False Discovery Rates:
Separating Skill from Luck in UK Unit Trust Performance

Tom Platonoff

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science by advanced study in Economics and Finance in the Faculty of Economics, Finance and Management.

University of Bristol, School of Economics, Finance, and Management

7th of September 2015

Student Number: 1113920

Word Count: 16,481

Abstract

I estimate the proportions of skilled and unskilled UK fixed income unit trusts controlling false discovery rates (FDR) in a multiple hypothesis testing framework. FDR controls for luck in fund performance, and I find that, gross-of-fees, 16.5% of funds are skilled and able to outperform their benchmark, 2.8% of funds are unskilled. I do not find significant evidence of abnormal performance in the remaining 80.7% of funds. I estimate the FDR at 13.9% for outperforming funds, and 48.7% for underperforming funds. This indicates that most funds beating their benchmark are skilled, whereas around half of underperforming funds are simply unlucky.

AUTHOR’S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University’s Regulations and Code of Practice for Taught Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, this work is my own work. Work done in collaboration with, or with the assistance of others, is indicated as such. I have identified all material in this dissertation which is not my own work through appropriate referencing and acknowledgement. Where I have quoted or otherwise incorporated material which is the work of others, I have included the source in the references. Any views expressed in the dissertation, other than referenced material, are those of the author.

SIGNED: DATE:

Table of Contents

List of Tables

List of Figures

1. Introduction

Unit Trusts (UTs) are a major institution in the UK financial industry, managing over £770bn in assets in December 2013. UTs are one of three sets of financial institutions in the UK (investment trusts and open-ended investment companies (OEICs) are the other two). UTs allow participants to buy and sell units/shares of a diversified portfolio of marketable securities, with no major restrictions on withdrawing capital invested, unlike, for example, pension funds, which are similar in most characteristics other than their illiquid nature.

Approximately 80% of UK funds are actively managed whereby managers are entrusted to pick the portfolio of securities. Evaluating the performance of actively managed funds is important to investors, who pay for the stock picking services of fund managers. The existence of abnormal performance is also important from a theoretical perspective testing the validity of the Efficient Market Hypothesis. A substantial body of empirical work exists assessing the performance of managed funds, although heavily weighted towards US studies, and relatively little towards the UK market.

Early works such as Jensen (1968) found that US fund managers demonstrate little or no stock picking abilities to deliver returns greater than a passive market index portfolio, especially net of management fees. More recent studies have found a wide range of results with more sophisticated analysis: Carpenter and Lynch (1999) claim that:

‘in general, empirical studies find that mutual fund performance is persistent’ ;

Whereas Bessler, Blake, Luckoff and Tonks (2010) inform the reader that:

‘it is widely recognised that equity mutual fund performance does not persist in the long term’.

Cross-sectional mean returns cast an unfavourable (or at least an uncertain) light on the value of active management, however the cross-sectional standard deviation of abnormal returns is high. This could indicate volatile returns around a zero mean, or that certain funds are consistently outperforming or underperforming the market benchmark. Individual investment decisions are revealing, and investors appear confident that they can identify outperforming actively managed funds.

In a fund universe where there exist unobservable proportions of skilled (positive abnormal performance), unskilled (negative abnormal performance), and funds that match the market, a cross-sectional average of returns will not tell us whether or not skilled or unskilled funds exist. It will just tell us what fund returns are on average. For this reason we require multiple hypothesis testing procedures. These are procedures which test each fund individually against a null hypothesis of no abnormal performance. However, multiple hypothesis testing procedures generate a certain proportion of type one errors, where the null hypothesis is incorrectly rejected, and we could say that the fund manager was simply lucky (or unlucky) in the testing period. These are false discoveries. I investigate methods to control the type one error rate..

I examine the cross-sectional distribution of abnormal returns, and using a method known as the false discovery rate (FDR), separate the truly skilled (unskilled) funds from those which have produced positive (negative) returns through sheer good (bad) luck.

I find significant evidence for the existence of non-zero minorities of outperforming and underperforming UK unit trusts between 1980 and 2014. Gross of fees, I find that 80.7% of funds do not exhibit significant abnormal returns, 16.5% are skilled, 2.8% are truly unskilled. Funds outperform the market by a cross-sectional average of approximately 0.2% per month, depending on the performance model used. As in other literature, there is substantial variation in performance.

I find that 13.9% of funds identified as outperformers by standard testing procedures are lucky rather than skilful, and that 48.7% of the small number of underperformers are unlucky. Contrary to net-of-fee literature, these findings show that before fees, very few fund managers underperform benchmarks due to a lack of skill, and a practically and statistically significant number outperform the market with skill.

1.1 Overview of procedure

To estimate the proportion of outperforming and underperforming funds, we must first estimate benchmark returns, then estimate the proportion of funds outperforming and underperforming the benchmark. Though simple in concept, this procedure contains numerous empirical difficulties, I attempt to address as many as possible within the constraints of this project. The main issue addressed in this paper is that of controlling the type one error rate.

To estimate benchmark returns, I use linear models of the form:

$$y_{it} = \ \alpha_i + x_{it}'\beta_i + \ \varepsilon_{it}$$

$$individuals\ i = 1, \dots, m \quad ; \quad time\ periods = 1, \dots, T$$

(1)

y_{it} is a vector of observed returns. α_i is a vector of fund-specific intercept terms. x_{it} is a vector of covariates which determine expected returns. β is a vector of fund specific coefficients. ε_{it} is vector of idiosyncratic errors, discrepancies between the observed y_{it} outcomes, and the returns explained by $\alpha_{it} + x_{it}'\beta$. This general model permits the intercepts and slope coefficients to vary over individual, but with a common vector of explanatory market variables. I separate the intercept from the vector of covariates for clarity.

The α_i intercept term is the parameter of interest and treated as a proxy for abnormal performance against a risk adjusted benchmark. I run the following hypothesis tests on all of the intercept terms:

$$H_{0,1}: \alpha_1 = 0, \qquad H_{A,1}: \alpha_1 \neq 0,$$

$$\dots.$$

$$H_{0,m}: \alpha_m = 0, \qquad H_{A,m}: \alpha_m \neq 0$$

(2)

Funds for which we reject the null hypothesis are identified as exhibiting abnormal performance.

Methodological difficulties in this procedure can be broken down into the following categories:

i) Controlling the type one error rate. Type one errors are the main issue addressed in this paper and I discuss luck in fund performance and methods to control type one errors in multiple hypothesis testing in section 2

ii) Selecting the covariates x_{it} which estimate the parameter of interest, α_i most accurately, under the most realistic assumptions, and whilst ensuring desirable properties of the coefficients β_i, error terms ε_{it}, and standard errors σ_{it}. I examine asset pricing models in section 3

iii) Collecting reliable data of the observed variables y_{it} and x_{it}, I discuss data selection in section 4

1.2 Organisation of the paper

The remainder of this paper is organised as follows: Section 2 discusses the methodology of the FDR and other methods for controlling for luck in a multiple testing framework. Section 3 discusses fund performance measurement. Section 4 describes the dataset and empirical difficulties. Section 5 performs empirical analysis on UK unit trust performance. Section 6 compares results in relevant literature. Section 7 discusses the practical and theoretical implications and explanations of results. Section 8 concludes.

2. Luck in Fund Performance

False discoveries, or type one errors, present a major problem in the identification of truly skilled funds. In this paper I use the BSW FDR method to estimate the proportions of truly skilled and unskilled funds in a population. This method is powerful, adaptive and scalable, asymptotically consistent and unbiased under weakly dependent test statistics and less conservative as the traditional Bonferroni family wise error rate procedure. The main disadvantages are its finite-sample assumptions of normally distributed and independent test statistics. These problems can be rectified with a bootstrap technique as in Kosowki, Timmermann, White, and Wermers (2006), but such methodology is beyond the constraints of this paper.

In section 2.1 I discuss the impact of luck in fund performance. In 2.2 I introduce false discoveries and the difficulties they present in fund performance analysis. In 2.3 I discuss the Bonferroni and Benjamini-Hochberg methods to control false discoveries. In 2.4 I discuss the BSW fdr control procedure. In 2.5 I discuss methodological difficulties and assumptions involved in the BSW method.

2.1 The impact of luck in fund performance studies

An abnormal return from trading stocks in a perfectly efficient market is analogous to a coin toss game of pure chance. 1 in 32 participants will throw five consecutive heads, and with an efficient market this is analogous to five consecutive years of beating (or losing to) the market. Further still, 1 in 1024 will throw ten heads in a row. Fund flows are strongly related to short term performance (Berk and Green (2004)) and 5 years of outperforming the market is certainly enough evidence for most investors to consider a manager talented. Short term outperformance causes the investment inflow to funds to increase substantially, and for managerial wage and client fee bargaining power to increase.

One can extend the coin toss analogy to a simultaneous move game with a mix of skilled and unskilled players. The unskilled players hope to match the moves of the skilled players, but can only play a random draw from buy/sell. A lucky manager will play exactly the same moves as a skilled manager, and with thousands of unit trusts currently operating, one can safely assume that there exists a non-zero minority who, even in a perfectly efficient market, through luck rather than skill will produce positive abnormal returns for several years in a row. In a world where skilled and lucky managers exhibit the same behaviour, one can clearly see the difficulties in identifying skill, if it exists at all.

2.2 Hypothesis testing

Applying luck in fund performance to a simple one-sided formal hypothesis test setting, we can propose a null hypothesis of zero abnormal performance, against an alternative hypothesis of abnormal performance. We measure abnormal performance as the fund's intercept, alpha term, against its benchmark returns (justified by its risk exposure). This testing procedure is performed under the assumption that there are two types of funds, with separate normal distributions: one centred around a mean of zero alpha; the other with mean positive alpha of x (BSW settle on +3.8% mean for the skilled funds).

True alphas are unobservable, so for any individual fund, we reject or fail to reject a null of zero alpha, or 'no abnormal performance' based upon the observed test statistic and a chosen significance level, and so the test statistic becomes our performance measure. In most literature, performance of a large number of funds are measured simultaneously in this way, with multiple hypothesis testing

In this one-sided test setting, represented by figure. 1, if a fund's observed abnormal returns are –y we fail to reject the null at any reasonable significance level, and can be confident that this fund certainly belongs in the null group of no abnormal performance. Similarly, if a fund's returns are y, then we can reject the null hypothesis and accept that this fund belongs to the group of abnormal performers.

2.2.1 The multiple testing problem: false discoveries

Most recent fund performance studies use multiple hypothesis testing. This is important in the context of a large cross section of funds but will generate a certain proportion of type one errors. In figure. 1, the shaded grey area represents the region in which observed returns could be from a fund from either distribution, causing the danger of rejecting a true null or failing to reject a false null (type one and two errors, respectively). For example, at significance level γ, testing returns from a very lucky but unskilled fund that has returns of x, we reject the null and incorrectly identify the fund as skilled.

Using a threshold of γ, the shaded red region represents the type one error rate. This illustration is used for simplicity, but as in BSW this can be extended to a two sided test where funds can exhibit positive or negative abnormal performance in addition to the null.

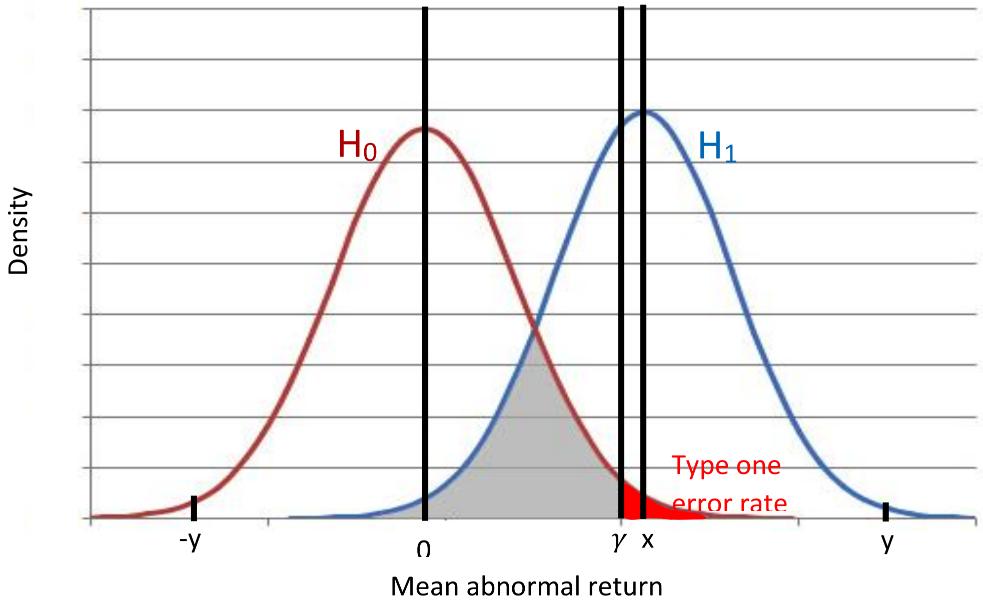


Figure 1. Distribution of returns from funds satisfying a null and alternative hypothesis

	Reject null	Do not reject null	Total
True null	V	U	m_0
False null	S	T	$m - m_0$
Total	R	m-R	m

Table 1. Possible outcomes from m hypothesis tests

Table 1 sets out the problem multiple hypothesis testing faces more clearly. m is the total number of hypotheses to simultaneously test. U, V, S and T are unobservable random variables. R is the total number of hypotheses rejected, an observable random variable which increases in γ, the selected significance level. The significance level is equivalent to the probability of V, a type one error or false discovery.

The proportion of falsely rejected null hypotheses is given by the unobserved random variable Q:

$$\mathrm{Q} = \frac{\text{True nulls rejected}}{\text{Total nulls rejected (both false and true)}} = \frac{\mathrm{V}}{\mathrm{V}+\mathrm{S}} = \frac{\mathrm{V}}{\mathrm{R}} \tag{3}$$

The FDR is given by $\mathbb{E}(Q)$, the expected value of Q, taken by plugging expected values into (3). If we set R to zero, hence rejecting all null hypotheses, Q is by definition also zero. But when we use a non-zero significance level which allows for nulls to be rejected, we also create the opportunity for V to occur with positive probability ie. for false discoveries to occur.

In a one-sided sample of 100 funds, with a test size γ = 0.05, we expect five true null hypotheses to be rejected. In other words, for 5 funds to be falsely identified as exhibiting abnormal performance.

2.3 Type one error rate control procedures

Benjamini and Hochberg (1995) first applied FDR methodology to fund performance. Storey (2002) adapted the FDR to improve its power especially with a large number of tests. In this paper we will use an FDR model, further extended for two-sided testing, as in BSW as the basis of analysis, trading some stringency for vastly improved test power.

2.3.1 BSW false discovery rate procedure

I use the same method and notation as BSW, which extends Storey's (2002) FDR by adapting it to a two sided setting to distinguish between good and bad luck. We use a significance level γ, and a population with proportions: π_0 zero alpha funds; π^+ skilled funds; and ; π^- unskilled funds. Such that $\pi_0 + \pi^+ + \pi^- = 1$. we can estimate the proportion of lucky ($\hat{F}_\gamma^+$) and unlucky ($\hat{F}_\gamma^-$) funds as:

$$\hat{F}_\gamma^+ = \hat{F}_\gamma^- = \frac{\pi_0 . \gamma}{2} \tag{4}$$

For example, in a population with 80% zero alpha funds, at 5% significance, we expect 2% of funds to be lucky and be identified as skilled with a positive alpha. Similarly, we can estimate T_γ^+, the proportion of truly skilled funds:

$$\hat{T}_\gamma^+ = S_\gamma^+ - \hat{F}_\gamma^+ \tag{5}$$

Where S_γ^+ is the total number of observed positively significant funds, the sum of skilled and lucky funds. The converse holds for $\hat{T}_\gamma^-$, the estimate of truly unskilled funds.

2.3.2 Family Wise Error Rates and the Bonferroni procedure

The FWER is the probability of one or more false discoveries occurring, and to control the FWER is to set a more stringent level of γ to reduce this probability. The classical approach to controlling the FWER is a simple linear transformation based upon m in the form of the Bonferroni approach, which controls the FWER in the following way:

Transform the selected significance level by dividing γ by the number of tests performed

$$\gamma_{Bonferroni} = \frac{\gamma}{m} \quad ; \quad \text{Reject } H_{0,i} \text{ if } p_i \leq \gamma_{Bonferroni}$$

(6)

$$\gamma_{Bonferroni} < \gamma \quad \text{where:} \quad m > 1$$

(7)

Clearly, for any multiple hypothesis test, m is greater than one. Therefore, the Bonferroni procedure reduces the size of the test, reducing the significance threshold for p values and we must observe even stronger evidence against the null to reject it. Whilst the guarantee of FWER control is appealing, the conservative new thresholds can result in low power. This increases the probability of type two errors, failing to reject a false null, especially in large samples.

2.3.3 Benjamini-Hochberg false discovery rate Procedure

The Benjamini Hochberg procedure is performed as follows: first, compute the individual p values of all m tests in sample; next order them in ascendance, and index as p(i), such that smallest p, the strongest evidence against the null, is indexed p(1) the second smallest p(2),… ,p(m); then, select a significance level eg. 5% or 0.05. Now compute an individual γ for each hypothesis test:

$$\gamma_{FDR,i} = \frac{\gamma . \mathrm{i}}{m}$$

(8)

For example, for the fourth smallest p value (i = 4)of a sample with population 100 (m = 100), and a significance level of 5% ($\gamma = 0.05$) , the transformed $\gamma_{FDR,4}$ = $\gamma_{FDR,4} = \frac{0.05(4)}{100} = 0.002$

For any given γ, find the largest *k* such that $P_{\mathrm{k}} \leq \gamma_{FDR,k}$

Then reject all $H_{0,i}$ from the range $i = 1, \dots, k$

This ensures that inequality (9) is satisfied

$$\mathbb{E}(Q) \leq \gamma$$

(9)

2.4 Properties and advantages of the BSW FDR

FDR procedures perform a similar function to the simpler Bonferroni, and the procedures are equivalent when all m nulls are true. However, FDR procedures are designed to control $\mathbb{E}(Q)$, the expected proportion of type one errors as in (1), whereas FWER controls the probability of any type one errors occurring at all. In many cases, the FWER is much too strict, especially when the number of tests is large, as in a mutual fund study. The FDR is a more liberal yet powerful approach.

In this sub-section I discuss the FDR's advantages in power, scalability and adaptivity, and the FDR's Bayesian Interpretation and its implications

2.4.1 Power

Sequential p-value methods, such as Bonferroni and Benjamini-Hochberg, fix the error rate and estimate the corresponding rejection region. The Storey (2002) method 'fixes' the rejection region and estimates the corresponding error rate. As shown in Storey (2002), this approach can yield an increase in power over 8 times that of the Benjamini-Hochberg procedure. The BSW method simply adapts the one sided Storey (2002) method into a two sided test and retains such properties.

If m = $m_{0,}$ then FDR and FWER are equivalent, in that FDR procedures also control the FWER. But when some false nulls exist, $m_0 < m$, the FDR is smaller and can be more powerful than FWER procedures at a given level.

The FDR is uniformly more powerful than the FWER (although the magnitude is unclear), and less conservative, allowing a higher absolute number of type one errors. In a simulation study by Benjamini and Hochberg the power of testing procedure decreases in m. However, the rate of decline of FWER procedures is higher. So the power advantages of FDR over the FWER increase in: *m*, the number of tests; and $m\text{-}m_0$ the number of non-nulls. So the loss of power as m increases is relatively small for FDR and this gives it an advantage in large sample studies.

2.4.2 Scalability and adaptivity

The FDR criterion is distinct to the FWER criterion as it controls the proportion of type one errors rather than the absolute number. In this sense, it is far more scalable when we consider the trade-off between type one and type two errors, as it controls the proportion of type one errors to a chosen level, adaptive to varying sample size.

Take the following example: In a sample with only 2 discoveries we may wish to control the absolute number of false discoveries with an FWER procedure as the cost of one or both discoveries being false may be high. However, in a sample with many discoveries, say 50, the cost of controlling absolute false discoveries to zero is likely to create an undesirably high type two error rate. Having 1 or 2 false discoveries out of 50 may be bearable.

The FDR also allows for datasets to be stratified and sub-divided, whilst retaining a low proportion of type one errors for each group of individuals.

For these reason FDR procedures are often more suitable for large sample fund studies, as in this paper.

2.4.3 Bayesian interpretation

Using Bayes' theorem and the population proportions:

$$G_i = \begin{cases} T & , \quad \Pr(G = T) = \pi_\gamma^+ \\ 0 & , \quad \Pr(G = 0) = \pi_0 \\ U & , \quad \Pr(G = U) = \pi_\gamma^- \end{cases}$$

G is a random variable which takes the value of T if a fund is truly skilled, 0 if zero alpha, and U if a fund is unskilled. The probabilities of skilled, zero alpha and unskilled are π_γ^+, π_γ, and π_γ^- respectively.

A fund's t statistic T_i is significant against the zero alpha null at significance level γ if it lies within the region $T_i \in (t_\gamma^A, \infty)$

Therefore we can define the Bayesian interpretation of the fdr as the probability that a fund has a true alpha of zero, given that its t statistic is significant (in positive abnormal performance) as in equation 10

$$\widehat{FDR}_{\gamma}^{+} = \Pr\left[\, G_i = 0 \;\middle|\; T_i \in \left(t_{\gamma}^{+}, \infty\right) \,\right]$$

(10)

$$= \frac{\Pr\left[\,\left(T_i \in \left(t_{\gamma}^{+}, \infty\right)\right) \middle| \, G_i = 0\,\right].\pi_0}{\pi_0.\Pr[T_i \in \left(t_{\gamma}^{+}, \infty\right) \,\middle|\, G_i = 0\,] + \,\pi_{\gamma}^{+}.\Pr\left[T_i \in \left(t_{\gamma}^{+}, \infty\right) \middle| G_i = T\right]}$$

$$= \frac{\pi_0.(Type\ 1\ error\ rate)}{\pi_0.(Type\ 1\ error\ rate) + \,\pi_1(Type\ 2\ error\ rate)}$$

$$= \frac{\pi_0.(Size)}{\pi_0.(Size) + \,\pi_{\gamma}^{+}(Power)}$$

(11)

(11) shows that the positive FDR increases in the true proportion of null hypotheses, and in the type one error rate (size), and decreases in the true proportion of type two errors (power). This proof relies upon the assumption of independent and identically distributed (iid) errors, which we must make in the FDR analysis, but would not in a bootstrap procedure. Bootstrap procedures are discussed in 2.5.3.

2.5 Methodological difficulties and limitations

2.5.1 Independence and Normality Assumptions

FWERs and FDRs concern comparisons of multiple treatments and families whose test statistics are independent and multivariate normal. In practice, many of the problems encountered are not of the multivariate type, and test statistics are not multivariate normal, and therefore the test may be misspecified. The assumption of independent test statistics across funds is very unlikely to hold in practice, as financial markets are subject to herding (Wermers 1999).

For the Benjamini Hochberg procedure we must assume components of iid. For this paper’s baseline analysis method, the BSW FDR we must make the assumption of normally distributed and independent errors. The limitations of the FDR procedure are such assumptions.

However we can be reassured that the FDR is asymptotically unbiased under weakly dependent test statistics, a more realistic assumption.

We can use the Bayesian interpretation to consider weakly dependent test statistics:

Under general dependence, it does not hold that $FDR_{\gamma}^{+} = \Pr[\, G_i = 0 \mid T_i \in (t_{\gamma}^{+}, \infty) \,]$

But this does asymptotically hold under weak dependence, providing that

$$\Pr\left(\sum_{i=1}^{m} \frac{1 - T_i - U_i}{m} \longrightarrow \pi_0 \right) = 1 \tag{12}$$

$$FDR_{\gamma} \longrightarrow G_0(\gamma) \tag{13}$$

$$\frac{\Pr\left[\, \left(T_i \notin (t_{\gamma}^{+}, \infty) \mid G_i \neq 0 \,\right].\pi_0}{\pi_0.\Pr[T_i \notin (t_{\gamma}^{+}, \infty) \mid G_i = 0\,] + \pi_{\gamma}^{+}.\Pr\left[T_i \notin (t_{\gamma}^{+}, \infty) \mid G_i = T\right]} \longrightarrow G_1(\gamma) \tag{14}$$

Where $G_0(\gamma)$ and $G_1(\gamma)$ are the asymptotic type 1 and type 2 error rates as a function of γ.

If the statistics are weakly dependent then the true proportion of false discoveries and the $\widehat{FDR}_{\gamma}$ converge for any γ. Conversely, if one is able is able to calculate or estimate G_0, G_1, π_0 then for large m these are good approximations for the observed FDR for all γ.

We may struggle to confidently make the assumption of independent test statistics, so this exercise is useful under the weaker and more realistic assumption of weak dependence

2.5.2 Estimating the true proportion of true nulls

S_{γ}^{+} can be observed as the number of significant p values for a chosen γ, so all that is left to estimate in (4) and (5) is π_0. The true value of π_0 is simply equation 15:

$$\pi_0 = \frac{m_0}{m} \tag{15}$$

However m_0 , the true number of true nulls, is unobservable, so we must apply an estimation technique.

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{m(1-\lambda)} \tag{16}$$

λ is a threshold for the proportion of p values, below which we can be confident that all nulls are false, and above which the p values are uniformly distributed and which is consistent with a range of true null hypotheses. λ* is the optimal choice of λ.

In this paper I estimate $\hat{\pi}_0(\lambda)$, the true proportion of true nulls in the population, by choosing λ^* to minimise the sum of squared errors (SSE) of $(\hat{\pi}_0(\lambda) \ -\pi_0)^2$. Therefore minimising the SSE of λ*.

$$\lambda^* = \lambda \in (0,1) which\ minimises\ :\ \int_0^1 (\lambda - \ \hat{\pi}_0(\lambda))^2 \tag{17}$$

BSW, and Cuthbertson, Nitzsche and O'Sullivan (2010a), estimate $\hat{\pi}_0(\lambda)$ in a similar method, with a bootstrap procedure on the mean squared error MSE. If the histogram of p-values is perfectly uniform to the right of our choice of λ, then the estimate $\hat{\pi}_0$ is independent of λ. If only the p-values of true nulls were included, $\hat{\pi}_0(\lambda)$ is an unbiased estimator of π_0 and hence the estimate of the FDR is unbiased. However if we erroneously include false nulls then $\hat{\pi}_0(\lambda)$ provides a conservative estimate of π_0 and the estimated FDR is conservative.

Alternative methods include: applying the smoothing method where we plot $\hat{\pi}_0(\lambda)$ against λ, then fit a cubic spline and take our estimate to be $\hat{\pi}_0(\lambda = 1)$; or the most simple method, which is to simply observe the λ past which the histogram of p values is uniformly distributed.

The SSE and MSE methods have the advantage of being data driven, and therefore less arbitrary in estimating results. The practical difference between this and other methods should be trivial, as BSW and Cuthbertson (2010a) show results to be robust to variations in estimates of π_0, and my estimate $\hat{\pi}_0(\lambda)$ falls within the region both papers find robust to changes (see 5.3.1).

2.5.3 Bootstrapping as a solution

Recent studies have addressed the false discovery issue with two broad categories of techniques. The first, adopt an Efron style fdr in a multiple hypothesis testing setting (BSW; Kim, In, Ji, Park (2014)). The second use cross-sectional bootstraps under a null of zero alpha to compare the distribution of simulated t-statistics from observed results to distinguish luck from skill (Kosowski (2006), Fama and French (2010)).

The cross-sectional bootstrap addresses many of the methodological weaknesses inherent in previous research on fund performances, most notably, the case of non-normality and dependence of test statistic distribution. Bootstrapping facilitates the explicit controlling of luck in a complex non-normal cross-sectional distribution of mutual fund alphas. Thus eliminating the need to specify an ex-ante parametric distribution from which returns are drawn, to estimate portfolio return correlations, or to explicitly control for ex post data snooping biases.

Horowitz (2003) shows that use of bootstrapping in Monte Carlo experiments can spectacularly reduce the difference between the true and nominal probabilities of correctly rejecting the null hypothesis. However the problem in bootstrapping is the difficulty in specification. An incredible amount of data needs to be processed in the mutual fund setting when performing a bootstrap operation, and for this reason the time constraints of this study will prohibit a reliable bootstrap from being performed.

Two advantages of FDR method are its simplicity in not imposing any arbitrary assumptions, and secondly, providing a more accurate partition of the universe of funds into positive, negative and zero alpha. The difference in results from using FDR procedures and inference using the bootstrap diminish in larger samples (m) and in a longer time series (t). I conclude that the BSW FDR is the most appropriate methodology to control for luck in fund performance in this paper, due to these factors, and the complexity of bootstrap specification in light of the constraints of this project.

2.5.4 Serial correlation and heteroskedasticity

It is widely acknowledged that market movements exhibit cross-correlations, time dependent variables, for example due to economic cycles, and herding as in Wermers (1999). The risk free rate and market rate variable should account for time dependence and market conditions. Additionally, I improve upon standard testing procedures using standard t-tests and critical regions by using Newey-West autocorrelation and heteroskedasticity robust standard errors.

3. Performance Measurement

In 1.1 I presented the linear regression model. In section 3 I discuss asset pricing models of this form which explain variation in UK fund performance. I must select market-wide factors as covariates in the model which best explain the observed variation in returns, hence accurately estimate our parameter of interest α_i.

For simplicity, I use unconditional models with time invariant fund-specific factor loadings. These models are interpreted as practically implementable, factor-mimicking portfolio strategies that any fund manager would expect to outperform. The models use exposure to risk to generate an expectation of fund performance, consistent with a model of market equilibrium.

Results will be biased in the if the asset pricing model is misspecified. For this reason, I consider four alternative well known unconditional factor models to evaluate performance, which are briefly described below. I select the Carhart 4 factor model as my baseline according to theory and evidence on UK asset pricing models as discussed in 3.13. I confirm its suitability in regression results in section 5. This model is then used to compute the FDR analysis. For robustness, I run regressions using the 4 models in 3.11 and 3.1.2.

In 3.1 I introduce the CAPM single index model, the Fama French 3 Factor model, the Carhart 4 factor model, and the Treynor Mazuy market timing variable. In 3.2 I discuss motivation and considerations in model choice. In 3.3 I discuss critiques of these simple factor models.

3.1 Linear factor models

3.1.1 CAPM Single Index, Fama French 3 factor and Carhart 4 factor models

$$r_{it} = \alpha_i + \beta_i\left(R_{mt} - r_{ft}\right) + \gamma_i SMB_t + \delta_i HML_t + \mu_i MOM_t + \varepsilon_{it} \tag{18}$$

Where: $$r_{it} = R_{it} - r_{ft} \tag{19}$$

r_{it} is the excess return of fund *i* over the risk free rate in month t, as shown in (15). $R_{mt}$is the gross return of fund i in month t. SMB HML and MOM are size (small minus big, market capitalisation stocks), value (high minus low beta stocks, or value versus growth) and momentum (lagged performance), respectively.

ε_{it}fund *i*s error term in month t. *i* terms refer to fund *i* across the cross section, *t* terms refer to month *t* along the time series. β_i is the sensitivity of excess returns of fund *i* to excess market returns. Coefficients γ_i, δ_i and $\mu_i$are the sensitivity of excess returns of fund *i* to the monthly size, value and momentum factors.

(18) represents the Carhart (1997) 4 factor model. I also use the Fama French 3 Factor model, obtained by setting $\mu_i = 0$. And the single index Capital Asset Pricing Model (CAPM), obtained by setting $\gamma_i = \delta_i = \mu_i = 0$. These are the three unconditional linear factor models which I consider, and which account for the vast majority of studies on fund performance.

I measure abnormal performance in fund *i* by its estimated constant intercept alpha $\widehat{\alpha_\iota}$ term, after estimating the portfolio's expected performance based upon its risk characteristics with the factor model. This measure is known as Jensen's alpha, after Jensen's 1968 seminal paper measuring abnormal returns in this way. For analysis of true skill, the test statistic $\widehat{\alpha_\iota}/\widehat{\sigma_{\alpha_\iota}}$ is our performance measure. $\hat{\sigma}_{\alpha_i}$ is the estimated standard deviation of $\widehat{\alpha_\iota}$, and from this we can calculate the p value.

3.1.2 Capturing market timing: Treynor-Mazuy model

I also capture market timing using Treynor and Mazuy's (1966) simple model of market timing, adding a quadratic term to a factor model to capture the curvature of market timing.

$$r_{it} = \alpha_i + \beta_i(R_{mt} - r_{ft}) + \gamma_i SMB_t + \delta_i HML_t + \mu_i MOM_t + \emptyset_i(R_{mt} - r_{ft})^2 + u_{it} \tag{20}$$

Market timing is measured by the direction and significance of $\emptyset_i$. A skilled manager will have a positive and significant $\emptyset_i$ term.

Manager skill can manifest itself as stock picking abilities, tilting portfolios to increase risk adjusted returns, which can be proxied by performance against the four-factor model. Or managers may be skilled in timing the market, profitably moving from one sector to another. Market timing can occur even in a confined investment universe, and it tests the ability of the manager to take an aggressive position and invest in high beta stocks in bull markets and switch to defensive positions by holding low beta stocks in bear markets. In other words, the ability of managers to sieze the opportunity of a bull market and do even better than expected, and to play a defensive strategy to cut losses and not do as badly as expected in a bear market.

3.2 Model motivation, literature and considerations

The Carhart four factor model is a widely used and empirically strong, pricing model, constructed by combining Fama and French's (1993) three factor model with the addition of Jegadeesh and Titman's (1993) one year momentum anomaly. This four factor model is commonly used in fund evaluation and as a basis for active management strategies. Empirical works on mutual fund performance, including BSW, Kosowski (2006), Tuzov and Viens (2010), and Chen, Chu and Leung (2012) use the Carhart (1997) model as a baseline. Additionally, most studies on UK UT managers use a three or four-factor model as a baseline with the other as comparison, including Clark (2013), Cutherbertson (2008, 2010a), Quigley Sinquefeld (2000), and Tonks (2005). Cuthbertson (2010a) find a 3 factor Fama and French factor model explains UK fund returns the best out of a range of factor models.

Treynor and Mazuy (1966), Merton and Henrikson (1981) amongst other authors have demonstrated that the Jensen alpha measure of performance is biased in the presence of manager market timing skills. Grinblatt and Titman (1994) proposed an extended factor model with the addition of a quadratic excess market return variable.

For these reasons I will test jointly for stock picking abilities and market timing skills, with the Treynor-Mazuy method, by adding a quadratic variable of excess market returns to a four factor model. In this case to create a five factor benchmark model as in (20).

3.3 Factor model critiques

We must assume our asset pricing model represents fundamental value and also that managerial incentives are aligned with maximising the factor model value. If this does not hold then the resulting estimators will be biased. The joint hypothesis test problem refers to the impossibility of comprehensively testing market efficiency without a proven pricing model. Fundamental value is unobserved and can only be estimated with arbitrary factor models. Identification of abnormal performance could reflect genuine abnormal returns, or normal returns and a misspecified model. Clearly there is not an ex-ante pricing model we can choose that is universally accepted to reflect fundamental value, or all stocks would be correctly priced and markets would be strong-form efficient, which there is much evidence against in the UK (King and Roell 1988). So we must rely upon the models for which there is most evidence, the model which ex-post fits the market data the best.

We must make these assumptions, but will briefly discuss the use of other models and critiques of the time invariant one, three and four factor models., Agei-Ampomah, Clare, Mason, and Thomas (2015) provide evidence that standard multi factor models misspecify managerial benchmarks and targets. They argue that we must not simply compare managerial performance against the entire cross-sectional universe of funds, but control for heterogeneous managerial style, constraints, conventions and incentives, and adopt style consistent benchmarks. By using style-consistent benchmarks they find that analysis using the standard factor models significantly underestimate managerial skill and overestimate luck due to misspecification of performance targets.

Suh and Hong (2011) use a 6 factor baseline asset pricing model, incorporating the Carhart (1997) four factor model and two bond market factors proposed by Fama and French (1993) from the consideration that many funds contain bonds in their portfolios. Ntozi-

Obwale, Fletcher, and Power (2008) show dummy variables, representing a strong and weak state of the economy, improve factor model performance in a multiple hypothesis test setting. The Ferson and Schadt (1996) method assumes that alphas may be linearly related to macroeconomic variables such as the interest rate and aggregate dividend yield. The inclusion of the conditioning variable to performance measurement investigates whether time variation in alphas is due to correlation between the precision of private managerial information and public information variables, in other words, whether the manager is simply using public information to select stocks.

Fund performance studies, including Kosowski (2011) and Fletcher (1995) in the US and UK respectively, use conditional time varying alpha and beta models using macroeconomic conditioning variables based on the Ferson Schadt (1996) model, and there is strong evidence that these variables add accuracy to pricing models.

There are many examples of additional factors proposed and proven significant to some extent in fund literature. However, the models I have chosen have been proven to explain results to a high degree of significance, and have become the benchmark in most fund performance literature in the UK and US.

This paper will attempt to make funds as homogenous as possible with restrictions on observable fund characteristics. But will not adjust for unobservable style consistent benchmarks or macroeconomic conditions, past the time varying market factors with time invariant fund specific coefficients, in the four factor model. Further research would benefit from these improvements to the pricing model.

4. Data

In section 3 I selected the Carhart 4 factor model as this paper's baseline asset pricing model. In section 4 I aim to collect reliable data on the observed variables y_{it} and x_{it} (as in model (1)). In 4.1 I describe the explanatory market variables x_{it}. In 4.2 I describe the outcome fund return variables y_{it}, and explain the restrictions, considerations and potential problems with available fund data.

Subtle differences in testing techniques, confidence intervals, netting of fees and controls on variables such as managerial style can quantitatively and qualitatively alter results. Therefore thorough consideration of testing techniques, variables and a meticulous approach to data issues is necessary for the results of such a study to carry any empirical weight.

4.1 Market data

Monthly market data was collected from the publicly available database on the University of Exeter Business School website, as described in Tharayan and Christidis (2013). These independent UK market variables are: the riskfree rate, the market return rate, size, value and momentum. The CAPM single index model only uses excess market return. Size and value variables are added in the three-factor Fama French model, and momentum is added in the Carhart four-factor model. The riskfree rate and market rate are proxied by the monthly return on three-month Treasury Bills and the FTSE 100 Index respectively.

4.2 Fund data

I used DataStream to collect a panel of monthly returns data for 73 actively managed UK fixed income unit trusts which were in existence for any 12 month (minimum) period between October 1980 and September 2014.

Fund returns are measured using bid-bid spreads, including reinvested dividends, gross of transaction costs and management fees. This study is interested in managerial talent, in particular, stock picking abilities, for which we only require the change in gross portfolio price, representative of net asset value in UTs. Fee policy is independent of managerial skill, and to take post-expense returns fails to separate the two.

The sample spans the financial crisis of 2008, which many studies avoid, often cutting short to samples ending in 2007, I see no good reason to exclude the financial crisis since the concept of beating the market remains the same during a shock. Cross-correlations of returns increase during times of market stress, but this should not have any practical implications on the results.

All market data from the Exeter database was taken from the first trading day of each month. Datastream fund data was taken on the first day of each month. The occasional difference between these should be trivial when using monthly panels.

Usefully, in the Exeter database, Tharayan and Christidis (2013) have done significant statistical legwork by cross-factor matching funds where year-ends are heterogeneous. Due to its empirical quality, the database is the basis of numerous UK performance analyses.

4.2.1 Restrictions and homogeneity considerations

In this section I consider the restrictions I make to the dataset to improve homogeneity.

The global unit trust fund universe on DataStream contains 447,489 funds. I restrict this to 56,972 UK funds trading in pounds sterling. I further restrict this to 184 fixed income funds. I then reduce filter out tracker and index funds and restrict to one share class per underlying portfolio to reduce the sample to the final size of 73 funds. I am satisfied with these homogeneity restrictions.

UK UTs are defined as undertaking 80% of their investment in UK stocks. This restriction on investment universe increases the homogeneity of the opportunities available to managers. Thus allows us to use more accurate benchmark factor portfolios to estimate risk adjusted abnormal performance.

Additionally, for the purposes of homogeneity, it is necessary to apply a restriction such as fixed income funds only. Characteristics such as fund strategies, targets, investment universe, taxes, and client and managerial expectations differ across the main sectors (equity, international, specialist, balanced, and fixed income) and therefore they do not have be perfectly comparable returns. I chose fixed income because most UK UT papers focus on equity and this improves the uniqueness of this paper’s contribution. Furthermore, fixed income funds are expected to exhibit consistency because of their relative homogeneity (Anderson and Ahmed 2005). Funds may change objectives across a time series, and this information is not clear from the Datastream data. Clark (2013) finds that objective changes are an uncommon occurrence and that ‘when fund sector changes do occur it is more likely to be within the same asset class than changing focus entirely’. Deliberate misclassification of funds, as highlighted in Mostovoy (2015), cannot be controlled for in this study and I will treat funds listed as fixed income as such.

After filtering UK fixed income unit trusts, I placed two additional restrictions, and reduced the sample of listed funds from 184 to 73. These were a restriction of only one share class per underlying portfolio, and the exclusion of index and tracker funds.

For the former, I excluded all but one fund within groups of funds under the same name but with a different share classes, for example keeping only one of “7im Aggressive Growth A”, and “7im Aggressive Growth B” (and C,D,E etc in some cases). I omitted these on the basis that the underlying portfolios are identical, and to include all of them would bias the results towards the funds with the most share classes or fee structures belonging to the same underlying portfolio. For the latter restriction, I omitted all funds with ‘tracker’ ‘index’ and variations such as ‘trkr’ and ‘indx’, and then inspected the remaining sample for any variations I had missed. Other studies such as Clark (2013) confirm the robustness of this simple identification method. These funds are passive market trackers rather than actively managed, so are not of interest to this study, but had failed to be filtered out by the DataStream active management filter.

We can rarely make such a bold assumption as homogenous error terms in panel data, certainly not in the likely presence of time varying market factors not entirely explained by the factor models. Therefore I use Newey-West heteroskedasticity robust standard errors.

4.2.3 Survivor bias

Survivor bias occurs when worst performing funds shut down, and cease to report figures, therefore removing their poor returns from analysis, leaving only the successful surviving funds. This creates a positive bias on observed active management returns with respect to true data.

My dataset includes some dead funds although an incomplete sample. It is also unclear which funds closed due to poor returns, and which were merged with others under the same umbrella fund. Such information has implications on survivor bias. Altogether, the dataset includes 8274 of a possible 29784 monthly observations from the funds studied. The missing data represents periods in which funds were not in operation (either yet to be created, or had died). This averages at 113.34 of 408 monthly periods, in other words that the average fund in the sample operated for 9.44 years of the 34 year sample, although there is substantial variation in lifetime. 45 of the 73 funds are live in the final period and assumedly lived beyond the end of the sample. Given the average 9.44 year life of a fund this illustrates the bias of the data towards live funds.

The CRSP database was developed by Carhart 1995, and is a US survivor-bias-free mutual fund database, offering information on live and dead funds, rather than just the current investment opportunity set. An equivalent publicly available dataset for the UK is not available, and this may offer a reason behind the comparatively small number of studies on UK funds, unit trusts and OEICs. Clark 2013 creates a UK Equity Unit Trust/OEIC Suvivor-Bias-Free Database from a collection of sources. However, the database is not publicly available, and the process of extensive matching and cross-checking funds across time periods from different datasets is beyond the time constraints of this study.

In the presence of heterogeneity of fund volatility across the cross section of returns, the best and worst performing funds are the most volatile. The worst performing funds do not survive, leaving the sample with volatile high performers (and the less volatile normal return portfolios). So conditional on survival, the highly volatile winners upwardly bias the sample since the highly volatile losers are removed.

Carpenter and Lynch (1999) simulate standard performance tests to identify and mitigate the effects of survivorship and attrition biases. The authors find evidence that the attrition of poor performers changes the composition of the sample. This effect causes upwardly biased estimates of performance and downwardly biased estimates of persistence. The authors recommend eliminating the poorest 3.6% of funds each year in line with Carhart's (1997) findings on fund disappearance.

Missing data can create survivorship bias, when the sampling methodology replaces missing months with the average performance of surviving funds. I avoid this practice and do not adjust missing values. Filling in missing data points is unnecessary in this ordinary least squares analysis, and the effect is trivial as, in the sample, I only identify 3 empty data points which occur during the lifespan of a fund.

Blake and Timmermann (1998) estimate average survivor bias in the UK of around 0.8% per year, with very strong evidence of underperformance averaging -3.3% in the last year of life. By any economic interpretation this is significant, and with this is mind, adjusting for survivor bias is important.

4.2.3 Minimum history bias

Minimum history bias is similar to survivor bias, and occurs where funds which underperform may not exist for long enough to be counted in the sample, causing an upward bias in returns. There is a trade-off between number of observations giving meaningful explanatory power in regressions, and minimum history bias.

I restrict the sample to funds alive for any period of 12 months or longer within the time frame. I have chosen a relatively small restriction with a view to decrease minimum history bias and increase the number of cross-sectional funds in the study.

4.2.2 Data shortfalls: managerial changes; fund misclassification; conditional variables

Managerial changes were not clear from the data. An extension to this paper could be to test Tonks' 2005 hypothesis that managerial changes are responsible for the lack of persistent abnormal performance. However in most cases this would require manually searching for each fund's managerial history, which the time constraints of this project do not allow.

Mostovoy (2015) highlights the prevalence of fund misclassification leading to deliberately misspecified market rate benchmarks, to improve the active management published returns. Mason (2015) highlights the importance of style consistent market benchmarks. Funds have different targets, styles (growth, income, aggressive growth etc) and ideally should be treated as heterogenous, or at least style controlled for with fund or at least sector specific benchmarks.

In this project I will not be able to test such claims and will assume a homogenous market benchmark proxied by the FTSE 100. The market rate given by the Exeter database is the FT All Share Index, I replace these values with the equivalent results from the FTSE 100 which represents a more appropriate large cap UK investment universe and therefore market index for UK fixed income UTs.

Unconditional factor models make the implicit assumption that managers use no information about the state of the economy to form expectations. Ferson and Schadt (1996) show unconditional measures to be biased, and advocate conditioning returns on predetermined publicly available information. Fletcher, Ntozi-Obwale and Power (2008) show conditional measures to be more accurate than unconditional ones when assessing UK funds. However, conditional versions of the factor data are not available on Datastream.

Clark (2013) finds a trivial difference between UK unit trust performance measurement using conditional and unconditional models (apart from timing coefficients). The constraints of this project mean that I focus on creating a quality unconditional dataset, rather than creating information conditioning variables. Thus conditional models are not used in this study.

5. Empirical Strategy and Results

In section 5 I aim to estimate the population proportions of skilled, unskilled, and market-matching funds. In 5.1 I outline the estimation procedure. In 5.2 I describe cross-sectional regression results. In 5.3 I describe the main results from this paper. In 5.4 I compare results from different type one error control procedures.

5.1 Estimation procedure

The estimation procedure is as follows: First I compute baseline fund returns based upon factor models. I define these as the zero-alpha null hypotheses of no abnormal returns, and against which I test the two sided alternative hypotheses. I then select the significance of the test and type one error rate, and apply two sided multiple testing procedures. I transform an m x 1 vector of p values according to the type one error controlling procedure. I then estimate the number and proportion of true skilled, unskilled and zero alpha funds. I repeat the procedure at different significance levels to identify where the truly skilled and unskilled funds are located within the tails of the distribution of fund alphas. In testing, I assume a normal distribution of independent or weakly dependent test statistic around a mean of the factor model generated baseline returns.

5.3.1 Key results

Using the BSW FDR procedure, I test 73 UK fixed income unit trusts operating at any time between 1980 and 2014, using the Carhart 4 factor model as a baseline. Gross of fees, I fail to reject the zero alpha null for 80.7% of funds at 10% significance. 16.5% of funds are truly skilled, and 2.8% are truly unskilled. I can reject that all funds are truly zero alpha, and also observe that the cross-sectional average alpha is positive, although small. These results are shown in table 2.

I find that 13.9% of nulls rejected for abnormal positive performance are false positive discoveries and due to luck, and 48.7% of nulls rejected for negative abnormal performance are false negative discoveries. These are the false positive and negative discovery rates. The false negative discovery rate is notably higher than the false positive rate, in contrast to existing literature (BSW, Cuthbertson (2010a)).

	Zero alpha $\hat{\pi}_0$	**Non-zero alpha**	**Skilled $\hat{\pi}^+_{0.1}$**	**Unskilled $\hat{\pi}^-_{0.1}$**
Proportion	**0.807**	**0.193**	**0.165**	**0.028**
Number	**59**	**14**	**12**	**2**

Table 2. BSW FDR procedure results, 10% significance, 3 decimal places

5.2 Cross-sectional regression results

Table 3 displays the coefficient estimates for four unconditional factor models using an equally weighted portfolio approach. Using each of the factor models, I run an ordinary least squares regression on time series data for each of the 73 funds, over the sample period October 1980 to October 2014 I use serial correlation and heteroskedasticity robust Newey-West standard errors.

There is statistically significant evidence of slight abnormal performance. In all four models, the cross-sectional alphas have a mean of approximately +0.2 and are positively skewed. These main findings hold over the similar but distinct models, which confirms the robustness of the results. This suggests that fund managers on average add some gross value.

The observed alphas in the benchmark model range from -0.64 to 1.63, with a standard deviation of 0.4767, demonstrating the wide range of ex-post observed alphas. The distribution of alphas is shown in figure 2. From a simple observation of figure 2, there do not appear to be separate distributions of skilled and unskilled funds, although one can possibly detect a positive skew, which could be supportive of a small group of skilled funds. I go on to apply the FDR procedures to test this formally. It is notable at this stage in the investigation that ideally the sample size would be larger, we should not be able to observe gaps in the histogram of alphas. The lack of sample size is the cost of choosing fixed income funds, for which there is only a small universe.

The histograms of alphas for the CAPM single index model and the Fama French 3 factor model also shown in figure 2. We can observe similar distributions and clustered alphas around the 0.1 to 0.2 region.

The market timing component, or the Treynor-Mazuy test for market timing, shows significant evidence of negative market timing. This result is consistent with Cuthbertson (2010b) and Clark (2013). I find that this negative coefficient result also holds across the one and three factor models, but do not report the results. The implications of a negative coefficient is that fund managers are either poor at timing the market, or have different incentives to maximising short term net asset value.

All four models explain notably little of the variation in observed returns, with the average R^2 in the single factor CAPM model at just 8%, ranging up to 24% in the five factor model. These R^2, the percentage of observed dependent variable variation explained by the independent variables, is not in line with other similar UK papers, for example Clark (2013), Quigley Sinquefield (2000) Cuthbertson (2010a) who all report R^2 statistics of 80% and above from three, four and five factor models in the UK.

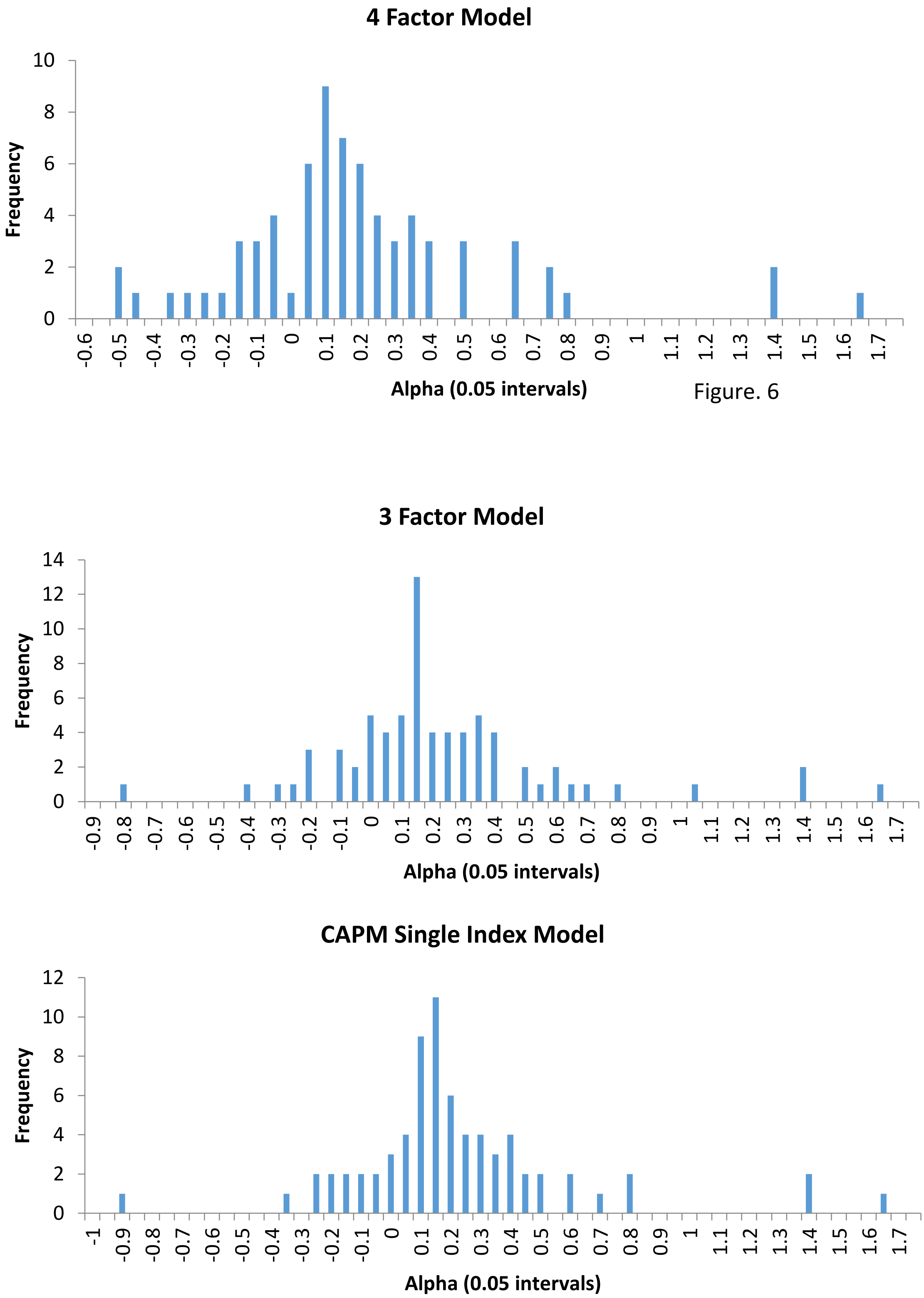
Despite the low explained variation, the distribution of the estimated alpha terms and the direction of the coefficients are in line with theory and UK evidence, which is supportive of the evidence presented in this paper.

Using survivor-bias-free data improved my results, as I have identified the two unskilled funds in the sample to have died before September 2014. Their exclusion would have qualitatively and quantitatively changed my results and conclusion.

Table 3: Summary statistics of UK Unit Trusts

	(1) CAPM	(2) 3 Factor (Fama French)	(3) 4 Factor (Carhart)	(4) 5 Factor Timing (Treynor Mazuy)
$\hat{\alpha}$	0.2052 (0.3900)	0.2192 (0.3932)	0.1889 (0.4134)	0.2190 (0.3537)
$\hat{\beta}$ **(Rm-rf)$_t$**	0.0086 (0.4784)	0.00931 (0.4876)	0.0142 (0.4532)	0.0173 (0.4428)
$\hat{\gamma}$ **SMB$_t$**		0.01573 (0.3879)	0.0261 (0.3812)	0.0248 (0.4032)
$\hat{\delta}$ **HML$_t$**		-0.0226 (0.5016)	-0.0230 (0.4296)	-0.0221 (0.4207)
$\hat{\mu}$ **Mom$_t$**			0.004796 (0.4212)	0.0015 (0.4439)
$\hat{\emptyset}$ **(Rm-rf)$_t^2$**				-0.0018 (0.2889)
R^2	0.08	0.17	0.19	0.24

N = 73. T range: 12-408, mean: 113.34. Performance measured with unconditional factor models using monthly data over the entire period 1980-2014. Each panel contains the estimated alpha ($\hat{\alpha}$), our measure of abnormal performance; and the estimated exposures to the market ($\hat{\beta}$), size ($\hat{\gamma}$), value ($\hat{\delta}$), momentum ($\hat{\mu}$), and market timing ability ($\hat{\emptyset}$). The adjusted R^2 relates to the mean equally weighted portfolio of all funds that are alive in each period. Figures in parenthesis denote the Newey-West heteroskedasticity and autocorrelation consistent estimated p value, calculated under a null hypothesis that the regression parameters are equal to zero.


4 Factor Model

3 Factor Model

CAPM Single Index Model

Figure 2. Distributions of fund alphas

The significance of results and alignment of size and direction of coefficients results from table 3 and the closeness of the distribution of alphas to a normal distribution in figure 2 supports my choice of the 4 factor model as baseline. We can also see empirical and theoretical strength behind the other two standard models. I am satisfied that these models are behaving correctly and are suitable for this investigation.

5.3 Barras Scaillet Wermers false discovery rate procedure

I find that 16.5% of funds are skilled and able to outperform their benchmark, 2.8% of funds are unskilled. I do not find significant evidence of abnormal performance in the remaining 80.7% of funds. I estimate the FDR at 13.9% for outperforming funds, and 48.7% for underperforming funds.

In this sub-section I describe the estimation results in the baseline procedure. In 5.3.1 I present the distribution of p values from the testing procedure, and estimate the true proportion of true nulls. In 5.3.2 I estimate the population proportions and fdr. In 5.3.3 I examine the tail distributions. In 5.3.4 I discuss the Bayesian interpretation of results. In 5.3.5 I consider the estimation of asymptotic results.

5.3.1 P value distribution, and estimating π_0

We can see the distribution of P values graphically in Figure 3. We can clearly observe a spike in p-values between 0 and 0.1. The distribution of p values is not perfectly uniform outside of the significance region from 0 to 0.1, however the sample of just 73 funds was perhaps not large enough to see the p values tend to their asymptotic distribution. Of the 18 zero-alpha nulls rejected at a 10% t test, 14 are positive abnormal returns indicating skilled or lucky funds, and 4 are negative indicating unskilled or unlucky funds.

I use the observed p values to estimate $\hat{\pi}_0$, the true proportion of true null hypotheses.

I estimate $\hat{\pi}_0$ = 0.534 using methodology described in section 3.6. The exact calculations are given in Appendix A2.

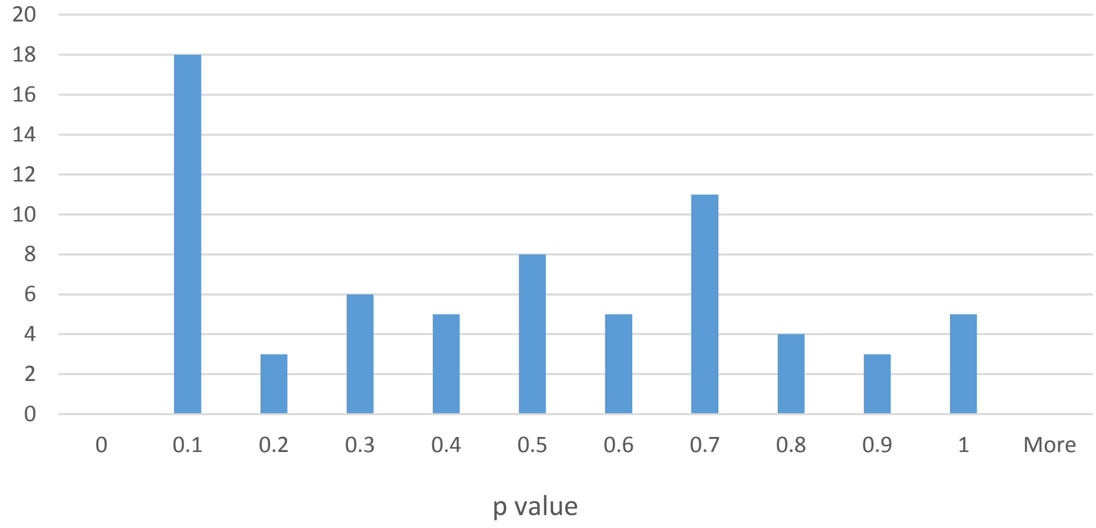


Figure 3. Distribution of p values, Carhart 4 factor model

BSW use a bootstrap methodology to calculate an estimate of 0.6, and results are robust to varying this to 0.5. Cuthbertson (2010a) actively choose a conservative estimate of 0.5, and also find robustness results. These results are in line with my SSE results.

The bias of the estimator $\hat{\pi}_0(\lambda)$ decreases in λ but its variance increases which represents a trade-off. The method exploits this trade-off and attempts to minimise the costs. The SSE method gives a relatively conservative estimate, with the majority of results in the right hand column of table A1 giving values over 0.534.

P value distributions of alternative models are given in Appendix section A1, figure A1.

5.3.2 Population proportion and false discovery rate results

$$\hat{\pi}_{0.1}^{+} = 16.51\% \qquad \widehat{FDR}_{0.1}^{+} = 13.93\%$$

$$\hat{\pi}_{0.1}^{-} = 2.81\%$$

$$\hat{\pi}_{0} = 80.68\% \qquad \widehat{FDR}_{0.1}^{-} = 48.72\%$$

13.73% of nulls rejected due to positive abnormal performance are false discoveries and should be considered lucky, when testing at 10% significance. This is the primary fdr result from my investigation. Of the 14 funds identified as positive abnormal performers, we should reject 2 of these as false positive discoveries, leaving 12 truly skilled performers, equal to 16.51% of the population.

Of the 4 nulls I reject in the sample for abnormal negative performance, I identify 48.72% of these as false negative discoveries, so conclude that 2 members of the population are truly unskilled, which is equal to 2.81%. The false negative discovery rate is notably higher than the false positive rate, in contrast to Cuthbertson (2010a) which found far higher false positive discovery rates in UK unit trusts.

Calculation details of these results are given in Appendix section A3 and A4.

5.3.3 Examination of tail distributions

From comparing the FDR at different significance levels, we can estimate where the truly significant funds are located in the distribution of alphas.

The underperformers suffer from low sample size, but in table 4 we can see the $\widehat{FDR}_{\gamma}^{-}$increasing in γ from 1% significance, where we can be sure with 99% confidence that a significant fund is a true underperformer, to 20% significance, where we can only be sure to around 3% confidence, with a 97% FDR.

The $\widehat{FDR}_{\gamma}^{+}$ increases in γ, from around 5% at 1% testing significance, to 22.9% at 20% testing significance. We have identified most outperforming funds at 5 % significance, and almost all at 10%. The $\widehat{FDR}_{\gamma}^{+}$ also increases significantly from 12.2% at 5% significance to 19.2% at 10% significance. This tells us that most true outperformers are located in the tail of the alpha distribution beyond p = 0.05, and certainly beyond p = 0.1. The results of Coles, Naveen and Nardari (2006) show that we have reasonably high power in detecting outperformance in the tails of the cross-section of the performance distribution. I conclude that an extremely low p value is a reliable indicator of a truly abnormal fund.

Significance level γ	$\hat{S}_{\gamma}^{-}$	$\widehat{FDR}_{\gamma}^{-}$	Number of true underperformers	$\hat{S}_{\gamma}^{+}$	$\widehat{FDR}_{\gamma}^{+}$	Number of true outperformers
1%	2	0.099	2	4	0.0485	4
5%	2	0.494	1	8	0.122	7
10%	4	0.487	2	14	0.139	11
15%	4	0.731	1	15	0.195	12
20%	4	0.972	0	17	0.229	13

Table 4: Comparison of false discovery rate at different significance levels

5.3.4 Bayesian interpretation

A Bayesian interpretation of the FDR is useful from an investor's perspective. Given observed t statistics and posterior beliefs about the true proportions of skilled unskilled and zero alpha funds $\Pr(G = i), i \in [T, 0, U]$, we can calculate the chance of a fund being skilled, given that it has significant ex-post outperformance. This is an important statistic for calculating expected returns.

$$fdr_\gamma^+ = \Pr[\, G_i = 0 \mid T_i \in (t_\gamma^+, \infty)\,] = \frac{\Pr[\,(T_i \in (t_\gamma^+, \infty) \mid G_i = 0\,].\Pr(G_i = 0)}{\Pr[T_i \in (t_\gamma^+, \infty)]} = \cdots$$

$$\ldots = \frac{\hat{F}_\gamma^+}{S_\gamma^+} = \hat{\pi}_0.\frac{\gamma}{2.S_\alpha^+} = \widehat{FDR}_\gamma^+$$

(21)

The Bayesian interpretation gives us a clear insight into the perspective of the investor, calculating the probability of a fund manager being truly skilled, conditional upon previously significant returns, or from the prospective of a fund CEO, deciding whether to headhunt a 'star performer'.

We can transform this into an investor loss function of the decision to invest in fund *i* as in Storey (2002) and BSW,

$$IL(t_\gamma^+) =$$

$$(1-\theta).\Pr[\,T_i \in (t_\gamma^+, \infty)].\Pr[\,G_i = 0 \mid T_i \in (t_\gamma^+, \infty)\,] + \cdots$$

$$\theta.\Pr[\,T_i \notin (t_\gamma^+, \infty)].\Pr[\,G_i = T \mid T_i \notin (t_\gamma^+, \infty)\,]$$

$$= (1-\theta).\Pr[\,T_i \in (t_\gamma^+, \infty)].fdr_\gamma^+ + \;=\; \theta.\Pr[\,T_i \notin (t_\gamma^+, \infty)].fnr_\gamma^+$$

(22)

Where $IL(t_\gamma^+)$ is the investor's loss as a function of the significance level. Θ is the cost of failing to detect a skilled fund, and can be interpreted as regret. fnr_γ^+ is the false nondiscovery rate, the probability of a skilled fund not being detected as significant. $IL(t_\gamma^+)$

can be minimised with a choice of significance threshold t_{γ}^{+}. The FDR as calculated in this paper is a necessary component of such an investment decision.

BSW find that a high FDR^{+} target is consistent with a high cost of regret of failing to identify a skilled fund, whereas a low FDR^{+} target is consistent with a low cost of regret.

5.3.5 Asymptotic adjustment of true p distribution

In 3.6.2 I showed how asymptotic results of the FDR are valid under weak dependence of test statistics. In this sub-section, I hypothetically adjust the results to their true asymptotic distribution and consider the effect on the $\widehat{FDR}$

To hypothetically correct for the non-normal distribution, we could take the same proportion of significant p values (0.247 at 10%), and make the assumption that the p values were otherwise normally distributed, as should be their true asymptotic distribution. I demonstrate this in Figure. 4. To then apply the same procedure as in 6.2.2 gives $\hat{\pi}_0(0.74) = 0.843$. This exercise produces nearly identical to our results from table 1, that the true proportion of true nulls 0.834 which is supportive of the results from my sample.

I obtain $\widehat{FDR}_{0.1}^{+} = 0.217$, $\widehat{FDR}_{0.1}^{-} = 0.769$. This leaves us with 7 truly skilled and 1 truly unskilled fund. These results, when taken with the sample results, do show quantitative and qualitative differences, with a smaller $FDR^{+}$and a very large FDR^{-}, further supporting a lack of truly unskilled funds. However, we can see similarities, and these hypothetical

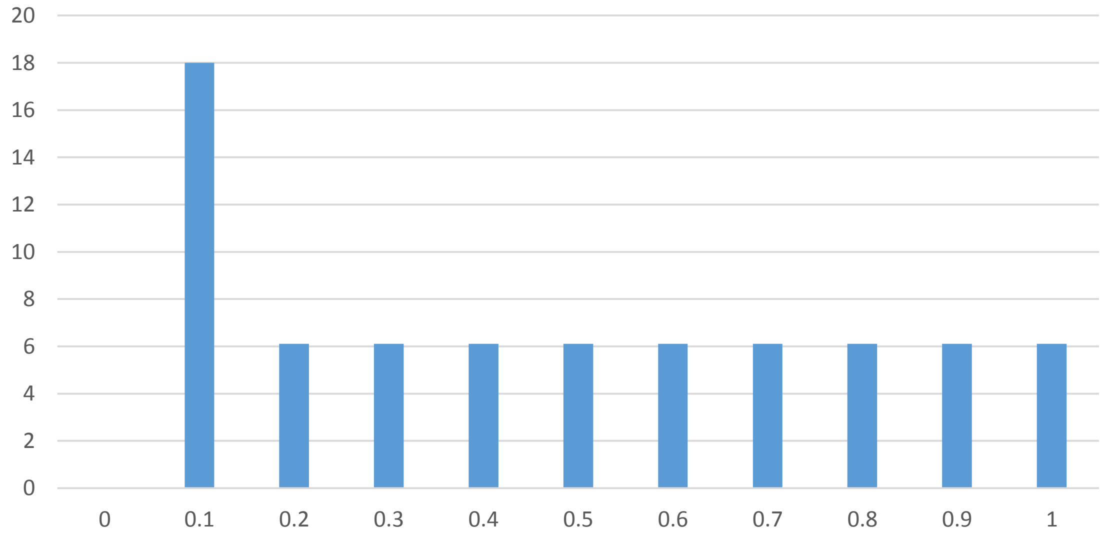


Figure. 4 Asymptotic Distribution of P Values

results may tend towards the true asymptotic FDR. This is a useful result to consider, but not to be taken as experimental results.

5.4 Comparison of type one error control procedures

Table 5 shows the proportion of funds that I find to be skilled, unskilled and zero-alpha, when applying different FDR and FWER control procedures.

Under the Benjamini and Hochberg (2001) FDR procedure at a 10% significance level. I find that, of the 73 funds in the study, I reject the zero-alpha null for just 6 funds. 4 of these rejected nulls have significantly positive abnormal returns, and 2 have negative abnormal returns. I fail to reject the null for the remaining 62 funds.

With the stringent Bonferroni FWER control procedure, we can only reject the null for one fund. The results make theoretical sense, with increasing stringency and fewer rejections from t-test to BSW to Benjamini Hochberg to Bonferroni procedures.

10% significance	**Zero alpha**	**Non-zero alpha**	**Skilled**	**Unskilled**
Standard t-test	**0.75** (55)	**0.25** (18)	**0.19** (14)	**0.06** (4)
Benjamini-Hochberg	**0.92** (67)	**0.08** (6)	**0.06** (4)	**0.03** (2)
BSW FDR	**0.81** (59)	**0.19** (14)	**0.16** (12)	**0.03** (2)
Bonferroni	**0.99** (72)	**0.01** (1)	**0.01** (1)	**0** (0)

Table 5. Comparative results of hypothesis testing procedures

All tests at 10 %. Results to nearest 2 decimal places.

Proportions of funds in bold (number of funds in brackets)

Table 6 shows the number of zero-alpha nulls we can reject using a standard t-test, the more conservative Benjamini Hochberg FDR and the more conservative still, Bonferroni FWER procedure. Additional funds are rejected at a decreasing rate as the significance level increases. The change from 1 to 10% significance sees a relatively large change in the number of rejections, 0 to 6 in the Benjamini-Hochberg FDR case, however 10-20% only generates 2 more rejections under FDR. This is supportive of the notion that in the presence of some abnormal funds, there exists a spike of p values in the significance region, and the remaining true-null funds should be normally distributed in the non-rejection region.

		Significance level				
		20%	**15%**	**10%**	**5%**	**1%**
Method	**Unadjusted t-test**	**21**	**19**	**18**	**10**	**6**
	Benjamini Hochberg FDR	**8**	**7**	**6**	**0**	**0**
	Bonferroni FWER	**3**	**2**	**1**	**0**	**0**
	BSW FDR	**13**	**13**	**14**	**8**	**6**

Table 6. Comparison of type one error control procedure results

6. **Comparison with Existing Literature**

This chapter discusses the empirical results of this paper in relation to other literature on fund performance. 8.1 compares with other studies separating luck from skill with fdr procedures. 8.2 compares with studies separating luck from skill with cross-sectional bootstraps, which accounts for the remainder of studies considering luck. 8.3 discusses other UK fund performance studies. 8.4 gives a brief overview of relevant global fund performance studies. 8.5 contains table 7, which details the key results and aspects of all studies in this chapter. For comparative ease, the results from this paper are included under Platonoff (2015).

6.1 Multiple hypothesis testing literature using false discovery rates

6.1.1 UK

The most comparable paper is Cuthbertson (2010a), which applies an FDR procedure to UK equity mutual fund data. Using net-of-fee returns from 1975-2002 the authors find that just 2% of funds display significant positive performance, that 78% of funds are zero alpha, and the remaining 20% are unskilled, at 10% significance. They find a 67% FDR in abnormal positive returns and 15.9% in negative returns, indicating that most positive performance is due to luck and not skill, whilst negative performance is due to lack of skill, in contrast to results from this paper.

The main difference in results is that I find more skilled, and fewer unskilled funds. The market, fund universe, factor model and FDR procedure are very similar, so I conclude with reasonable confidence that this difference is due to fees, as this paper uses gross-of-fee returns, whereas Cuthbertson (2010a) use net-of-fee returns. I identify the existence of managerial skill, and when we compare with Cuthbertson (2010a) we see these abnormal returns are not passed onto investors, but extracted via fee policy.

6.1.2 Other developed markets

BSW use US mutual fund manager returns from 1975-2006. The authors find that, before fees, 9.6% of managers are skilled, and 4.5% unskilled. However, net-of fees, just 0.6% return positive abnormal profits and 24% negative abnormal profits. This gives supporting evidence that before fees we can identify a minority of skilled managers, but these abnormal returns are almost entirely extracted by the fund and not passed onto investors.

BSW also uncover some notable time trends: net-of-fees, 2.4% of funds display positive abnormal returns when we treat each 5 year interval as a separate ‘fund’. This is supportive of managers exhibiting ‘hot hands’ over short periods. The authors find that the proportion of skilled managers has fallen from 14.4% in 1990 to 0.6% in 2006 . They find that skilled funds are concentrated in the extreme right tail of the cross-sectional alpha distribution, and therefore we can identify skilled funds by extremely low p values.

Kim (2014) examine the performance of fund managers in the Australian market between 1995 and 2009. They find that use of conditioning information improves fund performance. Using an unconditional four factor model, the authors find that 70.7% of funds are zero alpha funds, 27.5% are skilled and 1.8% are unskilled.

Using a Christopherson, Ferson, and Glassman (1998) conditional factor model, Kim (2014) find that 66.1% of funds are zero alpha, 33.7% are skilled, and just 0.3% are unskilled. They also find evidence of performance persistence in the unconditional but not in the conditional model, implying that successful techniques based upon smart utilisation of publicly available information are always available. However, the private information necessary to trade successfully in the conditional model is not always available.

Cuthbertson and Nitzsche (2013) investigate mutual fund performance, net-of-fees, in the German equity market from 1990 to 2009. The authors find that 0.5% of funds are skilled, 27% are unskilled, and 72.5% are zero alpha. Similar to findings in the UK and the US the authors find that the majority of positive discoveries are due to luck, whereas most poor performance is due to low skill (positive FDR of 80%, Negative FDR of 13%).

We observe similar results in UK, US, Australian and German markets (all developed markets), with minorities of skilled and unskilled managers, and 65-80% of true zero alpha null hypotheses. Gross of fees there is evidence that a non-zero minority of fund managers display stock picking abilities, however net-of-fee evidence suggests that in the vast majority of cases, almost none of this return is passed onto investors but is absorbed into managerial fees.

6.1.3 Developing markets

Suh and Hong (2011) examine Korean mutual fund returns between 2001 and 2009, and find that 60% of funds exhibit positive alphas, and 0% are unskilled, with the remaining 40% exhibiting zero-alphas, net-of-expenses. The Korean mutual fund industry is a developing market, and the authors correctly predict the market will show less market efficiency than the mature US mutual fund industry, and therefore exhibit arbitrage opportunities and significant positive alphas.

6.2 Multiple hypothesis testing literature using cross-sectional bootstraps

6.2.1 UK

Cuthbertson, Nitzsche and O'Sullivan (2008) investigate performance of UK unit trusts and OEICs between 1975 and 2002 using bootstrap methodology on the same dataset as Cuthbertson (2010a) to examine the tails of the alpha distribution. The authors find similar results to the 2010 paper using FDR. Net of fees, there is strong evidence of a proportion of around 5-10% of skilled funds (depending on factor model used). And that the bottom 40% of funds display significant negative alphas that can't be explained by bad luck alone. The authors reject the hypothesis that the best (worst) funds are simply lucky (unlucky). The authors find that the majority of funds with positive performance are lucky, whereas the majority of poorly performing funds are unskilled.

6.2.2 US

Kosowski (2006) find that, on average, US equity funds do not beat their four-factor benchmark. However the authors found, in contrast to BSW, that large sub-groups of funds exist in the tail extreme (top 10%), that is, with the most skilful managers, exhibiting significant alphas which are very unlikely to be explained by luck. These funds are notably concentrated in growth-oriented funds, whilst income-oriented funds exhibit underperformance.

Chen (2012) applies the bootstrap methodology to US enhanced-return index funds from 1996 to 2007 using conditional and unconditional Carhart four factor models. They find that positive alphas exist across the entirety of cross-sectional funds and that managers possess meaningful value-adding fund management skills that cannot be explained by luck or sampling variability. The authors find marginally higher alphas in unconditional models.

Chen (2012) find evidence that derivative-enhancement freedom creates larger positive alphas than stock-only enhancement strategies. This result could be explained by the higher level of investor freedom allowed in derivative-enhancement, and complements Eling and Faust (2010) who report superior performance in the alphas of hedge funds with higher degrees of investment freedom.

Agyei-Ampomah, Mason and Thomas (2015) control for luck in gross and net-of-fee US mutual fund returns from 1990-2011. Using style consistent benchmarks they find a sizable minority of positive and negative abnormal funds, and a majority of zero-alphas. The authors compare results to standard benchmarks, and stress the importance of carefully selected benchmarks in performance analysis to improve the precision of results.

Results using bootstrap procedures do not deliver clearly partitioned proportions of true positive negative and zero alpha funds within the fund universe like we see in FDR literature, so some results are not perfectly comparable to FDR results. Nonetheless, we see similar results of the existence of minorities of funds exhibiting true positive and negative abnormal performance in the highly developed US mutual fund market. Cuthbertson (2008) finds a larger proportion of unskilled funds in their bootstrap analysis, compared to using the FDR (2010a)

6.3 Other UK fund performance studies

Clark (2013) applies cross-sectional analysis to UK unit trust returns between 1980 and 2007 using a baseline Carhart four factor model, and using the 1 month FTSE All Share dividend yield as the risk free rate. Clark deals with the survivor bias problem by creating a survivor-bias-free UK unit trust database (not publicly available) using similar methodology to the CRSP database on US mutual fund data. Clark (2013) finds that in general there is very little significant evidence of abnormal performance, however only analyses the cross-section of alphas, not the tails of the distribution, so cannot make any conclusions about whether certain unit trusts consistently produce abnormal returns.

Fletcher, Ntozi-Obwale and Power (2008) examine whether UK unit trust fund managers exhibit security selection and market timing skills, using Fama's (1972) notion that performance can be broken down into stock selection and market timing (picking underpriced securities and predicting market movements),

Using a conditional four factor model including a market timing variable on data between 1988 and 2002, the authors found very little evidence that security selection skills of UK unit trust fund managers outperform the market and add any value to fund performance. However managers do demonstrate positive market timing ability. Growth and Income fund managers are the worst performing, with negative or zero selection skills conditional upon market conditions. Fletcher do not extend their analysis to controlling for false discoveries.

Blake and Timmermann (2002) analyse the cross section of all UK pension fund returns from 1991-1997 and find a negative (-0.7%) but insignificant alpha. The authors find no positive alphas (net-of-fees) significant at 5%, but 87% of funds displayed a negative alpha, and 15% of these were significant at 5%. The authors did not apply FDR analysis, but these results are similar to the equivalent UK cross-sectional regression results in Cuthbertson (2008, 2010a).

Table 7. Comparison of existing fund performance literature

Name	Period	Market	Data Source	Fees	Baseline	Luck control	Skilled	Zero Alpha	Unskilled
Platonoff 2015 (this paper)	1980-2014	UK 73 Fixed Income Unit Trusts	DataStream	Gross	4 Factor Unconditional	FDR	16.50%	80.80%	2.70%
Cuthbertson, Nitzsche and Sullivan 2010a	1975-2002	UK 675 equity unit trusts	Fenchurch	Net	3 Factor unconditional	FDR	2% (10% sig) (67% FDR)	78%	20% (15.9% FDR)
Barras, Scaillet and Wermers 2010	1975 - 2006	US 2076 equity funds	CRSP	Net	4 factor unconditional	FDR	0.6% (2006)	75.40%	24%
							14.4% (1990)	76.40%	9.20%
				Gross	"	"	9.6% (2006)	75.90%	4.50%
Kim, In, Inyeob and Park 2014	1995-2009	Australia 804 equity funds	Morningstar	Net	4 Factor unconditional	FDR	27.50%	70.70%	1.80%
					4 Factor Conditional		33.70%	66.10%	0.30%
Cuthbertson and Nitzsche 2013	1990-2009	Germany 550 equity funds	Bloomberg	Net	3 Factor Unconditional		0.50%	72.50%	27%
Suh and Hong 2011	2001-2009	South Korea 791 equity funds	Zeroin	Net	6 Factor unconditional	FDR	60%	40%	0%
Cuthbertson, Nitzsche and Sullivan 2008	1975-2002	UK 935 Equity unit trusts	Fenchurch	Net	Best fit of various 4 factor conditional and unconditional	Cross-sectional bootstrap	5-10%	50-55%	40%
Kosowski, Timmerman, White and Wermers 2006	1975-2002	US 2118 equity funds	CRSP	Net	4 Factor conditional and unconditional	Cross-sectional bootstrap	Significant evidence of outperformance and underperformance. 5-10% skilled enough to more than cover management fees		

Chen, Chu and Leung 2012	1996 - 2007	US 500 Enhanced index return funds	CRSP	Net	4 factor conditional and unconditional	Bootstrap	Significant evidence of outperformance		
Agyei-Ampomah, Mason and Thomas 2015	1990-2011	US 2384 Mutual funds	Morningstar	Gross & net	Style based Russel Indices in 4 factor model	Cross-sectional bootstrap	Significant evidence of outperformance and underperformance in a minority of funds		
Clark 2013	1980-2007	UK 973 equity Unit Trust	UK Equity Unit Trust/OEIC Survivor-Bias-Free Database	Gross	4 Factor unconditional		No evidence of significant outperformance of benchmarks		
Fletcher, Ntozi-Obwale and Power 2008	1988-2002	UK 432 equity and balanced unit trusts	FINSTAT, Money Management Periodicals,	Net	Conditional UK economic performance model		No evidence of outperformance in alphas. Negative timing ability.		
Blake and Timmerman 2002	1991-1997	UK 247 Pension international equity funds	WM Company, Edinburgh	Net	CAPM Single factor		0%	85%	15%
Quigley and Sinquefield 2000	1978-1997	UK 473 Equity Unit	Micropal	Gross	3 Factor Unconditional		No evidence of significant outperformance of benchmarks		
Fletcher 1995	1980-1989	UK 101 equity funds	DataStream	Net	Cross-product matrix of excess returns + own factor model		No evidence of significant outperformance of benchmarks		

7. Discussion

In Sections 7.1 I discuss unit trusts. In 7.2 and 7.3 I discuss the practical and theoretical implications of my results respectively.

7.1 Unit Trusts

Unit Trusts are a major component of the UK financial industry. In December 2013, UK unit trusts and OEICs managed approximately £770 billion in assets, an increase of 16% from the previous year. Over 2,000 trusts operate in the UK, accounting for approximately 15% of the £5 trillion total assets managed by IMA member funds. (2015 Investment Company Fact Book).

UTs were launched in the UK in 1931 by M&G to replicate US mutual funds, which had proved relatively robust to the 1929 Wall Street crash, due to their heavily diversified and open-ended nature, making them open to redemption on demand. (Prudential (2015), Fink (2008)). By 1939, there were around 100 UTs in the UK, and that has increased to well over 2000 operating today.

UTs are open-ended mutual funds, managed by asset management funds which pursue investment objectives specified in the trust deeds on behalf of the beneficial owners, and the trustees who act as custodians. UTs are legally established and subject to trust laws.

Approximately 80% of UK funds are actively managed (see figure. 5). Evaluating the performance of actively managed funds is important to investors, who pay for the stock picking services of fund managers. Considering the value of the UK Unit Trust Industry, the performance of active managers has real implications on a large number of individual and commercial investors, and on the economy as a whole.

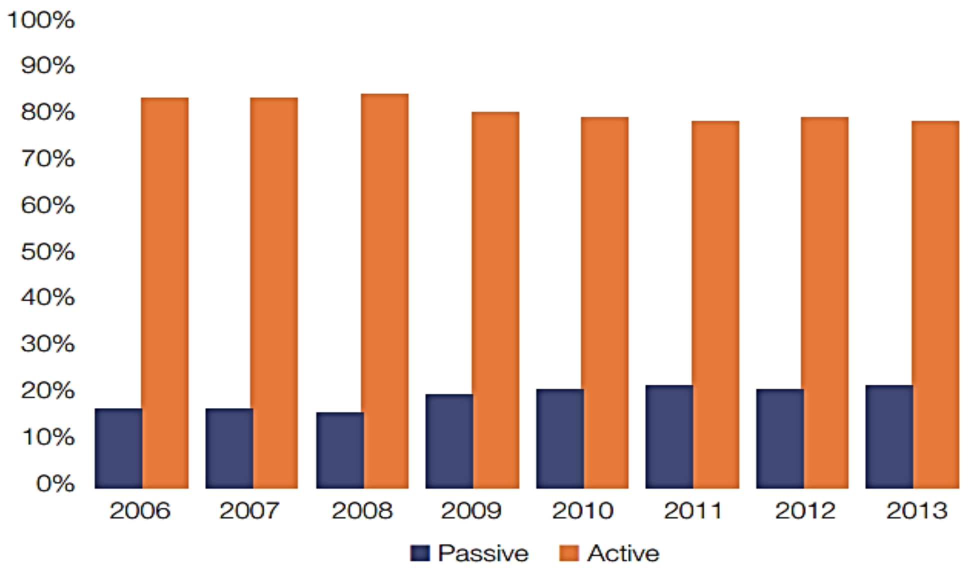


Figure. 5: Active and passive assets as a proportion of total UK assets under management (2006-2013)

(2015 Investment Company Factbook)

`

7.2 Practical implications of results

I have identified a minority of outperforming managers, and identified very low p values as a reliable indicator of a skilled manager. This has positive implications for the individual investor who wants to pick an active manager who outperforms a passive fund. However, 81% of managers simply match the market and furthermore, 2.8% reliably underperform. In the Bayesian interpretation of results as in 5.4, with a high cost of regret, investors may be disincentivised from investing in an industry with these proportions.

Despite these risk-adjusted performance results, managers may perform other functions. It is notable from literature including Jensen (1968) and Clark (2013) that managers do generally display strong risk control from skill in diversification, minimising idiosyncratic risk and performing a socially desirable function. High returns are one goal of investment management, but reducing risk, hence improving risk-return relationships is also paramount. A risk-averse investor will pay a premium for risk reduction services, and Rieman (2013) finds that the asset-weighted return volatility of actively managed funds is 14.14% annually, and 16.1% for passive funds, indicating that investors in active managers may receive slightly less return after fees, but are paying this price to reduce investment risk

7.3 Theoretical implications and explanations

The Berk and Green (2004) model of mutual fund flows and the efficient market hypothesis (EMH) both predict that fund managers will not outperform the market consistently. The Berk and Green (2004) model predicts that managers may outperform in the short term, but that fund inflows and decreasing returns to scale will lead to normal returns in the long run. The EMH can be interpreted in different strengths, but essentially tells us that financial markets are too informationally efficient for there to be any information not already incorporated into prices, and this prevents managers from obtaining an abnormal return by any means other than luck. I explain and discuss the EMH and Berk and Green equilibrium in more detail in 7.2.1 and 7.2.2 respectively.

This paper effectively tests both hypotheses by examining the performance of individual UK unit trusts (UTs), and investigating the proportions of funds that are truly skilled, unskilled, or produce no abnormal return, represented by abnormal returns against a risk adjusted market index.

My finding that there exist skilled managers who are able to reliably outperform the market over a long time period is evidence against both equilibrium theories in their strong form. However both equilibria could provide partial explanations for why the vast majority of managers struggle to beat the market despite presumed practice and expertise in fund management.

7.3.1 Efficient Market Hypothesis

In a fully efficient capital market, stock prices adjust instantaneously to new information and always reflect fundamental value (composed of present valued risk-return characteristics). Examining whether an investor can identify a portfolio that delivers returns above those justified by its fundamental risk-return characteristics is therefore also an examination of informational efficiency in security markets.

‘A market is efficient with respect to an information set Ω, if it is impossible to make economic profits by trading on the basis of information set Ω. By economic profits, we mean risk adjusted returns net of all costs’ Jensen (1978).

This definition means that within the constraints of the information set, there is complete absence of arbitrage. In this case it is not possible to make forecasts which yield risk adjusted profits. All information currently available is absorbed into prices, and therefore future movements in price are unpredictable and random.

The existence of alphas which are significant, positive, and due to skill, supports the existence of mutual fund ‘stars’, and contradicts the EMH.

7.3.2 Berk and Green Equilibrium (2004)

The competitive model of Berk and Green (2004) suggests that entry and exit of funds should ensure that in equilibrium there are neither funds with long-run positive nor negative abnormal performance

The Berk and Green (2004) equilibrium model explains lack of sustained abnormal fund performance with a mechanism of competition in fund flows. They argue that fund inflows are related to short term performance, and that returns are subject to decreasing returns to scale. Additionally, managerial changes are important to funds. Managers are aware that outperformance may be short term, and rationally cash in on their short term bargaining power by opportunely switching funds after strong performance, taking their skill with them. Conversely, underperformers are sacked.

Pollet and Wilson (2008) find supportive evidence of Berk and Green’s (2004) equilibrium by investigating the effect of size on mutual fund behaviour. Pollet and Wilson find that funds experience limits of scalability. When outperforming funds experience an inflow of investment, they increase investment into the same universe of stocks, rather than increasing the size of their investment universe. Increased fund size increases transaction costs and staff costs. Therefore funds experience decreasing returns to scale rather than the economies of scale exploited by production funds.

Fitting with the model of limits to talent (human capital) Pollet and Wilson find that in the year 2000, a doubling of fund size is only associated with an increase in number of stocks in the fund portfolio of 10%, and that the average large fund holds less than double the number of stocks of a fund less than one hundredth of its size. This proves the key finding in support of Berk and Green's (2004) equilibrium that liquidity costs cause decreasing returns to scale, and these are not counteracted by increasing number of stock holdings, as the managers' ownership shares increase at a decreasing rate in response to fund flows.

This model could provide a partial explanation of why the majority of funds fail to outperform the market in the long run in my study.

7 Conclusion

I use a multiple hypothesis testing framework to estimate the proportion of skilled, unskilled and zero-alpha UK fixed income unit trusts, using the false discovery rate approach. A standard t-testing approach identifies 24.7% of funds exhibiting significant abnormal performance, composed of 19.2% outperformance and 5.5% underperformance, against a Carhart (1997) 4 factor model, testing at 10% significance. However, I find an FDR of 13.9% for outperformers and 48.7% for underperformers. This indicates that some outperforming funds are lucky, and approximately half of funds underperforming against their benchmark are unlucky. Hence the proportion of funds exhibiting true skill is 16.5%, and only 2.8% of funds are truly unskilled.

These results cast doubt upon strong form efficient market hypothesis (EMH) as I conclude that, although the majority of funds neither outperform nor underperform their benchmark, there is a non-zero minority of funds that reliably 'beat the market'. This indicates the existence of informational inefficiencies: information or techniques regarding future market movements which are not factored into market prices and few managers are able to exploit.

These results use returns gross of management fees, and this appears to cause a contrast between the results of this paper and most other literature on fund management performance, which generally finds a high FDR for outperforming funds, and a low FDR for underperforming funds. This is likely to be because management fees reliably bring true zero-alpha funds to slightly below their benchmark. Taken with the results from other literature, looking at net-of-fee returns, I conclude that managers exhibit skill, but these abnormal returns are almost entirely extracted by fee policy.

I confirm that the Barras (2010) FDR controlling method reduces the number of null hypotheses rejected at all significance levels. I also confirm that the Barras (2010) FDR method is less stringent in rejections than the BH (1994) method, and that the Bonferroni FWER method is more stringent still. In a sample of 73 funds, at 10% significance I reject the zero-alpha null for 18 funds using a standard t-test, 14 using the Barras (2010) FDR method, 6 using the BH (1994) method, and just 1 using the Bonferroni FWER method.

Bibliography

2015 Investment Company Fact Book: A Review of Trends and Activities in the U.S. Investment Company Industry. (2015). 55th ed. Investment Company Institute.

Agyei-Ampomah, S., Clare, A., Mason, A. and Thomas, S. (2015). On luck versus skill when performance benchmarks are style-consistent. *Journal of Banking & Finance*, 59, pp.127-145.

Anderson, S. and Ahmed, P. (2005). *Mutual funds*. New York: Springer.

Barras, L., Scaillet, O. and Wermers, R. (2010). False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas. *The Journal of Finance*, 65(1), pp.179-216.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, Vol. 57(No. 1 (1995),), pp.pp. 289-300.

Berk, J. and Green, R. (2004). Mutual Fund Flows and Performance in Rational Markets. *Journal of Political Economy*, 112(6), pp.1269-1295.

Bessler, W., Blake, D., Lueckoff, P. and Tonks, I. (2010.). Why Does Mutual Fund Performance Not Persist? The Impact and Interaction of Fund Flows and Manager Changes. *SSRN Electronic Journal*.

Blake, D. and Timmermann, A. (1998). Mutual Fund Performance: Evidence from the UK. *Review of Finance*, 2(1), pp.57-77.

Blake, D. and Timmermann, A. (2002). Returns from active management in international equity markets: Evidence from a panel of UK pension funds. *J Asset Manag*, 6(1), pp.5-20.

Carhart, M. (1995). *Survivor Bias and Persistence in Mutual Fund Performance*. PhD. University of Chicago.

Carhart, M. (1997). On Persistence in Mutual Fund Performance. *The Journal of Finance*, 52(1), pp.57-82.

Carpenter, J. and Lynch, A. (1999). Survivorship bias and attrition effects in measures of performance persistence. *Journal of Financial Economics*, 54(3), pp.337-374.

Chen, A., Chu, Y. and Leung, M. (2012). The performance of enhanced-return index funds: evidence from bootstrap analysis. *Quantitative Finance*, 12(3), pp.383-395.

Christopherson, J., Ferson, W. and Glassman, D. (1998). Conditioning Manager Alphas on Economic Information: Another Look at the Persistence of Performance. *Review of Financial Studies*, 11(1), pp.111-142.

Clark, J.P. (2013). *Performance, Performance Persistence and Fund Flows: UK Equity Unit Trusts/Open-Ended Investment Companies vs. UK Equity Unit-Linked Personal Pension Funds.*. PhD. University of Exeter.

Coles, J., Daniel, N. and Nardari, F. (2006). Does the Choice of Model or Benchmark Affect Inference in Measuring Mutual Fund Performance?. *SSRN Electronic Journal*.

Cuthbertson, K., Nitzsche, D. and O'Sullivan, N. (2008). UK mutual fund performance: Skill or luck?.*Journal of Empirical Finance*, 15(4), pp.613-634.

Cuthbertson, K., Nitzsche, D. and O'Sullivan, N. (2010). False Discoveries in UK Mutual Fund Performance. *European Financial Management*, 18(3), pp.444-463.

Cuthbertson, K. and Nitzsche, D. (2013). Winners and losers: German equity mutual funds. *The European Journal of Finance*, 19(10), pp.951-963.

Eling, M. and Faust, R. (2010). The performance of hedge funds and mutual funds in emerging markets. *Journal of Banking & Finance*, 34(8), pp.1993-2009.

Fama, E. (1972). Components of Investment Performance*. *The Journal of Finance*, 27(3), pp.551-568.

Fama, E. and French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), pp.3-56.

Fama, E. and French, K. (2010). Luck versus Skill in the Cross-Section of Mutual Fund Returns. *The Journal of Finance*, 65(5), pp.1915-1947.

Ferson, W. and Schadt, R. (1996). Measuring Fund Strategy and Performance in Changing Economic Conditions. *The Journal of Finance*, 51(2), pp.425-461.

Fink, M. P. (2008) *The Rise of Mutual Funds: An Insider's View*. [Online]. Second Edition. Oxford University Press. [Accessed: 27 August 2015]. Available from: https://books.google.co.uk/books?id=6rk2NwWiAcwC

Fletcher, J. (1995). An Examination of the Selectivity and Market Timing performance of UK Unit Trusts. *Journal of Business Finance & Accounting*, 22(1), pp.143-156.

Fletcher, J. and Ntozi-Obwale, P. (2009). Exploring the Conditional Performance of U.K. Unit Trusts. *Journal of Financial Services Research*, 36(1), pp.21-44.

Gregory, A. Tharayan, R. And Christidis, A. (2013) 'Constructing and Testing Alternative Versions of the Fama–French and Carhart Models in the UK', Journal of Business Finance & Accounting, 40(1) & (2), 172–214, January/February 2013, 172-214.

Grinblatt, M. and Titman, S. (1994). A Study of Monthly Mutual Fund Returns and Performance Evaluation Techniques. *The Journal of Financial and Quantitative Analysis*, 29(3), p.419.

Horowitz, J. (2003). Bootstrap Methods for Markov Processes. *Econometrica*, 71(4), pp.1049-1082.

ICI (2015) *2015 Investment Company Fact Book: A Review of Trends and Activities in the U.S. Investment Company Industry* [Online]. 55th Edition. Available from: http://www.icifactbook.org/pdf/2015_factbook.pdf

Jensen, M. (1968). THE PERFORMANCE OF MUTUAL FUNDS IN THE PERIOD 1945-1964. *The Journal of Finance*, 23(2), pp.389-416.

Jensen, M. (1969). Risk, The Pricing of Capital Assets, and The Evaluation of Investment Portfolios. *The Journal of Business*, 42(2), p.167.

Jensen, M. (1978). Some anomalous evidence regarding market efficiency. *Journal of Financial Economics*, 6(2-3), pp.95-101.

Kim, S., In, F., Ji, P. and Park, R. (2014). False discoveries in the performance of Australian managed funds. *Pacific-Basin Finance Journal*, 26, pp.244-256.

Kosowski, Robert. (2011) 'Do Mutual Funds Perform When It Matters Most To Investors? US Mutual Fund Performance And Risk In Recessions And Expansions'. *Quarterly Journal of Finance* 01.03 (2011): 607-664. Web.

Kosowski, R., Timmermann, A., Wermers, R. and White, H. (2006). Can Mutual Fund "Stars" Really Pick Stocks? New Evidence from a Bootstrap Analysis. *J Finance*, 61(6), pp.2551-2595.

Mason, A., Agyei-Ampomah, S. and Skinner, F. (n.d.). Realism, Skill & Incentives: Current and Future Trends in Investment Management and Investment Performance. *SSRN Electronic Journal*.

Mustovoy, D. (2015). *Fund (mis)classification. Evidence based on style analysis. V2.0*. 1st ed. [ebook] London. Available at: http://www.northinfo.com/documents/641.pdf [Accessed 1 Aug. 2015].

Ntozi-Obwale, Patricia, Fletcher, Jonathan and Power, David (2008) Conditional performance in different states of the economy: evidence from U.K. unit trusts. Journal of Financial Transformation, 24. pp. 153-159

Henriksson, R. and Merton, R. (1981). On Market Timing and Investment Performance. II. Statistical Procedures for Evaluating Forecasting Skills. *The Journal of Business*, 54(4), p.513.

Pástor, Ľ. and Stambaugh, R. (2002). Mutual fund performance and seemingly unrelated assets. *Journal of Financial Economics*, 63(3), pp.315-349.

Pollet, J. and Wilson, M. (2008). How Does Size Affect Mutual Fund Behavior?. *The Journal of Finance*, 63(6), pp.2941-2969.

Prudential, (2015). *Business histories*. [online] Available at: http://www.prudential.co.uk/who-we-are/our-history/business-histories/m-and-g [Accessed 22 July. 2015].

Quigley, G. and Sinquefield, R. (2000). Performance of UK equity unit trusts. *J Asset Manag*, 1(1), pp.72-92.

Rieman, M. (2013). *Study: Only 24% of Active Mutual Fund Managers Outperform the Market Index - NerdWallet*. [online] NerdWallet Investing. Available at: http://www.nerdwallet.com/blog/investing/2013/active-mutual-fund-managers-beat-market-index/ [Accessed 22 Jul. 2015].

Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3), pp.479-498.

Suh, S. and Hong, K. (2011). Control of Luck in Measuring Investment Fund Performance*. *Asia-Pacific Journal of Financial Studies*, 40(3), pp.467-493.

Tonks, I. (2005). Performance Persistence of Pension-Fund Managers*. *The Journal of Business*, 78(5), pp.1917-1942.

Treynor, J., and Mazuy, F. 1966. Can mutual funds outguess the market? Harvard Business Review 44
(July-August): 131-36.

Tuzov, N. and Viens, F. (2010). Mutual fund performance: false discoveries, bias, and power. *Annals of Finance*, 7(2), pp.137-169.

Wermers, R. (1999). Mutual Fund Herding and the Impact on Stock Prices. *The Journal of Finance*, 54(2), pp.581-622.

Wermers, R. (2000). Mutual Fund Performance: An Empirical Decomposition into Stock-Picking Talent, Style, Transactions Costs, and Expenses. *J Finance*, 55(4), pp.1655-1703.

Yekutieli, D. and Benjamini, Y. (2001). The control of the false discovery rate in multiple testing Under Dependency. *Ann. Statist.*, 29(4), pp.1165-1188.

Appendix

A1 Alternative model distribution of p values

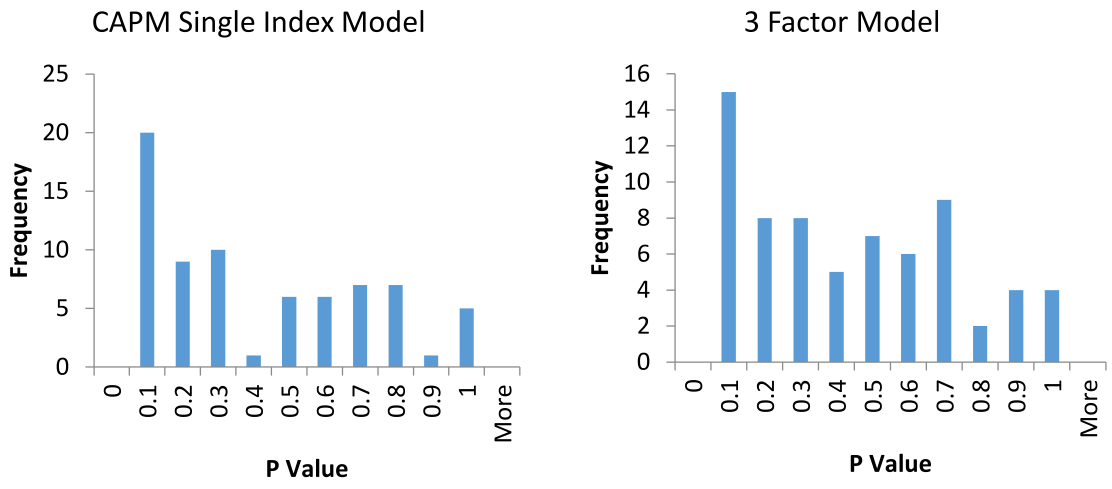


Figure A1. Alternative model distribution of p values

A2. Estimating λ^*

I estimate the true value of λ^* by taking the value of λ for which $\hat{\pi}_0(\lambda)$ gives the smallest sum of squared errors.

$$\lambda^* = \lambda \in (0,1) which\ minimises\ : \int_0^1 (\lambda - \hat{\pi}_0(\lambda))^2$$

(A1)

That is, the horizontal line with the smallest summed distance from the curve in figure. A3. In figure. A2 and table A1 I experiment with different values of λ, in intervals of 0.1, or 10% significance and obtain the result:

$$\lambda^* = 0.74\ (2\ d.p)$$

(A2)

Table A1 and figure A3 compares the sums of squared errors (SSE) for different values of λ. I find the λ_i for which the summed error is minimised is 0.74 (2 d.p). Plugging λ^* into (15) gives us:

$$\hat{\pi}_0(0.74) = 0.534$$

(A3)

λ	$\hat{\pi}_0(\lambda)$	$\int_0^1 (\lambda - \hat{\pi}_0(\lambda))^2$
0.1	0.7716	3.72632
0.2	0.81597	2.6731
0.3	0.8135	1.7998
0.4	0.8333	1.1065
0.5	0.7778	0.5933
0.6	0.7986	0.2600
0.7	0.5556	0.1068
0.8	0.5556	0.1335
0.9	0.6944	0.3403

Table A1. Estimating $\hat{\pi}_0$

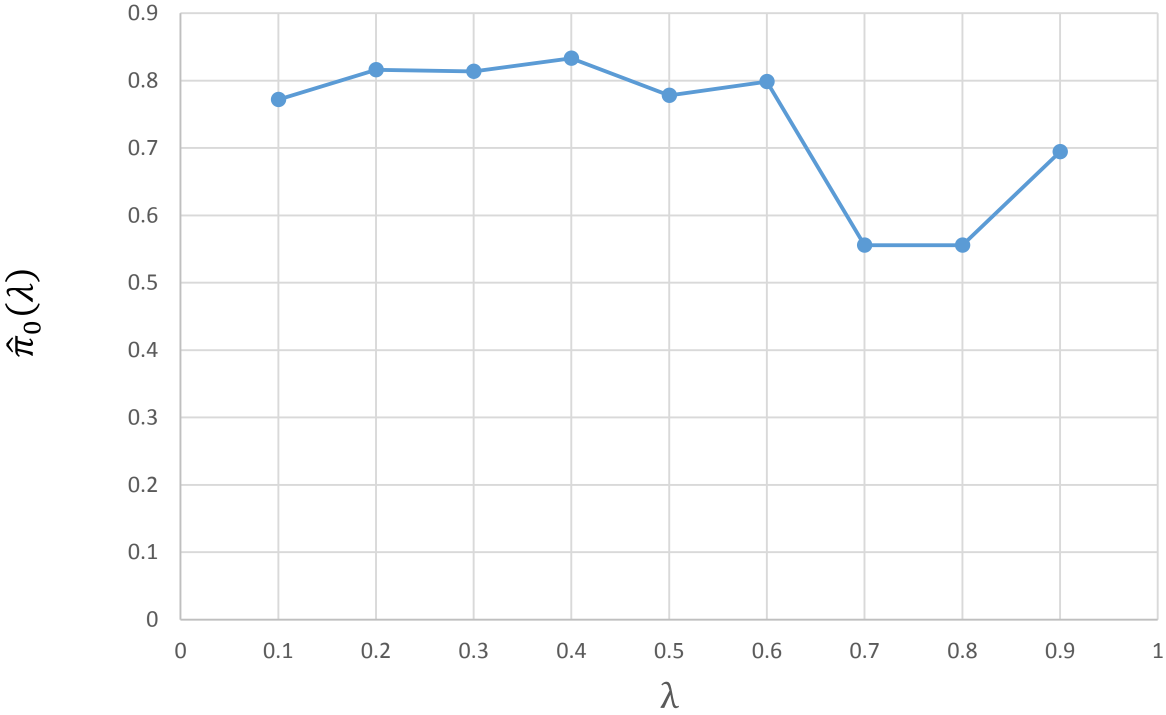


Figure A2. Estimating $\hat{\pi}_0$

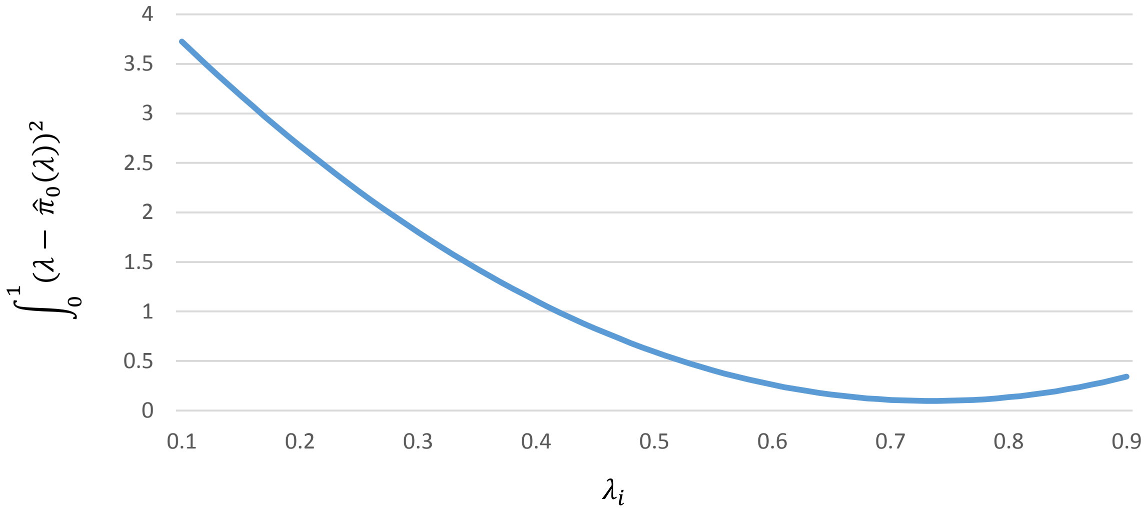


Figure A3. Sum of squared errors of $\hat{\pi}_0(\lambda)$

A3 BSW: calculating population proportions

Using (6) and (7), we can estimate the proportion of false discoveries $\hat{F}_\gamma^+$, and of skilled ($\hat{T}_\gamma^+$) and unskilled ($\hat{T}_\gamma^+$) funds, as 2.67%, 16.51% and 2.81% respectively (at 10% significance):

$$\hat{F}_\gamma^+ = \hat{F}_\gamma^- = \frac{\widehat{\pi_0}.\gamma}{2}$$

$$\frac{0.534\ (0.1)}{2} = 0.0267$$

(A4)

The proportion of truly skilled funds:

$$\hat{T}_\gamma^+ = S_\gamma^+ - \hat{F}_\gamma^+$$

$$\hat{T}_{0.1}^+ = 0.1918 - 0.0267 = 0.1651$$

(A5)

And the proportion of truly unskilled funds:

$$\hat{T}_{0.1}^{+} = \; 0.0548 - 0.0267 = 0.0281$$

(A6)

Giving us estimated population proportions

$$\hat{\pi}_{0.1}^{+} = \; 16.51\%$$

$$\hat{\pi}_{0.1}^{-} = \; 2.81\%$$

$$\hat{\pi}_{0} = \; 80.68\%$$

(A7)

A4. False discovery rate results

A4.1 False positive discovery rate

$$\widehat{FDR}_{\gamma}^{+} = \frac{\hat{F}_{\gamma}^{+}}{S_{\alpha}^{+}} = \hat{\pi}_{0}.\frac{\gamma}{2.S_{\gamma}^{+}}$$

$$= 0.534.\frac{0.1}{(2)0.1918} = 0.1392$$

(A8)

A4.2 False negative discovery rate

$$\widehat{FDR}_{\gamma}^{-} = \frac{\hat{F}_{\gamma}^{-}}{S_{\gamma}^{-}} = \hat{\pi}_{0}.\frac{\gamma}{2.S_{\gamma}^{+}}$$

$$= 0.534.\frac{0.1}{(2)0.0548} = 0.4872$$

(A9)