How Skilful are US Fixed-Income Fund Managers?

By

Andrew Clare,

Keith Cuthbertson,

Dirk Nitzsche

Faculty of Finance,
City, University of London, London. EC1Y 8TZ.

and

Niall O'Sullivan

Cork University Business School, University College Cork, Cork, Ireland

14 December 2020

Abstract
We develop a performance evaluation model that incorporates the factors proposed by Huij and Derwall (2008) and a fund-specific benchmark to analyse the performance of US fixed income funds. Using the full sample, and accounting for the possibility of false discoveries, we find that fund management companies extract most of any abnormal performance produced by their fund managers. Our sub-sample analysis indicates that after the Global Financial Crisis (GFC) there was a substantial increase in the number of bond funds with: both positive gross-of-fee alpha and positive net-of-fee alpha performance; and also a reduction in funds with negative-alpha performance. However, because the GFC was such a unique event, it would still be difficult to conclude that these managers offer value for money for investors compared to passive alternatives.

JEL: G11; G120

Keywords: Mutual Fund Bond Performance; False Discovery Rates

---

1. Introduction

The performance of equity mutual funds has been investigated extensively over the last few decades. Many papers document equity mutual fund alphas where excess returns are conditioned on both formal models such as the CAPM (Sharpe (1964)) and also on factor models that derive their validity from more empirical considerations, most notably the Fama-French three factor model (Fama and French 1992, 1993), its momentum-based enhancement (Cahart 1997) and its more recent incarnation, the Fama-French five factor model (Fama and French 2015). Other authors investigating equity mutual funds have focussed on: persistence in mutual fund performance (Carhart 1997); the impact of manager characteristics such as experience (see for example Porter and Trifts (2014)), gender (see for example Bliss and Porter (2002)) and education (see for example Gottesman and Morey (2006)); and the impact of factors such as the location of the fund manager (see for example Otten and Bams (2007)) and the “family status” of a fund (see for example Kempf and Ruenzi (2008)). Researchers have investigated many other aspects of the performance of equity mutual funds using both US and non-US data as they have sought to establish the value that the fund management industry provides.

By contrast, fewer papers have focussed on the performance of fixed income (bond) mutual funds. There may be a number of reasons for this. First, arguably, the focus of attention on equity fund returns reflects the higher, historic equity allocations in investor portfolios, particularly in the 1980 and 1990s during the “*cult of equity*”. By contrast bond investment was seen to be less relevant and perhaps less interesting. Second, there has been a general lack of agreement in the academic literature about the appropriate factor model to use for fixed-income funds, although neither the CAPM nor the APT are limited to explaining equity returns. However, in the lead up to, and in the wake of the Global Financial Crisis (GFC), financial innovation made more fixed-income asset classes available to investors. Some of these asset classes are complex, as pre-crisis investors in bank subordinated debt tranches will attest when they experienced equity-style losses. Further, the sums of money invested in fixed-income mutual funds is quite substantial, partly because of increased returns after the fall in global interest rates which has only just begun to reverse, but also because some investors have lost faith in equity markets after two major bear markets in the space of less than ten years. Although equity markets recovered in both cases, for anyone drawing an income from a predominantly equity-focused investment portfolio over these periods, portfolio values would have suffered permanent impairment as the perils of sequence risk became all too apparent. More conventional bond markets (that is, excluding investments in, for example, CDOs and subordinated bank debt) have given investors a smoother investment experience, which has in turn led to higher investor allocations into bond mutual funds. Of the $46trn invested in regulated open-end funds, equity mutual funds comprise 43% of this total while bonds and money market funds comprise 35% (2019 Investment Company Factbook, page 11[1]).

This paper contributes to the literature in a number of ways. First, we provide a comprehensive assessment of the performance of over 1,000 US bond mutual funds over the period from January 1998 to May 2018. This is an important segment of the mutual fund universe which has received much less attention in the academic literature in the past than equity mutual funds. Our performance analysis pays particular attention to the possible impact of the GFC on bond mutual fund performance, a dimension of bond mutual fund performance that, to our knowledge, has not been addressed in the literature before. Second, while the Fama and French three-factor and Carhart four-factor models have been long-established as being appropriate for the determination of equity mutual fund performance, there has been less attention on the

[1] https://www.ici.org/research/stats/factbook

most appropriate factor model for bond mutual funds. We investigate the performance of alternative factor models of mutual fund bond returns. Our preferred model, which combines factors initially proposed by Huij and Derwall (2008) with a fund-specific benchmark, performs particularly well. We believe that this model should form the basis of future performance evaluation studies of bond mutual funds. The third significant contribution of our paper is to apply the false discovery rate (FDR) methodology to the analysis of bond mutual fund performance. The FDR methodology corrects for estimated alpha-performance that may in part be due to luck when analysing performance across many funds (Barras et al. 2010). Accounting for false discoveries in overall fund performance may give different inferences from the standard approach – which normally involves a count of the number of funds that have statistically significant, non-zero abnormal performance (negative or positive), but takes no account of possible false positives.

Our key results are as follows. A four-factor model comprising a fund's benchmark index, a broad market bond index, a high-yield bond index and an index of mortgage-backed bonds captures around 80% of the variation in sectoral bond-fund returns. Over the whole sample period, after accounting for false discoveries, we find that around 25% of funds have positive gross-of-fee alphas but only around 5% of funds have positive net-of-fee alphas (at a 2.5% significance level). This result implies that fund management companies extract most of any abnormal performance produced by their fund managers. Turning to negative performance, very few bond-funds (3%) have truly negative gross-of-fee alphas but this rises to 30% of funds when using net-of-fee alphas, so again, an after-fees analysis leads to a substantial number of underperformers.

The performance of the bond funds, before and after the Global Financial Crisis (GFC), are clearly of interest given the Quantitative Easing (QE) programmes instigated by many of the world's most important Central Banks. After the GFC we document a substantial increase in the number of bond funds with both positive gross-of-fee alpha and positive net-of-fee alpha performance and also a reduction in funds with negative-alpha performance. However, because the GFC was such a unique event, it would still be difficult to conclude that these managers offer value for money for investors compared to passive alternatives.

The remainder of this paper is organized as follows. In section 2 we briefly discuss the related literature; in section 3 we present our models, methodology and data; while our empirical results and conclusions are in sections 4 and 5, respectively.

2. Related Literature

Most previous studies of net-of-fee mutual bond fund returns conclude that the funds do not generate positive alpha. Cornell and Green (1991) study the performance of US "low grade" bond funds defined as those bond funds that have at least two thirds of their holdings in bonds rated BAA or lower by Moody's, or BBB or lower by Standard and Poor's. Their aim was to use this data to investigate the claim of Drexel Burnham Lambert that the risk in holding "junk" bonds was more than offset by higher, risk-adjusted returns. However, to do so they estimate a model of bond mutual fund returns using contemporaneous and lagged values of both the level of the US T-Bill rate and the return on the S&P500. Although the focus of their paper was not alpha, close inspection of their results show that the low-grade bond funds did produce positive alphas. However, it would clearly be possible to argue that the model specification did not capture enough of the sources of risk and return that are typically faced by bond fund managers, such as the slope of the yield curve and by a variety of different credit premia.

Blake et al. (1993) estimate both single and multi-index models of bond mutual fund returns using factors that arguably better capture the sources of performance available to bond fund managers. They liken their single index model to the "market model" typically used to evaluate the performance of equity funds, but also imply that the single index essentially represents risk and return that could be achieved by a passive bond fund, that is, by simply holding the components of the index, in their index weights. The single index is the Lehman Brothers US government/corporate bond index. This factor (or index) is very broad indeed and captures, by definition, the broad trends in the market. However, it may produce misleading results for a manager that, for example, only invests in US Treasuries, or only in corporate bonds. In recognition of this, Blake et al. (1993) augment the single index model with indices that capture the opportunities to add value that arise from bond maturity and credit spreads between categories. They add the Lehman Brothers intermediate and long-term corporate indices to capture the maturity opportunities and the Lehman Brothers mortgaged backed securities index and the Blume/Keim high yield index to capture credit spread opportunities. They find that estimated net-of-fee alphas are indistinguishable from zero. Indeed, underperformance is found to equal the fees charged by the funds, indicating that bond gross-alpha fund performance does not exceed that of the fund benchmark

Blake et al. (1995) develop a model based upon the Arbitrage Pricing Theory (APT). In keeping with the requirements of APT, their model employs forecasts (prepared by economists and investment professionals) to measure unexpected changes in the fundamental economic influences that affect returns. The model comprises factors that capture three broad sources of risk and return. First, the excess return on the US stock market and the excess return on a broad index of bonds capture broader financial market risks. Second, Blake et al. (1995) include a measure of default risk, a measure of term risk and a measure of mortgage credit risk. This set of factors capture the risk-return opportunities available to bond fund managers. Finally, two factors that appear in Chen, Roll and Ross's (1986) APT model of equity returns are included in the model. These are unexpected changes in inflation and unexpected changes in economic growth. Using this equilibrium model approach Blake et al. (1995) find negative and statistically significant, net-of-fee alphas in all categories of bond funds examined.

Detzler (1999) examines the performance of a small set of *global* fixed-income mutual funds, and employs a wide set of factors. As well as estimating single index models as Blake et al. (1993) do, Detzler estimates three multi-index models incorporating foreign country bond indices and concludes that the sample of global bond funds did not outperform a broad-based US bond index. This suggests that expenses might have outweighed any diversification benefits that might have accrued to the fund over the sample period. A less generous interpretation of the results is that the managers detracted value with their exchange rate-related positions.

Ayadi and Kryzanowski (2011) use bootstrapping techniques to investigate the performance of Canadian mutual bond funds and find, amongst other things, that the performance of the best performing funds is due to "good luck" rather than to skill, and that "bad luck" could explain the performance of the worst performing funds (gross of fees).

A number of authors have investigated the persistence of bond mutual fund performance. Evidence of persistence is weak in earlier sample periods: Philpot et al. (1998) and Philpot et al. (2000) find short term persistence (over one year) for high yield, global and convertible funds based on contingency table tests, using Sharpe ratios. However, this finding is driven by funds ranked in the middle and lower end of the cross-sectional distribution of Sharpe ratios. Furthermore, the authors find no evidence of persistence over longer, five-year periods. A more recent study by Huij and Derwall (2008) however, does provide evidence of positive performance persistence in US bond mutual funds.

Huij and Derwall (2008) estimate alphas for a large sample of US bond mutual funds. Their multi-factor model comprises the excess returns on a broad investment grade corporate bond index, the excess return on an index of High Yield bonds, and the excess return on an index of Mortgage-backed securities. They argue that estimated alphas correctly measure abnormal returns that are not due passive fund management. Recursive portfolio formation tests indicate that decile net-alphas are generally monotonically decreasing from top to bottom deciles. While the difference between top and bottom decile alphas is a significant 3% p.a., the estimated alphas are generally negative. This indicates that while there is performance persistence, any *positive* persistence is not sufficient to produce returns that would reward investors for choosing an active bond fund manager.

In a recent paper, Clare et al. (2019) estimate alphas on a large sample of US bond mutual funds using a single index model and a parsimonious Blake et al. (1993) model. Clare et al.'s model comprises a broad index of bond returns, a measure of the US Treasury yield spread (i.e. US ten-year Treasury yield minus one-year US Treasury yield) and a measure of credit conditions (i.e. Baa-rated corporate bond yield minus the Aaa-rated bond yield). However, rather than only calculating alphas where the dependent variable is the return on the bond fund minus a proxy for the risk-free rate, the authors also estimate alphas where the dependent variable is the return on bond fund-j minus the return generated by bond-j's self-declared benchmark. Calculating excess returns in this way acknowledges that in practice fund manager performance is judged by both employers and by investors against self-declared benchmarks. The authors find evidence in support of long-run abnormal performance (alpha) in the top 10% of funds, but they do not find evidence that positive performance persists.

From our reading and knowledge of the existing literature on the performance of bond mutual funds, we believe that there are a number of areas relating to the performance of bond mutual funds that deserve further research attention. First, there is no broad agreement about the most appropriate factor model to use in assessing the performance of this sector of the mutual fund industry, in section 3 we explore the use of alternative factor models and develop a model that we believe describes the factor risks in these funds well. Second, the financial market turmoil associated with the GFC was arguably centred around fixed income instruments (mortgage-backed securities, etc), it is therefore worth considering the performance of fixed income fund managers over this period. In section 4.2 we consider the pre- and post-crisis performance of this section of the mutual fund industry, analysis of which has not been undertaken before to our knowledge. Third, while previous studies of bond fund performance have estimated fund alphas, such estimates have not been adjusted to take account of the possibility of false discoveries. In section 4.3 of this paper we address this gap in the existing bond performance literature by applying the False Discovery Rate methodology.

3. Factor Models, Methodology and Data

3.1 Factor Models

Clearly, an important element in performance measurement is the factor model used. We assess individual US bond fund performance for both gross-of-fees and net-of-fees returns, using the publicly-stated benchmark of the fund. However, since to add value there may be a 'tilt' away from these benchmarks (Sensoy 2009) we augment this approach with other relevant factors. More specifically, we use four factor models to calculate alphas. We then implement the FDR methodology to investigate the proportion of US bond mutual funds that produce performance that is driven by skill, as opposed to being the result of luck – good or bad.

We estimate bond fund alphas using two single-factor models; the three-factor Huij and Derwall (2008) model; and an enhanced version of the Huij and Derwall three-factor model. The first single-factor model ("market model") uses a broad index of US bond returns, incorporating both corporate bonds and Treasuries, analogous to the use of a broad index of equities when calculating alphas for equity mutual funds:

$$[1] \qquad (R_i - r_f)_t = \alpha + \beta_1 (R_m - r_f)_t + \varepsilon_{it}$$

where R_i is the monthly return on the mutual fund i for month t and r_{ft} is the risk-free rate (i.e. yield on US 30-day T-Bills); $(R_m - r_f)$ is the excess monthly return on a broad bond market index.

The second, single-factor model uses each fund's self-declared benchmark, as reported by Morningstar. Performance evaluation based on a benchmark-adjusted alpha has two key advantages. First, fund managers face a number of constraints in terms of the exposures that they can assume within the fund, for example, constraints relating to maturity categories and to credit ratings categories. These constraints are normally reflected in the fund benchmark. These benchmarks allow the fund manager to take appropriate risks and gives investors a way of evaluating fund performance. In addition, these benchmarks will also determine fund manager remuneration. Using a common, catch-all benchmark as in expression (1) could therefore result in very misleading conclusions about a fund manager's performance if it does not accurately reflect the constraints under which the manager is expected to operate (see Clarke et al. (2002), Kothari and Warner (2001) and Angelidis et al. (2013)). Second, for the case of equity benchmarks, Cremers et al. (2012) show that the benchmarks themselves can produce a non-zero alpha when compared against a broader index like the S&P 500 Composite index. If a manager "benchmark hugs" a benchmark with a non-zero alpha then this could lead to the conclusion that the manager has produced alpha when compared to the catch-all, broader index. It would not necessarily be correct to conclude, however, that the manager is skilful. Therefore, when we use a fund's self-declared benchmark-adjusted return the calculated alphas are unbiased in this regard. Few studies have focussed on benchmark-adjusted returns for bond funds. Given these issues our second model is specified as follows:

$$[2] \qquad \left(R_i - r_f\right)_t = \alpha + \beta_2\left(R_{bi} - r_f\right)_t + \varepsilon_{it}$$

where $(R_{bi} - r_f)$ is the excess monthly return on the designated-benchmark for bond fund-i.

Our third model, which we refer to as the 3FHD model, uses the Huij and Derwall (2008) three-factors, as shown in expression (3)

$$[3] \ \left(R_i - r_f\right)_t = \alpha + \beta_1\left(R_m - r_f\right)_t + \beta_2\left(R_{HY} - r_f\right)_t + \beta_3\left(R_{Mort} - r_f\right)_t + \varepsilon_{it}$$

where R_{HY} represents the return on an index of High Yield bonds; and a broad bond market index; a high yield bond index; and R_{Mort} represents the return on an index of mortgage-backed securities. For more information about these factors, see Table 2.

However, given the discussion above about fund-specific benchmarks we augment the 3FHD model with each fund's self-declared benchmark as a fourth factor. This 4-factor model we denote as "4FHD":

$$[4] \ (R_i - r_f)_t = \alpha + \beta_1(R_m - r_f)_t + \beta_2(R_{bi} - r_f)_t + \beta_3(R_{HY} - r_f)_t + \beta_4(R_{Mort} - r_f)_t + \varepsilon_{it}$$

3.2 Methodology: False Discovery Rate (FDR)

The impact of 'luck' in multiple hypothesis tests arises whenever we ask the question: '*How many of our statistically significant results are likely to be 'truly null'?* – that is 'false discoveries'. In this paper we estimate the false discovery rate, which measures the proportion of merely lucky funds amongst a group of funds, whose performance has been found to be statistically significant.

If we simply count the number of funds which are found to have a statistically significant performance measure, we run the risk of including funds which are truly null (i.e. Type I errors). For example, suppose the FDR amongst 20 statistically significant positive-alpha funds is 80%, then this implies that only 4 funds (out of the 20) have truly significant alphas - this is clearly useful information for investors. A key issue is whether this correction gives different inferences from the standard approach of simply counting the number of significant funds with non-zero abnormal performance.

The standard approach to determining whether the alpha of a single fund demonstrates skill or luck is to choose a rejection region and associated significance level γ and to reject the null of 'no outperformance' if the test statistic lies in the rejection region - 'luck' is interpreted as the significance level chosen. However, using γ = 5% when testing the alphas for each of M-funds, the probability of finding at least one non-zero alpha-fund in sample of M-funds is much higher than 5% (even if all funds have true alphas of zero)[2]. Put another way, if we find 20 out of 200 funds (i.e. 10% of funds) with significant positive estimated alphas when using a 5% significance level then some of these will merely be lucky.

In testing the performance of many funds a balanced approach is needed - one which is not too conservative but allows a reasonable chance of identifying those funds with truly differential performance. The false discovery rate attempts to strike this balance by classifying funds as 'significant' (at a chosen significance level) and then asks the question: *What proportion of these significant funds are false discoveries?* That is, those that are truly null (Benjamini and Hochberg 1995; Storey 2002; Storey, Taylor, and Siegmund 2004). The FDR measures the proportion of lucky funds among a group of funds which have been found to have significant (individual) alphas and hence measures luck among the pool of significant funds. Storey (2002) and Barras, Scaillet, and Wermers (2010) provide a detailed account of the FDR methodology, so it is only briefly summarized below. The null hypothesis that fund-i has no skill in security selection (alpha) and the alternative of either positive or negative performance is:

$$H_0 : \alpha_i = 0 \qquad\qquad H_A : \alpha_i > 0 \text{ or } \alpha_i < 0$$

A 'significant' fund is one for which the p-value for the test statistic (e.g. t-statistic on alpha) is less than or equal to some threshold $\gamma / 2$ $(0 < \gamma \leq 1)$. At a given significance level γ the probability that a zero-alpha fund exhibits 'good luck' is $\gamma / 2$. If the proportion of truly zero-

[2] This probability is the compound type-I error. For example, if the M tests are independent then Pr(at least 1 false discovery) = 1 – $(1-\gamma)^M$ = z_M , which for a relatively small number of M=50 funds and conventional γ = 0.05 gives z_M = 0.92 – a high probability of observing at least one false discovery.

alpha funds in the population of M-funds is π_0 then the expected proportion of false positives (or 'lucky' funds) is:

$$[5] \qquad E(F_\gamma^+) = \pi_0 \, (\gamma / 2)$$

If $E(S_\gamma^+)$ is the expected proportion of significant positive-alpha funds, then the expected proportion of truly skilled funds (at a significance level γ) is:

$$[6] \qquad E(T_\gamma^+) = E(S_\gamma^+) - E(F_\gamma^+) = E(S_\gamma^+) - \pi_0(\gamma / 2)$$

Varying γ allows us to see if the number of truly skilful funds rises appreciably with γ or not, which tells us whether skilled funds are concentrated or dispersed in the right tail of the cross-sectional distribution. The expected FDR amongst the statistically significant positive-alpha funds is:

$$[7] \qquad FDR_\gamma^+ = \frac{E(F_\gamma^+)}{E(S_\gamma^+)} = \frac{\pi_0(\gamma / 2)}{E(S_\gamma^+)}$$

The observed number of significant funds S_γ^+ provides an estimate of $E(S_\gamma^+)$. To provide an estimate of π_0, (i.e. the proportion of truly null funds in the population of M-funds), we use the result that truly alternative features have p-values clustered around zero, whereas truly null p-values are uniformly distributed, U(0,1). To estimate $\hat{\pi}_0(\lambda)$ we can simply choose a value λ for which the histogram of p-values becomes flat and use:

$$[8] \qquad \hat{\pi}_0(\lambda) = = \frac{W(\lambda)}{M(1-\lambda)} = \frac{\#\{p_i > \lambda\}}{M(1-\lambda)}$$

where $W(\lambda)/M$ is the area of the histogram to the right of the chosen value of λ (on the x-axis of the histogram) – see Figure 1. For example, suppose π_0 = 100% and we choose λ = 0.2. Then $W(\lambda)/M$ = 80% of p-values lie to the right of λ = 0.2 and our estimate of π_0 = 80%/ (1-0.2) = 100% as expected. For truly alternative funds (i.e. $\alpha_i \neq 0$), the histogram of p-values has a 'spike' near zero. But if the histogram of p-values is perfectly flat to the right of λ then our estimate of π_0 is independent of the choice of λ. So, if we were able to count only truly null p-values then [8] would give an unbiased estimate of π_0. However, if we erroneously include a few alternative p-values then [8] provides a conservative estimate of π_0 and hence of the FDR.

The bias in the estimate of $\hat{\pi}_0(\lambda)$ is decreasing in λ (as the chances of including non-zero alpha-funds diminishes) but its variance increases with λ (as we include fewer p-values in our estimation). Hence an alternative estimate of π_0 is to choose λ to minimize the mean-square error $E\{\pi_0(\lambda) - \pi_0\}^2$ (Storey 2002, Barras et al. 2010)[3] .

We use a bootstrap approach to calculate p-values of estimated t-statistics because of the non-normality in regression residuals (Politis and Romano 1994; Kosowski et al. 2006). The estimated factor model of returns is: $r_{i,t} = \hat{\alpha}_i + \hat{\beta}_i' X_t + \hat{e}_{i,t}$ for i = 1, 2, … M funds, where T_i

[3] Barras et al. (2010) use a Monte Carlo study to show that the estimators outlined above are accurate, are not sensitive either to the method used to estimate π_0 or to the chosen significance level γ. The estimators are also robust to the typical cross-sectional dependence in fund residuals (which tend to be low in monthly data).

= number of observations on fund-i, $r_{i,t}$ = excess return on fund-i, X_t = vector of risk factors, $\hat{e}_{i,t}$ are the residuals and $\hat{t}_i$ is the (Newey-West) t-statistic for alpha. We draw a random sample (with replacement) of length T_i from the residuals $\hat{e}_{i,t}$ and use these bootstrap residuals $\tilde{e}_{i,t}$ to generate an excess return series $\tilde{r}_{i,t} = 0 + \hat{\beta}_i' X_t + \tilde{e}_{i,t}$ under the null hypothesis $\alpha_i = 0$. Using $\tilde{r}_{i,t}$ the performance model is estimated and the resulting t-statistic for the alpha-performance measure, t_i^b is obtained. This is repeated B = 10,000 times and for a two-sided, equal-tailed test the bootstrap p-value for the alpha of fund-i is:

$$[9] \qquad p_i = 2.\min[B^{-1}\sum_{b=1}^{B} I(t_i^b > \hat{t}_i), \quad B^{-1}\sum_{b=1}^{B} I(t_i^b < \hat{t}_i)]$$

where $I(.)$ is a (1,0) indicator variable. The above 'basic bootstrap' uses residual-only resampling, under the null of no outperformance (Efron and Tibshirani 1993). A similar procedure is used for other hypothesis tests.

3.4 Data Description

We estimate our four competing factor models and implement the FDR methodology using monthly fund (total) return data (gross and net of fees), on 1,254 fixed income mutual funds from Morningstar, over the period from January 1998 to May 2018. The data set includes both surviving and non-surviving funds. To be consistent both across our sample and with other performance evaluation studies that use Morningstar data, we collect this monthly data on only the oldest share class of each fund, thereby ensuring that each fund in our database is unique. All of the funds in our sample are categorised by Morningstar as being part of the: Broad, Corporate, Government, Government/Corporate, Municipal, Securitized, Inflation-Protected and Government/Inflation categories. The Broad category is the largest group represented in our database, comprising 621 funds in total. The next three largest categories are Government/Corporate Funds, Government Funds and Corporate Funds comprising 241, 195 and 71 funds in the database, respectively. Overall, these four categories make up almost 90% of the funds in our sample. One of the issues that we explore in this paper is the appropriateness of the stated fund benchmark in capturing the returns generated by the funds. If the fund benchmarks are indeed meaningful representations of the portfolios constructed by the managers, then we should probably use these benchmarks as a means of identifying skilful from lucky fund managers. Each fund’s designated benchmark was also collected from Morningstar.

Table 1A presents descriptive statistics for our sample of bond funds. The statistics in Panel A are based on both total gross and net monthly fund returns in excess of the risk-free rate. The excess gross returns indicate that, compared to the return on T-bills, all fund groups have, on average, outperformed cash over the sample period. The corporate bond category produces the highest (excess), average monthly returns over this period of 0.36% per month, although the Govt/Corp sector produced the highest monthly Sharpe ratio of 0.28. The net-of-fee excess returns presented on the right-hand side of Panel A demonstrate the impact of fees. The difference between the gross and net-of-fee average monthly returns for all funds is 0.06% per month (about 0.72% p.a.). The equivalent figure for the category with the highest fees, the Corporate category, is 0.69% pa.

Panel B of Table 1A presents statistics analogous to those presented in Panel A, but the excess return is now defined as the fund’s return in excess of its benchmark. As would be expected, the average gross excess fund returns over the benchmark (Panel B) are lower than gross fund returns over the T-Bill rate (see Panel A). The average gross returns over the benchmark for all categories of bond funds is approximately 1 basis point per month. It is highest for the Govt/Corp category at 2.87 basis points per month, and is negative for the Corporate sector at -3.5 basis points per month. On average, all categories of bond fund produce net-of-fee benchmark-adjusted monthly returns that are negative. For all funds this is -5.06 basis points per month (see the right hand side of Panel B). Of course the benchmarks do not incorporate fees, but if equivalent passive vehicles (ETFs, or tracker funds) are available to investors at less than 5.06 bps per month, then on average investors would be better off with passive investment vehicles[4]. However, Table 1A presents average performance figures. One of our goals in this paper is to identify whether there are some funds that produce excess performance that is due to skill, after correcting for non-normality and for any false positives.

Another of our aims is to examine the pre- and post-GFC performance of the bond funds in our sample. To this end Tables 1B and 1C present net and gross performance statistics for the periods January 1998 to August 2008 and from April 2009 to May 2018 respectively. The most striking difference between the two periods is seen in the average return statistics for all of the funds. For example, in the pre-crisis period the net-of-fee performance of all funds in excess of their benchmark is around -7.5 basis points per month. However, the equivalent figure in the post-crisis period is just over 2 basis points per month. With the exception of the corporate bond category, all other categories of fund produce a positive net of fee, net of benchmark performance, with the Govt/Corp category providing the highest average of almost 4.7 basis points per month. It would seem that investors might have been better off, on average, with passive bond funds before the crisis, but with active bond funds after the crisis. This

[4]At the time of writing we find that both iShares and Vanguard ETFs, based upon the Barclays Capital US Aggregate Bond Index, are available to investors at an *annual* holding fee of 5 basis points.

performance difference, pre- and post-crisis is an issue that we explore using the factor models in the results section of this paper.

Finally, in Table 2 we present summary statistics of the four factors used. For each of the three Huij and Derwall factors we present the average yield difference between the factor and the risk-free rate (that is the unconditional premium represented by the factor) and the standard deviation of this premium, both for the full sample and for the two sub-samples presented in Table 1. While average, unconditional premiums on the first two factors look fairly similar across the sub-samples, the unconditional premium on the high yield factor looks very different over the pre- and post-crisis periods. The average high-yield unconditional premium is 0.16% prior to the crisis, but 0.92% after it. We include a fourth factor in Table 2, the TED spread, which is often used as a proxy for liquidity risk. Since the GFC was initially a liquidity crisis, it is included here as a cross-check for the impact on bond performance of liquidity (or illiquidity)[5]. Table 2 shows that the average value of the TED spread, which is defined here as the monthly percentage change in the difference between three-month USD- LIBOR and the yield on a three-month US T-Bills. The TED spread, falls marginally from 3 basis points per month in the pre-crisis period to 1 basis point in the post-crisis period.

4. Results

4.1 Factor Model Results

In Table 3 we present summary results generated by gross (before deduction of fees) and net (after fee) alphas, using the four different factors in equations (1) to (4), over the full sample period, from January 1998 to May 2018. Column 2 in Panels A and B of Table 3 show the average R-Squared of the four models for gross and net-of-fee returns. This shows that it is far

[5] We are grateful to an anonymous referee for suggesting this additional line of enquiry.

more appropriate to use the Benchmark Model (equation (2)) than to use the Market model (equation (1)). The former produces average R-Squared of almost 74%, while the latter produces a much lower average R-Squared of just over 54%. This result confirms the view of Clarke et al. (2002), Kothari and Warner (2001) and Angelidis et al. (2013) which is that a fund's benchmark is a far more appropriate way of capturing the risk and return profile of a fund than a more generic, market index. The highest average R-Squared is achieved with the 4FHD. The addition of the benchmark enhances the average R-Squared from 79% to 84%. These R-squared values are lower than those typically found when applying factor models to equity mutual fund returns. For example, Agyei-Ampomah et al (2015) examine the performance of US mutual funds, using "style-specific" benchmarks and find average R-squared of between 93% and 99%.

In columns 3, 4 and 5 in Panels A and B of Table 3 we present summary results relating to gross and net alphas respectively. Average gross-alphas for all four models are positive. The market model (where the factor is the excess return on the Barclays Capital US Aggregate Bond Index) produces an average gross-alpha of nearly 9 basis points per month, with an average R-Squared of around 54%, while the benchmark model produces an average gross-alpha of just over 3 basis points per month with an average R-Squared of 73%. The difference in the two sets of results should not be surprising. The manager is (or should be) focussed on their benchmark, since this is how their investors will evaluate their performance and how the fund manager's management team will also evaluate their performance. The estimated gross alpha falls further when we add more factors to the model. For example, the average alpha generated by the 4FHD model is just 1.5 basis points per month. Looking at columns 3 and 4 we see that when we apply the 4FHD model, almost 70% of the funds produce a positive alpha, of which just over 28% are found to be statistically different from zero using conventional inference. Just over 4% of the sample produce an alpha that is negative and significantly different from zero.

When we consider the net-of-fee alphas in Panel B of Table 2, the estimated average alphas all decline; now only the market model produces an average positive alpha of just under 3 basis points per month. The benchmark model tells a different story. The average net-alpha is almost *minus* 3 basis points per month. Given that the benchmark model represents the practical hurdle that the managers set themselves, this result paints the asset management industry in a less positive light. The augmented four-factor Huij-Derwall model (4FHD) shows that using only the benchmark model does not capture all of the sources of risk faced by investors in these bond funds. The average R-Squared of the 4FHD model is just over 84%, compared with around 74% for the one-factor, designated-benchmark model. For the 4FHD model the net-alpha is about -4.6 bps per month. Only 26.6% of the alphas estimated using the 4FHD model are now found to be positive and only 5.3% of the total are found to be positive and statistically different from zero, while 29.4% are found to produce a statistically significant negative alpha after fees. These results are qualitatively similar to those of Agyei-Ampomah et al (2015) who, when applying the one-factor CAPM to their sample of US equity mutual funds, find net alphas to average 3 basis points per month, a value that falls to between minus 4 and 5 basis points when using the four-factor Carhart model.

Table 3 also reports the number of funds in the sample that produce positive and significant alphas, as well as those that produce negative and significant alphas. The gross-of-fee results (Panel A) show that there are a noteworthy number of funds that do produce positive and significant alphas (at a 2.5% significance level). For example, for the 4FHD model we find that 353 (28.15%) of the 1,254 funds produce positive and statistically significant alphas.

However, this number falls to 67 (5.3%) for positive net-alphas. At the other end of the scale, the 4FHD model shows that 52 (4.15%) of funds have negative and statistically significant gross-alphas (Panel A) but a much larger number, namely 369 (29.4%) funds, have negative net-of-fees alphas (Panel B).

Panel C in Table 3 presents the average values for the factor loadings for each of the four models. In the interests of parsimony, we present only the average loadings for the gross-of-fee returns because the loadings are almost identical (to 3 decimal places) when net of fee returns are used. The beta on the market factor in the single factor model averages 0.7, compared with the beta on the benchmark factor in the benchmark model, that averages 1. This result demonstrates how much more appropriate this factor is for assessing risk and return than the wider, catch-all market factor. The benchmark factor dominates the market factor in the 4FHD model with the former having an average value of 0.87 and the latter having an average value of 0.01. Focussing on the 4FHD model, we see that the average exposure to the High Yield factor is positive while the average exposure to the Mortgage factor is negative.

In Table 4 we examine the gross and net-of-fee alpha-performance for the four bond fund sectors, as categorised by Morningstar, using the 4FHD model[6]. The Gov./Corp sector has the highest monthly gross alpha of 2 bps per month. For the Gov./Corp bond funds, 78 (32%) have positive significant gross-alphas and only 5 (2%) have statistically significant negative gross-alphas. Only 12 funds (5%) have statistically significant positive net-alphas, while 67 funds (28%) have statistically significant negative net-alphas. The largest category is the "Broad"

[6] We focus on the results from the 4FHD model as it produces the highest average R-squared. Results generated by the other models are available on request.

category which comprises 621 funds. We find that 158 (25.4%) of these bond funds produce positive and significant gross-alphas.

In Panel B of Table 4 we see that average net-of-fee alphas are negative for all four sectors. For example, the average net alpha is just under minus 5bps per month for Broad bond funds, compared to an average of just over 1bp per month gross-of-fees. The number of Broad category bond funds producing positive gross-of-fees alphas falls from 418 to 183 (29%) net of fees; while the number producing significantly positive alphas falls from 158 gross of fees to 32 (5%) net of fees. The number of Broad category funds that produce negative alphas rises from 203 gross of fees to 438 (70%) net of fees, of which 109 (17%) produce statistically significant, negative alphas. This pattern is broadly repeated in the other three sectors.

Turning to the average factor loadings in Panel C of Table of Table 4, we see again that the average factor loading on the broad market index is highest for the "Gov" category at 0.34 and, perhaps surprisingly, for the Broad category is negative at -0.15. For the Broad category the benchmark index provides an almost unitary average coefficient of 1.06. Unsurprisingly the average exposure of the Corporate sector to the High Yield factor is highest at 0.23 and lowest for the Government segment of the bond fund universe at 0.03. On average the Broad and Gov./Corp categories have a negative exposure to the Mortgage factor while the "Gov" and in particular the Corporate segment have positive exposure to this factor.

In Tables 3 and 4 the final column reports the proportion of funds that are found to have non-normally distributed residuals using the Bera-Jarque test which is consistently around 75% of funds across models and bond categories. This motivates our use of bootstrap p-values to determine statistical significance for alphas and factor model betas. We report results for the basic bootstrap procedure described in Section 3.3. Alternative bootstrapping procedures such as simultaneously bootstrapping the residuals and the independent variables or allowing for serial correlation (block bootstrap) or contemporaneous bootstrap across all (existing) funds at time t, produced qualitatively similar results. Hence, we only report results for the 'residuals only' bootstrap, although the results from the other bootstrap procedures are available on request.

4.2 Sub-sample analysis

In Table 1 we saw a difference in the unconditional performance of the funds in our sample, before and after the GFC, to the extent that investors might have been better off, on average, with passive bond funds before the crisis, but with active bond funds after the crisis. We now examine this result in more detail by estimating the 4FHD model with the inclusion of a step dummy variable to test for any significant performance difference and factor exposures between the two periods – Jan 1998 to Aug 2008, and April 2009 to May 2018. We estimate the following model:

$$\begin{aligned}[10] \qquad (R_i - r_f)_t = {} & \alpha + \beta_1 (R_m - r_f)_t + \beta_2 (R_{bi} - r_f)_t + \beta_3 (R_{HY} - r_f)_t + \\ & \beta_4 (R_{Mort} - r_f)_t + \alpha^* + \beta_1^* (R_m - r_f)_t + \beta_2^* (R_{bi} - r_f)_t + \\ & \beta_3^* (R_{HY} - r_f)_t + \beta_3^* (R_{Mort} - r_f)_t + \varepsilon_{it} \end{aligned}$$

Where the * superscript indicates that the regressor has been multiplied by a 0,1 dummy that takes the value of 0 between Jan 1998 to Aug 2008 and 1 for the period April 2009 to May 2018. The results of this regression are presented in Table 5, where we present the average coefficient values and proportion of funds for which these coefficients are statistically significant among the 748 fixed income mutual funds.

As indicated in Table 5, the average alpha is higher in the post-GFC period as shown by the positive average value on α^*. However, α^* is only statistically significant and positive for 16% of the funds, while it is statistically significant and negative for almost 10% of the funds. The dummy on the self-declared benchmark is also found to be positive, implying that there was, on average, a higher e exposure of funds to their own benchmark, in the post-GFC period. For nearly 20% of the funds this increase in exposure is statistically significant. Interestingly, for each of the other four factors the dummies indicate a decline, on average, to these risk factors. In terms of scale, the biggest decline is the exposure to the broad market factor by almost 0.17, which is statistically significant for just over 20% of the sample. A decline in exposure to the mortgage risk factor is statistically significant for around 7% of funds and an increase for around 11% of funds in the post-GFC period. Finally, in the post-GFC period, exposure to the High Yield factor increases for over 40% of the funds.

To summarise our results, it appears that post the GFC, funds tended to increase their exposure to their own benchmarks, and a sizeable proportion of the funds might have added value to their portfolios by increasing their exposure to the High Yield risk factor, whose premium was higher in the post GFC period (the High Yield factor premium in the post-GFC period was 0.92% compared to 0.16% before it).

4.3 Illiquidity considerations

One of the features of the Global Financial Crisis, an event that occurred in the middle of our sample, was the impact that it had on financial market liquidity. To test whether the 4FHD captures this risk adequately we include a proxy for liquidity risk namely, the TED spread (i.e. the difference between three-month Dollar LIBOR and the yield on a three-month US T-Bills). It represents the additional yield that investors demand for lending to commercial banks compared to lending to the US Treasury. Table 2 shows that the average TED spread pre- and post-GFC was very similar. However, at the peak of the liquidity crisis that engulfed the world's banking sector it increased dramatically. It is possible that a variable which reflects illiquidity during this crisis period might improve the factor model or change our conclusions[7].

To test this possibility Table 6 presents results using the benchmark model and the 4FHD model, each with and without the TED spread (Panels A and B respectively). The mutual fund returns are net-of-fees. For the benchmark model in Panel A, the addition of the TED spread has no noticeable impact on average fund exposure to its self-declared benchmark or on the number of funds that have a positive or negative significant exposure to it. The figures in parentheses show the proportion of coefficients that were found to be statistically significant and negative and statistically significant and positive (-ve/+ve). We find that, on average, the coefficient on the TED spread is negative and about 19% of funds have a statistically significant negative exposure to this proxy for liquidity risk. More notably, the addition of the TED spread leads to an average positive alpha, with around 19% that are statistically significant compared to only 9% when the TED spread is excluded from the factor model.

Similar results can be seen in Panel B of Table 6. The addition of the TED spread has only a marginal effect on the betas of the other four factors, but the average alpha is again found to be positive, compared with a negative average value without the liquidity proxy. However, the

[7] We are grateful to an anonymous referee for suggesting this additional line of enquiry.

proportion of positive significant alphas increases from 6% without the TED spread to 13.9% with its inclusion.

4.3 False Discovery Rates

In the analysis discussed above we have counted the number of statistically significant positive and negative alpha funds using bootstrapped p-values – but some of these outcomes could be false discoveries. The FDR approach adjusts for the *proportion* of false discoveries amongst those funds which are found to be statistically significant, based on individual (bootstrapped) t-alpha statistics.

Table 7 shows the results of applying the FDR methodology to the 4FHD model on all 1,254 funds over the whole sample period January 1998 to May 2018 (for various significance levels). The histogram of p-values (using the minimum mean square error criterion) determines the optimal λ which then gives the proportion of null (i.e. zero) gross-alpha funds, π_0 = 62.1% (Panel A) and a similar proportion of null net-alpha funds, π_0 = 53.9% (Panel B). Column S^+ presents the percentage (number) of statistically significant funds (at γ = 1%, 2.5%, 5% and 10% significance levels). For example, (at a significance level of 2.5%) the percentage of significant positive-alpha funds is S^+ = 26.4% (331 funds). The estimate of the false discovery rate $FDR^+ = \pi_0 (\gamma/2) / S^+$ = 2.94% is small, which implies that the proportion of the M=1,254 "significant positive alpha funds" that are false discoveries is very small at F^+ = 0.78%. Hence, the proportion of "truly" significant gross-alpha funds, $T^+ = S^+ - F^+$ = 25.6% (i.e. 321 funds out of 1,254), is only slightly diminished after adjustment by the FDR.

Applying the above analysis to the net-of-fee alphas, shown in Panel B of Table 7, gives the proportion of null funds π_0 = 53.9% and a count of statistically significant positive net-alphas of $S^+ = 5.3\%$ (67 funds) – at a 2.5% significance level. The $FDR^+ = 12.6\%$ which provides a moderate downward adjustment resulting in the proportion of truly significant positive net-alphas of $T^+ = S^+ - F^+ \approx 5.34\% - 0.67\% = 4.67\%$ (i.e. 59 funds). Hence, correcting for false positives using the FDR does not change our earlier qualitative results with regard to the number of positive gross and net-alpha funds: $T^+ = 25.6\%$ of fund managers show positive gross-alpha performance but in most cases this positive performance is not passed on to investors, as only $T^+ = 4.67\%$ of funds have positive net-alphas.

Applying the FDR to negative gross-alpha funds (Table 7, Panel A) gives a count of $S^- = 2.95\%$ (37 funds) with the $FDR^- = 26\%$, resulting in the proportion of truly significant negative gross-alpha funds $T^- = 2.17\%$ (i.e. 27 funds) - at a 2.5% level of significance. The proportion $S^- = 29.4\%$ (369 funds) of negative net-alpha funds (Table 4, Panel B) is substantial and as the $FDR^- = 2.3\%$ is very small, this results in a substantial proportion of truly significant negative net-alpha funds of $T^- = 28.8\%$ (361 funds). Overall therefore, even after correction for false discoveries, the vast majority of bond fund managers do not have negative gross-alphas, but negative net-alpha performance is much more prevalent at around 29% of funds.

4.4 Performance Before and After the 2008 Financial Crisis

When we look at the two sub-periods, before and after the 2008 financial crisis, the results for positive alpha-performance funds are very different. For the pre-financial crisis period (January 1998 – August 2008, Table 8 (Panel A)[8], the positive gross-alpha performance shows $T^+ = 12.36\%$ (122 funds with) truly significant alphas. In contrast, the post financial crisis period (Table 8, Panel A) reveals a much larger proportion $T^+=36.5\%$ (345) funds with truly significant positive gross-alphas. Looking at positive net-alpha funds (Table 8, panel B), these

[8] We report results in the main text using a 2.5% significance level.

also show an increase between the two periods. In the pre-crisis period, there are only T^+ = 1.45% of funds with truly significant positive, net alphas, but this increases to 10.12% in the post-crisis period.

Turning now to negative gross-alpha performance in the pre- and post-crisis periods, we find T^- = 1.99% and T^- = 0% respectively (Table 8, Panel A), so there are very few funds which underperform their risk factors, on a gross-return basis, in either period. Negative fund performance on a net-alpha basis (Table 8, Panel B) is relatively high in both periods, but it does fall quite dramatically from T^- = 40.8% to T^- = 17% between the two periods. This is mainly due to a substantial fall in the proportion of statistically significant negative net-alpha funds, S^- from 41.2% to 17.8% while the FDR remains small and fairly constant in both periods (Table 8, Panel B).

Overall, even after adjustment for false discoveries (using either gross or net returns), the number of truly significant positive-alpha bond funds (T^+) increased and the number of negative-alpha funds (T^-) fell, in the post-financial crisis period. This may be due to managers of bond funds positioning their portfolios, relative to their benchmarks for the falls in "official" interest rates immediately after the crisis and also for the repeated use of quantitative easing which both reinforced downward pressure on rates.

5. Conclusions

In this paper we evaluate the performance of US bond funds, based upon an enhanced Huij-Derwall (2008) by including each fund's, self-designated benchmark. Arguably, the fund-specific benchmark should be an important component in the evaluation of fund performance because it helps determine what index the manager is trying to outperform, how their remuneration is determined and how they are judged by both their employers and investors. Using this enhanced model and for the purposes of comparison, a single factor model and the original Huij and Derwall model, we estimate the alphas generated by a set of 1,254 US mutual bond funds over the period from January 1998 to May 2018. However, as residuals from factor models may be non-normal, we employ a bootstrap procedure from which we can draw more reliable inferences. Using bootstrapped p-values, we then further refine our results by accounting for the possibility of "false discoveries", namely that some funds may have statistically significant alphas, due to luck rather than skill.

After estimating robust p-values and accounting for any false discoveries, we find that around 25% of funds have positive gross-of-fee alphas but only around 5% of funds have positive net-of-fee alphas (at a 2.5% significance level). Also, we find that very few funds (3%) have truly negative gross-of-fee alphas, but that 30% of funds have negative net-of-fee alphas. These results should be of concern to the fund industry, to regulators and to investors. It implies that fund management companies extract most of any abnormal performance produced by the fund. So in answer to the question – *how skilful are US fixed-income fund managers?* – the answer is that there is evidence of skill, but that the industry itself extracts most of the rewards of this skill.

We also examined the pre- and post-GFC periods. Here the results are more encouraging. In the post-GFC period we find a substantial increase in the number of bond funds with both positive gross-of-fee alphas and positive net-of-fee alphas and also a reduction in the number of funds with negative-alpha performance (even after accounting for false discoveries). This result indicates that fund managers may have positioned their portfolios, relative to their benchmarks, for the falls in "official" interest rates immediately after the crisis and also for the repeated use of quantitative easing which both reinforced downward pressure on rates. However, the GFC was a very unique event, and we would not therefore be able to conclude that active US mutual bond fund industry can provide value for money for its investors in more normal periods.

References

Agyei-Ampomah, S., A. Clare, A. Mason, and S. Thomas, 2015. On luck versus skill when performance benchmarks are style-consistent, *Journal of Banking and Finance*, Vol. 59, pp. 1-55.

Andrikogiannopoulou, A. and Papakonstantinou, F., 2019. Reassessing false discoveries in mutual fund performance: skill, luck or lack of power? *Journal of Finance*, Vol. 74, No. 5, pp. 2667-87.

Angelidis, T., Giamouridis, D., Tessaromatis, N., 2013. Revisiting mutual fund performance evaluation. *Journal of Banking and Finance*, Vol. 37, No. 5, pp. 1759–1776.

Ayadi, M.A., and Kryzanowski, L., 2011. Fixed-income fund performance: Role of luck and ability in tail membership, *Journal of Empirical Finance*, Vol. 18, pp. 379–392

Barras, L., Scaillet, O., Wermers, R., 2010. False discoveries in mutual fund performance: Measuring luck in estimated alphas. *Journal of Finance*, Vol. 65, No. 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 Royal Statistical Society, Series B*, Vol. 57, No. 1, pp. 289-300.

Blake, C., Elton, E., Gruber, M., 1993. The performance of bond mutual funds. *The Journal of Business*, Vol. 66, No. 3, pp. 371–403.

Blake, C., Elton, E., Gruber, M., 1995. Fundamental economic variables, expected returns, and bond fund performance. *Journal of Finance*, Vol. 50, No. 4, pp. 1229–1256.

Bliss, R. and Potter, M., 2002. Mutual Fund Managers: Does Gender Matter? *The Journal of Business and Economic Studies*, Vol. 8, pp. 1-17.

Busse, J.A., Goyal, A. and Wahal, S., 2010. Performance and Persistence in Institutional Investment Management, *Journal of Finance*, Vol. 65, No. 2.

Carhart, M.M, On persistence in mutual fund performance, 1997. *Journal of Finance*, Volume 52, No. 1, pp. 57-82

Chen, N., R. Roll, and S. Ross, 1986. Economic forces and the stock market, *Journal of Business*, Vol. 59, pp. 383-404.

Clare, A., O'Sullivan N, Sherman M, Zhu S, 2019. The Performance of US mutual bond funds, *International Review of Financial Analysis*, Vol. 61, pp. 1-8.

Clarke, R., De Silva, H., and Thorley, S., 2002. Portfolio constraints and the fundamental law of active management. *Financial Analysts Journal*, Vol. 58, No. 5, pp. 48-66.

Cornell, B. and K. Green, 1991. The Investment Performance of Low-Grade Bond Funds, *Journal of Finance*, Vol. 46, No. 1, pp. 29-48

Cremers, M., Petajisto, A., Zitzewtz, E., 2012. Should benchmark indices have alpha? Revisiting performance evaluation. *Critical Finance Review*, Vol. 2, pp. 1–48.

Detzler, M. 1999. The performance of global bond mutual Funds, *Journal of Banking and Finance*, Vol. 23, pp. 1195-1217

Efron, B. and Tibshirani R.J., 1993. An Introduction to the bootstrap, Chapman and Hall, New York

Fama, E. F., French K, R., 1992. The cross-section of expected stock returns. *Journal of Finance*, Vol. 47, No. 2, pp. 427-465.

Fama, E. F., French K, R., 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, Vol. 33, No. 1, pp. 3-56.

Fama, E. F., French K, R., 2010. Luck versus skill in the cross-section of alpha estimates. *Journal of Finance*, Vol. 65, No. 5, pp. 1915–1947.

Fama, Eugene F., and Kenneth R. French, 2015. A Five-Factor Asset Pricing Model. *Journal of Financial Economics*, Vol. 116, No. 1, pp. 1–22.

Gottesman, A. and Morey M., 2006. Manager education and mutual fund performance, *Journal of Empirical Finance*, Vol. 13, No. 2, pp. 145-182

Huij, J.J, and Derwall, J., 2008. Hot Hands in Bond Funds. *Journal of Banking and Finance*, Vol. 32, No. 4, pp. 559–572.

Kempf, A. and Ruenzi, S., 2008. Tournaments in mutual fund families. *The Review of Financial Studies*, Vol. 21, No. 2, pp. 1013–1036.

Kosowski, R., Timmermann, A., Wermers, R., White, H., 2006. Can mutual fund 'stars' really pick stocks? New evidence from a bootstrap analysis. *Journal of Finance*, Vol. 61, No. 6, pp. 2551-2595.

Kothari, S. P., Warner, J. B., 2001. Evaluating mutual fund performance. *Journal of Finance*, Vol. 56, No. 5, pp. 1985-2010.

Otten, R. and Bams, D., 2007. The Performance of Local versus Foreign Mutual Fund Managers, *European Financial Management*, Vol. 13, No 4, pp. 702-720.

Philpot, J., D. Hearth, J. Rimbey, 2000. Performance persistence and management skill in nonconventional bond mutual funds, *Financial Services Review*, Vol. 9, pp. 247-258

Philpot, J., Hearth, D., Rimbey, J., and Schulman, C., 1998. Active management, fund size and bond mutual fund returns. *The Financial Review*, Vol. 33, pp. 115-126.

Politis, D.N. and J.P. Romano, 1994. The Stationary Bootstrap, *Journal of the American Statistical Association*, Vol. 89, pp. 1303-1313.

Porter, G.E. and Trifts, J.W., 2014. The Career Paths of Mutual Fund Managers: The Role of Merit, *Financial Analysts Journal*, Vol. 70, No. 4.

Sensory, B., 2009. Performance evaluation and self-designated benchmark indexes in the mutual fund industry, *Journal of Financial Economics*, Vol. 92, No. 1, pp. 25-39

Sharpe, W.F., 1964. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk, *Journal of Finance*, Vol. 19, No. 3, pp. 425-442.

Storey J. D., 2002. A Direct Approach to False Discovery Rates, *Journal of Royal Statistical Society, Series B*, Vol. 64, pp. 497-498.

Storey, J.D., Taylor, J.E and Siegmund, D. 2004. Strong control, conservative point estimates and simultaneous conservative consistency of false discovery rates : A unified approach, *Journal of Royal Statistical Society, Series B*, Vol. 66, No. 1, pp. 187-205

Figure 1: Distribution of p-values on alphas from bond funds

The histogram shows the distribution of (bootstrap) p-values under $H_0 : \alpha_i = 0$. The p-values of truly null funds are distributed uniformly, U(0,1). "Significant funds" (that is, skilled or unskilled funds) have p-values near zero. The proportion of truly null funds π_0 is estimated by the area under the horizonal line, shown in the figure.

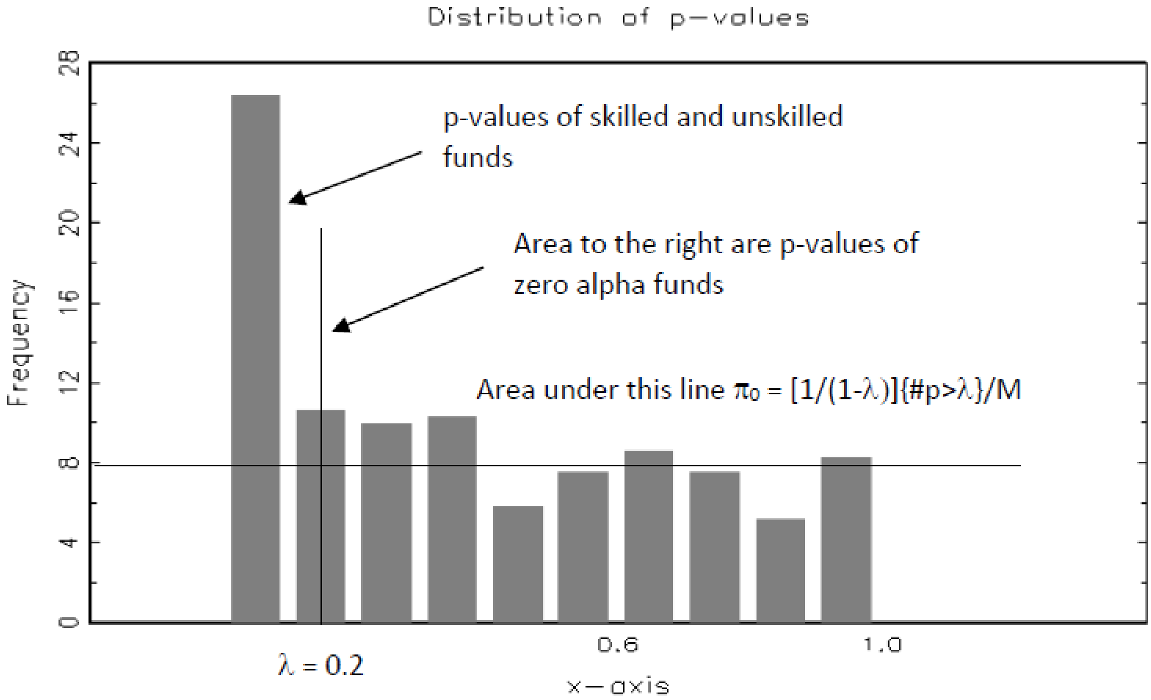

Table 1A: Gross and Net Fund Returns, Summary Statistics (January 1998-May 2018)

This table shows summary statistics for the full sample of 1,254 fixed income mutual funds, for the four largest investment styles. Panel A presents gross and net monthly bond fund returns net of the risk-free rate (r_f), while Panel B presents analogous statistics for bond fund returns in excess of each fund's benchmark return. "Average" is the average monthly percent excess return; SD is the average standard deviation of monthly returns; Sharpe is the average Sharpe Ratio of monthly fund excess returns; Min and Max are the average minimum and maximum percentage monthly fund returns; and Inform. Ratio is the Information Ratio.

Panel A: Gross and Net Returns over the risk-free rate

Category	# Funds	Gross Returns Average	SD	Sharpe	Min	Max	Net Returns Average	SD	Sharpe	Min	Max
All	1254	0.2468	0.2129	0.2425	-2.0367	0.9681	0.1861	0.2139	0.1757	-2.0831	0.9495
Broad	621	0.2634	0.2399	0.2248	-2.0367	0.9681	0.2011	0.2410	0.1689	-2.0831	0.9495
Govt/Corp	241	0.2224	0.1763	0.2810	-1.5610	0.8762	0.1652	0.1779	0.2004	-1.6212	0.8299
Govt	195	0.1772	0.1273	0.2492	-0.2997	0.8521	0.1168	0.1288	0.1481	-0.3310	0.8100
Corporate	71	0.3577	0.3147	0.2465	-1.0759	0.7980	0.2833	0.3205	0.2003	-1.1476	0.7434

Panel B: Gross and Net Returns over the Benchmark Return

Category	# Funds	Gross Returns Average	SD	Inform. Ratio	Min	Max	Net Returns Average	SD	Sharpe	Min	Max
All	1254	0.0101	0.1671	0.0663	-2.3141	0.8041	-0.0506	0.1711	-0.0877	-2.3861	0.6367
Broad	621	0.0023	0.1976	0.0511	-2.3141	0.8041	-0.0599	0.2005	-0.0884	-2.3861	0.6367
Govt/Corp	241	0.0287	0.1387	0.1057	-1.7174	0.5747	-0.0285	0.1412	-0.0631	-1.7775	0.4612
Govt	195	0.0241	0.0915	0.0871	-0.4151	0.6300	-0.0363	0.0984	-0.1011	-0.4867	0.5470
Corporate	71	-0.0350	0.2316	0.0209	-1.2326	0.2015	-0.1104	0.2400	-0.0909	-1.3436	0.1420

Table 1B: Gross and Net Fund Returns, Summary Statistics (January 1998-August 2008)

This table shows summary statistics for the full sample of 1,254 fixed income mutual funds for the four largest investment styles. Panel A presents gross and net monthly bond fund returns net of the risk-free rate (r_f), while Panel B presents analogous statistics for bond fund returns in excess of each fund's benchmark return. "Average" is the average monthly percent excess return; SD is the average standard deviation of monthly returns; Sharpe is the average Sharpe Ratio of monthly fund excess returns; Min and Max are the average minimum and maximum percentage monthly fund returns; and Inform. Ratio is the Information Ratio.

Panel A: Gross and Net Returns over the risk-free rate

Category	# Funds	Gross Returns Average	SD	Sharpe	Min	Max	Net Returns Average	SD	Sharpe	Min	Max
All	985	0.1616	0.1541	0.1689	-1.3753	0.7680	0.0983	0.1565	0.1006	-1.4227	0.6534
Broad	477	0.1608	0.1712	0.1479	-1.3753	0.7680	0.0957	0.1732	0.0901	-1.4227	0.6534
Govt/Corp	195	0.1470	0.1092	0.1960	-0.6520	0.4495	0.0872	0.1072	0.1119	-0.7146	0.4038
Govt	173	0.1575	0.1117	0.1945	-0.2997	0.6950	0.0948	0.1149	0.1057	-0.3310	0.6460
Corporate	47	0.1280	0.2583	0.0906	-0.8700	0.6514	0.0471	0.2618	0.0418	-0.9436	0.5556

Panel B: Gross and Net Returns over Benchmark Return

Category	# Funds	Gross Returns Average	SD	Inform. Ratio	Min	Max	Net Returns Average	SD	Inform. Ratio	Min	Max
All	984	-0.0118	0.1269	-0.0016	-1.5628	0.6620	-0.0752	0.1298	-0.1827	-1.6102	0.5485
Broad	476	-0.0199	0.1501	-0.0099	-1.5628	0.6620	-0.0850	0.1523	-0.1763	-0.9022	0.3217
Govt/Corp	195	-0.0063	0.0899	-0.0136	-0.7848	0.4022	-0.0661	0.0907	-0.2045	-0.8474	0.2974
Govt	173	0.0047	0.0788	0.0278	-0.4151	0.3781	-0.0581	0.0850	-0.1898	-0.4867	0.3291
Corporate	47	-0.0111	0.1975	0.0179	-1.0474	0.2783	-0.0920	0.2026	-0.0919	-1.1210	0.2156

Table 1C: Gross and Net Fund Returns, Summary Statistics (April 2009 - May 2018)

This table shows summary statistics for the full sample of 1,254 fixed income mutual funds, for the four largest investment styles. Panel A presents gross and net monthly bond fund returns, net of the risk-free rate (r_f), while Panel B presents analogous statistics for bond fund returns in excess of each fund's benchmark return. "Average" is the average monthly percent excess return; SD is the average standard deviation of monthly returns; Sharpe is the average Sharpe Ratio of monthly fund excess returns; Min and Max are the average minimum and maximum percentage monthly fund returns; and Inform. Ratio is the Information Ratio.

Panel A: Gross and Net Returns over the risk-free rate

Category	# Funds	Gross Returns Average	SD	Sharpe	Min	Max	Net Returns Average	SD	Sharpe	Min	Max
All	945	0.4279	0.3183	0.4354	-3.1776	2.5104	0.3702	0.3138	0.3594	-3.2413	2.4506
Broad	445	0.5151	0.3667	0.4429	-3.1776	2.5104	0.4552	0.3629	0.3839	-3.2413	2.4506
Govt/Corp	186	0.3475	0.1988	0.4960	0.0437	1.1231	0.2933	0.1989	0.3969	-0.0050	1.1227
Govt	132	0.2216	0.1562	0.3984	-0.2038	0.8071	0.1669	0.1535	0.2723	-0.2238	0.7898
Corporate	63	0.6960	0.3046	0.4563	-0.5274	1.5388	0.6248	0.3023	0.4039	-0.6522	1.4468

Panel B: Gross and Net Returns over Benchmark Return

Category	# Funds	Gross Returns Average	SD	Inform. Ratio	Min	Max	Net Returns Average	SD	Inform. Ratio	Min	Max
All	945	0.0798	0.1917	0.2143	-3.7545	0.8041	0.0220	0.1950	0.0371	-3.8182	0.6727
Broad	445	0.0934	0.2420	0.2019	-3.7545	0.8041	0.0335	0.2450	0.0539	-3.8182	0.6727
Govt/Corp	186	0.1007	0.1052	0.2997	-0.1553	0.05747	0.0465	0.1050	0.1010	-0.2381	0.4612
Govt	132	0.0835	0.1225	0.2723	-0.2809	0.6300	0.0289	0.1273	0.0541	-0.3369	0.5607
Corporate	63	-0.0313	0.2068	0.0200	-1.2189	0.2710	-0.1024	0.2167	-0.1247	-1.3436	0.1859

Table 2: Descriptive statistics of factors

In this table we present summary statistics for risk factors used, for the full sample and for the pre- and post GFC crisis samples. "Average" is the arithmetic average for the factor and SD is its standard deviation. r_f is the yield on a 1-month US T-Bill; R_m is the end month return on the Bloomberg Barclays US Aggregate index; R_{hy} is the end month return on the Bloomberg Barclays US High Yield index; R_{mort} is the end month return on the Bloomberg Barclays US MBS index; and the TED spread is month end percentage change in the difference between three-month dollar LIBOR and the yield on a three-month US T-Bill.

	Jan 98 to May 2018		Jan 98 to Aug 2008		April 2009 to May 2018	
	Average	SD	Average	SD	Average	SD
$R_m - r_f$	0.24%	0.99	0.20%	0.99	0.27%	0.82
$R_{hy} - r_f$	0.41%	2.56	0.16%	2.00	0.92%	2.15
$R_{mort} - r_f$	0.23%	0.73	0.19%	0.70	0.22%	0.68
TED Spread	0.04%	0.03	0.04%	0.03	0.03%	0.01

Table 3: Gross and Net Alphas and Betas for Alternative Factor Models (January 1998 - May 2018).

This table presents average results across all 1,254 funds for gross-alphas (Panel A) and net-alphas (Panel B) for the factor models described in Section 3.1 of the paper. The average R-squared across all funds is in column 2. Average alphas are in column 3. The proportion of funds with positive alphas are shown in column 4, the figures in parentheses represent the proportion of funds that have positive and significant alphas. The proportion of funds with negative alphas are shown in column 5, the figures in parentheses represent the proportion of funds that have negative and significant alphas. The proportion of funds with non-normal residuals (Bera-Jacque test) is shown in column 6. Statistically significant alphas use standard errors at a 2.5% critical value (one-tail test). In Panel C we present average beta values for the gross-of-fee fund returns. The figures in parentheses show the proportion of negative and significant betas and the proportion of positive and significant alphas, respectively.

Panel A: Gross-Alpha	**R-Sq**	**Alpha**	**+ve alpha**	**-ve alpha**	**% non-normal**
Market Model	54.2%	0.088	76.65% (37.29%)	23.35% (1.99%)	77.53%
Benchmark Model	73.6%	0.033	76.08% (33.89%)	23.92% (3.67%)	77.99%
3FHD Model	79.1%	0.017	69.40% (29.24%)	30.60% (3.67%)	75.78%
4FHD Model	84.0%	0.015	69.70% (28.15%)	30.30% (4.15%)	76.32%
Panel B: Net-Alpha	**R-Sq**	**Alpha**	**+ve alpha**	**-ve alpha**	**% non-normal**
Market Model	54.2%	0.028	54.74% (12.19%)	45.26% (16.18%)	77.69%
Benchmark Model	73.5%	-0.028	58.05% (22.57%)	41.95% (9.33%)	77.99%
3FHD Model	79.14%	-0.044	32.03% (6.61%)	67.97% (24.86%)	75.62%
4FHD Model	84.0%	-0.046	26.6% (5.3%)	73.4% (29.4%)	76.32%
Panel C: Betas (Gross)	$\boldsymbol{\beta_{Rm}}$	$\boldsymbol{\beta_{Rb}}$	$\boldsymbol{\beta_{HY}}$	$\boldsymbol{\beta_{RMort}}$	
Market Model	0.706 (0.07%/82.15%)				
Benchmark Model		1.02 (0%/96.2%)			
3FHD Model	0.638 (5.90%/82.15%)		0.237 (12.3%/65.8%)	-0.056 (18.7%/12.5%)	
4FHD Model	0.010 (14.3%/25.76%)	0.87 (1.67%/64.99%)	0.070 (9.19%/52%)	-0.025 (10.6%/10.2%)	

Table 4: Bond Fund Styles, Gross and Net Alphas and Betas, 4FHD Model (January 1998 - May 2018)

This table presents average results across all 1,254 funds for gross-alphas (Panel A) and net-alphas (Panel B) using they 4FHD factor model described in Section 3.1 of the paper. The average R-squared across all funds is in column 2. Average alphas are in column 3. The proportion of funds with positive alphas are shown in column 4, the figures in parentheses represent the proportion of funds that have positive and significant alphas. The proportion of funds with negative alphas are shown in column 5, the figures in parentheses represent the proportion of funds that have negative and significant alphas. The proportion of funds with non-normal residuals (Bera-Jacque test) is shown in column 6. Statistically significant alphas use standard errors at a 2.5% critical value (one-tail test). In Panel C we present average beta values for the gross-of-fee fund returns. The figures in parentheses show the proportion of negative and significant betas and the proportion of positive and significant alphas respectively.

Panel A: Gross Alpha	**R-Sq**	**Alpha**	**+ve alpha**	**-ve alpha**	**% non-normal**
Broad	86.11%	0.013	418 (158)	203 (19)	73.59%
Gov./Corp	80.85%	0.021	180 (78)	61 (5)	81.74%
Gov.	78.53%	0.010	130 (48)	65 (7)	72.31%
Corporate	85.06%	0.011	51 (22)	20 (2)	81.69%
Panel B: Net Alpha	**R-Sq**	**Alpha**	**+ve alpha**	**-ve alpha**	**% non-normal**
Broad	86.11%	-0.049	183 (32)	438 (109)	73.75%
Gov./Corp	80.83%	-0.036	64 (12)	177 (67)	81.74%
Gov.	78.49%	-0.050	31 (8)	164 (100)	72.31%
Corporate	85.06%	-0.063	29 (7)	42 (11)	81.69%
Panel C: Betas (Gross)	$\boldsymbol{\beta_{Rm}}$	$\boldsymbol{\beta_{Rb}}$	$\boldsymbol{\beta_{HY}}$	$\boldsymbol{\beta_{RMort}}$	
Broad	-0.155 (23.34%/19.48%)	1.059 (2.9%/54.3%)	0.060 (8.05%/57.6%)	-0.053 (14.17%/8.86%)	
Gov./Corp	0.1337 (5.81%/26.56%)	0.726 (0.41%/80.91%)	0.087 (8.71%/60.58%)	-0.043 (8.3%/9.13%)	
Gov.	0.3415 (1.54%/45.64%)	0.633 (0.51%/70.26%)	0.027 (13.33%/35.38%)	0.011 (8.72%/18.46%)	
Corporate	-0.057 (9.86%/5.63%)	0.600 (1.41%/61.97%)	0.233 (12.68%/32.39%)	0.091 (0%/5.63%)	

Table 5: 4FHD testing the significance of pre- and post-crisis.

This table presents the results from estimating equation (10) in the text. The base model is the 4FHD model, but variables have been multiplied by a dummy variable that takes the value of 0 between January 1998 and August 2008 and 1 for the period April 2009 to May 2018. Column 2 shows the average coefficients for the sample of funds. Column 3 show the proportion of coefficients that are negative and statistically significant. Column 4 shows the proportion of coefficients that are positive and statistically significant. The significance level used is 2.5% (one tail test).

Coefficient	Average	Proportion of negative significant	Proportion of positive significant
Alpha	-0.052	40.37%	2.28%
β (R_{bi})	0.774	1.87%	63.77%
β (R_m)	0.103	10.70%	29.14%
β (R_{HY})	0.016	14.97%	33.42%
β (R_{mort})	0.015	12.43%	12.30%
Alpha*	0.141	9.88%	16.18%
$\beta*$ (R_{bi})	0.013	5.08%	19.52%
$\beta*$ (R_m)	-0.169	20.32%	6.95%
$\beta*$ (R_{HY})	-0.001	6.15%	40.11%
$\beta*$ (R_{mort})	-0.040	7.35%	10.83%
R-Square	85.7%		
# Funds	748		

Table 6: Liquidity risk

This table presents the results from estimating the Benchmark model and the 4FHD model both with and without the addition of the TED spread. Average values for alpha and beta are shown in the respective columns. Figures in parentheses show the proportion of coefficients that are found to be negative and statistically significant and positive and statistically significant (-ve/+ve).

	# Funds	Alpha	β (R_i)	β (R_m)	β (R_{HY})	β (R_{mort})	β (TED)
Panel A							
Benchmark Model	1254	-0.0278	1.02				
		(22.57%/9.33%)	(0.8%/96.17%)				
Benchmark Model +	1254	0.0408	1.04				-1.4925
		(10.61%/19.38%)	(0.08%/97.29%)				(18.58%/4.94%)
Panel B							
4FHD Model	1252	-0.0456	0.88	0.0025	0.0701	-0.0246	
		(33.07%/5.99%)	(1.68%/65.1%)	(14.3%/25.72%)	(9.19%/52%)	(10.62%/9.98%)	
4FHD Model +	1252	0.0062	0.88	0.0119	0.0604	-0.0212	-1.3802
		(15.02%/13.90%)	(1.76%/65.66%)	(14.78%/25.72%)	(9.58%/52.16%)	(11.02%/10.62%)	(18.93%/5.59%)

Table 7: False Discovery Rate: 4FHD Model, (January 1998 to May 2018)

The augmented Huij Derwall Model (4FHD) includes a self-designated benchmark index, an aggregate bond index, a high yield bond index and a mortgage bond index. The figures reported are percentages (of the total number of funds). S+ = percent (number)of statistically significant positive-alpha funds (at 1%, 2.5%, 5% and 10% significance levels, one tail) based on bootstrap p-values, FDR+= percent of statistically significant funds that are false discoveries, F+ = percent of false positive alpha funds. 1254 funds are included in the analysis.

	Positive Alpha Funds				**Negative Alpha Funds**			
Panel A: Gross-Alpha π_0 = 0.6209								
Significance level	S^+	FDR^+	T^+	F^+	S^-	FDR^-	T^-	F^-
1%	20.02 (251)	1.55	19.71	0.31	1.52 (19)	20.49	1.20	0.31
2.5%	**26.40 (331)**	**2.94**	**25.62**	**0.78**	**2.95 (37)**	**26.30**	**2.17**	**0.78**
5%	31.18 (391)	4.98	29.63	1.55	5.34 (67)	29.05	3.79	1.55
10%	38.92 (488)	7.98	35.81	3.10	8.77 (110)	35.39	5.67	3.10
Panel B: Net-Alpha π_0 = 0.5388								
Significance level	S^+	FDR^+	T^+	F^+	S^-	FDR^-	T^-	F^-
1%	3.59 (45)	7.51	3.32	0.27	22.41 (281)	1.20	22.14	0.27
2.5%	**5.34 (67)**	**12.61**	**4.67**	**0.67**	**29.43 (369)**	**2.29**	**28.76**	**0.67**
5%	7.34 (92)	18.36	5.99	1.35	36.92 (463)	3.65	35.57	1.35
10%	10.13 (127)	26.60	7.43	2.69	44.82 (562)	6.01	42.12	2.69

Table 8: False Discovery Rate (4FHD Model), Pre- and Post-Crisis Periods, Gross and Net Alphas

The augmented Huij Derwall Model (4FHD) includes a self-designated benchmark index, an aggregate bond index, a high yield bond index and a mortgage bond index. The figures reported are percentages (of the total number of funds). S^+ = percent (number)of statistically significant positive-alpha funds (at 2.5% significance levels, one tail) based on bootstrap p-values, FDR^+= percent of statistically significant funds that are false discoveries, F^+ = percent of false positive alpha funds. The pre-financial crises period is January 1998 to August 2008 and the post-financial crisis period is April 2009 to May 2018.

		Positive Alpha Funds				**Negative Alpha Funds**			
	# funds	S^+	FDR^+	$\mathbf{T^+}$	F^+	S^-	FDR^-	$\mathbf{T^-}$	F^-
Panel A: Gross-Alpha									
Pre-Crisis π_0 = 0.7636	984	13.31 (131)	7.17	**12.36**	0.95	2.95 (29)	32.39	**1.99**	0.95
Post-Crisis π_0 = 0.5412	945	36.19 (342)	1.87	**36.51**	0.68	0.53 (5)	100	**0**	0.68
Panel B: Net –Alpha									
Pre-Crisis π_0 = 0.3846	984	1.93 (19)	24.90	**1.45**	0.48	41.26 (406)	1.17	**40.78**	0.48
Post-Crisis π_0 = 0.6259	945	10.90 (103)	7.18	**10.12**	0.78	17.78 (168)	4.40	**17.00**	0.78