left( \prod_{j=1}^{J} \left( 1 + \frac{DISTN_j}{RENAV_j} \right) \right)
\]

where \(NAV_t\) is the net asset value of the fund at end of period \(t\), \(J\) is the number of dividend or capital gains distributions during the period, \(DISTN_j\) is the \(j^{th}\) distribution in dollars and \(RENAV_j\) is the NAV at which the \(j^{th}\) distribution was reinvested. Net returns are therefore after deduction of all fund expenses and all security level transactions costs29. Gross fund returns (i.e. pre-expenses but post transactions costs) \(R_t^g\) are usually defined as \(R_t^g = R_t^{net} + TER\), where \(TER\) is the total expense ratio. In some studies ‘fund return’ is the return on the largest shareclass, while others use the value-weighted return of all individual shareclasses (see Wermers 2003a).

27 In the US, total commissions are reported to the SEC but not spread costs on the grounds that the latter are difficult to measure in a meaningful way. Funds do not publish commissions as the SEC believes this could be misleading if commissions and spread costs are inversely related.

28 In the US mutual funds are pass-through entities for tax purposes and the fund does not pay any taxes on its holdings – dividend and capital gains realizations are passed on equally to all the fund’s shareholders (regardless of when the shares were created). In the US, investors who purchase shares in a fund which has accrued but unrealized gains face the prospect of distributions of realized gains even if they have only held the fund for a short time – this is the ‘capital gain overhang’ (Bergstresser and Poterba 2002). Prior to the US Taxpayer Relief Act 1997, funds which realized more than 30% of their capital gains from positions held for less than three months did pay taxes. This act also removed restrictions against short sales (the ‘short-short’ rule) and derivatives trades. Mutual funds held in tax deferred form include those in IRA and 401K accounts.

29 This is the case for CRSP and Morningstar databases. Other US data bases include Lipper and some researchers compile their own survivorship-bias-free databases (Elton et al 1996a) or daily returns databases (Busse 1999). For UK returns data the main source is Standard and Poors Micropal - in calculating NAV, prices are measured bid-to-bid.
Comprehensive data on all mutual fund holdings and trades (buys/sells) is not usually available but by merging databases a reasonably long time series for funds’ stock holdings (and trades) is available quarterly - but only for the US. Since data on the identity of fund managers over time is somewhat sparse, studies which measure returns to specific fund managers are much less prevalent than those which measure returns to the fund itself. This is not a major drawback if a fund’s style is largely a group decision. However, it is clearly of interest for investors to assess whether performance is at the fund or fund manager level, since this may determine relative fund flows and the question of whether ‘money is smart’ (see sections 6.2 and 6.3).

3.2 SIZE, POWER AND FALSE DISCOVERIES

Whatever metrics we choose to measure fund performance, model error and the size and power of test statistics need to be assessed. This applies a fortiori if we have small samples or test statistics that do not follow standard distributions (e.g. due to excess kurtosis or skew in extreme winner or loser portfolios). Nevertheless, most tests of ex-post performance using $\hat{\alpha}$ and CS are based on standard t-tests and critical values. To assess size and power we can either simulate data from factor models or use stratified sampling to form random portfolios of stocks with similar characteristic (e.g. size, BMV, etc) to actual funds. For example, using the first approach for the market model with beta = 1, alpha = 1.8% p.a. (0.15% p.m.), R-squared = 0.9, and residual standard deviation of 1.5% p.m., then even with $T = 270$ months (> 22 years) of data, power (for a one sided test) is only 50%, whereas if alpha is as large as 3.6% p.a., power equal to 50% is achieved with only $T = 68$ months (5.7 years) of data – see figure 1.

[Figure 1 here]

Using the stratified sampling approach Kothari and Warner (2001) demonstrate that factor models using 36 observations to estimate $\alpha_{3F}$, $\alpha_{4F}$ or the CS(holdings) measure, have misspecification of around 0.5-1% p.a. in absolute terms and under the null, simulated funds reject $\alpha = 0$ too often (e.g. rejection rates vary between 13%-20% at 5% nominal test size, for the 3F and 4F alphas and characteristic based measures). Power is difficult to assess because

---

30 Prior to 2003, mutual funds were required by the SEC to report holdings twice a year but now must report quarterly holdings. However prior to 2003, most funds publicly disclosed portfolio holdings quarterly and these are available from private vendors such as Thompson Financial (formally CDA/Spectrum).
31 Reported t-statistics usually incorporate a Newey-West (1987) correction to standard errors for any serial correlation or heteroscedasticity.
32 For convenience, we assume a benchmark Sharpe ratio of zero, but this does not affect the general point made.
of these size distortions but is in excess of 80% for the 4F model when abnormal performance is greater than 3% p.a. and greater than 3 years of observations are used, while for the characteristics based measure power is in excess of 70% (and rises as more stocks are included in the simulated portfolios and as more observations are used).

To establish power when only stock *purchases* are used, Kothari and Warner add a simulated abnormal return of 3% over a 3 month horizon and assume *monthly* holdings data is available. The CS event study returns based on holdings data give a rejection rate of 100% whereas the 4F model has a rejection rate of 80%. Also, the holdings based data has (approximately) the correct size (at 5% significance level) whereas the 4F model has a nominal size of 13%. Hence the event study approach has better power when outperformance is over a short horizon but both methods have equally high power if the abnormal return is greater than 3% p.a. and accrues over longer horizons than 3 months. Also if we only have *quarterly* data on traded securities, the power of the event study approach declines dramatically, whereas power for the 4F approach remains high. Abnormal returns from a momentum strategy appear to be maximised over a 6-12 months horizon (Jegadeesh and Titman 2001, Rouwenhorst 1998) and therefore the 4F model should be reasonably good at picking up these effects.

Biases in estimating selectively (alpha) and market timing when the HM (TM) model is true but the TM (HM) model is estimated, are large and almost offsetting which suggests using an overall measure of performance (e.g. \( r_p = \frac{1}{T} \sum_{t=1}^{T} (\alpha_{mp} + \gamma_{p} \sigma \tilde{r}_{t+1}) \) or \( \alpha_{mp} + \gamma_{p} \sigma \tilde{r}_{t+1} \) for the TM model) which suffers from less bias (Coles et al 2006). However, note that Coles et al (2006) show that such model misspecification does not appreciably alter the power to detect successful selectivity or timing – it only affects the bias.

Overall these results suggests that it is only in the tails of the cross-section of the performance distribution that we might have reasonably high power in detecting outperformance - and also we should only include funds with \( T_i \geq 36\) observations, to avoid bias and size distortions. Note that the above results on size and power are based on standard test statistics but these are invalid if residuals are non-normal – this in turn, reinforces some of the points noted earlier namely, the use of bootstrap techniques because of possible non-normality and cross-correlation (in specific risk across funds) – issues we address in the next section.

There is one further statistical issue to be aware of when summarizing the performance of a large number of mutual funds. It is true that when testing alpha for a single fund (or a single portfolio, such as the average fund) then ‘luck’ is correctly measured by the significance level (\( \gamma \))
chosen. However, when we use $\gamma = 5\%$ for the alphas of each of M-funds, the probability of finding at least one lucky fund from a sample of M-funds is much higher than 5\% (even if all funds have true alphas of zero). For example, if we find 20 out of 200 funds (i.e. 10\% of funds) with positive estimated alphas at a 5\% significance level, then some of these will merely be lucky. The false discovery rate (FDR) measures the proportion of lucky funds among a group of funds which have been found to have significant (individual) alphas and hence ‘corrects’ for luck amongst the pool of ‘significant funds’. For example, suppose the FDR among our 20 winner funds is 80\% then this implies that only 4 funds (out of the 20) have truly significant alphas. As we see below this adjustment is frequently ignored although recently Barras et al (2005, 2009) apply the FDR to adjust for luck.

3.3 DATA ISSUES

In interpreting empirical results, data problems which may bias performance include omission bias, backfill bias, incubation bias, survivor bias and look-ahead bias. All databases contain errors and very often have to be ‘cleaned up’. For example, Elton et al (2001) report that the CRSP database suffers from omission bias, since some funds have monthly data, some annual and some have no returns data at all - and this affects the measurement of average alpha by around 40 bp per annum.

Backfill bias arises when a fund is added to a database only when its performance history is good (whereas a fund with poor performance is not included). Another form of backfill bias arises where an acquiring fund replaces its own return history with that of the acquired fund, if the latter’s performance is superior. Incubation bias occurs where a fund family initially operates a number of funds but subsequently only markets the top performing funds to the public, which are then included in the database, with their history intact. A “minimum history bias” arises when we use return data only for funds which have existed for a minimum length of time (e.g. in order to estimate alpha) - while this reduces estimation error in performance measures, it risks biasing findings upwards if the excluded funds are poor performing funds. Using recent US data (which includes alive and dead funds) Kosowski et al (2006) find that this effect does not appear to be large - over the period January 1975 - December 2002, including (domestic) equity funds with greater than 60 monthly observations (rather than all funds) gives a bias of around 20 bp per annum for net returns. Wermers (1999) for the 1975-99 period based on gross returns from stock holdings data reports a similar figure.

Survivorship bias occurs when performance is based only on funds which exist at the end of the sample period and hence ignores those possibly poorer performing funds which closed at an earlier point (i.e. ‘dead’ funds which may be excluded from the fund database). Survivorship
bias is also influenced by the choice of the minimum number of observations on fund returns required for estimation. Malkiel (1995) on US data (1982-90) finds survivorship bias of 1.4% p.a. (using value weighted fund returns) and a survivor premium of 6.5% p.a. while Elton et al (1996a) find survivorship bias is more concentrated in small funds and growth funds. Evans (2007) for US equity funds finds an incubation bias of between 1.9% p.a. and 3.3% p.a. for risk-adjusted (alpha) measures of performance. With average alphas of US funds measured at around minus 70 bp p.a. then incubation and survivorship bias are quantitatively important. Results on survivorship bias for UK funds is sparse but Quigley and Sinquefield (2000) using UK data over 1978-1997 (751 funds) report a survivor premium of 2.31% p.a. and a survivorship bias of 0.7% p.a., the latter being very close to that found by Blake and Timmermann (1998).

For tests of persistence, simulation of mutual fund returns shows that survivorship bias may result in errors in either direction, depending on whether a single period or multiperiod survival rule is imposed (Brown et al 1992, Carpenter and Lynch 1999, Carhart et al 2002b).

Look-ahead bias is a property of experimental design. When measuring performance persistence investors are assumed to hold a portfolio of funds between $t$ and $t+k$. Look-ahead bias arises when a fund is dropped from the analysis at $t$ if it does not exist for the full duration of the holding period. If such a fund “disappears” due to poor performance then this imparts a bias in performance measures. A partial solution to this problem is to include fund returns until the fund “disappears”, since then returns include some of the period of poor performance.

The above results demonstrate the importance of including all funds that have existed over the data period under study and more recent US and UK studies should not suffer from acute survivorship bias. Although other sources of potential bias remain, the coverage of the CRSP database of US funds and the Standard and Poors UK database are quite comprehensive.

### 3.4 US AND UK RESULTS

We now turn to the ex-post performance of US and UK mutual funds, both managed and index funds. In our survey of past work we frequently indicate the data period used, how many funds were included in the study, as well as some representative results, so the reader can better evaluate the relative contribution of past work to the key issues addressed. Statistical significance at the 10%, 5% and 1% significance levels are denoted by superscripts ‘*’, ‘**’, and ‘***’

---

33 Survivor premium is the difference between the annual compound raw returns of portfolios of surviving and nonsurviving funds. Survivorship bias is the difference between the annual compound returns of the surviving funds and the full set of both surviving and nonsurviving funds.
respectively and p-values are also reported. Unless stated otherwise ‘statistically significant’ refers to a 5% significance level (or better). Table 1 provides a summary of key results.

[Table 1 here]

Adjustment for risk is always a contentious issue for actively managed funds, so Elton, Gruber and Busse (2004) focus their attention on 52 US, S&P500 index funds (January 1996-December 2001) - they use the CAPM-alpha or simply the fund’s differential (net) return over the market index as measures of abnormal performance\(^{34}\). The (equally weighted) average index fund’s differential return is minus 0.485% p.a. and the CAPM-alpha performance is minus 0.410% p.a.\(^{35}\). The average TER is 0.444%, implying that underperformance arises because index funds incur advertising, rebalancing, cash flow and management fees in order to closely track the index. But given the wide spread in TERs for index funds of 0.06% to 1.35% p.a. (with 25\(^{th}\) and 75\(^{th}\) percentiles of 0.25% and 0.59%), it is also worth noting that it may be possible to find an index fund at near zero cost\(^{36}\).

For managed funds, early studies use Jensen’s alpha as a measure of risk adjusted performance. For example, Ippolito (1989) using a sample of 143 mutual funds finds that most earn abnormal returns sufficient to cover their expenses over the period 1965–1984 - results which were in contrast to some earlier studies (Friend, Blume and Crockett 1970, Jensen 1968, Sharpe 1966). However, Elton, Gruber, Das and Hlavka (1993) show that after correcting for non-S&P500 stocks in the benchmark market index, positive pre-expense alphas disappear - a result supported by Malkiel (1995) who finds that only a small number of mutual fund managers over the period 1971-91 have statistically significant (gross return) Jensen’s alpha - but there is no evidence that such ability exists at the net return level.

A key study by Chen, Jegadeesh and Wermers (2000) uses trade data to investigate stock picking skill. Based on raw returns they find stocks purchased by funds outperform stocks sold by them (over the next year) – this provides prima facie evidence for managed funds as a whole, having some skill in picking stocks (before expenses and trading costs). Also based on stock holdings data, Wermers (2000) examines the reasons for differences in returns between funds’ stock holdings and overall fund return, a difference of 2.3% p.a., and finds (over the 1974-

\(^{34}\) They also consider tracking error and tax efficiency of index funds versus the index itself – but we concentrate on performance results.

\(^{35}\) The maximum spread in alphas across index funds is quite considerable at -1.53% to 0.228% p.a.

\(^{36}\) The tracking errors across funds as measured either by the absolute value of $|\beta| - 1$ or the R-squared of the CAPM regressions, range from 0.005 to 0.021 and 0.9991-1.0000, respectively and hence are very small.
1994 period) that 0.7% is due to the lower returns of non-stock holdings, with expenses and trading costs also contributing roughly 0.7% p.a. each, to the reduction in funds’ gross returns.

So, although gross returns on the average funds’ stock holdings cover expenses and trading costs – other fund assets further reduce the overall return on the fund. The latter is taken up by Shukla (2004) who is concerned that measuring fund returns using asset holdings data ignores the interim trading costs between the two portfolio composition dates (which are usually quarterly) – this could bias results on performance using the CS index, given the high average turnover of fund assets (of around 100% in 2003). Regardless of management fees charged, any revisions to a portfolio would only add value for investors if the return on a revised portfolio (net of added trading costs) is higher than the return on a (buy-and-hold) passive portfolio. He shows that in the US, such a return differential is zero on average across funds (for up to 6 month horizons) – so investors do not obtain any additional return due to frequent portfolio revisions by traders, as a whole. However, there is some (weak) evidence that portfolio revisions by growth funds add value and there is also a wide dispersion in returns to trading across different funds – the return differential to trading is largest for small funds and those with more concentrated portfolios. So there may be some individual funds that are successfully trading at the margin.

In a novel approach Baker, Litov, Wachter and Wurgler (2007) examine the returns of fund holdings and trades around earnings announcement dates, arguing that ‘announcement returns’ are likely to be ‘abnormal’. In an event study framework they classify funds into those with weight increases (decreases) and examine subsequent returns for 3 days around the earnings announcement dates. Stocks in which funds have experienced increasing (decreasing) weights have 20 bp p.a. higher (21 bp lower) returns in the subsequent earnings announcement periods (relative to matched CS-returns). This difference in future CS returns of around 40 pb p.a. indicates stock picking skill (before costs) within characteristic groupings.

How do US funds perform on average when we consider net returns to investors rather than gross returns to the fund itself? On a CRSP net return (value weighted) basis Wermers (2000) finds $\alpha_{i,t}^{net} = -1.16\%$*** p.a. indicating underperformance by the average fund over 1974-94. Using more recent data (January 1975-December 2002) and around 1,700 funds, Kosowski

---

37 Transactions costs are estimated from the trade data using Keim and Madhavan’s (1997) estimates of institutional execution costs for stocks (commissions plus price impact), based on cross-section regressions.

38 Data used is from Morningstar Pricipia CD-ROM, for 458 funds, August 1995-December 2002. The return differential to trading (net of transactions costs) over the return on the passive portfolio is $R_{i,t}^{diff} = (R_{i,t} - TrC_{i,t}) - (R_{i,t}^{p,psv} - TrC_{i,t}^{p,psv})$, where $R_{i,t}$ = gross return from active trading, $R_{i,t}^{p,psv}$ = gross return on passive portfolio, with transactions cost on the passive portfolio of $TrC_{i,t}^{p,psv}$ - which are assumed to be zero.
et al (2006) find that the (equally weighted unconditional 4F) net return alpha is about minus 0.5% p.a. – so the ‘average managed fund’ underperforms its benchmarks. The cross-section standard deviation of these $\alpha_i$ across the 1,700 funds is high, indicating the possibility that some funds are performing very well (and others very badly).

Regime switching models provide an approach to modeling time varying parameters and, if robust, such models may increase our confidence that alpha is measured correctly. In an interesting study Kosowski (2006) finds that the average alpha (depending on the model used) is 3-5% p.a. higher in recessions than in expansions, demonstrating that investment in mutual funds (and particularly growth orientated funds) may provide “value added” in times when the marginal utility of wealth is high – this fits neatly into the view that part of the rationale for holding funds is to hedge changes in the MRS.

Much less empirical work on performance has been done on UK funds. Leger (1997) estimates the CAPM-alpha on 72 UK investment trusts in four non-overlapping five-year samples between 1974 and 1993 and finds little evidence of statistically significant ex-post performance. Quigley and Sinquefield (2000) examine the performance of all UK equity mutual funds (including non-survivors) between 1978 and 1997 (752 funds) using gross returns and finds an equally weighted portfolio funds gives $\alpha_i = -1\%$ p.a. ($t = -2.3$) and poor average performance is found across all four investment styles (i.e. growth, income, general equity and smaller companies). Similarly, Fletcher (1997) over 1980-1989 using an APT model for growth, income and general equity categories finds no statistically significant abnormal performance.

The above studies strongly suggest that the average US or UK fund (even within specific sectors) does not earn abnormal positive net returns. Of course, this does not preclude funds in the tails of the distribution having statistically significant alphas - an issue we now discuss.

### 3.5 Luck and Fund Performance: A Tale of Two Tails

39 Barras et al (2005) over the same time period find $\alpha_i = -0.44\%$ p.a. ($p=0.16$) but use 1472 funds whereas Kosowski et al (2006) have 1,788 funds (both studies use funds with $T_i \geq 60$ observations).

40 For example the top and bottom ranked funds have 4F net return alphas of 4.2% and minus 3.6% per month, respectively while the 95th and 5th percentile funds have alphas of 0.4% p.m. and minus 0.5% p.m. respectively (Kosowski et al 2006).

41 Over 1965-2002, Kosowski (2006) uses both a split sample technique (based on NBER recession/expansion periods) and a conditional regime switching model.

42 Benchmark portfolios are estimated by an asymptotic principal components technique as outlined in Connor and Korajczyk (1986).
As we have seen there is considerable evidence that the *average* managed fund under-performs its benchmark returns. However, it is also found that some subgroups of funds do seem to out-perform their benchmarks (e.g. US growth oriented funds, Chen et al 2000, Wermers 2000). Wermers (2003b) provides prima facie evidence that there may be a small number of funds taking 'big bets' which on average outperform funds taking small bets and the former also have positive risk adjusted returns\(^{43}\). But it is clear that isolating 'winner funds' is a difficult task and there are certainly some funds taking 'big bets' that do not outperform. So, can we be sure that such 'group' out-performance is not due solely to 'good luck'? Kosowski et al (2006) is the first paper to explicitly control for luck when measuring *individual* fund performance using a cross-section bootstrap methodology with US monthly net returns, while Barras et al (2005,2009) examine luck for a *group* of top performing funds using the false discovery rate. We discuss these in turn.

Funds in the extreme tails of the performance distribution are likely to exhibit non-normality in their idiosyncratic risk and they may also be 'short lived' funds - so standard asymptotic results do not apply. Under these circumstances bootstrapping procedures are required. The simplest approach is to apply the bootstrap on a fund-by-fund basis - but this excludes information in the cross-section of luck across *all* funds. To highlight the importance of a cross-section bootstrap consider whether the stellar performance by some funds even over a run of many years, is due to luck or skill. Are these 'stars' just the lucky ones amongst the cohort of all fund managers\(^{44}\)? For example, if we are told that a particular mutual fund has an abnormal average return of say 10% p.a. then we might well be impressed. But if we are told that this return of 10% p.a. was achieved by the best performing fund out of say 1,000 funds, we should be less impressed. This is because in a large universe of 1,000 funds there will always be some funds that perform well (badly), simply due to chance. The issue then arises as to how we can separate 'skill' from 'luck' for *individual* funds, particularly when idiosyncratic risks are highly non-normal – as is the case for funds in the extreme tails, in which investors are particularly interested.

Suppose we are interested in the performance of the best fund (in the ex-post data) and whether this is due to skill or luck. We could 'replay history' by only considering the idiosyncratic risk of the ex-post best performing fund. But when we replay history for the second or third etc. ranked fund in the ex-post data, it is quite conceivable that one of these funds now has the 'best' performance!

\(^{43}\) Funds undertaking 'big bets' are defined as those with a large *standard deviation* of market index adjusted returns, while abnormal returns are measured as either average market adjusted returns or Jensen’s alpha or the unconditional 4F-alpha. These variables are measured over 3-year non-overlapping periods, 1975-2000.

\(^{44}\) Peter Lynch of the Magellan fund is cited as a star manager (Marcus 1990) and the US based Schroder Ultra Fund earned 107% over 3 years ending in 2001 and was closed to new investors as early as 1998.
performance. Clearly, re-running history for just the ex-post best fund ignores the other possible distributions of luck encountered by all other funds – these other ‘luck distributions’ provide highly valuable and relevant information. Put more technically, in picking out the best fund (ex-post) we have ‘ordered’ it as ‘number one’ and because of that fact we need to compare its performance with all other funds which have the potential to be ‘number one’, if we are to separate skill from luck for the best fund – this is the theory of order statistics\textsuperscript{45}.

Kosowski et al (2006, Table II) apply the cross-section bootstrap (with a maximum of 1,704 US funds, January 1975-December 2002) using \( t_{\alpha} \) as the performance statistic\textsuperscript{46} and find that many funds ranked above the top 5\textsuperscript{th} percentile (i.e. a maximum of about 90 funds) are statistically significant (at a 5\% significance level) with \( \alpha_{1F}^{\text{net}} \) in excess of 2.0\% per month. The proportion of funds with positive alpha-skill is as high as 30-40\% in the 1975-89 period, but falls to around 5\% in the 1990-2002 period, when there are far more funds available\textsuperscript{47}. Increased competition because of expansion in the mutual fund sector as well as exploitation of anomalies by hedge funds are likely sources of reduced performance. Skilled funds are all found to be in either the aggressive growth or growth styles, rather than in ‘growth or income’, and ‘balanced-income’ sectors\textsuperscript{48}. Fama-French (2009) dispute the Kosowski et al (2006) results for positive alpha funds and find no evidence of superior (net return alpha) performance in the right tail but consistent with Kosowski et al find evidence of statistically significant negative-alpha funds. Fama-French (2009) attribute this difference to their use of an alternative cross-section bootstrap,

\textsuperscript{45} Unfortunately, analytic results from order statistics are only available for well defined distributions but since idiosyncratic risk across each fund is different and does not follow known distributions, we have to resort to bootstrap procedures (see Efron and Tibshirani 1993, Politis and Romano 1994).

\textsuperscript{46} \( t_{\alpha} \) has better sampling properties than \( \alpha \) - the obvious reason being that the former ‘corrects for’ high risk-taking funds (i.e. \( \sigma \) large), which are likely to be in the tails. Put another way, if the distribution of alpha for each fund is \textit{niid} (under the null) but each fund has a different \( \sigma \), then the cross-section distribution of the alphas \( f(\alpha) \), will \textit{not be normal}, but the distribution of \( f(t_{\alpha}) \) remains normal – this is the basis for \( t_{\alpha} \) (but not \( \alpha \)) being a ‘pivotal statistic’ and hence is the preferred performance metric. However, it is important to note that even when using \( t_{\alpha} \) we cannot generally infer what the tails of the cross-section distribution \( f(t_{\alpha}) \) will look like (e.g. if \( \epsilon \) are drawn from a mixture-normal distribution) – this is why we need the cross-section bootstrap (Hall 1986, 1992). Kosowski et al (2006) also find their results are invariant to potential serial correlation or contemporaneous correlation across the residuals of the funds which could be caused by herding into similar stocks).

\textsuperscript{47} Similarly, risk adjusted returns to hedge funds are higher in the early 1990’s than in the later 1990’s (Agarwal and Naik 2002).

\textsuperscript{48} Comer et al (2007) notes that for around 260 “hybrid funds” (i.e. balanced funds and asset allocation funds) which hold a substantial amount of bonds, alpha estimates from the Carhart 4F model may change substantially in the post 2001 period, when four additional bond return factors are added – this appears to be mainly due to a worsening of the alphas of poor performing funds (see table 5, p15).
rather than to different funds included, minimum data requirements or different data periods used.

Using a similar bootstrap approach to Kosowski et al (2006) on UK data (842 funds, 1975-2002), Cuthbertson, Nitzsche and O’Sullivan (2008) find that only 12 funds out of the top 20 UK funds are statistically significant (each at 10% significance level). As one moves further towards the centre of the performance distribution (i.e. at or below the top 3% of funds) there is no evidence of stock picking ability. They also show that ‘significant’ UK top performers are not necessarily those with an ex-post ranking right at the ‘top’. This makes it extremely difficult for the ‘average investor’ to pinpoint individual managed funds which demonstrate skill, based on their track records. For the UK, in contrast to the US results, skill appears to reside with equity income funds rather than ‘all company’ or small company funds.

Note that although Kosowski et al (2006) and Fama-French (2009) bootstraps give p-values of the ordered individual funds, a simple count of all funds with ‘significant’ p-values ignores the possibility of some significant funds being “false discoveries” (Storey 2002). The latter is taken up by Barras et al (2005, 2009) who focus on the ‘false discovery rate’ FDR - that is, the proportion of lucky funds among funds with significant (individual) alphas. The Barras et al (2005, 2009) procedure therefore deals with the proportion of ‘truly’ statistical significance funds within a portfolio of statistically significant funds, whereas Kosowski et al (2006) approach applies to the statistical significance of individual (ordered) funds.  

Kosowski et al (2007) find their results are largely invariant to use of an conditional/unconditional 4F model, to the minimum number of monthly observations used (18 < T_{min} < 120), sorting on alpha rather than t-alpha and to the type of bootstrap (e.g. residual only bootstrap, block bootstrap to account for serial correlation, or a bootstrap taking account of contemporaneous cross-fund correlations in idiosyncratic risk or bootstrapping both the factors and the residuals independently). The Fama-French (2009) bootstrap under the null of zero alpha uses “benchmark adjusted returns” that is returns less estimated alpha $R_{i} - \hat{\alpha}_{i}$, for funds $i = 1,2... M$. If observation $t$ is chosen in the bootstrap then benchmark adjusted returns at $t$ for all funds (with observations at $t$) are included – which allows for contemporaneous correlation. This is repeated for T=273 bootstrap observations across all M-funds and $t_{\alpha}$ is estimated for all M funds (for funds with a minimum of 8 observations). These M-values for $t_{\alpha}$ (under the null) are ranked. This is repeated 10,000 times to give 10,000 values of $t_{\alpha}$ for each percentile of the cross-section of M-funds, under the null. The null distribution for any particular percentile fund is then compared with the actual estimated $t_{\alpha}$ for the same percentile fund. The Fama-French bootstrap therefore takes account of correlated heteroscedasticity between factors and residuals but it is a little unnerving that this relatively small change produces substantially different inferences for positive alpha funds – although the high FDR reported below for US positive alpha funds indicates that inferences based on funds in the right tail are very uncertain.

Note that the Kosowski et al (2006) and Barras et al (2005) definitions of luck are very different. The former is based on the significance of an ordered fund’s individual alpha at a specific quantile of the multivariate performance distribution, for a given significance level, while the FDR can be used to calculate the number of genuinely significant funds from a portfolio of funds which have been found to be individually statistically significant (at a given significance level). We use the usual language and terminology found in the statistical literature on false discoveries and error rates. The use of the word “truly” or “genuinely” should not be taken to mean that we are 100% certain that a fund (or group of funds) outperforms. The Kosowski et al approach is still subject to type-I and type-II errors and the measured FDR even if it is found to be zero, is still subject to estimation error. Also note that the FDR says nothing about the statistical significance of the alpha of any particular individual fund - conceptually, the FDR only applies to a group of significant funds.

---

49 Kosowski et al (2007) find their results are largely invariant to use of an conditional/unconditional 4F model, to the minimum number of monthly observations used (18 < T_{min} < 120), sorting on alpha rather than t-alpha and to the type of bootstrap (e.g. residual only bootstrap, block bootstrap to account for serial correlation, or a bootstrap taking account of contemporaneous cross-fund correlations in idiosyncratic risk or bootstrapping both the factors and the residuals independently). The Fama-French (2009) bootstrap under the null of zero alpha uses “benchmark adjusted returns” that is returns less estimated alpha $R_{i} - \hat{\alpha}_{i}$, for funds $i = 1,2... M$. If observation $t$ is chosen in the bootstrap then benchmark adjusted returns at $t$ for all funds (with observations at $t$) are included – which allows for contemporaneous correlation. This is repeated for T=273 bootstrap observations across all M-funds and $t_{\alpha}$ is estimated for all M funds (for funds with a minimum of 8 observations). These M-values for $t_{\alpha}$ (under the null) are ranked. This is repeated 10,000 times to give 10,000 values of $t_{\alpha}$ for each percentile of the cross-section of M-funds, under the null. The null distribution for any particular percentile fund is then compared with the actual estimated $t_{\alpha}$ for the same percentile fund. The Fama-French bootstrap therefore takes account of correlated heteroscedasticity between factors and residuals but it is a little unnerving that this relatively small change produces substantially different inferences for positive alpha funds – although the high FDR reported below for US positive alpha funds indicates that inferences based on funds in the right tail are very uncertain.

50 Note that the Kosowski et al (2006) and Barras et al (2005) definitions of luck are very different. The former is based on the significance of an ordered fund’s individual alpha at a specific quantile of the multivariate performance distribution, for a given significance level, while the FDR can be used to calculate the number of genuinely significant funds from a portfolio of funds which have been found to be individually statistically significant (at a given significance level). We use the usual language and terminology found in the statistical literature on false discoveries and error rates. The use of the word “truly” or “genuinely” should not be taken to mean that we are 100% certain that a fund (or group of funds) outperforms. The Kosowski et al approach is still subject to type-I and type-II errors and the measured FDR even if it is found to be zero, is still subject to estimation error. Also note that the FDR says nothing about the statistical significance of the alpha of any particular individual fund - conceptually, the FDR only applies to a group of significant funds.
For 'all' US equity funds (1975-2006), Barras et al (BSW 2009) find no 'truly' positive-alpha skilled funds over the whole sample period but evidence of a small proportion (2 -2.5%) of skilled funds over "short-term" (non-overlapping 5-year) horizons, concentrated in the extreme right tail. Cuthbertson et al (2010) using the FDR technique on UK data (1975-2002) find a small proportion of both "long-term" and "short-term" positive-alpha skilled funds of around 5%. Because the estimated FDR amongst statistically significant positive-alpha funds is high in both the UK (35%) and US (in excess of 90%), a straightforward count of the number of statistically significant positive-alpha funds would give an extremely misleading measure of the number of truly skilled funds. This demonstrates the importance of the FDR technique in quantifying data snooping bias when assessing the overall performance of the mutual fund industry. Indeed, for both UK and US equity funds, around 75% of funds neither statistically beat nor are inferior to their benchmarks and therefore appear to do no better than merely tracking their style indexes.

For unskilled funds, US results in BSW (2009) and UK results (Cuthbertson et al 2010) are very similar - both studies find an FDR for statistically significant negative-alpha funds of around 30%, which results in strong evidence of a sizeable proportion of 'truly' negative-alpha unskilled funds of around 15-20%, spread throughout the left tail. BSW(2009) like Kosowski et al (2006) also find that the proportion of truly positive-alpha US funds has declined over the last 20 years while the proportion of unskilled funds has increased. So although the FDR approach applies to 'counts' of funds within a portfolio of 'significant funds', rather than to the statistical significance of specific ordered funds (as in Kosowski et al 2006, Fama-French 2009, Cuthbertson et al 2008), both approaches give a broadly similar picture for US and UK equity funds.

---

51 When recursive OLS (or the Kalman filter, random coefficients model) is applied to a portfolio of UK funds based on bootstrap rankings, Cuthbertson et al (2008) find considerable stability in the estimated portfolio alphas (as well as the market return and SMB factor loadings). This suggests that over several years, there is a constant and statistically significant outperformance amongst a few UK top funds and underperformance amongst many poorly performing UK funds. Also, for US funds at the bottom of the performance distribution, the (bootstrap) CS measure is not statistically different from zero, so these funds earn zero abnormal returns gross of fees and transactions costs, but as we have seen, negative net abnormal returns to investors (Kosowski et al 2006).
Note that the above results on top performing funds are broadly consistent with a competitive equilibrium since we expect to observe very few funds with positive risk adjusted returns over long horizons. This is because funds with genuine skill and high past returns have large inflows (see section 6.2) and with increasing marginal costs to active management, this should lead to zero long run abnormal returns to investors (Berk and Green 2004). The results for the left tail of the performance distribution are not consistent with the competitive model of Berk and Green (2004), since ‘bad skill’ should lead to large outflows from these funds and the return on such funds who subsequently survive should, in equilibrium, equal that on a passive (index) fund. The continued existence over long time periods, of a large number of funds which have a truly inferior performance (which cannot be attributed to bad luck), may indicate that many investors either cannot correctly evaluate fund performance or find it ‘costly’ to switch between funds, hence any approach to a competitive equilibrium appears to be very slow.

UNCERTAINTY AND EXTRANEOUS INFORMATION

Recent work has emphasized the use of information extraneous to a specific fund to help pick superior funds – this information can be purely data based or involve priors in a Bayesian framework (or both). The idea that badly measured (standard) fund alphas might be improved with the use of prior information on that fund and the formation of Bayesian alphas is well known. However, Jones and Shanken (2005) suggest that the alphas of all other funds may be used to improve estimates of an individual fund’s alpha. The intuition for this can be demonstrated as follows. Assume all funds’ alphas are distributed as $\alpha_i \sim \text{n iid} (\mu_a, \sigma_a)$ and you are concerned with the estimate of alpha-$XYZ$. If residuals are independent across funds and the number of funds $M$ is very large then we would have very accurate estimates of $(\mu_a, \sigma_a)$. In the absence of any information about $\alpha_{XYZ}$ it would seem reasonable to take $(\mu_a, \sigma_a)$ as an estimate of its alpha and precision – even though we do not know where $XYZ$ lies in the cross-section distribution. If we now have a (standard) estimate of $\alpha_{XYZ}$ it would seem sensible that our best estimate of $XYZ$’s alpha is a weighted average of $\alpha_{XYZ}$ and $\mu_a$ (with relative weights depending on the precision of the two estimates). This they refer to as ‘learning across funds’ since all other fund returns influence the estimate of any one fund’s alpha.

One further twist can be added to this approach. It follows from the above that priors about $(\mu_a, \sigma_a)$ should influence priors about an individual fund $XYZ$, which in turn influence that fund’s posterior alpha – so the priors for each fund are not independent. This avoids the problem
that with independence across residuals, then as the number of funds increases the maximum-alpha fund increases without bound even when all true alphas are zero – a standard result from the theory of order statistics (Baks, Metrick and Wachter 2005). But with prior dependence across funds this nonsensical outcome is precluded. This is because the maximum alpha-XYZ is shrunk towards $\mu_\alpha$ and if there is no skill across funds the latter will tend to zero – in addition, as the number of funds increases $\sigma_\alpha$ approaches zero and the relative weight given to $\mu_\alpha$ in the pooled estimate for alpha-XYZ also increases. The maximum posterior estimate for alpha-XYZ is therefore bounded$^{52}$. An example which illustrates such shrinkage is the Fidelity Magellan Fund which over 1963-2000 has an OLS alpha of 10.4%**p.a. Is this plausible? Using, ‘learning across funds’ and estimates $(\mu_\alpha, \sigma_\alpha) = (1.5, 1.5)$, the posterior alpha for Magellan is 4.8% p.a. - a substantial ‘shrinkage’ on the standard alpha$^{53}$. So pooling information from other funds seems to reduce the variability in (posterior) estimates of extreme fund alphas (for any set of priors) – this should help to avoid errors when choosing ‘extreme performers’.

Overall, what these latest studies demonstrate is the importance of looking at performance of funds in the tails of the distribution (rather than the average fund) and then making appropriate adjustments for idiosyncratic risk across all funds before making inferences. The clear message from recent UK and US results is that there are a few ‘top funds’ who demonstrate skill but the majority have either no skill and do well because of luck or, perform worse than bad luck and essentially waste investors time and money.

### 3.6 MARKET TIMING

Our discussion in section 2.2 demonstrates that consistent estimation of market timing using factor models is fraught with difficulty and we have little idea of the magnitude of bias in alpha when market timing is truly successful. In addition “skill” is reflected in both alpha and the market timing coefficient – a joint hypothesis – and it is almost impossible to separate skill in security selection from skill in market timing when using parametric models. As well as the theoretical problems noted above, statistical problems such as heteroscedasticity (Breen at al 1986) or skewness (Jagannathan and Konarajczyk 1986, Goetzmann et al 2000) give biased

---

$^{52}$ There is a useful parallel here with the cross-section bootstrap procedure of Kosowski et al (2006) which also compares the maximum ex-post alpha with the cross-section of risk across all funds (which may not be independent), thus avoiding misleading inferences by just bootstrapping on a fund’s own residuals.

$^{53}$ The Bayesian approach incorporates uncertainty in estimates of $(\mu_\alpha, \sigma_\alpha)$ by using Gibbs sampling techniques. Also, portfolio allocation (based on power utility from next periods terminal wealth) is far less sensitive to priors when ‘learning across funds’ is used since without the latter, portfolio allocations are very sensitive to ‘high’, ‘some’ and ‘no’ skepticism in the priors – another example of ‘shrinkage’.
estimates of the market timing parameter $\gamma_{pm}$ and the test is poor both in terms of size and power.

Notwithstanding the theoretical and empirical difficulties in parametric estimation of market timing from factor regressions, there is overwhelming empirical evidence from TM or HM factor regressions (with allowance for alphas and betas depending on public information) which suggest no positive market timing for US funds and even negative market timing when using monthly returns data (Wermers 2000, Ferson and Schadt 1996, Becker et al 1999, Goetzmann et al 2000). The above applies even after conditioning on public information (Ferson and Warther 1996, Ferson and Schadt 1996, Becker et al 1999). For the US even when using daily data, the evidence for successful market timing (in the TM and HM type regressions) is weak (see Bollen and Busse 2001 and Comer 2006 for “hybrid equity funds” which have substantial bond holdings). The debate about the relative merits of daily versus monthly data in factor models is not yet resolved. Higher power from using daily data may be offset by increased bias in estimation of $\gamma_{pm}$ (Coles, Daniel and Nardari 2006) or too many false positives because of market microstructure problems (Mamaysky et al 2008). Kalman filter modeling of time varying factor loadings also fails to identify market timing using daily data, but does find some evidence of positive timing using monthly data - both the “unobserved factor” and observable macro-variables both play a role, although in terms of increased R-squared the former has more impact than the latter (Mamaysky et al 2008). Non-parametric tests on $\theta$ (using monthly data) suggest little evidence of successful market timing in US equity funds and slightly stronger evidence of negative market timing.

There has been relatively little research carried out on the market timing skills of UK funds. Byrne et al (2006) apply the conditional timing tests of Becker et al (1999) to 421 funds but find no evidence of positive timing ability over the period 1988 – 2002. Fletcher (1995) evaluates the market timing of 101 mutual funds between 1980 and 1989 using the HM and TM models but produces similar findings and in fact the results suggest that funds, on average, reduce their market exposures when subsequent market returns are high (and vice-versa). Leger (1997) using UK equity investment trusts between 1974 and 1993 also finds negative and

---

54 Using a daily data set Busse (1999) examines the allied concept of volatility timing. Do funds reduce their market betas when conditional volatility is higher than average and hence enhance risk adjusted returns ? Although the Sharpe ratio is higher for those funds which reduce market exposure in times of abnormally high volatility, there is no relation between the strength of volatility timing in the previous six months and fund performance in the next six months (in terms of either the 4F-alpha or the Sharpe ratio) – so volatility timing is not an exploitable strategy for investors (see also, Busse 2001).

55 Also, funds do not appear to be able to time style factors such as SMB or HML (Chan, Chen and Lakonishok 2002).
statistically significant timing (using the TM model). Cuthbertson et al (2010) using non-parametric approaches find a relatively small number (around 1%) of UK mutual possess significant positive market timing skill, while around 19% are shown to miss-time the market. But after controlling for publicly available information, there is very little evidence of successful market timing based on private information by mutual funds. (Similar results apply to UK pension funds, Cuthbertson et al 2010).

Lack of strong evidence supporting successful market timing by managed equity funds may be due to a genuine lack of skill in predicting benchmark returns and the latter is certainly consistent with evidence on daily/monthly predictability and parameter instability in time series forecasting equations for stock market returns – see for example, Ang and Bekaert (2007).

Another reason for not finding evidence of successful market timing may be due to the “dilution effect”. Funds experience an increase in investor cashflows during periods when the market return is relatively high (Warther 1995, Edelen and Warner 2001), hence increasing the fund’s cash position, leading to a concurrent lower overall portfolio return56 (Bollen and Busse 2001). Poor market timing is then the price investors pay for liquidity provision. However, if these cash flows are mainly into and out of no-load (rather than load) funds or, if institutional investors are thought to be more informed than retail investors, one would expect to see differential market timing skills between these categories – but Jiang (2003) finds that this is not the case. Using data on all trades of Canadian mutual funds, cash inflows result in flat or negative returns on stocks purchased and positive or flat returns on stocks sold – so transaction costs consequent on cash inflows lead to low fund returns (Christoffersen, Keim and Musto 2006).

STALE PRICES

The above results do not rule out the possibility that some types of “timing strategies” might be successful, for example, “late trading” and “market timing” of fund purchases/sales based on stale prices. Consider late trading. Funds’ net asset values (NAV) in the US are determined at 4pm Eastern Standard Time, EST. If positive (negative) information about share prices arises immediately after 4pm but some market participants (e.g. large brokers, hedge funds) are allowed to undertake buy (sell) orders for their clients after 4pm, this may represent a (risky) profitable arbitrage opportunity known as late trading (which is illegal in the US).

“Market timing” also exploits “stale prices”. Consider a US mutual fund which invests in Japanese stocks. The 4pm fund NAV is based on closing prices of Japanese stocks at the prior

56 Note that this would require investors in managed funds to time the market ahead of the fund managers themselves – a somewhat tenuous hypothesis.
1am (EST) price. For example, a US investor who notes that the Nikkei-225 *futures* price has increased between 1am and 4pm (because of the arrival of good news about the Japanese economy), knows that the “true” NAV of the “Japanese mutual fund” is now higher. She can purchase the US-based Japanese equity fund on-line at zero transactions cost (from the fund complex using a money market fund), *at its stale price* just before 4pm EST. She can subsequently sell the Japanese mutual fund at any time before 1am (NY time) the next day, when the NAV is determined in Japan. There are fundamental risks in this strategy including exchange rate risk.

A number of studies have documented successful daily market timing based on stale prices, for US–based international equity funds (Goetzmann, Ivkovic and Rouwenhorst 2001, Greene and Hodges 2002, Boudoukh et al 2002) and there is weaker evidence for exploitable profits for domestic equity funds, particularly small cap funds (Chalmers, Edelen and Kadlec 2001) and bond funds (Zitzewitz 2003, 2006). These profitable “timing strategies” would be highly unlikely to be picked up by the above market timing measures, using monthly data.

Late trading and market timing based on stale prices have an externality effect on long-term fund investors since net inflows or outflows of funds by day traders cause a “dilution effect”. Greene and Hodges (2001) show that flows into US-based international funds (but not into domestic equity funds) by day traders exploiting stale prices, has a dilution impact of around 0.5% p.a. (because short-term cash flows are not immediately invested), which is borne by passive non-trading shareholders in the fund. Similarly, Zitzewitz (2006) finds a dilution effect of 3.8 bp p.a. from late trading (up to 2003) in US-based international funds. However, Goetzmann et al (2001) report that long term shareholders have not, *on average*, been “seriously” affected by day traders exploiting pricing errors although the effect varies considerably across different fund classifications and the authors cite this as the source of their differing findings *vis-à-vis* those of Greene and Hodges (2001). Academic research has undoubtedly informed debate in this area and has led to policy changes to curb these practices – action by Elliot Spitzer, the New York Attorney General resulted in mutual fund sponsors having to pay around $1.5bn in penalties.

**TECHNICAL TRADING**

We cannot rule out the successful use of technical trading rules over very short horizons (daily and intraday) many of which are based on “market timing” signals such as moving

---

57 Dickson, Shoven and Sialm (2000) demonstrate a negative externality in US funds when outflows trigger capital gains realizations which are then passed through to existing investors in the fund but inflows convey a positive externality as they dilute unrealized capital gains. Cai, Chan and Yamada (1997) also report a possible dilution externality on performance for Japanese funds – but this is due to a special tax arrangement whereby inflows can be purchased at NAV less accrued taxes (per share), so that the overall NAV per share is diluted by the inflow.
averages, relative strength indices and filter rules — again these are extremely unlikely to be picked up using regression techniques as they are often based on intraday trading. Menkhoff and Taylor (2007) argue that these techniques yield profits in foreign exchange markets (even after adjustments for risk) which might influence the performance of international mutual funds (but not domestic funds). Direct evidence of the success of technical trading rules in US equity markets using intraday data is not conclusive (Brock, Lakonishok and Le Baron 1992, Blume, Easley and O’Hara 1994, Lo, Mamaysky and Wang 2000, Kavajecz and Odders-White 2004). However, using daily data and after allowance for “data snooping” Sullivan et al (1999) find no evidence for the success of technical trading rules for US broad market equity indices or for calendar rules (Sullivan et al 2001). In any case it seems unlikely that any intraday trading using technical analysis would severely bias estimates of alpha since the resulting \( \text{cov}(\beta_m^*, r_{d+1}) \) would be small in monthly data.

HOLDINGS DATA

Tests of market timing using portfolio holdings data has yielded broadly similar results to those reported above. Analyzing the relationship between cash and equity holdings to see if the manager increases (decreases) her exposure to the market return just before a rise (fall) in the market index, Graham and Harvey (1996) and Becker et al (1999) find no evidence of successful market timing although Chance and Hemler (2001) using daily data find “professional market timers” exhibit skill even after transactions costs. Using the characteristic timing measure \( \text{CT} \), studies on US data find no evidence of successful market timing over a horizon of one year (Wermers 2000) – funds therefore cannot successfully time the characteristic benchmarks (e.g. accurately forecasting when returns on small stocks will be higher than those on large stocks).

Directly regressing fund betas (calculated using fund holdings and estimated stock betas) on the HM and TM market return variables \( \beta_i = \beta_0 + \lambda f [r_{m,i+1}] \), Jiang et al (2007) find stronger evidence for relatively “long-horizon” positive market timing over 3-month and 6-month horizons (over the period 1980-2003), based on private information (i.e. after correcting for the influence of publicly available information). For the median fund the economic value of market timing (based on Henriksson-Merton 1981 contingent claims approach) is around 0.6% p.a.\(^{58}\) – but note that the latter figure is the contribution to fund performance before transactions costs.

\(^{58}\) The value of the contingent claim is \( V = (1 + r_f)^{-\frac{1}{2}} E_r [\lambda e^{\sum r_{m,i}^2}] = (1 + r_f) \lambda (e^{\sum r_{m,i}^2} - 1) \). They also show that estimates of \( \lambda \) are not subject to artificial timing or interim trading bias, have more power than the return-based \( \gamma_{pm} \) measure on monthly fund returns (because estimates of stock-betas use a long time series of daily stock returns) and remain statistically significant for some funds after (bootstrap) corrections for non-normality and contemporaneous correlation across fund betas (e.g. due to herding by funds).
Tests of market timing have been addressed in a variety of ways in the mutual fund literature. Apart from timing based on stale prices, the evidence for positive market timing and volatility timing is relatively weak and is unlikely to provide investors with profitable strategies on a net return, risk adjusted basis. Overall, there is some good news here. If, subject to all the caveats and difficulties noted above, we are willing to accept that there is little or no market timing ability (i.e. either very few funds attempt to time the market or if they do they are unsuccessful) then for most funds, 4F-alpha may provide a "consistent" and reasonably accurate measure of security selection. This is not to deny that time variation in alphas and betas may need to be modeled for example, either based on publicly available information (i.e. conditional alpha-beta models) or "unobserved factors" or due to changing state variables (e.g. regime switching models) and that issues such as non-normality and contemporaneous correlation of specific risks still have to be dealt with – as noted above.

4. PICKING WINNERS: PERSISTENCE

Ex-post alphas (and CS) measure a fund’s average abnormal performance over some past data period. However it is also important to assess whether there are ex-ante rules which can be used to choose funds which subsequently earn statistically and economically significant abnormal returns, either for the fund or for the investor (i.e. after deduction of all costs) – in short whether there is persistence in fund performance. This section evaluates alternative methods used to measure persistence from a ‘statistical’ and economic viewpoint and reports results for US and UK mutual funds. For investors as a whole to benefit from persistence we also have to establish that they are ‘smart’ and allocate additional funds to ‘winners’ and withdraw funds from ‘losers’ – this is discussed in section 6.2. Of course, if we cannot establish persistence then investors may be wasting resources in chasing potential ‘winner funds’.

STATISTICAL PREDICTABILITY

Predictability and persistence are slightly different concepts. Persistence implies funds ranked as ‘past winners’ (losers), tend to stay winners (losers) in the future, so there is a positive correlation between past and future performance as funds maintain their relative positions – this is the “hot hands” effect (Hendricks, Patel and Zeckhauser 1993). Predictability allows for the latter but also for past winners to become future losers (i.e. reversals, implying negative

---

59 As we have seen Berk and Green (2004) demonstrate that if investors chase potential winners, the market may be efficient even though investors do not earn an equilibrium return above that on a passive fund. Picking winners usually involves a changing set of funds at each rebalancing period, so the finding of short-run persistence is not necessarily inconsistent with Berk and Green’s equilibrium result.
correlation) and also allows variables other than past performance to influence future performance.

Tests for persistence/predictability fall into two broad categories: statistical predictability or economically significant predictability or both. Statistical measures of persistence rank funds over some past horizon and measure the association between past performance and future performance. This approach measures the average association between the relative orderings of funds in the pre-sort and post-sort periods using correlation, regression or contingency tables. A weakness in such studies, however, is that although such tests may provide evidence of persistence it is often not clear how this may be exploited by investors. For example, rank correlations or a regression of pre-sort and post-sort alphas may establish statistical predictability but all post sort alphas could be negative. Alternatively, measured persistence could be due mainly to repeat losers rather than repeat winners. In addition (Spearman) rank correlations treat each point in the ranking equally and lack power against the hypothesis that predictability in performance is concentrated in the tails of fund performance. So while these common approaches have been used to establish predictability, investors are presumably more interested in the future absolute risk adjusted performance of both winners and losers taken separately.

RECURSIVE PORTFOLIOS

The recursive portfolio methodology of Hendricks et al (1993), Carhart (1997) and others allows a direct assessment of the economic as well as statistical significance of persistence. For example, using monthly data we might classify funds into deciles at time-\(t\), based on any fund attribute and form (equally weighted or value weighted) decile portfolios. The portfolio holding period (\(t, t+h\)) is then established (e.g. \(h=12\) months) and the monthly returns noted, after which rebalancing takes place and new decile portfolios are formed. This gives rise to a sequence of monthly ex-ante ‘forward looking’ (or ‘post-sort’) returns \(R^f_{i}(t,T)\) - where \(t = t+1, t+2, \ldots T\). If the ranking criterion is based on a fund’s alpha then the recursive portfolio method allows (past) factor loadings to change over time. If the complete (concatenated) monthly time series \(R^f_{i}(t,T)\) is used to estimate the ‘forward looking’ post-sort \(\alpha^f_{i}\) then we assume the ‘forward

---

60 When a fund dies sometime over the forward looking horizon, it is usually included in the portfolio until it dies and the portfolio is then rebalanced amongst the remaining ‘live’ funds, until the next rebalancing period. If we do not implement this procedure then ‘lookahead bias’ ensues.

61 In some studies there is a gap between the rebalancing dates and the measurement of post-sort returns. For example, at each rebalancing date we might track future returns only over horizons from \(t+3\) to \(t+15\) rather than from \(t\) to \(t+12\) – this allows a test of longer horizon persistence, without confounding the results with short-horizon persistence (e.g. see Teo and Woo 2001).
looking’ factor loadings (and alpha) over \((t, T)\) are constant - but this should be tested\(^{62}\). One method to mitigate the latter problem is to estimate the factor model using a moving window with a minimum of (say) successive 36 monthly values of \(R^f_i(t, T)\) - here we assume \(\alpha^f_i\) (and factor loadings) for fractile portfolios are constant over each 36 month period. Clearly, testing for ‘short horizon persistence’ and time-varying factor loadings, can only be achieved by using relatively high frequency data (e.g. daily).

There is an added danger when the model used to assess ex-ante performance is the same as that used in the (ex-post) ranking criterion, since any bias in the performance measure will apply to both periods, which may lead to an erroneous inference of persistence (Brown, Goetzmann, Ibbotson and Ross, 1992). Also any transactions costs (load fees, advisory fees) associated with high portfolio turnover may well eliminate any profit. Finally, most studies which use recursive portfolios examine very large decile or quartile portfolios of funds – this could disguise possible persistence among smaller portfolios of funds.

When constructing the CS measure from the \(R^f_i(t, T)\), one potential advantage over the parametric alpha method is that persistence on a risk adjusted basis can be measured over any horizon which is an integer multiple of the frequency of stock holdings or trade data – the latter minimum period is usually one quarter. This is because ‘risk adjustment’ using characteristics of the stocks in the portfolio, only requires data for the benchmark returns (and stock holdings).

It is worth noting that sorting may involve any rule that is thought to separate funds into future ‘winners’ and ‘losers’ (e.g. the number of letters in the name of the fund) but the performance metric will usually be a risk adjusted measure. Single or multiple sorts (e.g. by past return and by fund size) are possible but clearly there are data limitations on how far one can undertake multiple sorts – however as we see below, recent literature has concentrated on more ‘economically informed’ sorting rules which often involve multiple criteria. Whatever ranking criteria is chosen, then within any fractile, funds may be quite heterogeneous (e.g. the top return decile may contain funds with quite a large variation in returns, fund size, styles, turnover, etc.). Sorting into finer fractiles (e.g. top 1%) may involve a trade off between power (due to high and non-normal idiosyncratic risk) versus the heterogeneity of funds within the fractile. Here

\(^{62}\) If we use recursive past alphas to rank the funds but then use all the post-ranking data \(R^f_i(t, T)\) to estimate ex-ante performance we are being somewhat inconsistent since we are implicitly assuming that past fund factor loadings and alphas maybe time varying but future fractile portfolio parameters are constant. Elton et al (1996b) note that the composition of the top and bottom portfolios will very likely change over time and Carhart (1997) reveals this to be strongly the case.
bootstrapping across either individual funds or a cross-section bootstrap may be desirable – see section 5.2.

DATA SNOOPING BIAS

All tests are of course subject to data snooping bias. This arises when a given data set is used more than once for inference or model selection. In any finite set of data, a search over a large number of models (or trading strategies) will unearth some successful ‘outcomes’ (i.e. positive abnormal returns) purely by chance – and these are the ones that may be reported. Indeed with enough permutations we can find a successful mechanical trading rule on a set of random numbers - provided that we can test our rule on the same set of random numbers which we use to discover the rule. Add to this the likelihood of survivor bias in rules, (i.e. rules that do not work on new data are discarded) then the probability of finding at least one successful rule in even a long time series of data, may be quite high (Sullivan et al 1999, 2001).

Data snooping bias arises because tests are usually only conducted on a subset of surviving rules and not on all the other trading rules which did not survive. The dangers of data snooping bias in tests of persistence of mutual fund performance are high because of the large number of permutations across the different rules for portfolio formation, the rebalancing period chosen and the horizon over which future returns are evaluated. Guarding against these biases requires that successful ‘rules’ remain successful when confronted by new data.

5. PERSISTENCE: EVIDENCE

Having examined empirical results on ex-post performance in section 3 we now turn to evidence of possible ex-ante strategies whereby investors can ‘beat the market’ - in short, an analysis of the EMH applied to mutual funds. In section 5.1 we present evidence on ‘predictability’ (i.e. statistical association between past and future performance) and in section 5.2 we consider the economic significance of persistence (i.e. the size as well as the statistical significance in abnormal performance). We define short (long) horizon predictability/persistence as involving statistically significant effects over a period of less (greater) than one year.

5.1 STATISTICAL MEASURES

Key early studies find evidence of statistical predictability, over 1-month, 1-year and 2-year horizons (Grinblatt and Titman 1992, Hendricks, Patel and Zeckhauser 1993, Goetzmann and Ibbotson 1994). In general these early studies use databases that include only surviving funds but Brown et al (1992) pointed out that survivorship bias can result in a level of persistence
as found in the above studies, even though none exists. Brown and Goetzmann (1995) and Elton, Gruber and Blake (1996b) use a database free of survivorship bias (over 1970s to 1990’s) and find evidence of statistical predictability for rebalancing periods of 1 and 3-years, when using a number of sorting rules (i.e. past 1-year and 3-year alphas and t- alphas, from a three factor model plus a bond return).

The influential paper by Carhart (1997) uses a comprehensive database of over 1,800 US equity mutual funds (1963 – 1993), ranks funds into deciles based on past one-year net returns (or past 3-year, 4F- alphas) and finds some evidence of one-year persistence for the top and bottom decile ranked funds using a contingency table approach. He then tracks each decile fund’s gross returns over the subsequent 1-5 years and finds persistence of up to 3 years occurs for the lowest decile ranked fund but for all other deciles there is little or no evidence of persistence. More recently Teo and Woo (2001) use the contingency table approach after classifying funds into the nine styles used by Morningstar and find evidence of statistical predictability over horizons of up to three years, for funds ranked by style adjusted returns \((SAR_{i,t} = R_{i,t} - R_{i,S})\), where \(R_{i,t}\) is the equally weighted net return on all funds with style S). Blake and Morey (2000) and Morey and Gottesman (2006) also study the Morningstar 5-star rating service as a predictor of US domestic equity mutual fund performance and find that the change in Morningstar ratings methodology that took place in 2002 does predict future relative risk adjusted returns over the next 3 years - so 5-star funds do better than lower ranked funds - but the future 4F-alpha of the 5-star funds is negative.

Leger (1997) evaluates performance persistence for UK funds by simply counting the number of funds with positive or negative Jensen’s alphas in non-overlapping 5 year sub-periods (1984-93) and finds no evidence of persistence – a result reinforced by the contingency table approach, for one-year rebalancing using raw returns and Jensen’s alpha measures with persistence clearly driven by repeat losers (Lunde et al 1999, Allen and Tan 1999, Fletcher and Forbes 2002). Overall, studies of predictability on US and UK data using statistical measures, find evidence that poor performance persists for up to 3 years while there is mixed evidence that

\[\text{63} \quad \text{Also, Gorman (2003) over the 1986-2000 period, ranks 35 US small-cap funds on one, two, or three-year past conditional alphas. Using a regression of future alphas on past alphas (only), he finds evidence of positive persistence over the next year and reversals in years two and three.}\]

\[\text{64} \quad \text{Results are based on a regression of the 3 year forward looking 4F-alpha (over 2002-2005) on (0, 1) dummy variables for the Morningstar ratings (1 to 4) given in 2002. The intercept is statistically significant with a value of about -0.15, implying that the average 5-star rated fund has a forward looking alpha of about -1.8% p.a. Other performance metrics, such as the Sharpe ratio, Jensen’s alpha and a conditional beta model (measured both with and without returns adjusted for loads) are also used and give qualitatively similar results.}\]

\[\text{65} \quad \text{Similar results are obtained using a regression of last periods ranked (abnormal and raw) returns on next periods returns or using the Spearman rank correlation coefficient.}\]
winners repeat over periods in excess of one year. But as we have noted such statistical measures of persistence do not necessarily imply an exploitable trading strategy – an issue we take up below.

5.2 RECURSIVE PORTFOLIO METHOD
LONG HORIZON

Elton, Gruber and Blake (1996b) was one of the first studies to examine the economic significance of persistence in a survivorship bias free sample of US funds (188 funds, 1977-1993) using a 3F model plus a bond return. For the top decile funds (over a 1-year or 3-year horizon), they find evidence of a small positive forward looking alpha of around 0.5% p.a. and for the bottom decile a forward looking alpha of between -2.4% and -5.4% p.a. (depending on the ranking and rebalancing periods used) – none of the top and bottom decile forward looking alphas are statistically significant taken individually, but their difference is statistically significant at the 5% significance level or above. Similar result are reported by Brown and Goetzmann (1995).

There was something of a watershed when Carhart (1997) using all US equity funds (1963 – 1993) suggested that persistence found in earlier studies using the CAPM or 3F-alpha, may be a manifestation of the momentum effect in stocks which are ‘accidentally held’ by funds, rather than funds actively choosing stocks with a high loading on the momentum factor. He finds $\alpha^{f,net}_{4F}$ is negative for all decile portfolios (significantly so for the bottom 3 deciles) – so poor performance persists but good performance does not66. Carhart also ranks funds by their loading on the momentum factor but funds with high past $\beta_{MOM,i}$ are found to underperform in the post-ranking period, using $\alpha^{f,net}_{4F}$. Thus funds which have high momentum factor loadings are not actively pursuing a momentum strategy but instead “accidentally” happen to hold last year’s winning stocks and by doing so, enjoy a high average (raw) return over the next year but a negative risk adjusted return.

Kosowski et al (2006) apply a cross-section bootstrap and update the Carhart (1997) data. Using the whole universe of domestic equity funds (and ranking on past 36 month alpha) they find the top decile exhibits persistence over one year with a (net return) alpha of 1% p.a. (p=0.05), with growth funds largely responsible for persistence in ‘winner’ funds67. At the bottom

---

66 The top decile has a 4F-alpha of -1.4% p.a. (t=1.6) and for the bottom three deciles alpha is -1.6% (t=2.5), -2.4% (t=3.1) and -4.8% (t=4.3), respectively. Similar qualitative results ensue when ranking into deciles based on past 4F-alphas (estimated recursively, using the previous three years of data and rebalancing annually).

67 They initially examine persistence over 1978-2002 using recursive (equally weighted) portfolios, ranked on the past 36 month (unconditional) 4F-alpha, with one year rebalancing periods. When ranking over the shorter horizon of past 1-year, 4F-alphas and using a 1-year rebalancing period they find the top two winner deciles have significant persistence.
of the performance distribution (for all funds), deciles 6-10 have significantly negative abnormal performance (of about -1% p.a. for deciles 6-9 and -3.5% for decile-10).

Persistence in performance can be examined by controlling the FDR at the portfolio formation date, in order to limit the proportion of "lucky funds" in our ex-ante portfolios. For example, as the significance level for t-alpha is increased, we will obtain more "significant funds" in our portfolio but if this is accompanied by an increase in the FDR, many of these significant funds may be merely lucky – in forming portfolios of funds it may therefore be prudent to include a small number of significant funds which have a low FDR, rather than form a larger portfolio of significant funds having a higher FDR. This allows us to recursively identify a subset of funds for which the FDR is less than some chosen value, (say 10%) and provides the investor with a subset of significant funds to include in a fund-of-funds portfolio, for which she has set the FDR at an acceptable level. We may then estimate a "forward looking" alpha for this changing portfolio of funds. This would appear to be an intuitively better way to form ex-ante portfolios than the standard method of including funds in a given fractile (e.g. top-decile t-alpha), many of which could be false discoveries (i.e. lucky funds).

Unfortunately in practice it is difficult to 'hit' the desired FDR = 10% target when forming positive-alpha recursive portfolios annually. In practice, even attempting to get close to hitting the positive FDR=10% portfolio, tends to merely result in the inclusion of a relatively small number of funds (compared with say the number in the top-decile t-alpha portfolio) all of which have very low p-values. Perhaps it is therefore not too surprising when such a portfolio has a higher ex-ante alpha than a relatively large "top-decile" portfolio. Using the FDR approach BSW (2009) find evidence of positive persistence for an FDR target of 10%, with $\alpha' = 1.45\%$ p.a. ($p = 0.04$) for US equity funds – they do not report results on negative persistence. For UK funds there

$\text{with 4F-alphas of 1.4\% p.a. (p<0.01) for the top decile and 0.84\% p.a. (p<0.01) for the second best decile fund, and again deciles 6-10 have statistically negative average performance of around } -1.8\% \text{ p.a. Note that Carhart (1997) finds no persistence for the top decile funds and the difference between the two studies may be attributed to the different sample period and more importantly to the non-normality in the specific risk of the top decile funds, which is taken into account in the bootstrap but not necessarily with Carhart's parametric p-values. Also, an earlier study by Chan, Chen and Lakonishok (2002) using a 3 x 3 sort on size and BMV with annual rebalancing, finds that only small-cap growth funds give a statistically significant 4F alpha.}$

$\text{68 BSW (2009, table V) find that attempting to hit a target } FDR^* = 10\% \text{ results in portfolios that for 6 years out of their 27 portfolio rebalancing periods has a minimum estimated } FDR^*_n \text{ of over 70\%, for a further 6 years the } FDR^*_y \text{ is between 30\% and 60\% and for the remaining 14 years is between 10\% and 30\%.}$
is neither positive nor negative persistence for FDR=10%, winner or loser portfolios (Cuthbertson et al. 2010).

Many studies of persistence rank funds using past raw returns. Teo and Woo (2001) argue that this may not produce performance persistence because past raw return winners contain funds with different styles from year to year, so that ranking on past returns may not pick up managerial skill within a style. For example, past ‘winners’ may be mainly growth funds in one year, while in the next year value funds predominate. We should be looking for the best managers within each style. Note that ranking on the basis of 4F alphas should mitigate this problem but Teo and Woo (2001) suggest that an ex-ante portfolio formation rule based on a directly measured style adjusted return (SAR) provided by Morningstar may be more appropriate when trying to pick future winners.

Sorting funds into deciles (1984-99) on past style adjusted returns $SAR_1's$ and forming recursive portfolios, Teo and Woo use the spread between the past winner-loser decile forward looking returns $SPRD_{w-l}^f(t,T) = SAR_w^f(t,T) - SAR_l^f(t,T)$ as the dependent variable in the 4-factor model. They find statistically significant 4F-alphas of around 0.3%*** p.m. (3.6% p.a.), for portfolios formed on either past 1-3, 2-4, 3-5, or 4-6 years $SAR_1's$. But there are no results reported for the separate post-formation alphas of past winners and losers, only for the difference between the winner and loser deciles. But if you cannot short-sell mutual funds (or successfully use ETFs or replicate the portfolio composition of the loser funds and short sell their constituent shares) then it is difficult to see how in practice one might exploit the estimated “spread alphas”. However, the reported cumulative SAR’s over 70 months (i.e. about 6 years) for past winner and loser portfolios (sorted on past 1-3 year SAR’s) are 2% and minus 16% respectively, so the economic impact of ‘persistence’ seems to lie with the past losers rather than the past winners and as noted above, ranking on 4F alphas also appears to adequately pick up “style effects”.

Overall it appears that the evidence for economically significant long-run persistence for US past loser funds is well established, whereas evidence for persistence of past winners is less well established. We now turn to evidence of short-run persistence based on monthly or quarterly rebalancing.

---

69 Teo and Woo (2001) also find that persistence (based on 4F alphas) for top minus bottom deciles also holds when ranking portfolios by raw returns for formation periods of 1-3 years, 2-4 years and up to 3-5 years – note that this squares with Carhart’s (1997) results since the persistence found in the top minus bottom decile excess returns in Teo and Woo appears to be predominantly due to the bottom decile.
SHORT HORIZON

In our discussion of market timing we noted the theoretical difficulties in establishing performance attribution using parametric factor models and discussed the possibility of time varying parameters due to publicly available information. Mamaysky, Spiegel and Zhang (2007) question both the use of unconditional constant factor models and ranking on badly mis-measured variables such as past alphas. They argue that it is unlikely that any fixed parameter model of performance can adequately capture the myriad of portfolio risks and the diverse range of trading strategies pursued by the universe of mutual funds. Therefore, ranking by estimated alphas might result in the top (bottom) deciles containing genuine winners (losers) but also a substantial number of funds with the highest estimation errors – particularly if less than 60 monthly observations are used. To overcome this problem they suggest using a time-varying parameter model (e.g. recursive OLS or Kalman filter) coupled with a back-testing technique in which the factor model must exhibit some past predictive success before a fund may be included in a given portfolio. When the rebalancing period is every 12 months there is little evidence of persistence but for monthly rebalancing the top decile portfolio has a statistically significant abnormal (net return) alpha of 2.5% (t=4.0) to 4.5% p.a. (t=3.2) depending on the model used. (Similar results apply for the top-5, top-10 and top-20 funds.) There is also significant evidence of negative abnormal performance for the bottom five deciles.

We noted that the “power” may be greater when abnormal returns accrue over short horizons and trade data is available monthly (Kothari and Warner 2001) – unfortunately the latter is usually only available quarterly. In an attempt to increase power and to find a more economically informative signal of future performance Kacperczyk, Sialm and Zheng (2006) investigate whether a fund’s actions between portfolio disclosure dates provides incremental information which could be used by investors to ‘pick winners’. They measure the ‘unobserved actions’ of funds by the ‘return gap’ – the difference between the observed quarterly net return and the quarterly buy and hold return (using previously reported portfolio weights). The ‘return gap’ measures any benefits of interim trades less any hidden costs (e.g. transactions costs), over each quarter. Sorting funds on the past return gap either over the previous 4 or 15 months gives

---

They suggest two sequential filter rules to identify potential winner funds. For example, a successful strategy for including a fund at time t is first to estimate the factor model with 60 observations prior to time t-2 and then (i) only accept a fund if the forecast made at t-2 of the fund’s alpha for t-1 has the same sign as its excess return over the market \( R_{t-1} - R_{m,t-1} \), and (ii) re-estimate the factor model using 60 months of data up to t-1 and only accept a fund if its estimated alpha and market beta are in the ranges ±2% p.m. and 0 to 2, respectively. This group of ‘accepted’ funds are then sorted into deciles based on their estimated alphas.
\[ \alpha_{4F}^{f, \text{net}} = -4.0^{***} \text{ % p.a.} \] for the past loser decile with monthly rebalancing. There is no persistence in the past winner decile unless we also use a ‘predictive filter’ which results in \[ \alpha_{4F}^{f, \text{net}} = 2.52^{**} \text{ % p.a.} \]. The latter is due in part to higher “hot” IPO allocations being allocated to high return gap funds (Reuter 2006), which might also be funds affiliated to the lead underwriter – the nepotism hypothesis - Ritter and Zhang (2007). These IPOs are the source of the ‘unobserved actions’ of funds.

Instead of using a predictive filter to identify ‘winner funds’ Bollen and Busse (2005) seek to refine the choice of ‘winners’ by allowing factor loadings and alpha to change each quarter, by using daily data. They find the top decile portfolio has a statistically significant abnormal net return alpha (using bootstrap standard errors) in the post-ranking quarter of 1.3%*** p.a. while deciles 6-10 have statistically significant negative abnormal performance of between minus 0.8% to 3.2%*** p.a. Because the post-ranking recursive regressions are estimated quarterly they effectively mimic (non-parametrically) either a conditional model or a time varying parameter model (e.g. Mamaysky, Spiegel and Zhang 2007).

**EXTRANEOUS AND PRIOR INFORMATION**

In section 3.5 we noted that the use of extraneous information may improve inference about ex-post performance of funds in the extreme tails. Can this methodology improve sorting rules to identify ex-ante future winners and losers? Additional precision is achieved in estimating individual fund alphas when extraneous data on returns on “seemingly unrelated” passive assets is used (Stambaugh 1997, Pastor and Stambaugh 2002, Busse and Irvine (2006)). The procedure is based on the idea that truncating a set of returns so all return series are of equal length is inefficient – but this is exactly what occurs in standard estimation of fund alphas. The intuition behind this approach is that returns on passive non-benchmark assets may be correlated

---

71 Further evidence of an incremental effect of the return gap on short-term persistence is confirmed in panel regressions of \( \alpha_{4F}^{f, \text{net}}(i,t) \) on the lagged return gap (and other fund characteristics, including the lagged excess fund return).

72 The ‘predictive filter’ used is similar in spirit to that of Mamaysky, Spiegel and Zhang (2007) and is to only include funds where the sign of the average excess return (over the market return) equals the sign of the return gap.

73 They use daily data on 230 mutual funds (January 1984 to December 1995) and form decile recursive portfolios each quarter ranking on 4F-alpha (augmented by the HM or TM market timing variables). Their results are not inconsistent with Carhart (1997) who finds no positive persistence. This is because Carhart ranks funds on raw returns, uses longer pre- and post-sort horizons and concatenates the post-ranking returns when estimating alphas. Bollen and Busse (2004) simulate the latter effects with their daily data and consistent with Carhart, also find no positive persistence over horizons longer than one quarter.

74 Passive assets are those that do not appear in the factor model for fund returns. So, for example the SMB factor would constitute passive assets for the CAPM factor model, while industry returns would constitute passive assets for the 4F-model. Pastor and Stambaugh (2002) show that incorporating a long time series of passive asset returns can substantially alter the estimate of alpha and its precision (compared with the standard approach) and also that differences in alpha estimates across models (e.g. CAPM, 3F-model) are attenuated when using passive asset returns.
with fund holdings, so incorporating a long data series for non-benchmark returns increases precision for alpha. Busse and Irvine (2006) apply the methodology to tests of persistence, using daily data (1985-95) on 230 equity fund returns and data on non-benchmark assets (from 1968 onwards). Using a variety of alternative factor models they rank funds into deciles based on both standard alpha, a frequentist alpha (which incorporates non-benchmark returns) and a Bayesian alpha. With quarterly rebalancing they demonstrate increased predictability (based on Spearman rank correlations of pre- and post-alphas), for both the ‘non-benchmark’ frequentist-alphas and the Bayesian alpha rankings, compared with the standard frequentist alpha rankings. This provides prima facie evidence that extraneous information may help identify portfolios which yield performance persistence.

The idea that not all information about a fund’s future performance is encapsulated in a fund’s past alpha is also taken up by Cohen, Coval and Pastor (2005) but their ‘additional information’ uses the holdings of other successful funds, to help predict a particular fund’s performance. They draw on the ‘home-bias’ mutual fund literature which notes that physical proximity may facilitate relevant information transmission, which results in a concentration of fund assets in geographically ‘nearby companies’ or fund managers in the same city having similar portfolios (Coval and Moskowitz 1999, Hong, Kubik and Stein 2005). Specifically, Cohen, Coval and Pastor (2005) provide an alternative ranking metric for fund performance based on ‘commonality’, that is how closely a particular fund’s stock holdings currently mimic the stock holdings of funds which have performed well (based on their past alphas)\(^75\). The idea is that additional precision about fund-i’s performance is improved if information across all funds is pooled. They use data from April 1982-September 2002 (with a maximum of 1,502 funds) and rebalance quarterly. For example, based on a double sort (into 25 quintiles) they find that the “top-alpha, top-commonality” ranked portfolio has \(\alpha_{5,5}^{(5,5)} = 4.61\%\) p.a. and the bottom ranked portfolio has \(\alpha_{1,1}^{(1,1)} = -3.79\%\) p.a. with the long-short portfolio yielding 8.4\%*** p.a.\(^76\). Once again there is evidence for the incremental predictive power of extraneous information but note that reported alphas are gross of TERs and therefore represent abnormal returns to the fund.

---

\(^75\) Essentially fund-i’s ‘skill’ (\(s_{k,i}\)) is a weighted average of other funds’ alphas, with the weights depending on the covariances between fund-i’s portfolio weights and the current weights of the other managers. If fund-i holds only stocks that are held by no other manager then \(s_{k,i}\) collapses to \(\alpha_i\), otherwise \(s_{k,i}\) is high if it has portfolio weights which are similar to portfolio weights of other funds with high alphas. The analogy they use is that after observing a group of basketball players you note that the average score is 8/10 for the two-handers, but only 4/10 for the one-handers. Then if two players, one one-hander and one two-hander are observed, each currently with a 4/5 score, then you would bet that the two-hander is more likely to have a higher score out of 10 - the track records of the other two-handers are better than the one-handers, so you assume the current two-hander has a better technique and the one-hander is more likely to have been lucky with his first 5 shots.

\(^76\) Separate t-statistics for the (5,5) and (1,1) portfolios are not given, so we cannot infer if the strategy gives statistically significant abnormal returns solely to the ‘winner’ portfolio, which would remove the problem of short-selling mutual funds. Also, the figures reported above are the largest found for the various sorts, across a variety of models.
rather than to investors – as rebalancing is quarterly, transactions costs (load and advisory fees) may be high.

So, as far as the use of filter rules, extraneous information and Bayesian approaches are concerned they do appear to provide incremental predictive power for future fund performance when rebalancing is frequent (monthly or quarterly) – so, for picking ex-ante winners these approaches are promising but as yet are far from definitive, given the unknown impact of stochastic rebalancing costs (due to loads, advisory fees and search costs)\(^{77}\).

### 5.3 STOCK HOLDINGS AND TRADE DATA

We now turn to the question of whether the use of stock holdings and trade data provide more definitive results for persistence in performance amongst US equity funds. An early study by Grinblatt and Titman (1993) uses a performance measure based on the covariance between security returns and portfolio weights and using a sample of 155 mutual funds they report positive persistence among both top and bottom performing funds, based on fund rankings in two sub-periods.

Chen, Jegadeesh and Wermers (2000) examine the important question of whether ‘winning’ funds actively follow momentum strategies or whether as described by Carhart (1997), winning funds accidentally hold the previous periods winning stocks and hence benefit from the well documented momentum effect\(^{78}\). They find that past CS returns (of the ‘holdings’ or ‘buys’) of the winning funds are higher than the past CS returns of the losing funds by around 2%** p.q. - this seems to point to active momentum investing by winning funds relative to losing funds. However, only the ‘CS(all holdings)’ of winning funds subsequently outperform the CS returns of losing funds (by around 0.5%** p.q.) and only in the current and next quarter - so momentum in gross returns is short lived and is not due to an active momentum strategy. An interesting question addressed by Wermers (2003b) is whether persistence in gross returns results in positive net return alphas. For all equity funds, the top minus bottom decile (W-L) has \(CS_{W-L}\) (all

\(^{77}\) At each rebalancing date a different set of funds constitutes the ‘winner portfolio’ but it is impossible to know in advance the turnover of funds – this gives rise to stochastic rebalancing costs in the form of load fees and any advisory fees. There is also an externality effect on long-term fund holders as they have to bear the commissions, bid-ask spreads and price impact costs consequent on fund flows caused by investors who switch funds.

\(^{78}\) They use stock holdings and buy and sell trades for all (equity) mutual funds (1975-1995, quarterly rebalancing), where ‘winners’ and ‘losers’ are the top and bottom quintile of funds ranked on past one year net returns.
holdings) = 0.83%** p.a.\textsuperscript{79} – but this does not carry over to $\alpha_{3F}^{\text{net}}$ and $\alpha_{4F}^{\text{net}}$ measures, which are negative (but usually not statistically significant). So although one year persistence in risk adjusted gross returns (CS) for stocks is evident, indicating skill, this does not carry over to abnormal net returns for investors.

**FUND STYLES AND HERDING**

There is evidence that herding can affect stock prices (Lakonishok, Shleifer and Vishny 1992) and that fund purchases themselves may impact on prices particularly at reporting dates (Carhart et al 2002a) or due to the allocation of ‘hot’ IPOs to particular funds within a fund family (Gaspar, Massa and Matos, 2006). Wermers (1999) follows up these ideas for different fund styles. Using stock holdings data and quintile sorts based on several measures of herding he finds that herding influences gross abnormal fund returns over the following 6 months\textsuperscript{80}. This effect is more pronounced among small stocks and growth orientated funds and the effect is strongest in the first 10 year period 1975-1984 rather than the last 10 years 1985-94, where only herding in small stocks effects subsequent returns. So herding by mutual funds particularly in small stocks may be a source of the momentum effect in stock returns – but given large bid-ask spreads (Lesmond, Schill and Zhou 2004) it is an open question whether such effects are exploitable. In a later paper Wermers (2003a) follows up the idea that return persistence may be stronger in particular fund styles and finds persistence in a sample of “winner minus loser” decile sorted growth funds over a 4 year horizon ($CS_{W-L}$,(holdings) = 2%*** p.a.). This arises mainly from ‘same stock’ purchases rather than purchases of new stocks thus any short run herding by past ‘growth fund’ winners is into stocks already held, which tends to push up their prices reinforcing any underlying momentum effect.

Overall, our conclusions from studies based on stock holdings and trades of all funds is that past US winner funds do (‘accidentally’) hold more momentum stocks than past losers and persistence is relatively short lived - even before trading costs and management fees. There is stronger evidence of long run persistence for growth funds, due to herding into existing momentum stocks.

**5.4 OTHER SOURCES OF PERSISTENCE**

\textsuperscript{79} The characteristic timing CT is positive for past winners and losers and for W-L, but CT is not statistically significant, indicating that past winning funds do not increase their holdings of benchmark stocks with high future returns – they cannot market time the stock characteristics.

\textsuperscript{80} Here, abnormal returns are returns on stocks in excess of their (equal weighted) average-size quintile return. Grinblatt, Titman and Wermers (1997) also find evidence of herding by funds and Musto (1997, 1999) finds evidence of herding around reporting dates.
The above studies have examined persistence based on ranking funds on some measure of past performance but some studies have also investigated predictability based on ‘sorting rules’ such as turnover, industrial concentration of fund holdings, managerial activity and incentive fees, deviations of stock holdings from their benchmark index and the utility derived from the chosen portfolio. We briefly discuss these below.

**TURNOVER**

It is possible that funds with well informed traders have high turnover (= ‘winners’), which then results in higher future abnormal returns. Conversely, there may be many uninformed traders who ‘churn’ stocks in order to look skilled but in fact are not. Wermers (2000) for the period 1975 – 1993 finds the W-L turnover ranked decile has a positive risk adjusted gross CS measure but all (decile turnover) funds have an $\alpha_{4F}^{fs, net}$ of about minus 1% p.a. - so ranking by prior turnover does not provide a trading rule which gives positive abnormal net returns to the investor$^{81}$.

**INDUSTRIAL CONCENTRATION**

It has been argued that funds that have relatively high levels of industrial concentration ICI (relative to the market index) may have higher future performance, because these ‘winner’ funds obtain an informational advantage by carefully analyzing just a few industries$^{82}$. Using a recursive portfolio approach (quarterly rebalancing, equally weighted) and decile sorts based on ICI, Kacperczyk et al (2005) find that for the decile ‘winner’ portfolio, $\alpha_{4F}^{fs, net} = 2.1\%$ p.a. but $\alpha_{4F}^{fs, net}$ and the CS (holdings) performance measures are both not statistically significant.

**ACTIVE MANAGEMENT AND INCENTIVES**

Cremers and Petajisto (2006) find stronger support for persistence after ranking funds into (5x5) quintiles based on a double sort on an index of active management, (which they call Active Share, AS)$^{83}$ and the previous years’ average benchmark-adjusted return ($\bar{R}_i - \bar{R}_b$).

---

$^{81}$ Funds are ranked into deciles (also repeated for quintiles) by their levels of turnover during the previous year and portfolios are rebalanced annually. Most of the W-L gross return of 4.3% p.a. is due to AS$\bar{R}_{W-L} = 2.2\%$*** p.a., rather than due to stock picking or market timing skills since CS$\bar{R}_{W-L} = 1.22\%$ p.a. and CT$\bar{R}_{W-L} = 0.23\%$ p.a. which are not statistically significant (Wermers 2000, table VI).

$^{82}$ Each share held by the mutual fund is allocated to one of 10 industrial classifications and the industrial concentration index: $ICI_{i,k} = \sum_{k=1}^{10} (w_{i,k} - \bar{w}_{k,k})^2$ where $\bar{w}_{i,k}$ is market (value) weight of industry-k in the market index and $w_{i,k}$ is the weight of fund-i’s holdings of stocks in industry sector k.

$^{83}$ The Active Share index is similar to that for ICI, namely $AS_{i,k} = \sum_{k=1}^{N} |w_{i,k} - \bar{w}_{i,k}|$ where $\bar{w}_{i,k}$ is (value) weight of stock-k in the fund’s benchmark index and $w_{i,k}$ is the weight of fund-i’s holdings of stocks. Note that for AS the absolute
Rebalancing annually they find the ‘past winner’ quintile portfolio (i.e. high AS and high past return) has $(\bar{R}_i - \bar{R}_b)^{\text{net}} = 3.69\%$** p.a. and $\alpha_{4F}^{f,\text{net}} = 2.29\%$* p.a., with ‘winner minus loser’ portfolio having $\alpha_{4F}^{f,\text{net}} = 3.09\%$*** p.a. Massa and Patgiri (2008) define “high incentive funds” as those with a linear compensation structure and “low incentive funds” as those where the percentage advisory fee declines as total assets increase. Sorting funds annually using past 4F-alpha (as in Carhart 1997 but using quintiles) and then into quintiles using an index of the degree of managerial incentive, Massa and Patgiri (2008) find that last years “high return-high incentive” funds have $\alpha_{4F}^{f,\text{net}} = 4.8\%$ p.a. (t=2.36) over the period 1996 – 2003. Persistence in past high incentive winners is found to be due to an active strategy since “unobserved actions” rather than buy and hold returns are the main contributor to the winner funds performance.

Overall, ranking funds on the basis of industrial concentration or turnover does not produce winners that persist. Double sorts on “active Share and past returns” or “high return-high incentives” provides stronger evidence of positive future abnormal return performance but the question remains as to whether this is exploitable after transactions costs such as loads and advisory fees.

DISCRETIONARY AND LIQUIDITY TRADES

As we have seen studies that use trade data to measure fund returns argue that the approach is more likely to detect skill because an active decision is required to change one’s asset holdings. But this argument ignores the possibility that some trades may be due to exogenous changes in net cash flows. A number of studies have tried to isolate “discretionary trades” from “liquidity trades” on the grounds that the former are more likely to involve genuinely active bets. Edelen (1999) finds that discretionary trades have a positive effect and liquidity trades a negative effect on subsequent performance. This is because liquidity trades unexpectedly alter the funds cash holdings (“dilution effect”), move the fund from its target portfolio and cause managers to undertake non-discretionary trades which are likely to lose money.

value of the difference in weights is used and the index is the ‘style index’ of the fund, whereas for ICI the square of the weights is used with only a single market index - see Cremers and Petajisto (2006) for a discussion of the relative merits of these two measures.

$^{84}$ It is really this two-way sort on past returns and past Active Share that gives relatively large ‘forward looking’ alpha estimates for the past ‘winner’ portfolio. If we merely sort on past Active Share, then the highest quintile has $\alpha_{4F}^{f,\text{net}} = 1.36$ (t=1.43). The persistence of ‘past winners’ is even stronger if there is a three-way sort by size, AS and past returns, with the smallest size quintile (but with highest AS and past return) giving $\alpha_{4F}^{f,\text{net}} = 5.63\%$ p.a. (t=3.66). In a broadly similar vein Fang and Kosowski (2006) report that stocks sorted on the basis of mimicking top analysts stock recommendations give portfolios with higher post-sort 4F-alphas than sorts based on mimicking average analyst’s recommendations – demonstrating the superior information content of “semi-private” information over public information.
Perhaps if we could more carefully isolate discretionary trades from liquidity trades we might find evidence of skill. Alexander, Cicci and Gibson (2007) attempt this by taking 324 US equity funds (January 1997-December 1999) and use a double sort into 25 quintiles based on net flows and the dollar value of trades. ‘Valuation motivated trades’ are defined as large dollar-buys (sells) which take place when there are heavy net outflows (inflows) while ‘liquidity motivated trades’ are funds where small dollar buys (sales) are accompanied by large inflows (outflows). They find that valuation motivated trades earn substantially more on a risk adjusted gross CS basis over the subsequent year than do liquidity motivated trades - even after trades resulting from possible tax loss selling or mandated reporting months (which might involve window dressing) are excluded. So there is some evidence of skill for ‘valuation motivated trades’.

UTILITY

While all of the above studies of persistence involve portfolios of funds, the portfolios of past winners and losers are formed on fairly ad-hoc grounds. In contrast, Avramov and Wermers (2006) chose optimal portfolio weights at each monthly rebalancing date, in order to maximize next periods quadratic utility (which therefore depends on mean and variance). Does this result in a better ex-post abnormal performance than say, decile sorts on the basis of past winner raw returns (‘hot hands’) or past 4F-alphas, which give at best small positive values for $\alpha_{4F}^{f,net}$? This approach results in a portfolio which yields an $\alpha_{4F}^{f,net}$ that is statistically significant, ranging between 9-12% p.a. – providing investors utilize the predictability in the factors which are driven by macroeconomic variables such as the dividend yield, default and term spreads and the interest rate. In the model investors are Bayesians and take account of estimation risk in choosing the optimal weights (Baks et al 2005, Barberis 2000). Clearly, the ex-post abnormal return to this ‘recursive optimal portfolio strategy’ is substantial but it appears that investors would have to mimic holding up to 1,300 funds and rebalance every month – an issue that needs further investigation particularly with respect to transactions costs.

There are far fewer studies of persistence using UK data and in the main they are consistent with US results namely, that although past winners may persist in terms of pre-

---

85 Funds include actively managed funds, index funds, sector funds and exchange traded funds (ETFs). They allow investors to hold (positive) weights in these 1,301 (no-load domestic equity) mutual funds over the period December 1979 to November 2002. In this mean-variance framework there are no hedging demands – the importance of the latter see for example, Ait-Sahalia and Brandt (2001), Campbell et al (1999, 2003), Viceira (2001) and Cuthbertson and Nitzsche (2004).

86 The model of fund returns is the 4F conditional alpha-beta model, where the 4-factors then each depend on the lagged macroeconomic predictor variables $z_{t-1}$ and the latter are themselves forecast using a VAR model – a set up similar to Barberis (2000).
expense 3F-alphas they do not persist on a net return basis over a 1-year horizon or more, while past losers have statistically significant negative (gross) 3F-alphas of around -3.5% p.a. (Fletcher 1997, Fletcher and Forbes 2002, Quigley and Sinquefield 2000)\textsuperscript{87}.

**SUMMARY**

The evidence on persistence is voluminous, yet attempting to give a brief yet balanced summary of performance persistence across US and UK studies is difficult. Different studies examine different sample periods, some use purely ‘statistical measures’ (e.g. correlation, rank correlation, contingency tables) and others use economic measures of performance such as raw returns (gross or net) or adopt different measures of risk adjusted abnormal return. Furthermore, different studies rank funds on different fractiles and use different rebalancing and holding periods. Finally, not all studies control for survivorship and look ahead bias or cover the whole universe of funds and with so many sorting rules used, data snooping bias is an issue.

Even after noting the above caveats, it seems that there is some persistence amongst the top decile of all US funds ranked by several characteristics including past raw returns, 4-factor alphas or CS measures. Using a risk adjusted gross returns (CS) measure persistence may last up to four years for a small number of growth funds and for up to one year when the top decile is formed from ‘all funds’. Persistence amongst past winners does not seem to last longer than 1-year when we use an abnormal net returns metric such as the 4-factor alpha. For example, in Kosowski et al (2006), ranking on past 3-year, 4F-alpha and rebalancing annually gives a top decile $\alpha_{4F}^{f,net} = 1\%$ p.a. Unless investors can mimic, with a small number of funds, the performance of the top decile portfolio (which may currently contain over 180 funds) and avoid load fees to minimize stochastic rebalancing costs, it is doubtful that a significant exploitable ‘persistence anomaly’ exists. In contrast, there is strong evidence that poor performance persists fairly uniformly across deciles 5-9 with $\alpha_{4F}^{f,net}$ around -1% p.a. and the bottom decile has $\alpha_{4F}^{f,net} = -3.6\%$ p.a. Broadly similar results apply to the relatively few comprehensive UK studies on persistence.

Recent work using filter rules (Mamaysky, Spiegel and Zhang 2007, Cremers and Petajisto 2006), extraneous information (Pastor and Stambaugh 2002, Cohen, Coval and Pastor 2005), managerial incentives (Massa and Patgiri 2008) or incorporating the predictability of factors in an optimal portfolio (Avramov and Wermers 2006) have demonstrated a prima facie case for successful ex-ante strategies – but these may only be feasible for rather sophisticated

\textsuperscript{87} The exception here is Blake and Timmermann (1998) who with one-month rebalancing find statistically significant positive and negative persistence in alpha-performance, particularly for smaller company funds - but results vary depending on the risk adjustment model used.
investors and any predictability would need to be shown to be robust over time and outweigh any transactions costs such as loads and advisory fees, due to frequent rebalancing. Overall our analysis of persistence provides useful insights for investors about which funds to avoid - but offers much less certainty about which funds to purchase.

6. FUND FLOWS AND PERFORMANCE

In section 6.1 we examine characteristics which might influence a fund’s abnormal return, concentrating particularly on relative costs. In section 6.2, we ask the question “Is Money Smart?” If there is persistence in performance of past winner or loser funds and money is smart, then we expect to observe investors switching from loser to winner funds, with the former either ceasing to exist (or changing to a successful strategy) and the latter giving rise to positive abnormal returns – at least over the short-run. If performance persists at the fund manager level, rather than at the fund level, then the ‘smart money’ should follow successful managers – this is examined in section 6.3.

6.1 PERFORMANCE AND FUND CHARACTERISTICS

Do large funds perform better than small funds, or do high turnover or high cost funds provide a better return than low cost or low turnover funds? In short, what type of fund characteristics influence performance. Since each fund’s abnormal return and the relationship between the abnormal return and fund characteristics may vary over time a Fama-MacBeth (1973) rolling regression is often adopted:

\[ \alpha_{i,t} = \theta_i + \delta_i X_{i,t-k} \]

where \( \delta_i = (\delta_{1i}, \delta_{2i}, \ldots, \delta_{mi}) \) and \( X_{i,t-k} \) is the m-vector of fund characteristics at time t-k. Note that one of the regressors in \( X_{t-k} \) could be the fund’s previous performance, in which case we are measuring persistence, after accounting for other fund characteristics. An alternative is to estimate [23] using a suitable estimator for an unbalanced panel (usually with time fixed effects) but this method is less popular, in part because it allows fewer parameter estimates to be time varying (Petersen 2005 provides a comparison of the two methods).

Using a monthly rolling Fama-McBeth regression of 4F-net alphas of individual funds on fund characteristics, Carhart (1997) finds the total expense ratio (\( \text{TER} \)), turnover (\( \text{TURN} \)) and
load fees (LOAD)\(^{88}\) all have a negative impact on (abnormal net return) performance - of particular interest is the coefficient on the expense ratio: for every 100 basis point increase in TER, the net return alpha falls by 1.54%. Carhart also notes that the negative coefficient on load fees contradicts the oft-cited claim that such managers are more skilled than those of no load funds (see also, Chen et al 2004). Similar results are found by Chalmers, Edelen and Kadlec (1999) who rank funds over the 1984-1991 period into quintiles (rebalanced quarterly) using either total costs, trading costs, expense ratios or turnover. For all three cost ranking criteria, there is a strong negative and statistically significant relationship between costs and future performance, while turnover is also negatively related to performance but is not statistically significant (see also Warther 1995)\(^{89}\).

The most persuasive study on the impact of trading costs on performance is Edelen, Evans and Kadlec (2007) who directly estimate trade costs (commissions, spreads and price impact) rather than using “turnover” or indirect estimates as in earlier studies (Grinblatt and Titman 1989, Kacperczyk, Sialm and Zheng 2006). They find strong evidence that relative trade-size\(^{90}\), supplants fund-size and turnover as determinants of performance. While this result is qualitatively consistent with the Berk and Green (2004) diseconomies of scale argument, the fact that funds using relatively large trade-size incur losses is not, since active funds should switch to indexing before this situation arises. Edelen et al (2007) provide evidence that some (but not all) of “excess” trading is due to investor flows which result in liquidity trades rather than skilled discretionary trades – this may result in diseconomies as assumed on the Berk and Green (2004) model.

The impact of family strategies on performance is investigated by Massa (2003) who finds that the greater the degree of product differentiation (in terms of the dispersion of either fees or returns across a fund family) the lower are fund returns – this is because funds compete not only on return but also on providing a low cost switching option for fund investors within the same fund family\(^{91}\).

---

\(^{88}\) TURN and TER are monthly averages of annual figures and are contemporaneous while LOAD and TNA are lagged one year. TURN is reported turnover plus 0.5 times the percentage change in TNA (adjusted for investment returns and mergers). The rolling cross-section regression is from July 1966 to December 1993.

\(^{89}\) Using daily data Rakowski (2002) finds that returns are strongly influenced by the standard deviation of fund flows as well as turnover but not by expense ratios and loads, in contrast to studies using monthly data.

\(^{90}\) Relative trade size (TRADESIZE) takes a value of 1 for trades that are large relative to the average of funds in the same market capitalization strategy (and zero otherwise). The variable used in the Fama-MacBeth (1973) cross section regressions is (TRADESIZE*TRADECOST), where TRADECOST is the estimated sum of the percentage commission, spread and price impact costs per unit trade.

\(^{91}\) Given the difficulty in obtaining relevant UK data, relatively little work has been conducted on the relationship between winner and loser funds and fund characteristics. However, Fletcher and Forbes (2002) examine whether annual
FUND RESTRICTIONS, INCENTIVES AND PERFORMANCE

There is a wide dispersion in the legal restrictions applied across funds – for example in their use of derivatives, margin purchases, short-selling, borrowing and categories of restricted (usually illiquid) stock. However, such restrictions do not appear to influence abnormal returns (Almazan, Brown, Carlson and Chapman 2004). Funds which employ incentive fees (based on a benchmark such as the S&P500) might be expected to motivate existing managers or attract better managers. In 1999, only 108 US funds (1.7% of all funds covering 10.5% of total fund assets) used incentive fees and these had an average alpha (over the 1990-1999 period) of about 0.6% p.a., which exceeded that on non-incentive fee funds who had an average alpha of -0.4% p.a. – this differential performance is due to the lower expense ratio of the incentive-fee funds (Elton, Gruber and Blake 2003). Massa and Patgiri (2008) report that high incentive contracts lead managers to take relatively more risk but although high incentive funds have higher 4F-alphas than low incentive funds, both sets of alphas are negative.

INDUSTRIAL CONCENTRATION AND PUBLIC INFORMATION

As noted above Kacperczyk et al (2005) by sorting on a fund’s industrial concentration (in stock holdings) show that this helps predict future performance because manager skill is focused on a few sectors. This is also found to be the case when quarterly decile sorts on three alternative forward looking performance measures \( PERF_f(t,T) = \{ \alpha_{4F}^{f,net}, \alpha_{4F}^{f,s}, CS_f^{f,s} \} \) are regressed on fund characteristics as well as \( ICI \) and they find \( ICI \) is statistically significant but its economic impact is not large. For example, one standard deviation increase in \( ICI \) increases \( \alpha_{4F}^{f,net} \) by 0.5% p.a. whereas a 1% increase in the TER leads to a fall in \( \alpha_{4F}^{f,net} \) by 1.6% p.a. The idea that skill resides with specific activities of funds is taken up by Kacperczyk and Seru (2007) who develop a theoretical model whereby skilled funds rely less on publicly available charges, load charges and fund size are correlated with quartile (raw return) ranked portfolios of funds (rebalanced annually). The authors report very little cross-sectional variation in these characteristics and hence the authors suggest that such characteristics do not explain the “winner minus loser” abnormal returns.

The restrictions listed are combined into an index and the analysis then proceeds in the usual manner either by sorting based on the index (above and below average) and then forming long-short portfolios that are re-balanced annually or using a Fama-McBeth cross-section regression of the rolling value of \( \alpha_{4F} \) on the score index (and other fund characteristics). They use 324, US domestic equity funds from January 1997 to December 1999.

The performance model is the 3F model with additional factors for bond returns and an international index. The non-incentive fee funds were paired with the incentive fee funds by size and investment objective.

They use (time) fixed effects, unbalanced panel estimation (quarterly, 1984-1999). Variables include the age of the fund \( \ln(AGE) \), \( NCF \) is the previous quarters net cash flow and other control variables are \( TER, TURN \) and \( \ln(TNA) \). All variables are lagged one quarter except for \( TER \) and \( TURN \) which are lagged one year due to data availability. Variables \( ICI, \ln(AGE) \) and \( TER \) are statistically significant in the net return 4F-alpha regressions over the whole period and sub-periods 1987-93 and 1994-1999.
information (because it is already be compounded in prices) and more on private signals. The “reliance on public information” RPI of a fund is measured by the R-squared of a fund’s change in portfolio stock holdings on lagged values of the change in the consensus analysts forecast for specific stocks in the fund’s portfolio. They undertake a cross-section regression of fund performance (e.g. alpha or CS) on RPI and control variables (i.e. size, turnover, expenses and fund flows) and find that a one standard deviation increase in RPI reduces fund performance by about 0.5% p.a. It would be interesting to see if this result holds when the past performance of the fund is also included in the regression – then we could separate out the incremental effect of RPI from the documented performance persistence effect noted in other studies. Ranking funds on RPI and rebalancing monthly gives a statistically significant $\alpha_{F}^{f, net}$ of 2.16% p.a. for the long-short portfolio (top minus bottom 30% of ranked funds) – but it is not clear how much of this is due to holding the high-skill ranked funds (i.e. low RPI funds) or shorting the low-skill funds and as the latter is problematic it leaves open the question of whether such a strategy is exploitable.

LOAD AND NO-LOAD FUNDS

The presence of load fees has been attributed to fund managers trying to separate investors with different liquidity needs (Chordia 1996, Nanda, Narayanan and Warther 2000). But what is the relative performance of load and no-load funds? With the exception of Ippolito (1989), who finds that load funds earn rates of return that plausibly offset the load charge, most studies generally find that there is no significant difference between the performance of load and no-load funds (even before the former are adjusted for load charges) – see inter alia, Elton et al (1993), Grinblatt and Titman (1994), Droms and Walker (1994), Gruber (1996), Fortin and Michelson (1995) and Morey (2003). In addition, Morey (2003) examines relative performance within load funds using a number of risk adjusted measures (e.g. Jensen and 4F alphas) and finds there is little significant difference in abnormal returns between high load funds and low load funds. The above empirical results contradict the predictions of the theoretical model of Nanda, Narayanan and Warther (2000), where investor returns in load funds exceed those in no-load funds and investors in funds with high-load fees earn a higher return than those in low-load funds. (See Khorana et al 2008 for a detailed study explaining the differences in mutual fund fees world-wide.)

To summarize, for funds as a whole their net abnormal returns seem to be strongly and negatively related to costs such as TER, load fees and possibly trading costs. There is some evidence supporting the view that differences in ‘industrial concentration’, reliance on public information, the number of restrictions applied to funds and the use of incentive fees has some effect on future performance. But the quantitative impact of such effects does not appear to be

---

95 Houge and Wellman (2006) also note that load funds have higher expense ratios than non-load funds (50 bp difference over 2000-2004).
very large and is likely to be outweighed by switching costs. As far as load/no-load funds are concerned the evidence strongly suggests that abnormal returns on load funds do not cover the additional load fees charged. The latter may account for the decline in both the number of funds and cash under management in load funds (in the US) and is also consistent with (no-load) funds trying to recoup charges in higher 12-1b fees (Mahoney 2004). Here the evidence is clear, investors should choose no-load funds and funds with low expenses if they wish to increase the probability of higher net abnormal returns.

6.2 INVESTMENT FLOWS AND PERFORMANCE: IS MONEY SMART?

In a competitive market we might expect active investors to re-allocate cash away from past poor performers and towards past winners, in the expectation that this will increase future returns. Key areas for investigation are first, the relationship between past fund performance and subsequent fund flows and second, whether fund flows provide an investment signal which can be used to give economically significant future returns – that is, whether money is smart. These propositions can be tested in an event study framework by sorting funds into appropriate portfolios (e.g. on past flows or performance) and following their subsequent returns. Alternatively one can use a (Fama-MacBeth) cross-section regression approach with either performance or flow metrics as the dependent variable, with their lagged values as independent variables - plus other fund characteristics as control variables.

PAST RETURNS AND FUTURE FUND FLOWS

After allowing for the influence of fund characteristics (e.g. size, age, return volatility, affiliation with “star” fund families), an asymmetric relationship whereby superior past performance (for a number of alternative performance measures) attracts disproportionately large net inflows, is well established empirically (Ippolito 1992, Gruber 1996, Chevalier and Ellison 1997, Sirri and Tufano 1998, Massa 2003, Nanda, Wang and Zheng 2004). Given manager compensation is linked to the size of the fund this is consistent with favourable allocations of “hot” IPO’s to funds affiliated to lead underwriters (Ritter and Zhang 2007) and to strategic allocation of IPOs amongst funds in fund families (Gaspar et al 2006).

As we have seen this scenario is the basis of the equilibrium model of Berk and Green (2004), while the possibility of ‘strategy switching’ by poorly performing funds in the Lynch and Musto (2003) model predicts relatively low outflows from poorly performing funds.

Chevalier and Ellison (1997) note that the convex performance-flow relationship provides an incentive for managers who are performing worse than the market in the first part of the year, to increase the variance of their returns in the second part of the year, since they obtain a very large increase in fund inflows (and hence fees) if they are successful but do not suffer large inflows if they are unsuccessful. This is similar to the Goetzmann, Ingersoll and Ivkovich (2000) idea of performance manipulation. Evans (2007) notes that some (but not all) of the large impact of past good performance on inflows is due to incubation bias, since incubated funds tend to have high past returns and their percentage inflow is also high (because they are relatively small funds).
Using Fama-McBeth (quarterly and monthly) rolling cross-regressions approach, Barber, Odean and Zheng (2004) re-examine the impact of past returns and different types of fees on future fund flows. They find that (quarterly) net inflows into individual funds are positively related to past (excess) fund returns, returns squared and 12b-1 (advertising) fees but are negatively related to front-end loads - however, funds with high operating expenses do not experience reduced inflows. They argue that this demonstrates greater sensitivity of flows to advertising and load fees which are ‘visible’, rather than to operating expenses which are less ‘visible’ (see also Wilcox 2003 and Del Guercio and Tkac 2002, Ivkovic and Weisbenner 2009). In a similar vein, Sirri and Tufano (1998) and Jain and Wu (2000) find that funds which spend more in 12b-1 fees to advertise their recent good performance, experience higher inflows (relative to good funds who do not advertise) – which can be interpreted as funds minimizing the search costs for investors about past good performing funds. The latter is taken up by Huang, Wei and Yan (2007) who show that the slope of the flow-performance relationship depends on “participation costs” (i.e. expense ratios and load fees, star affiliation and fund family size).

Fund flows amongst poor performers are examined by Lynch and Musto (2003) who find that flows are less sensitive to past performance when past performance is relatively poor. Other key US studies also find evidence that net cash flow is less sensitive to poor performance – for a variety of performance measures (e.g. raw returns, style adjusted returns, returns in excess of the market, Jensen’s alpha – see Sirri and Tufano 1993, Chevalier and Ellison 1997, Del Guercio and Tkac 2002). Once again for the UK, data deficiencies imply that the relationship between fund performance and fund flow is comparatively unexplored.

There is also evidence that tax considerations have differential effects on inflows and outflows. US funds with relatively large unrealized capital gains (i.e. the ‘tax overhang’) experience lower inflows and outflows (so investors avoid realizing the gains) and inflows are

---

98 The quarterly cross-section regressions begin in 1970Q1 and end in 1999Q3. Control variables include ln(NAV), AGE and fund return volatility.

99 Similar results are reported for flows into the fund family (Gallaher, Kaniel and Starks 2006) and in addition a convex flow relationship also applies to advertising expenditure, but only for the highest performing fund families (see also, Cronqvist 2006). An anonymous referee also points out that funds which have been successful in the past, due to good luck, have an incentive to spend more on advertising to attract “dumb dentists” who wrongly believe that performance persists when in reality it does not.

100 Chevalier and Ellison (1999b) also report that poorly performing funds who change managers, suffer less outflow than those that retain their managers.

101 However, for the UK Fletcher and Forbes (2002) examine whether cash flow is linked to past performance. Ranking funds recursively into quartiles annually on past year excess returns reveals that the highest performing quartile experiences the largest cash inflow during the year. The worst quartile experience the least cash inflow, but do not suffer an absolute cash outflow, which suggests little penalty for their relatively poor performance. This is corroborated by Keswani and Stolin (2008), who have monthly data separately for inflows and outflows and where past performance is measured using 4F-alpha (estimated over the previous 36 months).
more sensitive to after-tax returns than pre-tax returns (Barclay, Pearson and Weisbach 1998, Bergstresser and Poterba 2002). Ivkovic and Weisbenner (2009) using separate figures on inflows and outflows (rather than net flows) find that inflows respond to *relative* fund performance but outflows only respond to “absolute” performance (as well as to high expense ratios) – this they attribute to tax-loss selling. Also, investors are more reluctant to sell funds that have appreciated in value and more likely to sell funds that have fallen in value – so potential tax burdens appear to influence fund flows (see also Ivkovic, Poterba and Weisbenner (2005)\textsuperscript{102}.

The above studies clearly show that inflows into managed funds respond very strongly to good past performance, to high advertising expenditure and low costs in a rational way - and this is also found to be the case for *index* funds (Elton, Gruber and Busse 2004)\textsuperscript{103}. The slope of the flow-performance relationship is also influenced by search and participation costs. The *relative* lack of fund outflow from past poor performers is partly explained by Lynch and Musto’s (2003) “change of strategy” and subsequent higher returns of really poorly performing funds. But the lack of a significant outflow from poorly performing funds as a whole (many who remain subsequent poor performers) is worrying for the equilibrium model of Berk and Green (2004), although recent evidence in Ivkovic and Weisbenner (2009) that outflows are related to *absolute* poor performance mitigates this conclusion.

**FLOWS AND FUTURE RETURNS**

Evidence that investors’ cash chases past winners is clear, but we now examine if this ‘new money’ results in higher future returns. Gruber (1996) finds that the average (net return) alpha on ‘new cash’ is small at 29 bp per annum and the average investor saves 22 bp per annum, by removing their capital from poorly performing funds\textsuperscript{104}. In a broadly similar study but using about 1,800 funds Zheng (1999), over January 1970-December 1993, finds that high cash

\textsuperscript{102} Of course, funds may also experience increased cash inflows in response to a rise in *market* returns (as well as to relative fund returns). However, the relation between *market* returns and fund flows is questioned in Edelen and Warner (2001) who use daily and intra-day data to demonstrate a causal effect from fund flows to market returns.

\textsuperscript{103} They use 52 US (S&P500) *index* funds (Jan 1996-Dec 2001) and find that cash flow responds to past relative performance (measured either as the CAPM alpha or excess fund returns over the S&P500 index). ‘Cash flow’ is actually the *unexpected* cash flow and is the residual from a regression of actual cash flow on fund size. The cross section regression has annual ‘year’ intercept dummies and two control variables: a load dummy and the number of funds in the same family, which both have positive and statistically significant effects on future fund flows. The former they interpret as stronger selling incentives to brokers after receiving the load fee which outweighs the direct higher costs to investors and the latter they interpret as the convenience of easy inter-fund transfers and record keeping for the investor. These results on fund flows are of added importance because the correction for risk of an index fund is relatively uncontroversial. Risk measures such as \((\beta - 1)\) or the tracking error (measured by the CAPM, R-squared of the fund relative to the average R-squared across all funds) are found to be statistically insignificant determinants of fund flows.

\textsuperscript{104} The factor regression has market, size and value indices plus a bond return index. Cash inflows into a fund in quarter \(t\) are multiplied by the risk adjusted return of the fund in (a number of) subsequent periods. Returns are then aggregated over all funds and all time periods.
inflow funds subsequently outperform low inflow funds (using conditional and unconditional $\alpha^{f, net}_{3F}$) but neither alpha is individually statistically significant and therefore the strategy requires shorting low inflow funds, which may not be feasible\textsuperscript{105}.

Gruber (1996) and Zheng (1999) do not use a momentum factor so their documented flow-future performance link may be due to passive short-term momentum effect from existing stocks in the fund (Carhart 1997) or from new purchases of existing momentum stocks (Wermers 2000). Sapp and Tiwari (2004) test this proposition by repeating Zheng’s portfolio sorts on new money flows and find that abnormal performance using $\alpha^{f, net}_{3F}$ does not carry over when using the 4F alpha. This suggests that money is not ‘smart’, because it does not chase funds with active momentum styles but merely chases past (raw) return winners and the latter strategy does not earn positive future abnormal returns based on $\alpha^{f, net}_{4F}$\textsuperscript{106}.

Due to paucity of data, Keswani and Stolin (2008) is the only UK study which links new cash inflows and outflows to future performance (measured by 4F-alpha) over the period 1992-2000, using around 500 funds. With monthly portfolio rebalancing they find that ‘new money’ flows earn a higher abnormal return than ‘old money’ - but in each case the abnormal 4F return is negative.

From the above it can be seen that early studies suggest that money is ‘smart’, in the limited sense that most cash inflows are into past winner funds who subsequently experience higher future returns than past losers – note that this implies a relatively better outcome but does not imply that investors can earn positive abnormal returns. We would rather retain the word ‘smart’ for ex-ante investment strategies that earn positive abnormal returns – investors switching funds to improve their relative position but still earning negative abnormal returns, might be better described as ‘less dumb than average’ strategy. Indeed, more recent studies find that investors’

\textsuperscript{105} For example funds are sorted each quarter into two portfolios based on NCF\textsubscript{i} > 0 and NCF\textsubscript{i} < 0 or alternatively sorts are undertaken using median cash flow as the ‘break point’ – in total, Zheng (1999) uses 6 ‘new money’ sorting rules. Using later data 1980-2003, Frazzini and Lamont (2005) sorting funds on relative flows, find significant positive 3F alphas over 3 and 6 month horizons for a long-short portfolio of high inflow minus low inflow funds (quintiles) and weaker evidence of a negative 3F alpha over a three year horizon.

\textsuperscript{106} They also test whether investors’ net cash flows are determined by an active momentum strategy rather than investors blindly chasing past returns. A Fama-MacBeth cross-section regression for NCF\textsubscript{i}, is repeated quarterly and they find that ($\beta^{NOM}_i$), is not statistically significant - but past returns and previous NCF’s are significant and positive. Control variables used are \{InTNA, TURN, TER, LOAD\}. They also test this proposition by sorting funds each quarter into decile portfolios based on ($\beta^{NOM}_i$), and find that there is little difference in future cash inflows across these deciles in any of the next 4 post-ranking quarters, whereas future cash flows do blindly follow past raw return winner decile funds. An anonymous referee has pointed out that this evidence would imply that money is smart, if momentum is not a risk factor.
cash blindly follows past raw return winner-funds (rather than funds with an active momentum strategy) and such funds do not have positive future abnormal returns after correcting for the momentum effect – so on our definition, money is not ‘smart’.

The above conclusion is reinforced by Cooper, Gulen and Rau (2005) who examine the flow-performance relationship (1994-2001) for 332 funds which changed their names to reflect a current ‘hot style’. Suppose it is the case that when a fund changes its name to a hot style (e.g. from ‘growth’ to ‘value’, or ‘small’ to ‘large’) but does not actually change its style this results (ceteris paribus) in a large cash inflow, yet the subsequent abnormal performance of this fund is poor\textsuperscript{107}. We would not then infer that “money is smart” - this is exactly the conclusion reached by Cooper et al (2005). They find that funds which change their names to ‘hot styles’ are either funds which have been doing badly (based on 3F-alpha), or are established funds or are funds with low advertising and low recent cash inflows\textsuperscript{108}. The subsequent extra cash inflow attributable simply to the (cosmetic) name change is a substantial 25% after one year (in excess of flows to matched funds with no name change). But the subsequent returns and 3F-alpha performance of the cosmetic name change funds is worse than their pre-name change performance and worse than funds with no-name changes. For example, the pre and post name change \( \alpha_{3F}^{\text{net}} \) are minus 0.11%* p.a. and minus 0.23%*** p.a. respectively, while their average raw returns are 1.42% to 0.33% p.a., respectively.

Of course, some fund managers are ‘smart’, since simply by undertaking a cosmetic name change they can attract additional funds from investors (which are enhanced by additional advertising of the name change). But these investors pay around 3.75% transactions cost on average (for loads, expenses and fees), yet they subsequently earn no extra return. At a minimum this implies that more disclosure of fund holdings may be required so that investors can make more informed decisions. But somewhat pessimistically, it may also imply that investors do not use what knowledge is available in a sensible manner\textsuperscript{109}.

\textsuperscript{107} Cosmetic name changes are those which do not result in a change in style as measured by the change in factor loadings on SMB and HML. If the fund’s style factor loading in the 3F-model (measured over the 2 years after the name change) does not exceed that of the quintile BMV and size sorted control portfolio ‘break points’, then the name change is ‘cosmetic’. Flows (3 months after the name change) also respond positively to changes in past advertising (12b-1 fees), performance (e.g. 3F alpha, net returns decile rankings relative to all equity funds) and negatively to changes in expenses and load fees (not significant) and to fund size (lnTNA) – as found in earlier studies. The ‘hot style’ is defined by a \( (0, 1) \) dummy, taking the value 1 when the corresponding style premium (e.g. \( R_{\text{AHML}} \)) is ‘up’ and zero otherwise.

\textsuperscript{108} ‘Name change’ is a \( (0, 1) \) dependent variable in a logit regression on fund characteristics.

\textsuperscript{109} Frazzini and Lamont (2005) argue that non-financial firms are also smart since they issue more stock (via seasoned issues, merger finance) when investor inflows into mutual funds push up the prices of their stock (above fundamental value) – but investors who move into high inflow mutual funds are dumb since they experience subsequent low returns. For example, over a 3 year horizon, top quintile inflows into mutual funds exceed those in the bottom quintile by around 12% of shares outstanding and this results in these firms issuing issuing 3% more shares. Hence,
Index funds are one of the simplest investment products, so Elton, Gruber and Busse (2004) investigate whether investors can use simple rules to move into index funds with relatively high future returns. First they show that past performance (using the differential return over the market or CAPM-alpha) or TERs over either 1 or 3 years, has high predictive power for future one and 3-year performance\textsuperscript{110} – in contrast to studies of statistical prediction for actively managed funds. They then measure the actual performance of index funds over one and three year horizons\textsuperscript{111} and compare this with returns to two types of ‘alternative’ index portfolios. The first are naïve portfolios (i.e. equally or value weighted) and the second are ‘smart’ investor portfolios (i.e. top deciles of index funds based on highest past returns or, highest past CAPM-alphas or lowest past total expense ratios, TERs).

The average return earned from actual net inflows into index funds are generally worse than any of the above alternative portfolios. For example, actual investors in index funds as a whole, do 15 bp per annum worse than if they had purchased the top 10% of funds based on past returns (for 1-year holding periods). True, this is not a particularly large differential to ‘active search’ but the future ‘winner-loser’ differential return is 92 bp per year – which is economically (and statistically) significant.

Elton et al (2004) suggest that the lower returns from actual flows into index funds (compared with the above alternatives) is due to higher marketing costs - as funds actually held have higher loads and 12b-1 fees and also have higher expenses. Of course, higher expenses could lead to more advice from brokers but what is clear is that such compensation seems incompatible with the interests of investors, since it leads to inferior performance compared with simple mechanical rules for investing in index funds. So, financial advisers and brokers benefit from relatively high fees from the funds recommended but there is no extra return for investors relative to the index itself, or to simple strategies based on past performance of index funds. It is also the case that poorly performing index funds also receive substantial cash inflows even though their subsequent performance is relatively poor. Once again we have evidence of inertia or ignorance on the part of many investors who are investing in the simplest mutual fund product, while advisers have an economic incentive to sell inferior products. The counter-argument is that

\begin{align*}
\text{Actual (cash flow weighted) returns are } R^{\text{A}}_{t+j} &= \sum_{i=1}^{N} w_i R_{t+j}^i, \quad \text{where } w_i = \frac{CF_i}{\overline{CF}_i} \text{ and } \overline{CF}_i = \text{mean cash inflow over all } N \text{ index funds. (If cash inflow is negative, the fund is excluded for that month.) Funds are rebalanced monthly.}
\end{align*}

\textsuperscript{110} The R-squared in various alternative regressions of ‘3 year performance’ on ‘past 3 year performance’ are in the range 0.77 to 0.88 and the coefficients are close to 1 for differential returns on past differential returns, alpha on past alpha and are close to -1 for returns or alpha on past TERs. Similar results apply for 1 year horizons.

\textsuperscript{111} Increase in supply of stocks by firms which are experiencing strong investor sentiment (via mutual fund inflows) meets around 25% of the increase in demand.
differential fees provide differential information which is valued by investors – even though they could do better with simple rules (see Hortacsu and Syverson 2004).

### 6.3 MUTUAL FUND MANAGERS

Skill and persistence in performance may reside at the fund manager level rather than at the fund level. If so, investors should ‘chase’ past top performing managers not necessarily top funds. Using a cross-section Fama-MacBeth approach, Chevalier and Ellison (1999b) for a sample of 492 mutual fund managers (1988 – 1994) find that managers with higher undergraduate SAT scores obtain higher risk adjusted returns (after controlling for other fund characteristics). They attribute this outperformance to better natural ability, education and professional networks associated with having attended a higher SAT score undergraduate institution. The impact of SATS scores on (risk adjusted) performance appears to hold only for funds domiciled outside of major US cities, whereas for funds located in major cities performance is influenced by relative manager tenure ("experience") in the city and with the fund (Christoffersen and Sarkissian 2009). However, Baks (2003) who tracks managers as they move between funds, concludes that the fund typically has a greater influence on future performance than the manager.

A recent study by Khorana, Servaes and Wedge (2007) investigates the impact of manager holdings in their own funds on future performance. Because the relevant disclosure rule by the SEC only came into force in 2004, the impact on performance is only assessed for 2005. Other things equal, one would expect managers who invest more of their own capital in their funds would have either greater incentives or superior information about future performance\(^\text{112}\). Using a cross-section regression on 1400 funds they find that the percent ownership by managers in a fund (in 2004) does influence (risk adjusted) performance in 2005 – even after controlling for a wide range of other fund characteristics. However, the authors note some caveats to this statistical relationship. First the effect does not appear to be particularly large since an increase in ownership proportion from the 25th to the 75th percentile fund, increases risk adjusted performance by only 21 bp per annum. Also to make this an exploitable strategy the "high ownership" funds chosen would also have to be large, have low load fees and low turnover.

---

\(^{112}\) The authors note that fund managers with ownership hold on average about $100,000-$150,000 in the fund (0.04%-0.08% of the total assets) with 10% of all managers holding above $160,000-$405,000 (0.09%-0.22% of total assets). It is this percentage holding in the fund that is used as a measure of "manager ownership", whereas one might think that it is the manager's holding in the fund relative to her own (financial) wealth (and perhaps baseline salary) that would more closely reflect the manager's incentives.
– since these characteristics also influence performance\textsuperscript{113}. There is also evidence that better performance influences managerial rewards in the form of future promotions and demotions of managers (Evans 2006, Khorana 1996, Chevalier and Ellison 1999, Ding and Wermers 2005) – so corporate governance does in part support investor objectives.

Do we get similar results on manager skill and performance when using holdings and trade data on stocks held by funds? Ding and Wermers (2004) rank funds (each quarter into deciles) based on either cumulative years of experience of the fund manager or the cumulative value of the manager’s CS index and then track the returns and CS index over the next year. Top managers, based on their past cumulative CS index (but not on cumulative years experience) gives a statistically significant $CS_{t+4Q} = 2.2\%$ p.a. (December 1985-December 1999) – but the latter is statistically insignificant when the data period is extended to December 2002 (Ding and Wermers 2005)\textsuperscript{114}. Although net cash inflows are mainly into past winner managers’ funds (28% p.a. increase) rather than past losers (7.8% inflow), the past winners do not earn a statistically significantly higher net return or conditional or unconditional (Ferson-Schadt 1996) values for $\alpha_{4F}^{net}$.\textsuperscript{115}

Fund managers appear to have skill in picking some types of stock but what is the source of this skill? Cohen, Frazzini and Malloy (2007) suggest that it is the superior information flow in previously established social networks between fund managers and three senior officers of firms (ie. CEO,CFO,Chairman). They classify stocks held by mutual funds into “connected stocks”

\textsuperscript{113} In the cross-section regression, the positive effect of ownership on performance is relative effect – but the authors do report that the average fund with managerial ownership outperforms its benchmark return by 144 bp p.a., while the zero ownership funds only outperform by 29 bp.

\textsuperscript{114} There are problems when forming portfolios, these include the assumption of homogeneity within deciles, the possible impact of other fund characteristics being correlated with the characteristic used to sort, and when using factor models the assumption of constant factor loadings over time. Hence Ding and Wermers (2004, 2005) also undertake a Fama-MacBeth cross section regression (repeated annually, 1986-2000) of $CS_{t+1}$ on fund characteristics such as years of experience and the cumulative past performance of the manager CumCS. These variables, are found to be statistically significant when using “all funds” but with data extended to 2002, Ding and Wermers (2005) results are less robust, with past cumulative performance remaining statistically significant only for growth oriented funds, while years of experience is statistically significant but only for managers of large funds. But an I(0) variable, $CS_{t+1}$ regressed on a potentially I(1) variable CumCS may be problematic.

\textsuperscript{115} Similar results are reported for the performance of pension fund mandates over 1993-2003, with a total of 9,581 decisions over 3,715 plan sponsors examined (Goyal and Wahal 2008). Fund managers that underperform their benchmarks over the last 3 years are fired and new mandates given to past winners. However, in the 3 years before hiring these managers earn excess returns (over the benchmark) of 14.1% (s.e.=2.3) but in the post-hiring 3 years they earn only 1.6% (s.e.=1.6). For managers fired for bad performance (as opposed to say a change of mandate) the pre-firing 3-year excess return is -5.3% (s.e.=1.2%) but in the subsequent 3 years is 6.3% (se=2.9). Consistent with these figures, using a matched sample of round trip firing and hiring decisions, there is a net loss in terms of return to the change in mandate – as well as the transition costs of around 2%. Put another way, plan sponsors cannot successfully market time changes in mandates - and the only benefit appears to be maintaining incentives among incumbent funds, due to the threat of the loss of the mandate.
based on educational background. For example, the strongest measure of “connectedness” is when a fund manager and at least one of the firm’s senior officers attended the same university school, overlapped in years attended and did the same degree. Using (quarterly) recursive rebalancing (1990-2006), the returns on “connected stocks” held earn a statistically significant risk-adjusted return (e.g. CS measure, multifactor alpha) of between 2% and 8.7% p.a. (depending on the precise definition of “connectedness”), whereas the “unconnected stocks” do not earn positive abnormal returns. The positive returns on connected stocks occur around corporate news announcements, so mutual fund managers “front run” announcements of “good news” (but it is also found that they cannot time “bad news” announcements on connected stocks since they do not sell ahead of bad news announcements). These effects appear to be systematic since they are pervasive across different fund styles, across local and distant locations of funds (Coval and Moskowitz 2001) and across Ivy League versus non-Ivy League universities. This study provides a compelling account of how information can be used successfully by “connected” fund managers to earn gross abnormal returns on part of their portfolio but it does not set out to address the overall performance of the fund – which also depends on the performance of the “non-connected” assets and on fund expenses.

Overall, it appears that choosing funds on the basis of the past skill characteristics of their managers has some statistical predictive power for future abnormal returns. However, the link from differential skill to future abnormal CS return on stock holdings is not particularly large, applies mainly to growth funds and does not carry over to higher net return alphas for all funds. Thus any skill that resides at fund manager level seems to accrue to those running the fund and not to investors.

So, “Is Money Smart?” Our overall conclusion from the above studies must be that most (but not all) money is pretty dumb. Investors blindly chase past winners (both active funds and index funds), chase funds with cosmetic name changes, respond strongly to fund advertising and they do not chase funds with a high loading on the momentum factor. Past winner funds do earn positive risk adjusted gross but not net returns, in subsequent periods – fund managers therefore expropriate the returns to their skills, as in the equilibrium model of Berk and Green (2004). However, what is also clear is that at the other end of the performance distribution, the absence of large cash outflows from poorly performing funds probably inhibits the competitive process, since most of these funds continue to earn persistently poor abnormal returns (Kosowski et al 2006, Barras et al 2009). So, some money may be smart but it is at the lower end of the performance distribution that money is really dumb – in contrast to the predictions in the Berk and Green (2004) model.
7. CONCLUSIONS

In this survey we noted the wide range of innovative studies used in measuring the performance of mutual funds and these methodologies have also been applied to pension funds and more recently to hedge funds and even to venture capital funds (Cochrane 2005). Work on mutual funds has revealed a great deal about their performance and the reasons behind this performance. What does this literature reveal about investment decisions?

First consider the simplest investment product, index funds. The statistically significant ex-ante differential return between past decile winner and loser index funds is 92 bp per year and there is a substantial 2% p.a. differential between the best and worse performing US index funds. Investors therefore appear to ignore simple ex-ante strategies for increasing performance and clearly with expense ratios as high as 1.35%, many investors do not avoid high cost index funds.

Turning now to managed funds, the average (US or UK) fund underperforms its benchmarks by around 1% p.a. (Kosowski et al 2006, Barras et al 2009, Cuthbertson et al 2008) but the average US fund does earn a positive risk adjusted return in recession periods (Kosowski 2006), which is consistent with an intertemporal hedging demand\textsuperscript{116}. The large dispersion in the cross-section of fund abnormal returns in both the US and UK suggests it is worth examining funds in the tails of the performance distribution and assessing the role of luck versus skill. In terms of ex-post performance recent US and UK studies find that at most around 5% of funds in the extreme right tail have ‘truly’ positive net return alphas and around 15-20% of funds spread throughout the left tail have ‘truly’ poor performance (Kosowski et al 2006, Barras et al 2009, Cuthbertson et al 2008)\textsuperscript{117}. For example, the net return alpha for the US fund (1975-2002) at the 95\textsuperscript{th} percentile is 4.8% p.a. and in the left tail at the 5\textsuperscript{th} percentile is minus 6% p.a. both of which are statistically significant (Kosowski et al 2006, Table II, panel A) - but the former is disputed by the Fama-French bootstrap. US data reveal that the top performers are in growth and aggressive growth styles, while the top ‘growth and income’ and ‘balanced or income’ funds do not beat their 4-factor benchmarks. In contrast, in the UK, skilled funds tend to be in the income style rather than in growth or small cap funds.

\textsuperscript{116} Post 2000, some US mutual fund companies have introduced funds that emulate hedge funds and Agarwal, Boyson and Naik (2008) find that on average these ‘hedged mutual funds’ outperform traditional mutual funds by at least 3% p.a. (after controlling for expenses, past performance, risk and fund characteristics).

\textsuperscript{117} For the US there is evidence of more winners in the 1975-1989 period relative to the 1990-2002 period but widespread poor performance is prevalent in both periods – Kosowski et al (2006).
There are substantial methodological difficulties in separating security selection from “market timing” but apart from strategies based on stale prices, there is little evidence that funds successfully time the market and provide positive abnormal returns to investors – the caveat here is that we have little or no evidence on intraday timing.

What causes the differential performance across funds? US studies show that the main influences on the cross-section of managed fund’s abnormal net return alphas are the strong negative impact of fees (e.g. 12b-1 fees, TER’s, loads) and to a lesser extent, high turnover.

What about ex-ante rules for picking winners? Statistical measures of persistence such as contingency tables and cross-section regressions of current on past abnormal returns indicate relatively weak short-term persistence among the past winners (which is strongest over horizons of one year or less) and longer horizon persistence (up to 3-5 years) for past losers – for both US and UK studies. However, what is important to investors is the economic significance of any persistence. Using the recursive portfolio approach, US studies show statistically significant persistence in risk adjusted (CS) gross returns for the top decile of ‘all funds’ of around 1% p.a. over a 6 month horizon and persistence up to 4-years ahead of around 2% p.a. for growth funds (using data 1975-93, Wermers 2003b). Short-term persistence among past winners is weaker when using abnormal net returns to the fund as a whole. For example, US evidence indicates some persistence over one year for the ‘winner’ decile ($\alpha_{4F}^{f, net} = 1\%$ p.a., $p =0.05$)\textsuperscript{118} but what is absolutely clear is that it is worth avoiding funds in the bottom 4 ‘loser’ deciles since these have persistent negative net return alphas (e.g. for the bottom decile $\alpha_{4F}^{f, net} = -3.5\%$ p.a., $p <0.01$ - Kosowski et al 2006, Fama-French 2009).

Studies using manager performance (rather than fund performance) find a statistical relationship between various measures of management skill and future performance. But the effect is not very large and is unlikely to yield sorting rules which give substantive abnormal net returns - the same goes for ranking by variables such as turnover and industrial concentration. While ‘filter rules’ and the use of extraneous fund information may help in picking winners the evidence here is relatively new and involves relatively sophisticated ranking procedures (Cohen, Coval and Pastor 2005), so current evidence suggests that only sophisticated investors should pursue an active investment strategy of trying to pick winners - and then with much caution.

\textsuperscript{118} Bollen and Busse (2004) find a similar figure with quarterly rebalancing, over 1984-94.
It may also be worth noting at this point that broadly speaking, US bond fund performance is similar to that found for domestic equity funds. Early studies using around 300 bond funds find little evidence of persistence and expenses are the key determinant of ex-post relative performance (Blake, Elton and Gruber 1993, Elton, Gruber and Blake 1995). However more recently, Huij and Derwall (2008) using over 3,500 US bond funds (1993-2003) find statistically significant negative persistence in forward looking alphas (of between -1% and -3% p.a.) but no positive persistence, when rebalancing monthly. But when the top decile portfolio weights are based on ‘modern portfolio theory’,119 they find forward looking alphas of 1.27%, 1.13% and 0.53% p.a. for monthly, quarterly and annual rebalancing, respectively and these figures are around 0.5% higher for no-load bond funds. So there is some evidence that (ignoring loads and investors’ search costs), top decile-alpha bond funds exhibit short-run persistence and stronger evidence that losers persist.

Because of space constraints and (often) data deficiencies, our survey does not cover the performance of mutual funds in Europe (ex-UK), Australia and the Far East. Our broad conclusions for the US and UK of few winner funds with little persistence in performance but substantially more loser funds which persist, appears to apply to most funds and fund styles in other European countries (Otten and Bams 2002) and in Australia (Vos, Brown, and Christie 1995, Hallahan, 1999, Hallahan and Faff 1999, 2001, Gallagher 2003, Holmes and Faff 2004, Bilson, Frino, and Heaney 2005). For Japanese mutual funds a key issue in assessing performance is the dilution effect arising from investors’ tax position when holding the fund - this accounts for some but not all of the observed strong negative abnormal performance (Cai, Chan and Yamada 1997, Brown, Goetzmann, Hiraki, Otsuki and Shiraishi 2001, Goetzmann et al 2003).

Performance persistence results for US funds ignore load fees when rebalancing funds and any advisory fees, search costs and investor’s taxes120 - so is chasing a net return alpha of around 1% p.a. (p=0.05) for the top decile winner portfolio worth it when faced with stochastic rebalancing costs (Kosowski et al 2006)? Even ignoring estimation and model error, many practical difficulties need to be addressed. Most investors would have to mimic the performance of the winner decile (of around 180 funds at the end of 2002) with a smaller number of funds - and when rebalancing (at least annually) would need to avoid any funds with load/advisory fees, and finally, consider any tax implications of fund purchases/sales. (The average load fee is around 3.6% and applies to about 56% of funds – Mahoney 2004).

119 Portfolio weights are based on the appraisal ratio and alpha must be positive for the fund to be included in the portfolio.
Picking winners from equity mutual funds appears to be a difficult task. A key area for further research is assessing the performance of funds in the tails of performance distribution. Recent statistical advances such as separating skill from luck for individual funds when funds are ranked by performance (i.e. the theory of order statistics and appropriate bootstrapping techniques) and the possibility of detecting “false discoveries” in a set of ‘significant’ funds are interesting new developments in this area. Bootstrapping techniques can be used to mitigate problems of non-normality and cross sectional correlations among fund returns (possibly due to herding). The increased use of Bayesian approaches and sophisticated “filter rules” when trying to “pick winners” is also a promising area of research. Concentrating more on individual fund performance or the performance of small ‘fund-of-fund’ portfolios is important, since it brings academic work closer to the real world choices of investors and industry practitioners. We also note something of a lacuna over the question of the impact of stochastic switching costs when “chasing winners” - we have to know the number and dollar value of specific funds purchased/sold at each rebalancing date – this is a key area for future research if the “smart money” debate is to be resolved.

Benchmark error remains a problem in estimating performance using factor models, yet performance measures avoiding benchmarks are also problematic. Problems of data snooping and false discoveries also arise and recent papers have begun to explore these interrelated issues (Sullivan et al 1999, 2001, Barras et al 2009, Cuthbertson et al 2010). Gaming of standard performance measures is likely to become of increasing importance as mutual fund managers adopt some hedge fund strategies and further research into manipulation proof performance measures is required (Goetzmann et al 2007).

Is money smart? It turns out that for investment in US equity mutual funds, a substantial amount of cash flows into past winner funds and the winner decile portfolio earns positive gross abnormal returns - yet this only translates into future abnormal net returns (before switching costs) of around 1% p.a. at best. Thus the Berk and Green (2004) equilibrium model is broadly correct for winners – fund managers expropriate any rents from their skill and the return to smart investors after all transactions costs is probably rather small. However, the strong

---

120 Dickson and Shoven (1993) show that capital gains realizations can have a major impact on US investors after-tax returns.
121 When holdings data is available Lo (2007), building on the approach in Grinblatt and Titman (1993), suggests a simple measure of active management that does not use a benchmark. Since $E_{p} = \sum_{n=1}^{N} E[w_{p}r_{p}] = \sum_{n=1}^{N} \text{cov}(w_{p}, r_{p}) + \sum_{n=1}^{N} E[w_{p}]E[r_{p}] = \delta_{p} + v_{p}$, he suggests using $\theta = \delta_{p} / (\delta_{p} + v_{p})$ as a measure of active management. But one limitation noted is that the sampling interval for asset holdings must be at least as fine as the decision interval of the actual investment process, otherwise the measure is severely biased.
persistence in poorly performing funds together with low positive inflows (rather than large outflows) suggest that the Berk and Green model does not apply at the negative end of the performance distribution. Although very poor funds may change their strategy and improve their performance in the future (Lynch and Musto 2003), the prevalence of a large number of persistent poor performers indicates that a lot of money is dumb - and any move towards a competitive equilibrium among loser funds, appears to be relatively slow.

Why any mutual fund, particularly a long-lived fund, which truly underperforms would be permitted to survive in a competitive market is puzzling. In part it may be that for some investors, interpreting performance measurement statistics is a difficult task and for precision requires a long fund life-span. Various rational explanations for the continued existence of poorly performing funds include investors being 'locked in' (e.g. pension plans) or having accrued capital gains (Gruber 1996). Other reasons include biases in investor information sets, the influence of advertisements, blindly following brokers recommendations as well as inertia, ignorance, actual or psychological costs (e.g. disappointment aversion, disposition effect) – in short, an element of irrationality, if your baseline model is one of informed decisions in relatively frictionless and low information cost markets (see for example, Elton, Gruber and Busse 2004, Sapp and Tiwari 2004, Cooper, Gulen and Rau 2005).

The EMH requires marginal investors to remove pricing anomalies. In view of the relatively small number of funds with successful ex-post positive abnormal performance together with the large number of funds which have a zero abnormal performance, our analysis implies that the market for managed funds is reasonably efficient. However, a few funds do earn positive abnormal returns and the literature suggests this may be due to lower expenses and fees and (more contentiously) skill in exploiting anomalies. But recent evidence suggests that even this limited “success” has been eroded over the last 15 years (Wermers 1999, Kosowski et al 2007). At the other end of the performance distribution there are a substantial number of funds that are consistent poor performers, yet still stay in business. Given the difficulties in assessing performance this may not be entirely due to irrational investor behaviour - but neither can the latter be ruled out. The above conclusions are reinforced by the difficulties in finding winner funds that persist, despite the large number of “sorting rules” investigated and such strategies generally earn relatively small abnormal returns. On the other hand the existence of persistent losers is well established and appears to violate market efficiency.

Instances of ‘non-rational’ behavior for other financial decisions is widely documented in the behavioural finance literature - an excellent recent survey is Barberis and Thaler (2003). Note that poor ex-post performance for even the worst ranked individual hedge funds is not statistically significant and is therefore due to ‘bad luck’ – nevertheless, investors should switch into hedge funds with statistically significant positive alphas (Kosowski et al 2006).
On a public policy level, the academic literature cited above suggests there is certainly scope for more transparency about individual fund costs, trades and performance measures and also a strong argument for a more impartial educative role directed towards retail fund investors. Overall, academic work demonstrates that there are relatively few mutual funds which have genuinely positive alphas and picking ex-ante winners is very difficult when one considers potential data snooping bias, model/estimation error and possible transactions costs of rebalancing (i.e. load, advisory fees and information costs). In contrast, persistence among past loser funds is well established. The evidence on fund performance suggests that any ‘national’ savings schemes (e.g. 401K schemes in the US and Turner’s (2006) ‘BritSaver’ scheme) should warn against trying to ‘pick winners’ and seek to provide impartial, independent information on fund performance. Sensible advice for most investors would be to hold low cost index funds and avoid holding past ‘active’ loser funds.

123 For example, see Mahoney (2004) on market timing scandals in the US and the literature on choices in 401K pension schemes (e.g. Benatzi and Thaler 2001, Huberman and Jiang 2006). Sandler (2002) and Turner (2004, 2005) analyze private long-term savings in the UK (including the design of compulsory versus voluntary schemes) and OECD (2005) provides a survey of nascent programmes in developing financial education in member states. Wermers (2001) analyses the issues associated with more frequent disclosure of asset holdings.

124 Given the lack of consensus on the provision of long-term savings in the UK, it has been suggested in evidence to the Turner Commission (Cuthbertson et al 2005) and in response to arguments about regulatory capture and bankers’ bonuses in the wake of the financial crisis (Cuthbertson et al 2009), that one should set up an independent financial body, the National Institute for Financial Excellence (NIFE) along the lines of the Bank of England’s Monetary Policy Committee and the UK’s National Institute of Clinical Excellence (NICE). NIFE would act as an independent organization focused on impartial long term advice and education concerning generic principles and ‘good practice’ in assessing the performance of long term savings vehicles including mutual, pension and alternative investments.
References


Baks, Klaas P., A. Metrick, and J. Wachter, 2005, Should Investors Avoid All Actively


Dickson, Joel M., John B. Shoven and Clemens Sialm, 2000, Tax Externalities of Equity Mutual Funds, SSRN working paper.


Evans, Richard, 2207 The Incubation Bias, University of Virginia, Working Paper, SSRN.


Fletcher, Jonathan, 1997, An Examination of UK Unit Trust Performance Within the Arbitrage Pricing Framework, *Review of Quantitative Finance and...*

Frazzini, Andrea and Owen A. Lamont, 2005, Dumb Money: Mutual Fund Flows and the Cross-Section of Stock Returns, NBER-WP, 11526.


Frazzini, Andrea and Owen A. Lamont, 2005, Dumb Money: Mutual Fund Flows and the Cross-Section of Stock Returns, NBER-WP, 11526.


Houge, Todd and Jay Wellman, 2006 The Use and Abuse of Mutual Fund Expenses, University of Iowa, Working Paper, January.


Ivkovic, Zoran and Scott Weisbenner, 2009, Individual Investor Mutual Fund Flows,


Morey, Matthew, R., 2003, Should You Carry the Load? A Comprehensive Analysis


Teo, M., and Sung-Jun Woo, 2001, Persistence in Style-Adjusted Mutual Fund Returns, Manuscript, Harvard University. Available at SSRN.


Table 1: Fund Performance and Persistence

<table>
<thead>
<tr>
<th>Study</th>
<th>Period</th>
<th>Funds $^{1}$</th>
<th>Methodology</th>
<th>Empirical Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ippolito (1989)</td>
<td>1965–1984 yearly</td>
<td>143 funds</td>
<td>Single factor model</td>
<td>Abnormal returns sufficient to cover expenses</td>
</tr>
<tr>
<td>Study</td>
<td>Period</td>
<td>Funds</td>
<td>Data Source</td>
<td>Findings</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------------</td>
<td>-------</td>
<td>-------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Period</td>
<td>Sample Size</td>
<td>Methodology</td>
<td>Findings</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------</td>
<td>-------------</td>
<td>------------------------------------------------------------------------------</td>
<td>----------</td>
</tr>
</tbody>
</table>
### Evidence of Negative Performance

Fama and French (2009) examined monthly performance from 1984-2006 with 660-3156 equity funds. The cross-section bootstrap methodology identified no funds with positive abnormal returns, with many exhibiting negative returns.

#### Panel B: UK Studies (Selective)

<table>
<thead>
<tr>
<th>Study</th>
<th>Period</th>
<th>Funds Description</th>
<th>Methodology Description</th>
<th>Empirical Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keswani and Stolin (2008)</td>
<td>1992-2000 monthly</td>
<td>500 equity funds</td>
<td>Sort winners and losers on fund flows</td>
<td>New money’ flows earn a higher abnormal return than 'old money' - but in each case the abnormal 4F return is negative.</td>
</tr>
</tbody>
</table>

**Notes:**

i). The number of funds used varies within each study, therefore figures quoted are illustrative.
Figure 1: Power Function for Alpha

Power: alpha = 1.8% p.a. and alpha = 3.6% p.a. (---)