JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS Vol. 52, No. 3, June 2017, pp. 1279–1299
COPYRIGHT 2017, MICHAEL G. FOSTER SCHOOL OF BUSINESS, UNIVERSITY OF WASHINGTON, SEATTLE, WA 98195
doi:10.1017/S0022109017000229

New Evidence on Mutual Fund Performance: A Comparison of Alternative Bootstrap Methods

David Blake, Tristan Caulfield, Christos Ioannidis, and Ian Tonks*

Abstract

We compare two bootstrap methods for assessing mutual fund performance. The first produces narrow confidence intervals due to pooling over time, whereas the second produces wider confidence intervals because it preserves the cross correlation of fund returns. We then show that the average U.K. equity mutual fund manager is unable to deliver outperformance net of fees under either bootstrap. Gross of fees, 95% of fund managers on the basis of the first bootstrap and all fund managers on the basis of the second bootstrap fail to outperform the luck distribution of gross returns.

I. Introduction

Evidence collected over an extended period on the performance of open-ended mutual funds in the United States (Jensen (1968), Malkiel (1995), and Wermers, Barras, and Scaillet (2010)) and unit trusts and open-ended investment companies (OEICs)[1] in the United Kingdom (Blake and Timmermann (1998), Lunde, Timmermann, and Blake (1999)) has found that, on average, a fund manager cannot outperform the market benchmark and that any outperformance is more likely to be due to luck rather than skill.

More recently, Kosowski, Timmermann, Wermers, and White (KTWW) (2006) report that the time-series returns of individual mutual funds typically

*Blake (corresponding author), d.blake@city.ac.uk, Cass Business School, City University of London; Caulfield, t.caulfield@cs.ucl.ac.uk, University College London; Ioannidis, c.ioannidis@aston.ac.uk, Aston Business School, Aston University; and Tonks, i.tonks@bath.ac.uk, School of Management, University of Bath. The data set used in this paper was constructed while Tonks was an Economic and Social Research Council (ESRC) Business Fellow at the United Kingdom's Financial Services Authority (FSA) in 2009 (RES-186-27-0014), and Tonks is obliged to the FSA's Economics of Regulation Unit for hosting this visit. We are grateful for comments and discussion from Peter Andrews, Alok Bhargava, Stephen Brown (the editor), Qun Harris, Allan Timmermann, and Russell Wermers (the referee).

We direct your attention also to our Internet Appendix (available at www.jfqa.org) that provides robustness tests of our findings.

[1] These are, respectively, the U.K. and European Union terms for open-ended mutual funds. There are differences, however, the principal one being that unit trusts have dual pricing (a bid and an offer price), while OEICs have single pricing.

exhibit nonnormal distributions.[2] They argued that this finding has important implications for the luck-versus-skill debate and that there was a need to reexamine the statistical significance of mutual fund manager performance using bootstrap techniques. They applied a bootstrap methodology (Efron and Tibshirani (1993), Politis and Romano (1994)) that creates a sample of monthly pseudo excess returns by randomly resampling residuals from a factor benchmark model and imposing a null of zero abnormal performance.[3] Following the bootstrap exercise, KTWW determine how many funds from a large group one would expect to observe having large alphas by luck and how many are actually observed. Using data on 1,788 U.S. mutual funds over the period Jan. 1975–Dec. 2002, they show that, by luck alone, 9 funds would be expected to achieve an annual alpha of 10% over a 5-year period, but in fact, 29 funds achieve this hurdle. KTWW note,

> This is sufficient, statistically, to provide overwhelming evidence that some fund managers have superior talent in picking stocks. Overall, our results provide compelling evidence that, net of all expenses and costs (except load charges and taxes), the superior alphas of star mutual fund managers survive and are not an artifact of luck ((2006), p. 2553).

Applying the same bootstrap method to 935 U.K. equity unit trusts and OEICs between Apr. 1975 and Dec. 2002, Cuthbertson, Nitzche, and O'Sullivan (2008) find similar evidence of significant stock-picking ability among a small number of top-performing fund managers. Blake, Rossi, Timmermann, Tonks, and Wermers (2013) show that fund manager performance improves if the degree of decentralization (in the form of increasing specialization) is increased.

However, these results have been challenged by Fama and French (FF) (2010), who suggest an alternative bootstrap method that preserves any contemporaneously correlated movements in the volatilities of the explanatory factors in the benchmark model and the residuals. They calculate the Jensen (1968) alpha for each fund, then compute pseudo returns by deducting the Jensen alpha from the actual returns to obtain benchmark-adjusted (zero-alpha) returns, thereby maintaining the cross-sectional relationship between the factor and residual volatilities (i.e., between the explained and unexplained components of returns). Their sample consists of 5,238 U.S. mutual funds over the period Jan. 1984–Sept. 2006, and following their bootstrap calculations, they conclude that there is little evidence of mutual fund manager skill.

There are three differences between the KTWW (2006) and FF (2010) studies. First, although both studies use data for U.S. domestic equity mutual funds, KTWW use data from 1975 to 2002, whereas the data set in FF covers the more recent 1984–2006 period. Second, the studies use different fund-inclusion criteria: KTWW restrict their sample to funds that have a minimum of 60 monthly

[2] KTWW ((2006), p. 2559) attributed this to the possibilities that i) the residuals of fund returns are not drawn from a multivariate normal distribution, ii) correlations in these residuals are nonzero, iii) funds have different risk levels, and iv) the parameter estimation error results in the standard critical values of the normal distribution being inappropriate in the cross section.

[3] One of the earliest applications of this methodology is that by Brown and Warner (1985). They employ a block resampled bootstrap for the evaluation of event-study measures of investment performance.

observations, whereas FF restrict theirs to funds that have a minimum of 8 monthly observations. Third and most important, with respect to the bootstrap method used, for each bootstrap simulation, the former simulate fund returns and factor returns independently of each other, whereas the latter simulate these returns jointly.

It is therefore important to identify whether the different results from the two studies are due to the different time periods analyzed, different inclusion criteria, or the different bootstrap methods used. We use a data set of U.K. domestic equity mutual fund returns from Jan. 1998 to Sept. 2008 to assess the performance of mutual fund managers. We also compare the two different bootstrap methods using the *same* sample of funds over the *same* time period and with the *same* fund-inclusion criterion.

It is well known that the Jensen (1968) alpha measure of performance is biased in the presence of fund manager market-timing skills (Treynor and Mazuy (1966), Merton and Henriksson (1981)). Grinblatt and Titman (1994) have suggested a total performance measure that is the sum of the Jensen alpha and market-timing coefficients in an extended factor-benchmark model. Allowing for market timing exacerbates the nonnormality of standard significance tests, and an additional contribution of this paper is to assess the significance of the total performance measure in the KTWW (2006) and FF (2010) bootstrapped distributions.

The structure of the paper is as follows: Section II reviews the approach to measuring mutual fund performance and shows how this approach has recently been augmented through the use of bootstraps. Section III discusses our data set. The results are presented in Section IV, and Section V provides a summary of additional robustness checks. Section VI concludes.

II. Measuring Mutual Fund Performance

A. Measuring Performance Using Factor-Benchmark Models

Building on Jensen's (1968) original approach, we use a 4-factor benchmark model to assess the performance or excess return over the riskless rate ($R_{it} - rf_t$) of the manager of mutual fund i obtained in period t (out of a total of T possible periods):

$$(1)\quad R_{it} - rf_t \quad = \quad \alpha_i + \beta_i(R_{mt} - rf_t) + \gamma_i \mathrm{SMB}_t + \delta_i \mathrm{HML}_t + \lambda_i \mathrm{MOM}_t + \varepsilon_{it},$$

where the 4 common factors are the excess return on the market index ($R_{mt} - rf_t$); the returns on a size factor, SMB_t; a book-to-market factor, HML_t (Fama–French (1993)), and the return on a momentum factor, MOM_t (Carhart (1997)). The genuine skill of the fund manager, controlling for these common risk factors, is measured by alpha (α_i), which is also known as the selectivity skill.[4]

Under the null hypothesis of no abnormal performance (i.e., no selectivity skill), the expected value of $\hat{\alpha}_i$ should be equal to 0. For each fund, we could test

[4] Ferson and Schadt (1996) suggest a conditional version of this 4-factor benchmark model that controls for time-varying factor loadings. However, KTWW (2006) report that the results from estimating the conditional and unconditional models are very similar, and in the remainder of this paper, we follow them and consider only the unconditional version of equation (1).

the significance of each $\hat{\alpha}_i$ as a measure of that fund's abnormal performance relative to its standard error. We might also test the significance of the average value of the alpha across the N funds in the sample (Malkiel (1995)). Alternatively, we could follow Blake and Timmermann (1998) (and also Fama and French (2010), Table II) and regress an equal-weighted (or a value-weighted) portfolio p of the excess returns $(R_{pt} - rf_t)$ on the N funds on the 4 factors in equation (1) and test the significance of the estimated $\hat{\alpha}_p$ in this regression.

The original Jensen (1968) approach made no allowance for the market-timing abilities of fund managers when fund managers take an aggressive position in a bull market (by holding high-beta stocks) and a defensive position in a bear market (by holding low-beta stocks). Treynor and Mazuy (1966) tested for market timing by adding a quadratic term in the market excess return in the benchmark model to capture the "curvature" in the fund manager's performance as the market rises and falls. To test jointly for selectivity and market-timing skills, we estimate a 5-factor benchmark model:

$$(2) \qquad R_{it} - rf_t = \alpha_i + \beta_i(R_{mt} - rf_t) + \gamma_i \text{SMB}_t + \delta_i \text{HML}_t + \lambda_i \text{MOM}_t + \eta_i\,(R_{mt} - rf_t)^2 + \varepsilon_{it}.$$

Market-timing ability is measured by the sign and significance of $\hat{\eta}_i$. To capture both selectivity and timing skills simultaneously, we use the Treynor–Mazuy total performance measure (TM$_i$) averaged over T periods:

$$(3) \qquad \text{TM}_i = \alpha_i + \eta_i \text{Var}(R_m - rf).$$

This was derived by Grinblatt and Titman ((1994), App. B, p. 441), and its significance can be assessed using a t-statistic based on its standard error.

B. Assessing Performance Using Bootstrap Methods

On account of nonnormalities in returns, bootstrap methods need to be applied to both of the factor benchmark models (1) and (2) to assess performance. To apply the KTWW (2006) bootstrap in equation (1), we first obtain ordinary least squares (OLS) estimated alphas, factor loadings, and residuals using a time series of monthly excess returns for fund i in equation (1). We then construct a sample of pseudo excess returns by randomly resampling residuals with replacement from $\{\hat{\varepsilon}_{it}, t = T_{i0}, \ldots, T_{i1}\}$ while preserving the historical ordering of the common risk factors and imposing the null of zero abnormal performance ($\alpha_i = 0$):

$$(4) \qquad (R_{it} - rf_t)^b \equiv \hat{\beta}_i\,(R_{mt} - rf_t) + \hat{\gamma}_i \text{SMB}_t + \hat{\delta}_i \text{HML}_t + \hat{\lambda}_i \text{MOM}_t + \hat{\varepsilon}_{it}^b,$$

where b is the bth bootstrap, and $\hat{\varepsilon}_{it}^b$ is a drawing from $\{\hat{\varepsilon}_{it}, t = T_{i0}, \ldots, T_{i1}\}$. By construction, this pseudo excess return series has zero alpha. For bootstrap $b = 1$, we regress the pseudo excess returns on the factors:

$$(5) \qquad (R_{it} - rf_t)^b = \alpha_i + \beta_i(R_{mt} - rf_t) + \gamma_i \text{SMB}_t + \delta_i \text{HML}_t + \lambda_i \text{MOM}_t + \tilde{\varepsilon}_{it},$$

and we save the estimated alpha. We repeat for each fund, $i = 1, \ldots, N$, to arrive at the first draw from the cross section of bootstrapped alphas $\{\tilde{\alpha}_i^b, i = 1, \ldots, N; b = 1\}$

and the corresponding t-statistics $\{t(\tilde{\alpha}_i^b), i=1,\ldots,N; b=1\}$. We then repeat for all bootstrap iterations $b=1,\ldots,10{,}000$. It is important to reiterate that the common risk factors are not resampled in the KTWW (2006) bootstrap: Their historical ordering is not varied across simulation runs. It is only the residuals that are reordered with this bootstrap.

We now have the cross-sectional distribution of alphas from all the bootstrap simulations $\{\tilde{\alpha}_i^b, i=1,\ldots,N; b=1,\ldots, 10{,}000\}$ that result from the sampling variation under the null that the true alpha is 0. The bootstrapped alphas can be ranked from smallest to largest to produce the "luck" (i.e., pure chance or zero-skill) cumulative distribution function (CDF) of the alphas. We have a similar cross-sectional distribution of bootstrapped t-statistics $\{t(\tilde{\alpha}_i^b), i=1,\ldots,N; b=1,\ldots,10{,}000\}$, which can be compared with the distribution of actual $\{t(\hat{\alpha}_i), i=1,\ldots,N\}$ values once both sets of t-statistics have been reordered from smallest to largest. We follow KTWW (2006), who prefer to work with the t-statistics rather than the alphas because the use of the t-statistic "controls for differences in risk-taking across funds" (p. 2555).[5]

FF (2010) employ an alternative bootstrap method. They calculate alpha for each fund using the time-series regression in equation (1), as do KTWW (2006). But FF do not resample the residuals of each individual fund as do KTWW; rather, they resample with replacement over the full cross section of returns, thereby producing a common time ordering across all funds in each bootstrap. The historical ordering of the common risk factors is therefore not preserved in this bootstrap. In our study, we resample from all 129 monthly observations in the data set, and we impose the null hypothesis as do FF by subtracting the estimate of alpha from each resampled month's returns.[6] For each fund and each bootstrap, we regress the pseudo excess returns on the factors:

$$\begin{aligned}(6)\qquad \left[(R_{it}-rf_t)-\hat{\alpha}_i\right]^b &= \alpha_i+\beta_i\,(R_{mt}-rf_t)+\gamma_i\text{SMB}_t\\ &\quad +\delta_i\text{HML}_t+\lambda_i\text{MOM}_t+\tilde{\varepsilon}_{it},\end{aligned}$$

and we save the estimated bootstrapped alphas $\{\tilde{\alpha}_i^b, i=1,\ldots,N; b=1,\ldots,10{,}000\}$ and t-statistics $\{t(\tilde{\alpha}_i^b), i=1,\ldots,N; b=1,\ldots,10{,}000\}$. We then rank the alphas and t-statistics from lowest to highest to form the FF (2010) "luck" distribution under the null hypothesis.

The most important difference between the two methods is that within each bootstrap run, the FF (2010) bootstrap takes into account the cross-sectional distribution of the residuals, conditional on the realization of the systematic risk

[5] KTWW ((2006), p. 2559) note that the t-statistic also provides a correction for spurious outliers by dividing the estimated alpha by a high estimated standard error when the fund has a short life or undertakes risky strategies.

[6] To illustrate, for bootstrap $b=1$, suppose that the first time-series drawing is month $t=37$; then, the first set of pseudo returns incorporating zero abnormal performance for this bootstrap is found by deducting $\hat{\alpha}_i$ from $(R_{i,37}-rf_{37})$ for every fund i that is in the sample for month $t=37$. Suppose that the second time-series drawing is month $t=92$; then, the second set of pseudo returns is found by deducting $\hat{\alpha}_i$ from $(R_{i,92}-rf_{92})$ for every fund i that is in the sample for month $t=92$. After T drawings, the first bootstrap is completed. This contrasts with the KTWW (2006) bootstrap in which for $b=1$, the first drawing for fund 1 might be that for month $t=37$ (assuming it is in the sample for this month), whereas the first drawing for fund 2 might be for month $t=92$ (assuming it is in the sample for this month), and so forth.

factors, whereas the KTWW (2006) bootstrap uses the unconditional distribution of the residuals and assumes both that there is independence between the residuals across different funds and that the influence of the common risk factors is fixed at their historical realizations.[7]

There is one other potentially important difference between the two bootstrap methods as implemented in the two studies. KTWW (2006) include funds in their analysis with more than 60 monthly observations in the data set, whereas the fund-inclusion criterion with FF (2010) is 8 months. The different inclusion criteria involve a trade-off between the low estimation precision that is associated with estimating a model with a small number of degrees of freedom and the potential look-ahead bias associated with estimating a model that requires funds to be in the data for some time. Carhart, Carpenter, Lynch, and Musto (2002) discuss sample biases in mutual fund performance evaluation and distinguish between "survivor biases" (evaluation only of the selected sample of funds still in existence at the end of the time period) and "look-ahead biases" (evaluation of funds by considering only funds that survive for a minimum length of time). Survivor bias is regarded as a property of the data set, whereas look-ahead bias results from any test methodology imposing a minimal survival period. In order to assess the sensitivity of these sample-selection criteria on the look-ahead bias, we construct separate subsamples based on including funds with at least 8, 15, 20, 40, and 60 monthly observations. As the minimum number of monthly observations increases, the number of funds included in the subsample decreases.

FF (2010) report that the distribution of actual $t(\hat{\alpha}_i)$ values is to the left of that of the "luck" distribution of the bootstrapped $t(\tilde{\alpha}_i^b)$ values, particularly for funds with negative alphas but also for most funds with positive alphas. FF conclude that there is little evidence of mutual fund manager skills. This contrasts with KTWW (2006), who conclude that there are a small number of genuinely skilled "star" fund managers.

FF (2010) point out a common problem with both methods. By randomly sampling across individual fund residuals in the KTWW (2006) method and across individual time periods in the FF method, any effects of autocorrelation in returns is lost. KTWW (p. 2582) performed a sensitivity analysis of this issue by resampling in time-series blocks up to 10 months in length. They found that the results changed very little.

III. Data

The data used in this study combine information from data providers Lipper, Morningstar, and Defaqto and consist of the monthly returns on a full sample of 561 U.K. domestic equity open-ended mutual funds (unit trusts and OEICs) over the period Jan. 1998–Sept. 2008, a total of 129 months. The data set also includes information on annual management fees, fund size, fund family, and relevant Investment Management Association (IMA) sectors.[8] We include in our sample

[7] FF (2010) argue that the KTWW (2006) bootstrap's "failure to account for the joint distribution of joint returns, and of fund and explanatory returns, biases the inferences of KTWW towards positive performance" (p. 1940).

[8] In 2014, the IMA changed its name to the Investment Association.

the following primary sector classes for U.K. domestic equity funds with the IMA definitions: UK All Companies, UK Equity Growth, UK Equity Income, UK Equity & Growth, and UK Smaller Companies. The sample is free from survivor bias (Elton, Gruber, and Blake (1996), Carpenter and Lynch (1999)) and includes funds that both were created during the sample period and exited due to liquidation or merger. In order to assess the degree of look-ahead bias in the alternative bootstrap methodologies, we construct 5 subsamples of the data by imposing the restriction that funds in the sample must have at least 8, 15, 20, 40, and 60 consecutive monthly returns. These criteria result in subsamples of 552, 535, 516, 454, and 384 funds, respectively. We perform our bootstrap analysis on each of these 5 subsamples separately.

"Gross" returns are calculated from bid-to-bid prices and include reinvested dividends. These are reported net of on-going operating and trading costs, but before the fund management fee has been deducted. As reported by Khorana, Servaes, and Tufano (2009), operating costs include administration, record-keeping, research, custody, accounting, auditing, valuation, legal costs, regulatory costs, distribution, marketing, and advertising. Trading costs include commissions, spreads, and taxes. We also compute "net" returns for each fund by deducting the monthly equivalent of the annual fund management fee. We have complete information on these fees for 451 funds. For each of the remaining funds, each month, we subtract the median monthly fund management fee for the relevant sector class and size quintile from the fund's gross monthly return. Following KTWW (2006) and FF (2010), we exclude initial and exit fees from our definition of net returns.

Table 1 provides some descriptive statistics on the returns to and the size of the mutual funds in our data set. We compare the distributional properties of the gross and net returns for two of the subsamples based on the selection criterion of 8 and 60 consecutive monthly observations. The average (equal-weighted) monthly gross return across the 552 funds with at least 8 consecutive monthly observations in the data set is 0.45% (45 basis points (bps)), compared with an

TABLE 1

Descriptive Statistics on U.K. Equity Mutual Funds (1998–2008)

Table 1 reports average monthly gross and net returns (in units of one hundredth of a percent) from Feb. 1998 to Sept. 2008 (129 months) for the case of 552 funds with a minimum of 8 consecutive observations and for the case of 384 funds with a minimum of 60 consecutive observations. It also reports the monthly total standard deviation for these cases and the between-fund and within-fund standard deviation. The former is the average over time of the cross-sectional standard deviation of fund returns, and the latter is the average across funds of the time-series standard deviation of returns. The table also reports key percentiles of the distribution of returns. Finally, it reports average monthly fund management fees over the same period and the size of funds at the end of the sample period.

Descriptive Statistics	Gross Returns (≥8 months)	Gross Returns (≥60 months)	Net Returns (≥8 months)	Net Returns (≥60 months)	Fund Management Fee (≥8 months)	Size at Sept. 30, 2008 (≥8 months, in £millions)
Mean	0.0045	0.0049	0.0033	0.0038	0.0011	234.64
Std. dev.	0.0482	0.0478	0.0482	0.0478	0.0002	644.24
Between-fund std. dev.	0.0081	0.0030	0.0082	0.0030	0.0002	
Within-fund std. dev.	0.0479	0.0477	0.0479	0.0477	0.0001	
10%	−0.0587	−0.0580	−0.0598	−0.0592	0.0008	7.86
25%	−0.0186	−0.0183	−0.0198	−0.0194	0.0010	25.87
50%	0.0128	0.0132	0.0117	0.0121	0.0012	63.30
75%	0.0330	0.0333	0.0318	0.0322	0.0012	196.18
90%	0.0525	0.0532	0.0514	0.0520	0.0012	503.77
No. of obs.	48,030	42,255	48,030	42,255	48,030	299
No. of funds	552	384	552	384	552	299

average monthly return over the same period of 0.36% for the Financial Times Stock Exchange (FTSE) All-Share Index.[9] The overall monthly standard deviation of these returns is 4.82%. In the case of 384 mutual funds with a minimum of 60 consecutive observations, the gross mean return is marginally higher and the variance marginally lower. The mean monthly net return for the larger subsample of 552 funds is 0.33%, implying that the monthly fund management fee is 0.11%. The mean return is now very close to the mean return of 0.36% for the FTSE All-Share Index. This provides initial confirmation that the average mutual fund manager cannot "beat the market" (i.e., cannot beat a buy-and-hold strategy invested in the market index) once all costs and fees have been taken into account.

Table 1 also shows that the within-fund standard deviation is much larger than the between-fund standard deviation, implying that fund returns tend to move together in any particular month but are more volatile over time. Furthermore, the between-fund volatility in the case of a minimum sample size of 8 monthly observations is much higher than in the case where the minimum sample size is 60 monthly observations. This is because samples involving a minimum of 8 consecutive observations are more likely to be drawn from the tails of the distribution of returns than those involving a minimum of 60 consecutive observations. Funds with only 8 observations in the data set are likely to have been closed down due to very poor performance.[10]

The final column of Table 1 shows that the distribution of scheme size is skewed. Whereas the median fund value for a subsample of 299 funds for which data on fund size are available in Sept. 2008 is £63.3 million, the mean value is much larger at £234 million. It can also be seen that 10% of the funds have values above £503.8 million.

IV. Results

We now turn to assessing the performance of U.K. equity mutual funds over the period 1998–2008. The results are divided into four sections. The first section looks at the performance of equal- and value-weighted portfolios of all 561 funds in the full sample against the 4- and 5-factor benchmark models over the whole sample period. The second section examines the properties of the moments of the actual, KTWW (2006), and FF (2010) CDFs for both the $t(\hat{\alpha}_i)$ and $t(\widehat{\mathrm{TM}}_i)$ performance measures. The third section compares the alpha performance of all the funds based on the actual t-statistics $t(\hat{\alpha}_i)$ from the factor models with the simulated t-statistics $t(\tilde{\alpha}_i^b)$ generated by the bootstrap methods of KTWW and FF described previously. We report the results for both gross and net returns and for the different fund-selection criteria. The fourth section conducts a total performance comparison based on the actual and simulated t-statistics, $t(\widehat{\mathrm{TM}}_i)$ and $t(\widetilde{\mathrm{TM}}_i^b)$, for the two bootstraps, again using both gross and net returns.

[9] Note that the FTSE All-Share Index return is gross of any costs and fees.

[10] Evidence to support this conjecture is contained in Table 1. The 384 funds (with a minimum of 60 observations) have a mean gross monthly return of 0.0049, whereas the 552 funds (with a minimum of 8 observations) have a lower mean gross return of 0.0045. The latter group of funds have a higher standard deviation than the former, and the quantiles of the CDF also show that these funds have much poorer returns throughout the distribution. These results are not affected by the censoring of the data at the beginning and end of the sample period.

A. Performance against Factor-Benchmark Models

Following Blake and Timmermann (1998), Table 2 reports the results from estimating the 4- and 5-factor models in equations (1) and (2) across all $T = 129$ time-series observations for the full sample of 561 funds, where the dependent variable is, first, the excess return on an equal-weighted portfolio p of all funds in existence at time t, and, second, the excess return on a value-weighted portfolio p of all funds in existence at time t, using starting market values as weights.[11] For each portfolio, the first two columns report the loadings on each of the factors when the dependent variable is based on gross returns, whereas the second two columns report the corresponding results using net returns. The loadings on the market portfolio and on the SMB_t factor are positive and significant, whereas the loadings are negative but insignificant on the HML_t factor. The factor loadings are positive but insignificant on the MOM_t factor.[12]

The alphas based on gross returns differ from the corresponding alphas based on net returns by the average level of fund management fees. However, the most important point is that the alpha (α_p) is not significant in the 4-factor model, and

TABLE 2

Estimates of Factor Models for U.K. Equity Mutual Fund Portfolios

Results in Table 2 use the 4-factor model without market timing (equation (1)) and the 5-factor model with market timing (equation (2)), based on data for 561 funds. The dependent variable is either the excess return on an equal-weighted portfolio or on a value-weighted portfolio (p) of all funds in existence at time t. The dependent variable is measured both gross and net of fund management fees. The total performance measure (equation (3)) is also reported. Relevant t-statistics estimated from White's (1980) robust standard errors are reported in parentheses below each parameter estimate. ** and *** indicate significance at the 5% and 1% levels, respectively.

	Equal-Weighted				Value-Weighted			
Independent Variables	Gross Returns	Gross Returns with Market Timing	Net Returns	Net Returns with Market Timing	Gross Returns	Gross Returns with Market Timing	Net Returns	Net Returns with Market Timing
α_p	0.0002	0.0016**	−0.0010	0.0005	−0.0002	0.0014	−0.0013	0.0003
	(0.20)	(1.71)	(−1.27)	(0.49)	(−0.21)	(1.414)	(−1.62)	(0.27)
$(R_{mt} - rf_t)$	0.9490***	0.9167***	0.9485**	0.9168***	0.9380***	0.9037***	0.9379***	0.9038***
	(41.53)	(40.36)	(41.46)	(40.3)	(41.39)	(43.29)	(41.39)	(43.29)
SMB_t	0.2526***	0.2522***	0.2528**	0.2524***	0.1832***	0.1828***	0.1834***	0.1829***
	(9.96)	(10.88)	(9.96)	(10.88)	(7.35)	(8.33)	(7.36)	(8.34)
HML_t	−0.0298	−0.0318	−0.0298	−0.0318	−0.0068	−0.0090	−0.0068	−0.0089
	(−1.27)	(−1.40)	(−1.26)	(−1.40)	(−0.30)	(−0.42)	(−0.30)	(−0.42)
MOM_t	0.0178	0.0136	0.0178	0.0135	0.0031	−0.0015	0.0031	−0.0015
	(0.99)	(0.78)	(0.98)	(0.78)	(0.17)	(−0.09)	(0.17)	(−0.09)
$(R_{mt} - rf_t)^2$		−0.8117**		−0.8102		−0.8725**		−0.8694
		(−2.16)		(−2.16)		(−2.15)		(−2.15)
TM_p		0.0002		−0.0010		−0.0001		−0.0012
		(0.24)		(−1.28)		(−0.18)		(−1.63)
R^2	0.964	0.966	0.964	0.966	0.958	0.961	0.958	0.961
No. of obs.	129	129	129	129	129	129	129	129

[11] We use the monthly FTSE All-Share Index as the market benchmark for all U.K. equities. We take the excess return of this index over the U.K. Treasury bill rate. SMB_t, HML_t, and MOM_t are U.K. versions of the other factor benchmarks as defined by Gregory, Tharyan, and Huang (2013).

[12] Note that the estimated factor loadings for the models where the dependent variable is based on gross returns are very similar to those in the corresponding models where the dependent variable is based on net returns. This is because the fund management fee is fairly constant over time. Although this will lead to different estimates of the intercept (α_p) in a regression equation, it will not lead to significant changes in the estimates of the slope coefficients.

the total performance measure ($\text{TM}_p = \alpha_p + \eta_p \text{Var}(R_m - rf)$) is not significant in the 5-factor model. In the latter case, although α_p can be significant (as in the case of the equal-weighted portfolio using gross returns, at the 10% level), this is more than compensated for by the significantly negative loading on $(R_{mt} - rf_t)^2$. This holds whether the portfolio is equal-weighted or value-weighted[13] and whether we use gross returns or net returns. A particularly interesting finding in Table 2 is that the estimate of α_p in the 4-factor model is very similar in size to the estimate of TM_p in the corresponding 5-factor model, even though both estimates are not statistically significant.[14] Again, this is true whether we compare on the basis of gross or net returns, or an equal- or value-weighted portfolio. This can happen, of course, only if the estimate of α_p in the 5-factor model is lower than the estimate of α_p in the corresponding 4-factor model by an amount approximately equal to the size of $\eta_p \text{Var}(R_m - rf)$.

The implication of these results is that the average equity mutual fund manager in the United Kingdom is unskilled in the sense of being unable to deliver outperformance (i.e., unable to add value from the two key active strategies of stock selection and market timing) once allowance is made for fund manager fees and for a set of common risk factors that are known to influence returns, thereby reinforcing our findings from our examination of raw returns in Table 1. But what about the performance of the best and worst fund managers? To assess their performance, we turn to the bootstrap analysis.

B. Moments of Actual, KTWW, and FF Cumulative Distribution Functions

We estimate the 4- and 5-factor benchmark models (1) and (2) across a range of subsamples ($N = 552$, 535, 516, 454, and 384) of mutual funds corresponding to the sample selection criteria of 8, 15, 20, 40, and 60 consecutive monthly time-series observations between 1998 and 2008. For each subsample, we then have a cross section of t-statistics on alpha that can be ranked from lowest to highest to form a CDF of the $\{t(\hat{\alpha}_i), i = 1, \ldots, N\}$ statistics for the actual fund alphas. We also generate 10,000 KTWW (2006) and FF (2010) bootstrap simulations for each fund as described in Section II. For each bootstrap, this will generate a cross section of t-statistics on alpha and TM (1966), assuming no abnormal performance. For the 5 subsamples, there will be 5.52 million, 5.35 million, 5.16 million, 4.54 million and 3.84 million respective t-statistics that can also be ranked from lowest to highest to create a CDF of bootstrapped "luck" $\{t(\tilde{\alpha}_i^b), i = 1, \ldots, N; b = 1, \ldots, 10{,}000\}$ statistics for each bootstrap. In Figures 1 and 2, we plot these CDFs of the t-statistics on alpha for each percentile point of the distribution for the subsamples constructed from the 552 funds with a minimum of 8 observations. The solid line in the center of Figures 1 and 2 shows the actual distribution of $t(\hat{\alpha})$ estimated for gross and net returns, respectively,

[13] The lower values of α_p and TM_p in the value-weighted regressions compared with the corresponding equal-weighted regressions indicate diseconomies of scale in fund management performance.

[14] Grinblatt and Titman ((1994), p. 438) report the same finding in their data set and argue that "the measures are similar because very few funds successfully time the market. In fact, the measures are significantly different for those funds that appear to have successfully timed the market."

FIGURE 1
CDFs: Gross Returns $t(\hat{\alpha})$

Figure 1 shows the results based on the 4-factor model in equation (1) ($i = 1, \ldots, 552$), where the dependent variable is the excess gross return. The 552 funds in this sample have at least 8 monthly observations of returns. Figure 1 shows the cumulative distribution function (CDF) of the averaged values of the actual $t(\hat{\alpha})$ statistics for the estimated alphas in this regression. The figure also shows the CDF of the averaged values of $t(\hat{\alpha})$ from 5.52 million simulations of the KTWW (2006) and FF (2010) bootstraps, together with the 5% and 95% confidence intervals. The darker gray shaded area denotes the 5%–95% confidence interval at each percentile point from the KTWW chance distribution. The lighter gray shaded area denotes the 5%–95% confidence interval at each percentile point from the FF chance distribution. The solid line denotes the CDF of the actual/estimated $t(\hat{\alpha})$, and the dashed and dotted lines are the CDFs generated by the KTWW and FF chance distributions, respectively.

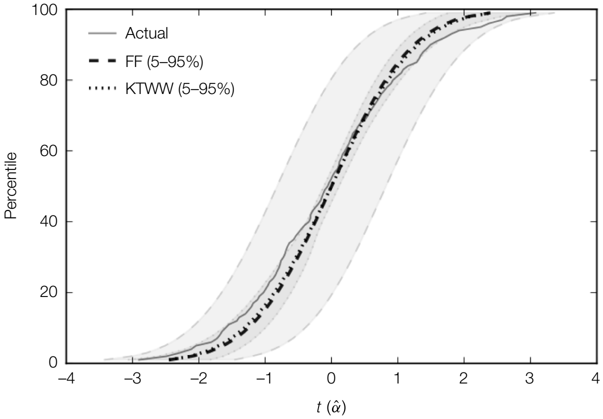
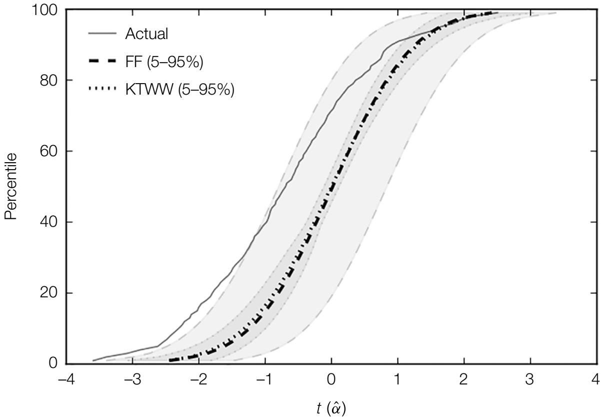


FIGURE 2
CDFs: Net Returns $t(\hat{\alpha})$

Figure 2 shows the results based on the 4-factor model in equation (1) ($i = 1, \ldots, 552$), where the dependent variable is the excess net return. The 552 funds in this sample have at least 8 monthly observations of returns. Figure 2 shows the cumulative distribution function (CDF) of the averaged values of the actual $t(\hat{\alpha})$ statistics for the estimated alphas in this regression. The figure also shows the CDF of the averaged values of $t(\hat{\alpha})$ from 5.52 million simulations of the KTWW (2006) and FF (2010) bootstraps, together with the 5% and 95% confidence intervals. The darker gray shaded area denotes the 5%–95% confidence interval at each percentile point from the KTWW chance distribution. The lighter gray shaded area denotes the 5%–95% confidence interval at each percentile point from the FF chance distribution. The solid line denotes the CDF of the actual/estimated $t(\hat{\alpha})$, and the dashed and dotted lines are the CDFs generated by the KTWW and FF chance distributions, respectively.



using the 4-factor model. The heavy dashed and dotted lines in these figures show the CDFs for the average $t(\hat{\alpha})$ values across the 10,000 simulations for the KTWW and FF bootstraps, respectively.

The moments of the actual $t(\hat{\alpha})$ and $t(\widehat{\text{TM}})$ distributions, together with key percentiles of the corresponding KTWW (2006) and FF (2010) bootstrap distributions, are shown in Table 3 for 2 subsamples: funds with a minimum of 8 and 60 monthly observations, with sample sizes of 552 and 384 funds, respectively. The factor models generate similar distributions for both gross and net $t(\hat{\alpha})$ and $t(\widehat{\text{TM}})$, with standard deviations in the 1.2–1.3 range, modest positive skewness around 0.5, and kurtosis in the 8–9 range. The KTWW bootstrap also generates similar distributions for both $t(\hat{\alpha})$ and $t(\widehat{\text{TM}})$. The distributions have (approximately) unit variance. They are also fairly symmetric and have a modest degree of excess kurtosis compared with the normal distribution. By contrast, the FF bootstrap distribution has a larger variance and much fatter tails (especially in the case of $t(\widehat{\text{TM}})$, where the left-skew is also more prominent). It also has a high level of kurtosis in the subsample formed from a minimum of 8 consecutive monthly observations, implying that the sample-selection criterion that gives the largest number of funds included in the analysis induces fat tails in the FF bootstrap simulations. Again, the most likely explanation is very poor performance prior to closure.

Table 3 also reports for each distribution the p-value from applying a Jarque–Bera (1980) test against the null hypothesis of normality. For the "actual" distribution, the hypothesis of normality was rejected in 4 out of 8 cases at the 5%

TABLE 3

Moments of CDFs of $t(\hat{\alpha})$ and $t(\widehat{\text{TM}})$: Actual, KTWW, and FF Bootstraps

Table 3 shows key moments of cumulative distribution function (CDF) for $t(\hat{\alpha})$ and $t(\widehat{\text{TM}})$ statistics from 4-factor and 5-factor models and KTWW (2006) and FF (2010) bootstraps for both gross and net excess returns. For each distribution, the table also reports p-values from applying a Jarque–Bera (JB) (1980) test against the null hypothesis of normality (a p-value below a specified significance level indicates rejection of normality at that significance level).

	$t(\hat{\alpha})$				$t(\widehat{\text{TM}})$			
	Gross Returns		Net Returns		Gross Returns		Net Returns	
Moments	Min. 8 Obs.	Min. 60 Obs.	Min. 8 Obs.	Min. 60 Obs.	Min. 8 Obs.	Min. 60 Obs.	Min. 8 Obs.	Min. 60 Obs.
Panel A. Actual Method								
No. of obs.	552	384	552	384	552	384	552	384
Mean	−0.039	0.033	−0.671	−0.702	−0.039	0.054	−0.680	−0.687
Std. dev.	1.242	1.222	1.304	1.303	1.264	1.232	1.325	1.313
Skewness	0.524	0.569	0.500	0.533	0.439	0.587	0.426	0.554
Kurtosis	8.131	7.752	9.189	9.284	8.615	7.925	9.669	9.496
JB p-value	0.004	0.007	0.055	0.103	0.023	0.009	0.172	0.100
Panel B. FF Method								
No. of obs.	5.52m	3.84m	5.52m	3.84m	5.52m	3.84m	5.52m	3.84m
Mean	−0.005	−0.001	0.012	0.003	−0.025	−0.021	−0.024	−0.025
Std. dev.	1.102	1.064	1.102	1.069	1.184	1.075	1.192	1.073
Skewness	−0.298	−0.041	0.725	−0.046	5.169	−0.093	23.657	−0.097
Kurtosis	18.130	4.345	140.882	4.419	1,679.743	4.619	12,309.389	4.591
JB p-value	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Panel C. KTWW Method								
No. of obs.	5.52m	3.84m	5.52m	3.84m	5.52m	3.84m	5.52m	3.84m
Mean	−0.006	−0.006	−0.005	−0.007	−0.009	−0.011	−0.008	−0.011
Std. dev.	1.037	1.021	1.038	1.021	1.042	1.021	1.042	1.021
Skewness	−0.024	−0.047	−0.024	−0.045	−0.096	−0.061	−0.106	−0.061
Kurtosis	4.215	3.495	4.222	3.493	9.569	3.504	7.737	3.531
JB p-value	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

significance level and in 6 out of 8 cases at 10%. In addition, this same test also clearly rejects the normality of both the KTWW (2006) and FF (2010) average $t(\hat{\alpha})$ distributions, which means that we cannot simply use the 5% and 95% confidence intervals from the entire KTWW and FF distributions to detect significant under- or outperformance. Instead, the accumulated statistical evidence suggests that we need to apply the 5%–95% confidence intervals of the KTWW and FF distributions *at each percentile point* of the actual CDF to determine abnormal performance.

C. Alpha Performance Using KTWW and FF Bootstraps

For each subsample, we compare the averaged values at selected percentiles of the CDF of the t-statistics on the actual alphas ($t(\hat{\alpha})$) with the distribution of the t-statistics derived from the KTWW (2006) and FF (2010) bootstrap simulations ($t(\tilde{\alpha}^b)$) in the same percentile ranges. We report the results of the analysis first using gross and then net returns.

1. Alpha Performance: Gross Returns

Panel A of Table 4 reports key percentiles of the CDF of the $t(\hat{\alpha})$ statistics of the cross section of funds in the subsample of 552 funds formed from a minimum of 8 observations for gross returns using the distribution of ranked t-statistics for all such funds. Figure 1 shows the same results graphically. It can be seen that the left tail of the CDF of the actual t-statistics lies to the left of that of both bootstraps. For example, in the 1st percentile range, the actual t-statistic of the worst-performing 1% of funds is −2.9043, whereas the KTWW (2006) and FF (2010) t-statistics for the same point on the distribution are −2.4516 and −2.4490, respectively. This suggests that those funds in the bottom 1% of the distribution are there as a result of poor skill rather than bad luck. This holds for most of the distribution of returns. Only for percentiles of the CDF above approximately 70% is it the case that the actual t-statistics begin to exceed those from either simulation method. For example, at the 95th percentile, the actual t-statistic is 2.2522, whereas the corresponding KTWW and FF t-statistics are 1.6830 and 1.6158. This means that those funds above the 70th percentile outperform their luck distribution, providing evidence of skill in terms of gross returns.

We can also assess the significance of the actual t-values at each percentile point of their distribution. For every percentile point of the chance distribution generated by each of the two bootstrap methods, we calculate the 5%–95% confidence intervals (CIs). This allows us to test whether the actual $t(\hat{\alpha})$ lies within the CI of each chance distribution. If the actual $t(\hat{\alpha})$ lies to the right (left) of the CI at a given percentile point, this provides robust evidence of managerial outperformance (underperformance) at that percentile point. The confidence intervals at each percentile point are reported in Table 4 in parentheses below the mean values of the KTWW (2006) and FF (2010) bootstrap values. It can be seen that the actual $t(\hat{\alpha})$ at the 1st percentile point of −2.9043 lies within the CI of both the KTWW (−3.0689, −1.8342) and FF (−3.4300, −1.4680) chance distributions, and therefore we cannot reject the null of no underperformance for the worst-performing 1% of funds. However, at the other end of the distribution, the actual $t(\hat{\alpha})$ value at the 99th percentile of 3.0773 lies to the right of the KTWW CI (1.7724, 3.0630)

TABLE 4

Percentiles of CDFs of $t(\hat{\alpha})$: Actual, KTWW, and FF Bootstraps (Selection Criterion of at least 8 Monthly Observations)

Table 4 shows percentiles of CDFs of $t(\hat{\alpha})$ estimated from the 4-factor model in equation (1) with ($i=1,\ldots,552$), where the dependent variable in these regressions is the excess gross (net) return in Panel A (Panel B). The 552 funds in this sample have at least 8 monthly observations of returns. The table shows the averaged values for selected percentiles (PCT) of the CDF of the actual $t(\hat{\alpha})$ statistics for the estimated alphas (ACT) in this regression. For the same percentiles, the table also shows the averaged values of $t(\hat{\alpha})$ from 5.52 million simulations of the KTWW (2006) and FF (2010) bootstraps (sim(KTWW) and sim(FF)), together with the 5% and 95% confidence intervals (in parentheses).

	Panel A. Gross Returns			*Panel B. Net Returns*		
PCT	ACT	sim(KTWW) (5%, 95%)	sim(FF) (5%, 95%)	ACT	sim(KTWW) (5%, 95%)	sim(FF) (5%, 95%)
1	−2.9043	−2.4516 (−3.0689, −1.8342)	−2.4490 (−3.4300, −1.4680)	−3.5897	−2.4493 (−3.0738, −1.8247)	−2.4369 (−3.3976, −1.4762)
5	−2.0434	−1.7043 (−2.0850, −1.3236)	−1.6336 (−2.5118, −0.7553)	−2.6194	−1.7043 (−2.0892, −1.3195)	−1.6172 (−2.4815, −0.7529)
10	−1.5990	−1.3149 (−1.6302, −0.9997)	−1.2577 (−2.1149, −0.4005)	−2.3073	−1.3154 (−1.6347, −0.9960)	−1.2413 (−2.0892, −0.3934)
20	−1.0258	−0.8604 (−1.1096, −0.6112)	−0.8172 (−1.6571,0.0227)	−1.7972	−0.8603 (−1.1120, −0.6085)	−0.8013 (−1.6340,0.0313)
30	−0.7248	−0.5353 (−0.7277, −0.3429)	−0.5076 (−1.3364,0.3212)	−1.3425	−0.5363 (−0.7328, −0.3399)	−0.4907 (−1.3144,0.3331)
40	−0.3404	−0.2589 (−0.3781, −0.1398)	−0.2446 (−1.0685,0.5794)	−0.9795	−0.2596 (−0.3812, −0.1380)	−0.227 (−1.0453,0.5905)
50	−0.0587	0.0003 (−0.1139,0.1145)	0.0003 (−0.8210,0.8215)	−0.6939	0.0006 (−0.1143,0.1155)	0.0178 (−0.7986,0.8342)
60	0.2022	0.2551 (0.1358,0.3743)	0.2440 (−0.5772,1.0651)	−0.3818	0.2549 (0.1351,0.3746)	0.2611 (−0.5566,1.0788)
70	0.5485	0.5281 (0.3854,0.6708)	0.5053 (−0.3182,1.3289)	−0.0599	0.5291 (0.3844,0.6737)	0.5221 (−0.3009,1.3450)
80	0.9258	0.8495 (0.6642,1.0347)	0.8139 (−0.0149,1.6427)	0.3639	0.8497 (0.6653,1.0342)	0.8297 (−0.0006,1.6599)
90	1.4930	1.2982 (1.0486,1.5479)	1.2485 (0.4064,2.0905)	0.9103	1.2987 (1.0479,1.5496)	1.2633 (0.4177,2.1089)
95	2.2522	1.6830 (1.3515,2.0144)	1.6158 (0.7532,2.4784)	1.6392	1.6837 (1.3479,2.0196)	1.6316 (0.7659,2.4974)
99	3.0773	2.4177 (1.7724,3.0630)	2.3966 (1.4288,3.3645)	2.5013	2.4184 (1.7830,3.0537)	2.4147 (1.4414,3.3880)

but within the FF CI (1.4288, 3.3645). The implication is that the top 1% of funds significantly outperform the KTWW chance distribution but not the FF chance distribution.

As can be seen from Figure 1, the 5%–95% CI at each percentile point is much wider for the FF (2010) bootstraps than the KTWW (2006) bootstraps. Further, the range of the 5%–95% CIs is relatively constant over the entire distribution of the FF bootstraps. In contrast, the 5%–95% CIs for the KTWW bootstraps are narrower over the entire distribution, and they narrow considerably around the median (which is the point of zero abnormal performance under the null). The wider CIs for the FF bootstrap are a consequence of using the same time-series observations for all funds to "capture the cross-correlation of fund returns and its effects on the distribution of $t(\hat{\alpha})$ estimates" (FF (2010), p. 1925), whereas the narrower CIs for the KTWW bootstrap are attributable to "pooling over time" (Fitzenberger and Kurtz ((2003), p. 357). Within each KTWW bootstrap, some funds' excess returns for a given time period (under the null of no abnormal performance and conditional on the realization of the common risk factors) will be drawn from a period in the data sample when there was a bull market, whereas other funds' excess returns will be drawn from a period when there was a bear market. This will result in a narrowing of the distribution of abnormal returns when averaged across a large number of bootstraps. The CIs for the KTWW bootstrap widen slightly in the tails of the distribution because in this region, the bootstrap will pick up more extreme outliers, and hence the pooling effect is reduced. By contrast, within every FF bootstrap, all funds' excess returns for a given time period (under the null of no abnormal performance) will be drawn from the same randomly selected historical period, which could be either a bull market or a bear market. With the FF bootstrap, there is no pooling over time. This results in a wider distribution of abnormal returns under the FF methodology when averaged across *the same number of bootstraps* compared with the KTWW methodology.[15]

We also investigate the effect of the sample selection criteria on the detection of significant abnormal performance. In Panel A of Table 5, we report the

TABLE 5

Percentiles of CDFs of $t(\hat{\alpha})$: Actual, KTWW, and FF Bootstraps (Selection Criterion of at least 60 Monthly Observations)

Table 5 shows percentiles of CDFs of $t(\hat{\alpha})$ estimated from the 4-factor model in equation (1) with ($i=1,\ldots,384$), where the dependent variable in these regressions is the excess gross return in Panel A and the excess net return in Panel B. The 384 funds in this sample have at least 60 monthly observations of returns. The table shows the averaged values for selected percentiles (PCT) of the CDF of the actual $t(\hat{\alpha})$ statistics for the estimated alphas (ACT) in this regression. For the same percentiles, the table also shows the averaged values of $t(\hat{\alpha})$ from 3.84 million simulations of the KTWW (2006) and FF (2010) bootstraps (sim(KTWW) and sim(FF)), together with the 5% and 95% confidence intervals (in parentheses).

	Panel A. Gross Returns			*Panel B. Net Returns*		
PCT	ACT	sim(KTWW) (5%, 95%)	sim(FF) (5%, 95%)	ACT	sim(KTWW) (5%, 95%)	sim(FF) (5%, 95%)
1	−2.7686	−2.4262 (−2.9963, −1.8560)	−2.2587 (−3.3744, −1.1430)	−3.5897	−2.4294 (−3.0005, −1.8582)	−2.2598 (−3.3830, −1.1366)
5	−1.9076	−1.6735 (−2.0627, −1.2843)	−1.5292 (−2.5601, − 0.4983)	−2.7294	−1.6742 (−2.0596, −1.2888)	−1.5295 (−2.5628, − 0.4962)
10	−1.4281	−1.3055 (−1.6442, − 0.9669)	−1.1892 (−2.2035, − 0.1750)	−2.3424	−1.3066 (−1.6417, − 0.9715)	−1.1888 (−2.2045, − 0.1730)
20	−0.9844	−0.8607 (−1.1439, − 0.5774)	−0.7820 (−1.7800,0.2161)	−1.8070	−0.8614 (−1.1376, − 0.5851)	−0.7795 (−1.7805,0.2214)
30	−0.6723	−0.5328 (−0.7645, − 0.3011)	−0.4817 (−1.4744,0.5110)	−1.4191	−0.5321 (−0.7514, − 0.3128)	−0.4789 (−1.4740,0.5161)
40	−0.2997	−0.2584 (−0.4036, − 0.1133)	−0.2323 (−1.2212,0.7565)	−1.0309	−0.2591 (−0.4002, − 0.1179)	−0.2290 (−1.2215,0.7634)
50	0.0071	0.0013 (−0.1324,0.1350)	0.0065 (−0.9814,0.9944)	−0.6945	0.0007 (−0.1313,0.1327)	0.0104 (−0.9795,1.0002)
60	0.2737	0.2535 (0.1164,0.3907)	0.2386 (−0.7490,1.2261)	−0.4239	0.2523 (0.1168,0.3878)	0.2423 (−0.7468,1.2315)
70	0.5920	0.5241 (0.3579,0.6903)	0.4869 (−0.5021,1.4760)	−0.0432	0.5233 (0.3590,0.6876)	0.4909 (−0.4989,1.4807)
80	0.9790	0.8489 (0.6407,1.0570)	0.7835 (−0.2118,1.7789)	0.3295	0.8483 (0.6402,1.0565)	0.7884 (−0.2045,1.7814)
90	1.5600	1.2876 (1.0283,1.5469)	1.1848 (0.1792,2.1905)	0.8548	1.2869 (1.0239,1.5500)	1.1906 (0.1854,2.1957)
95	2.2804	1.6474 (1.3349,1.9600)	1.5144 (0.4943,2.5346)	1.6854	1.6470 (1.3319,1.9622)	1.5223 (0.5020,2.5425)
99	2.9835	2.3783 (1.8809,2.8757)	2.2157 (1.1201,3.3113)	2.5013	2.3775 (1.8783,2.8768)	2.2251 (1.1325,3.3178)

[15] Note that KTWW ((2006), p. 2583) also consider a "block bootstrap" that samples across funds during the same time period to preserve any cross-sectional correlation in the residuals.

actual and bootstrapped $t(\hat{\alpha})$ statistics of the cross section of funds in the subsample formed from a minimum of 60 observations for gross returns. The effect of increasing the minimum number of observations for inclusion in the subsample is to shift the actual $t(\hat{\alpha})$ CDF to the right compared with Table 4. For example, at the 50th percentile point, the actual t-statistic is 0.0071 compared with -0.0587 in Table 4. This shift to the right in the $t(\hat{\alpha})$ distribution is consistent with a positive look-ahead bias in the gross returns in the more restrictive sample of 384 funds with at least 60 consecutive observations.[16] The effect of the look-ahead bias on the distribution of the KTWW (2006) and FF (2010) bootstraps is also apparent: For both bootstraps, the range of the 5%–95% CIs widens. For example, at the 10th percentile, the KTWW range widens slightly from 0.631 to 0.677, but the FF range widens noticeably more from 1.714 to 2.029. In both cases, the widening of the CIs is explained by having a smaller number of funds in this bootstrap compared with the one in Panel A of Table 5 (i.e., 384 against 552), reducing the precision of our estimates of the parameters of the underlying distribution.

2. Alpha Performance: Net Returns

Assessing alpha performance using net returns rather than gross returns raises the performance hurdle because we are now assessing whether fund managers are able to add value for their investors after covering their operating and trading costs *and* their own fees. Subtracting fees from gross returns to derive net returns will reduce the values of both the actual alphas and their t-statistics. Figure 2 shows the consequences of this graphically: the CDF of the actual t-statistics of the alphas shifts significantly to the left.[17] This is confirmed by Panel B of Table 4, in the case where the selection criterion requires a minimum of 8 monthly observations. For example, at the 5th percentile, the actual t-statistic is -2.6194, down from -2.0434 in Panel A. By contrast, there is little or no change in either the KTWW (2006) t-statistic at -1.7043 (unchanged) or the FF (2010) t-statistic at -1.6172 (up from -1.6336). Figure 2 and Panel B of Table 4 clearly show that once fund manager fees are taken into account, the actual $t(\hat{\alpha})$ either lies to the left of the CIs of the 2 chance bootstrap distributions or within the CIs themselves, but it never lies to the right, implying that no fund in our sample generated significant outperformance.

Turning to the effect of the sample selection criteria, Panel B of Table 5, where the selection criterion requires a minimum of 60 monthly observations, shows that the distribution of the actual t-statistics on net returns is not greatly affected by the increase from 8 to 60 observations, in contrast to the results for gross returns, with slight movements to the right or left at different points along

[16] This implies that tests requiring a minimum of 8 observations are more stringent than those requiring a minimum of 60 observations.

[17] The CDFs for the averaged values of both the KTWW (2006) and FF (2010) bootstrap simulations do not move significantly when there is a switch from gross to net returns. In the case of the KTWW bootstrap, this can be seen if we set $\alpha_i = 0$ in equation (5) for both gross and net returns because no other variable on the right-hand side of equation (5) changes when we make an allowance for fund manager fees. In the case of the FF bootstrap, the influence of fees is broadly cancelled out in the dependent variable $[(R_{it} - rf_t) - \hat{\alpha}_i]^b$ in equation (6) because R_{it} will be lower by the ith manager's fee, and $\hat{\alpha}_i$ will be lower by the average fee across the sample, which will be of similar size. Figures 1 and 2 show the same result graphically.

the distribution. With respect to the distributions of the 2 bootstraps, as with the gross returns, the ranges of both 5%–95% CIs widen, slightly for KTWW (2006) and more so for the FF (2010) bootstrap. For example, at the 10th percentile point, the ranges of the KTWW bootstrap widens from 0.639 to 0.670, and the range of the FF bootstrap widens from 1.696 to 2.032.

D. TM Performance Using KTWW and FF Bootstraps

We now repeat the analysis of the previous subsection, but we use the 5-factor benchmark model in equation (2) and focus on the TM (1966) total performance measure instead of alpha. Using the case of the subsample constructed on the basis of a minimum of 8 consecutive monthly observations, we report the results of the analysis first using gross and then net returns.[18]

1. TM Performance: Gross Returns

Panel A of Table 6 looks at TM (1966) performance based on gross returns. A comparison of the "ACT" column in this table with that in Panel A in Table 4

TABLE 6

Percentiles of CDFs of $t(\widehat{\text{TM}})$: Actual, KTWW, and FF Bootstraps

Table 6 shows percentiles of CDFs of $t(\widehat{\text{TM}})$ estimated from the 5-factor model in equation (2) with ($i = 1, \ldots, 552$), where the dependent variable in these regressions is the excess gross return in Panel A and the excess net return in Panel B. The 552 funds in this sample have at least 8 monthly observations of returns. The table shows the averaged values for selected percentiles (PCT) of the CDF of the actual $t(\widehat{\text{TM}})$ statistics for the estimated alphas (ACT) in this regression. For the same percentiles, the table also shows the averaged values of $t(\widehat{\text{TM}})$ from 5.52 million simulations of the KTWW (2006) and FF (2010) bootstraps (sim(KTWW) and sim(FF)), together with the 5% and 95% confidence intervals (in parentheses).

	Panel A. Gross Returns			*Panel B. Net Returns*		
PCT	ACT	sim(KTWW) (5%, 95%)	sim(FF) (5%, 95%)	ACT	sim(KTWW) (5%, 95%)	sim(FF) (5%, 95%)
1	−3.0142	−2.4717 (−3.1903, −1.7531)	−2.5900 (−3.6829, −1.4971)	−3.5963	−2.4756 (−3.2328, −1.7184)	−2.5874 (−3.6558, −1.5191)
5	−2.0419	−1.7128 (−2.1084, −1.172)	−1.6994 (−2.5981, −0.8007)	−2.7332	−1.7135 (−2.1143, −1.3128)	−1.6938 (−2.5786, −0.8091)
10	−1.6153	−1.3205 (−1.6410, −1.0001)	−1.3064 (−2.1778, −0.4351)	−2.3454	−1.3198 (−1.6393, − 1.0003)	−1.3023 (−2.1614, −0.4432)
20	−1.0288	−0.8642 (−1.1168, −0.6117)	−0.8504 (−1.6982, −0.0026)	−1.7912	−0.8634 (−1.1149, −0.6119)	−0.8476 (−1.6858, −0.0095)
30	−0.7274	−0.5395 (−0.7386, −0.3403)	−0.5327 (−1.3659,0.3004)	−1.3916	−0.5374 (−0.7336, −0.3413)	−0.5305 (−1.3567,0.2957)
40	−0.3523	−0.2619 (−0.3885, −0.1353)	−0.2646 (−1.0884,0.5593)	−0.9965	−0.2603 (−0.3830, −0.1376)	−0.2631 (−1.0822,0.5560)
50	−0.0477	−0.0016 (−0.1183,0.1150)	−0.0153 (−0.8345,0.8040)	−0.6899	−0.0006 (−0.1167,0.1156)	−0.0142 (−0.8297,0.8013)
60	0.2012	0.2529 (0.1321,0.3736)	0.2319 (−0.5855,1.0494)	−0.3813	0.2538 (0.1328,0.3747)	0.2323 (−0.5825,1.0470)
70	0.5541	0.5268 (0.3751,0.6785)	0.4962 (−0.3212,1.3135)	−0.0267	0.5272 (0.3791,0.6753)	0.4953 (−0.3202,1.3107)
80	0.9515	0.8481 (0.6553,1.0408)	0.8073 (−0.0131,1.6278)	0.3761	0.8485 (0.6576,1.0394)	0.8063 (−0.0138,1.6265)
90	1.5173	1.2961 (1.0379,1.5543)	1.2463 (0.4186,2.0739)	0.9453	1.2958 (1.0408,1.5508)	1.2449 (0.4151,2.0746)
95	2.3152	1.6794 (1.3309,2.0278)	1.6212 (0.7755,2.4668)	1.6624	1.6798 (1.3340,2.0257)	1.6196 (0.7740,2.4652)
99	3.0057	2.4118 (1.7161,3.1075)	2.4541 (1.4642,3.4439)	2.4855	2.4135 (1.7151,3.1119)	2.4474 (1.4628,3.4321)

[18] In the case of the FF (2010) bootstrap, the dependent variable in equation (6) becomes $[(R_{it} - rf_t) - \hat{\alpha}_i - \hat{\eta}_i (R_{mt} - rf_t)^2]^b$.

shows some similarity in the values of the t-statistics for the TM and alpha gross return performance measures at the same percentiles.[19] Both tables demonstrate that it is only for percentiles of the CDF above approximately 70% that the actual t-statistics exceed those from either simulation method. For example, at the 95th percentile point, the actual t-statistic is 2.3152 (compared with 2.2522 when the performance measure is alpha), whereas the KTWW (2006) average t-statistic is 1.6794 (compared with 1.6830), and the FF (2010) average t-statistic is 1.6212 (compared with 1.6158). Above the 95th percentile, the actual t-statistic significantly outperforms the KTWW chance distribution[20] but not the FF chance distribution. The regression analysis in Section IV.A produces a similar finding. We therefore have the same interpretation of this finding, namely, that only a minority of funds are able to generate returns from stock selection and market timing that are more than sufficient to cover their operating and trading costs, let alone the fund manager fee.

2. TM Performance: Net Returns

Panel B of Table 6 examines TM (1966) performance based on net returns. A comparison of the "ACT" column in this table with that in Panel B of Table 4 shows the same pattern in the values of the TM and alpha net return performance measures that the previous subsection found when looking at gross returns. There is significant underperformance at the lower end of the distribution for both bootstraps, and funds never significantly outperform either bootstrap at the upper end of the distribution. This is also shown in Figure 3.

V. Robustness Tests

An Internet Appendix (available at www.jfqa.org) provides a series of robustness tests of our findings. In particular, we report the results from varying the selection criterion from a minimum of 8 observations, through 15, 20, and 40 observations, to 60 observations. These differing sample selection criteria result in 5 subsamples of funds, with the size of the subsamples equal to 552, 535, 516, 454, and 384 funds. The bootstrap distributions are generated for both definitions of returns (gross and net) and for the 4- and 5-factor models ($t(\hat{\alpha})$ and $t(\widehat{\text{TM}})$).

In general, we find that as we increase the minimum number of observations (and reduce the number of funds) for inclusion in the analysis, the actual distribution of gross returns shifts to the right slightly. This is consistent with look-ahead bias: Funds with greater average gross abnormal performance stay longer in the data set (and vice versa). As we increase the required minimum number of observations (and reduce the number of funds) for inclusion in the analysis, both the FF (2010) and KTWW (2006) 5%–95% CIs widen, most particularly in the case of the FF bootstrap. The number of funds included in the analysis falls, reducing the precision of our estimates of the parameters of the underlying distribution and hence widening the CIs.

Under the generation of the chance distribution using the KTWW (2006) methodology, we find evidence of abnormal performance for the top-performing

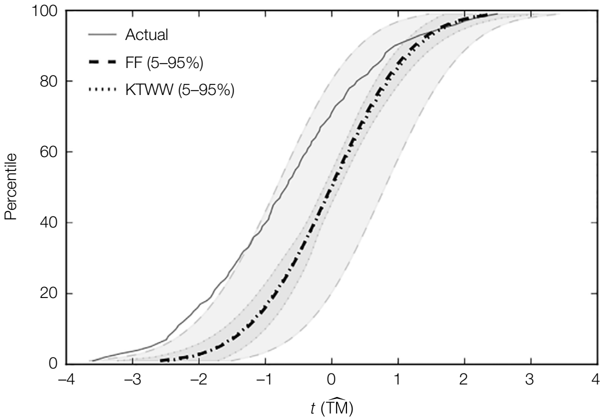
[19] For the same reason given by Grinblatt and Titman ((1994), p. 438) in footnote 14 above.

[20] Except at the 99th percentile.

FIGURE 3

CDFs: Net Returns $t(\widehat{\text{TM}})$

Figure 3 shows the results based on the 5-factor model in equation (2) ($i = 1, \ldots, 552$), where the dependent variable is the excess net return. The 552 funds in this sample have at least 8 monthly observations of returns. Figure 3 shows the CDF of the averaged values of the actual $t(\widehat{\text{TM}})$ statistics for the estimated TM values from this regression. The figure also shows the CDF of the averaged values of $t(\widehat{\text{TM}})$ from 5.52 million simulations of the KTWW (2006) and FF (2010) bootstraps, together with the 5% and 95% confidence intervals. The darker gray shaded area denotes the 5%–95% confidence interval at each percentile point from the KTWW chance distribution. The lighter gray shaded area denotes the 5%–95% confidence interval at each percentile point from the FF chance distribution. The solid line denotes the CDF of the actual/estimated $t(\widehat{\text{TM}})$, and the dashed and dotted lines are the CDFs generated by the KTWW and FF chance distributions, respectively.



funds in terms of gross returns for all selection criteria for both $t(\hat{\alpha})$s and $t(\widehat{\text{TM}})$s. In contrast, for the FF (2010) methodology for gross returns, there are no instances, irrespective of either the selection criteria or the factor model employed, of rejection of the null hypothesis of no abnormal performance. Under both methodologies, when it comes to examining net returns, there is no evidence of (positive) abnormal performance using any assessment criterion.

VI. Conclusions

Our paper contributes to the literature in two ways. First, we use a new data set of U.K. equity mutual funds to assess both the Jensen (1968) alpha and TM (1966) total performance measures of mutual fund manager skills using factor-benchmark models. TM is superior to an assessment based on alpha alone because it includes market-timing skills as well as selectivity skills; most existing studies, including KTWW (2006) and FF (2010), examine only selectivity. Second, we directly compare the KTWW and FF bootstrap methods for assessing mutual fund manager performance (both alpha and TM) using the *same* funds selected using the *same* inclusion criteria over the *same* sample period.[21] We conduct the analysis for both gross and net (of fund manager fee) returns. On the basis of a data set of

[21] FF (2010) do not reproduce the KTWW (2006) bootstrap method with their data set, although they use their own bootstrap method with the KTWW inclusion criterion and sample period to assess the KTWW method.

equity mutual funds in the United Kingdom over the period 1998–2008, we draw the following conclusions.

First, the average equity mutual fund manager in the United Kingdom is unable to deliver outperformance from either stock selection or market timing once allowance is made for fund manager fees and for the set of common risk factors known to influence returns. There is some evidence that when considered against the KTWW (2006) criterion, the top-performing fund managers do outperform in terms of gross returns. However, there is no evidence that any fund manager significantly outperforms with respect to either gross or net returns on the basis of the FF (2010) bootstrap. The TM (1966) results yield similar conclusions and indicate that the vast majority of fund managers are very poor at market timing. There is some evidence that the top-performing fund managers outperform with respect to gross returns when using the KTWW bootstrap. Any selectivity skills that fund managers might possess (and at best, only a very small number of them do) are wiped out both by their attempts to time the market and by their fees.

Our results suggest that the evaluation of fund manager performance depends crucially on the bootstrap methodology employed. In the case of gross returns, the KTWW (2006) bootstrap identifies a number of fund managers whose performance produces significant abnormal returns (as indicated by the alpha t-statistics) at certain percentiles. However, when the CIs are calculated for the FF (2010) bootstrap at the same percentiles, there is no evidence of outperformance; the CDF of the actual alpha t-statistics lies well within the FF confidence interval. For net returns, neither methodology produces any evidence of significant abnormal performance.

The explanation for this difference in findings is that within each bootstrap simulation, the KTWW (2006) bootstrap simulates fund returns and factor returns independently of each other, which means that for a given time period, some returns will be drawn for a period in the data sample when the market was bullish and some from a period when the market was bearish, whereas the latter simulates these returns jointly and hence draws all returns from the same historical time period. As a result, over a large number of simulation trials, the KTWW bootstrap will be affected by “pooling over time,” which leads to much narrower CIs than the FF (2010) bootstrap.

Taken together, these results provide powerful evidence that the vast majority of fund managers in our data set were not simply unlucky; they were genuinely unskilled. Although a few “star” fund managers appear to have sufficient skills to generate superior gross performance (in excess of operating and trading costs), they extract the whole of this superior performance for themselves via their fees, leaving nothing for investors.

References

Blake, D.; A. Rossi; A. Timmermann; I. Tonks; and R. Wermers. “Decentralized Investment Management: Evidence from the Pension Fund Industry.” *Journal of Finance*, 68 (2013), 1133–1178.

Blake, D., and A. Timmermann. “Mutual Fund Performance: Evidence from the UK.” *European Finance Review*, 2 (1998), 57–77.

Brown, S., and J. Warner. “Using Daily Stock Returns: The Case of Event Studies.” *Journal of Financial Economics*, 14 (1985), 1–31.

Carhart, M. M. “On Persistence in Mutual Fund Performance.” *Journal of Finance*, 52 (1997), 57–82.

Carhart, M. M.; J. N. Carpenter; A. W. Lynch; and D. K. Musto. “Mutual Fund Survivorship.” *Review of Financial Studies*, 15 (2002), 1439–1463.

Carpenter, J. N., and A. W. Lynch. “Survivorship Bias and Attrition Effects in Measures of Performance Persistence.” *Journal of Financial Economics*, 54 (1999), 337–374.

Cuthbertson, K.; D. Nitzche; and N. O’Sullivan. “UK Mutual Fund Performance: Skill or Luck?” *Journal of Empirical Finance*, 15 (2008), 613–634.

Efron, B., and R. J. Tibshirani. *An Introduction to the Bootstrap*. New York, NY: Chapman and Hall (1993).

Elton, E. J.; M. J. Gruber; and C. R. Blake. “Survivor Bias and Mutual Fund Performance.” *Review of Financial Studies*, 9 (1996), 1097–1120.

Fama, E. F., and K. R. French. “Common Risk Factors in the Returns on Stocks and Bonds.” *Journal of Financial Economics*, 33 (1993), 3–56.

Fama, E. F., and K. R. French. “Luck versus Skill in the Cross-Section of Mutual Fund Returns.” *Journal of Finance*, 65 (2010), 607–636.

Ferson, W. E., and R. W. Schadt. “Measuring Fund Strategy and Performance in Changing Economic Conditions.” *Journal of Finance*, 51 (1996), 425–461.

Fitzenberger, B., and C. Kurtz. “New Insights on Earnings Trends across Skill Groups and Industries in West Germany.” *Empirical Economics*, 28 (2003), S479–S514.

Gregory, A.; R. Tharyan; and A. Huang. “Constructing and Testing Alternative Versions of the Fama–French and Carhart Models in the UK.” *Journal of Business Finance and Accounting*, 40 (2013), 172–214.

Grinblatt, M., and S. Titman. “A Study of Monthly Mutual Fund Returns and Performance Evaluation Techniques.” *Journal of Financial and Quantitative Analysis*, 29 (1994), 419–444.

Jarque, C., and A. Bera. “Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals.” *Economics Letters*, 6 (1980), 255–259.

Jensen, M. C. “The Performance of Mutual Funds in the Period 1945–1964.” *Journal of Finance*, 23 (1968), 389–416.

Khorana, A.; H. Servaes; and P. Tufano. “Mutual Fund Fees around the World.” *Review of Financial Studies*, 22 (2009), 1279–1310.

Kosowski, R.; A. Timmermann; R. Wermers; and H. White. “Can Mutual Fund ‘Stars’ Really Pick Stocks? New Evidence from a Bootstrap Analysis.” *Journal of Finance*, 59 (2006), 2551–2595.

Lunde, A.; A. Timmermann; and D. Blake. “The Hazards of Mutual Fund Underperformance: A Cox Regression Analysis.” *Journal of Empirical Finance*, 6 (1999), 121–152.

Malkiel, B. G. “Returns from Investing in Equity Mutual Funds 1971 to 1991.” *Journal of Finance*, 50 (1995), 549–572.

Merton, R. C., and R. D. Henriksson. “On Market Timing and Investment Performance II: Statistical Procedures for Evaluating Forecasting Skills.” *Journal of Business*, 54 (1981), 513–534.

Politis, D. N., and J. P. Romano. “The Stationary Bootstrap.” *Journal of the American Statistical Association*, 89 (1994), 1303–1313.

Treynor, J., and K. Mazuy. “Can Mutual Funds Outguess the Market?” *Harvard Business Review*, 44 (1966), 131–136.

Wermers, R.; L. Barras; and O. Scaillet. “False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas.” *Journal of Finance*, 65 (2010), 179–216.

White, H. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, 48 (1980), 817–838.