Pilbeam, Keith; Preston, Hamish

Article

An empirical investigation of the performance of Japanese mutual funds: Skill or luck?

International Journal of Financial Studies

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Pilbeam, Keith; Preston, Hamish (2019) : An empirical investigation of the performance of Japanese mutual funds: Skill or luck?, International Journal of Financial Studies, ISSN 2227-7072, MDPI, Basel, Vol. 7, Iss. 1, pp. 1-16, https://doi.org/10.3390/ijfs7010006

This Version is available at:
https://hdl.handle.net/10419/195758

International Journal of
Financial Studies

MDPI

Article

An Empirical Investigation of the Performance of Japanese Mutual Funds: Skill or Luck?

Keith Pilbeam [1,*] and Hamish Preston [2]

[1] Department of Economics, City, University of London, Northampton Square, London EC1V 0HB, UK
[2] S&P Dow Jones Indices, 55 Water Street, New York, NY 1004, USA; hamish.preston@spglobal.com
* Correspondence: k.s.pilbeam@city.ac.uk; Tel.: +44-20-7040-0258

Received: 23 October 2018; Accepted: 9 January 2019; Published: 17 January 2019

Abstract: This paper assesses the performance of 355 actively managed Japanese Equity Mutual Funds between April 2011 and April 2016. The equal weight portfolio and Jensen's alpha measures of active management provide strong evidence that Japanese Mutual Funds fail to outperform the benchmark four-factor capital asset pricing model. When it comes to market timing, the Treynor and Mazuy measure shows that 33 funds have significant positive market timing ability which is largely offset by 31 funds with significant negative timing ability. To ensure the statistical inference is robust to the non-normality found in 33 funds we employ Fama and French's cross-sectional bootstrap. The results show that a large proportion of funds fail to outperform a hypothetical world with no skill. On the persistence of skill we find that there is stronger persistence for poor performing funds than for strong performing funds.

Keywords: mutual fund performance; bootstrap; Jensen's alpha; Fama and French model

JEL Classification: C15; G11

1. Introduction

Mutual funds allow investors to benefit from diversified portfolios by grouping investments together and these investment groupings typically differ according to investment style. Owing to the purported skill involved in actively managing investors' money, active mutual funds typically charge higher fees than exchange-traded funds (ETFs) that seek to mimic the market returns. Despite a plethora of literature being available for US mutual funds, relatively little focus has been given to Japanese mutual funds[1].

This paper contributes to the literature by using popular performance measures to infer the active managers' skill level in Japan[2] by analyzing Japanese mutual fund performance over the period April 2011 and April 2016, This paper considers gross returns that include trading costs, management fees, and expenses, therefore the tests of skill enables us to determine if fund managers had sufficient skill to beat passive, diversified benchmarks.

The equal weight (EW) measure[3] of fund performance provides an analysis of the average fund returns and gives an indication of the abnormal performance of the Japanese mutual fund market

[1] For clarity, Japanese mutual funds are defined as unit trusts on Datastream owing to the different legal structure in place. However, this difference does not materially impact the objective of active managers therefore we refer to the Japanese unit trusts as Japanese mutual funds for the remainder of the paper.
[2] Some research has been done into Japanese mutual fund performance, notably Cai et al. (1997), but such papers focus on returns from earlier periods and do not use some of the performance tests in this paper.
[3] All the performance measures, except the persistence measure, use *t*-statistics to discern between performance driven by luck or skill.

relative to the stated benchmark.[4] The measure is also useful for understanding the factors driving cross-sectional returns, thus establishing the four-factor CAPM model as a valid benchmark.

This paper adds to the existing literature by discussing the interaction between stock-picking and market timing abilities. The stock picking ability of each fund is analyzed using the Jensen's alpha measure, defined as the "average incremental rate of return on the portfolio per unit of time which is solely due to the managers' ability to forecast future security prices" Jensen (1968). Forecasting security prices is analogous to stock picking since managers would not buy stocks if they forecast them to perform poorly. If a fund's alpha is positive (negative) this implies that managers exhibit good (poor) stock selection ability. The greater granularity of this measure compared to EW also means the distribution of funds' alphas can be examined.

To measure the market timing ability of funds we use the Treynor and Mazuy (1966) test on each fund; excess fund returns are expected to be a convex function of the market's excess returns if a fund manager has timing ability. The rationale for this follows from Bollen and Busse (2001) definition:

> Market timing refers to the dynamic allocation of capital among broad classes of investments, often restricted to equities and short-term government debt. The successful market timer increases portfolio weight on equities prior to a rise in the market and decreases the weight prior to a fall in the market.

An additional contribution to the existing literature on the performance measures is the use of Fama and French (2010) cross-sectional bootstrap. No studies to our knowledge have thus far used this method to assess Japanese funds. This methodology is used because the *t*-statistics in the aforementioned tests rely on the regression residuals being normally distributed. There are many reasons why this assumption fails for mutual fund returns; Kosowski et al. (2006) offer several explanations such as many fund managers holding large positions in specific stocks/sectors use dynamic risk strategies in response to a change in their relative rankings. Fund managers' incentive to change risk is dependent on their year-to-date returns. In their well-known study, Chevalier and Ellison (1997) provide strong evidence of non-normality, it is for this reason that we employ the Royston (1982) version of the Shapiro and Wilk (1965) test for normality.

Analyzing the persistence of fund returns enables one to draw conclusions about whether skill can be identified based on prior returns. This is important because investors may believe they can overcome a scarcity in active managers' skill levels if they are able to identify the best skilled managers based on previous performance and also avoid poor performance if there is persistence in poor performing funds. We focus on the extreme 60 funds (the 30 best and 30 worst performing by Jensen's alpha measure) across different formation periods.[5]

The rest of this paper is structured as follows: Section 2 provides a literature review, Section 3 outlines the dataset used in this study and Section 4 looks at the key equal weight regression analysis employed in our empirical research. Section 5 looks at Jensen's alpha and Section 6 looks at the Treynor and Mazuy measure for market timing. In Section 7 we undertake tests for normality while in Section 8 the Fama and French (2010) bootstrap method is outlined along with the associated results for our selection of Japanese funds. Section 9 looks at the issue of persistence in fund performance and Section 10 concludes.

2. Literature Review

Jensen (1968) created the Jensen's alpha measure to determine whether active managers had skill in forecasting security prices. In his study, the average returns of 115 mutual funds between 1955

---

[4] Throughout this paper, abnormal performance and alpha are used interchangeably.

[5] One might question why the paper departs from using *t*-statistics. The reason is the underlying non-normality in many funds' residuals makes inferences from *t*-statistics potentially misleading, whereas the parameter estimates themselves do not depend upon normality to be unbiased and consistent.

and 1964 showed no such skill, although this result may be impacted by survivorship bias. Jensen also concluded that there was little evidence of any individual fund being able to do significantly better than the random buy-and-hold the market strategy.

Building on the market timing test methodology developed by Treynor (1965), Treynor and Mazuy (1966) analyzed a sample of 57 mutual funds between 1953 and 1962. The authors found that that only one fund had the desired convex curvature in characteristic lines consistent with market timing; the other 56 fund managers appeared to have no market timing ability. Whilst Treynor and Mazuy mention that this does not rule out managers delivering positive abnormal performance, their results indicate that this performance derives from forecasting security prices rather than timing the market. The results from Treynor and Mazuy (1966) are robust to the time period studied and the test used. Henriksson (1984) used the market timing measure of Henriksson and Merton (1981) to show only 3 funds out of 116 exhibited significant positive market timing between 1968 and 1980. Similarly, Graham and Harvey (1996) used the suggested allocations between cash and equity from 326 distinct funds' investment newsletters and found no significant evidence of timing ability over the period 1983 to 1995.

A more recent development involves increasing the frequency of the data used in the tests. Goetzmann et al. (2000) explain that monthly data may not capture market timing ability because decisions on market exposure are likely made over shorter horizons for many funds. Bollen and Busse (2001) show that the use of daily data reveals that more funds exhibit timing ability, raising concerns about the aforementioned papers' market timing results, all of which used monthly (or annual) data.

Another issue for the research on market timing is that the estimation techniques did not allow for variation in mutual funds' risks and risk premia. Jensen et al. (1972) recognized this and examined changes to intercept and slope parameters over several different sub-periods. Such studies, including Grant (1977), interpreted any changes in parameter estimates as reflecting superior information or market timing ability. Ferson and Schadt (1996) question this assessment; they argue that alpha pertaining to an active portfolio strategy that can be replicated using readily available public information should not be interpreted as superior performance. By incorporating conditional models into the analysis of mutual fund performance and timing ability for 67 mutual funds between 1968 and 1990, Ferson and Schadt (1996) find that the perverse market timing evidence was removed and the average abnormal performance became centered around zero rather than negative.

With respect to Japanese mutual fund performance, Cai et al. (1997) show that Jensen's alpha results are not sensitive to the use of conditional factor models (or to the choice of benchmark). Their results show that funds' alpha remained negative and significant in all cases over the period 1981–1992. Whilst this underperformance of the average active manager supports the equilibrium accounting discussed in Fama and French (2010), which in turn is based on Sharpe (1991), some evidence does exist of outperformance. Otten and Bams (2002) use both conditional and unconditional four-factor models to assess the net returns of 506 mutual funds from the five largest European countries covering 85% of European mutual fund assets. They find evidence of outperformance in their survivorship bias free study in four of the countries namely, France, the UK, Netherlands, and Italy and a negative alpha only in the case of Germany.

More recently, Fama and French (2010) and Kosowski et al. (2006) have raised concerns that all inferences arising from the significance of parameter estimates rely on the assumption of normality. With many return profiles failing the normality assumption, a bootstrapping methodology has been developed to assess whether skilled funds exist without identifying the particular funds. This ensures an ex-ante distribution is not imposed on the returns distribution. A further development has been to allow parameter estimates to change over time; by incorporating such changes in the Fama and French (2010) bootstrap methodology. Ercolani et al. (2018) find substantial evidence of time variation in US mutual funds' parameter estimates and their results reveal little evidence of performance differing from the benchmark.

Some researchers have highlighted the need to be cautious when using test statistics, simulated or otherwise. Cuthbertson et al. (2012) analyze the false discovery rate (FDR) of UK funds based upon

the significance of alpha. They find that a large percentage of the outperforming funds delivered significant performance by chance. By correcting for so-called 'false positives' their method provides another way to assess the mutual fund industry as a whole.

An additional assessment of active management incorporates the persistence of funds' abnormal returns relative to their peers; if there is low persistence then abnormal returns are likely driven by luck than skill. Whilst the majority of evidence shows past relative performance is a poor predictor of future fund performance, Carhart (1997) found evidence of return persistence in US mutual funds. In light of the general poor predictability of returns, many academics have concluded this supports market efficiency, see for example, Malkiel (1995). That many investors continue using active management, when skill appears to be scarce and unpredictable, remains largely unexplained. One possible set of explanations relies upon behavioral issues of investors outlined by Kahneman and Riepe (1998) and Cuthbertson et al. (2016).

3. The Data

The fund returns in this paper are derived from the monthly adjusted-close prices (in US dollars) between April 2011 and April 2016 on Japanese Mutual Funds from the DataStream mutual fund database. The use of adjusted-close prices ensures the returns include management fees and trading costs, and the returns account for splits and/or mergers of funds across our sample period. Hence, we are able to directly determine if managers' skill is sufficient to outperform the stated benchmark without having taken into account the impact of each fund's fee structures on performance relative to the benchmark—Fama and French (2010) point out that trading cost estimates may be subject to large errors owing to the time variation in costs.

The data is restricted to those funds that are based in Japan and invest primarily in Japanese equity markets.[6] To ensure we focus on active funds whose objective it is to outperform the Japanese equity market, tracker funds are removed from our study. Failing to remove these funds would skew our results towards not finding significant outperformance.[7]

One of the issues with the DataStream database is survivorship bias; funds that cease to exist during the sample period are not included in the dataset. To ensure this paper's results are not distorted by funds with small returns histories, all funds contained in the dataset are alive for the entire sample period of 61 months. This left us with 355 funds to analyze[8].

The four-factor capital asset pricing model (CAPM) is used as the benchmark because including Fama and French (1993) size and book-to-market, along with Carhart (1997) momentum factors have been shown to capture a large fraction of Sharpe's (1964) single-factor CAPM anomalies. Using a four-factor model ensures fund performance does not arise from fund managers exploiting these well-known anomalies.[9] The monthly returns (in US dollars) on this benchmark are constructed from the Japanese data on Kenneth French's online factor library; details of which can be found on the factor library website.

For those concerned that tests of performance are sensitive to the chosen benchmark, as found in papers such as Grinblatt and Titman (1992), it should be noted that Cai et al. (1997) results on Japanese mutual fund performance were robust to the choice of benchmark.

---

6 Datastream defines equity funds as those with more than 50% of its assets invested in equities.

7 To remove such funds, names including "Index", "IDX", "ETF", "tracker fund" and any other permutations are highlighted. A search of these funds' objectives was run to verify their omission from the dataset.

8 Another possible issue is incubation bias, which arises from how funds may be "created" prior to opening a fund for public investment. The funds with the best returns are the ones most likely to be made available to the public. Despite the returns profile seeming impressive compared to a benchmark, there is a good chance that the performance arose from luck rather than skill. As Fama and French (2010) find only trivial effects of incubation bias and recent papers such as Barras et al. (2010) and Cuthbertson et al. (2008) make no particular adjustment for it we do not make an adjustment either.

9 As Bollen and Busse (2001) point out, an alternative method to deal with the issue of benchmark efficiency would be to use the stochastic discount factors as in Chen and Knez (1996); Dahlquist and Söderlind (1999); and Farnsworth et al. (2002). We leave it to future research to determine if using this alternative method leads to a material difference in the results.

4. The Equal Weight Portfolio

The first measure to assess the performance of Japanese mutual funds relative to the four-factor benchmark is the equal weight (EW) portfolio. The EW portfolio returns are calculated by averaging the 355 fund returns in each month and subtracting the corresponding one-month US T-bill rate.[10] This yields the average fund return in excess of the (US) risk free rate in each month, t.

The following time series regression was run:

$$R_t = \alpha + b(R_{Mt} - R_{ft}) + s(SMB_t) + h(HML_t) + m(MOM_t) + \varepsilon_t$$

In each month, t, R_t is the EW average excess return; R_{Mt} – R_{ft} is the excess market return; and SMB_t and HML_t are the Japanese size and value factors, respectively, calculated from Fama and French (1993). MOM_t is the Japanese momentum factor from Carhart (1997). Finally, α is the average fund return that is left unexplained by the benchmark model. ε_t is the regression residual. The alpha or intercept in the above regression is assumed to be constant which is consistent with the model of manager skill developed by Berk and Green (2004).

An explanation of the three factors we use in our regression is as follows: Small minus big (SMB)—this factor is also referred to as the "small firm effect," or the "size effect," where size is based on a company's market capitalization. In other words, if a portfolio has more small-cap companies in it, it should outperform the market over the long run. High minus low (HML)—'high' refers to companies with a high book value to market value ratio while 'low' refers to companies with a low book value to market value ratio. This factor is also referred to as the "value factor" or the "value versus growth factor" because companies with a high book to market ratio are typically considered value while companies with a low market to book value are typically growth stocks. Research has shown that value stocks tend to outperform growth stocks in the long run. So, over the long run, a portfolio with a large proportion of value stocks should outperform one with a large proportion of growth stocks. Momentum (MOM)—refers to funds with strong past performance continue to outperform funds with poor past performance in the next period. It is calculated by taking the average return on the two high prior return portfolios minus the average return on the two low prior return portfolios when the portfolios have been divided up into 6 portfolios based upon prior 12 month returns.

Following other papers assessing Japanese mutual fund performance, such as Cai et al. (1997), it is expected that the b coefficient will be positive; investors are rewarded for the non-diversifiable systematic risk they hold. Research into the interaction between momentum and value in Japan by Asness (2011) and Fama and French (2012) suggests that m will be negative and h will be positive; historically momentum has not been rewarded in Japan and there exists an historical negative correlation between momentum and value. It is also expected that s will be negative given that Fama and French (2010) find the value premia in average stock returns increase with size in Japan. If these results should materialize, it indicates that the average fund excess return is positively driven by the market and value premia but is negatively influenced by the momentum anomaly and the size factor. This outcome could suggest there were more managers employing value and reversal strategies, along with those strategies based on market risk premia, and that fund managers tended to invest in large companies compared to smaller ones.

There is debate over whether the average SMB, HML, and MOM returns at each time period exist as rewards for risk or mispricing, we do not take a view on this matter. As in Fama and French (2010), the interpretation of the four-factor returns included on the right-hand side of the regression is that they represent returns from diversified portfolios that capture the patterns of average returns between April 2011 and April 2016.

---

[10] The monthly US T-bill rate is also taken from Kenneth French's factor library.

To determine good or bad performance this paper invokes Dybvig and Ross (1985) analysis. Theorem 5 in their paper states that if α is positive (negative), a portfolio that places a positive (negative) weight on the benchmark portfolio of funds along with a positive weight on the passive benchmark portfolios will generate returns with a higher Sharpe ratio than simply having a positive weighting on the portfolios of benchmark returns. Since the Sharpe ratio discounts excess returns by the non-diversifiable risk taken by an investor in achieving those excess returns, α therefore indicates if the equal weight portfolio, on average, produces risk-adjusted returns different from what would be expected by simply having exposure to the benchmark portfolios. Table 1 reports the results over the entire sample period; the constant term refers to α.

Table 1. Summary statistics for equal weight regression.

	R_M-R_f	*SMB*	*HML*	*MOM*	*Constant*
Coefficient	0.5936	0.2646	−0.4059	−0.2910	−0.0028
t-statistic	8.55	2.18	−2.77	−3.08	−1.12
Adjusted R-squared	0.5418				
F-test statistic	18.74				

Of the four-factors in the benchmark model, all are significant at the 5% level,[11] with the market excess return factor being especially significant with a *t*-statistic of 8.55. All the factor coefficients have the expected sign, indicating that the average excess returns are driven as in other papers. The fact the F-test statistic is significant at all levels indicates that the four-factor model is reasonable at describing the average excess fund returns in our sample period.

The negative alpha indicates that the EW portfolio has delivered average monthly excess returns below what was obtained from the benchmark portfolios. As the corresponding *t*-statistic is not significant, the performance of the EW portfolio did not differ significantly from the four-factor model in the sample. This result is similar to that of Cai et al. (1997); their EW regressions always delivered a negative alpha, although the estimated alphas in their paper were always significant and the adjusted R^2 reported for their EW regressions between 1981 and 1992 were higher. The similarities in the results suggests that the average fund performance still delivered monthly returns below what could be achieved from the benchmark portfolios. The difference in significance on the alphas may suggest that the portfolio of funds became slightly better in delivering average excess returns, which were not statistically different from the benchmark.

5. Jensen's Alpha Measure

As outlined by Jensen (1968), the setup is very similar to the EW portfolio. For each fund i, the following regression is run over the entire 61 months, correcting the residuals for the observed heteroskedasticity[12]

$$R_{it} = \alpha_i + b_i(R_{Mt} - R_{ft}) + s_i(SMB_t) + h_i(HML_t) + m_i(MOM_t) + \varepsilon_{it}$$

R_{it} is the excess return for fund i at time t. The factors remain the same as in the EW portfolio regression; the subscripts on the parameter estimates reflect the 355 sets of estimates that are obtained. Since Jensen's alpha assesses each fund separately, the interpretation of α becomes whether fund managers have an ability to forecast security prices. Of the 355 estimates for alpha, 53 are positive and 302 are negative. Table 2 below shows the distribution of these alphas, along with how many alpha estimates are significant for each range of alpha estimates.

[11] The 5% level is based on conducting five separate two-tail tests on whether each coefficient is equal to zero.
[12] Calculating robust standard errors is necessary since 23 of 355 funds' residuals exhibit heteroskedasticity at the 5% level across the entire sample period. These results come from undertaking a Breusch-Pagan test for heteroskedasticity.

Table 2. Distribution of Jensen's alpha estimates across 355 funds between April 2011 and April 2016.

Estimated Alpha	Frequency	Number of Significant Alpha Estimates by Category
$\alpha \leq -0.024$	0	0
$-0.024 < \alpha \leq -0.022$	1	1
$-0.022 < \alpha \leq -0.020$	1	1
$-0.020 < \alpha \leq -0.018$	1	1
$-0.018 < \alpha \leq -0.016$	3	3
$-0.016 < \alpha \leq -0.014$	6	4
$-0.014 < \alpha \leq -0.012$	10	7
$-0.012 < \alpha \leq -0.010$	6	4
$-0.010 < \alpha \leq -0.08$	8	2
$-0.08 < \alpha \leq -0.06$	19	2
$-0.06 < \alpha \leq -0.04$	34	2
$-0.04 < \alpha \leq -0.02$	117	0
$-0.02 < \alpha \leq 0.00$	96	0
$0.00 < \alpha \leq 0.02$	35	0
$0.02 < \alpha \leq 0.04$	12	0
$\alpha > 0.04$	6	0
Total	**355**	**27**

This indicates that funds were more likely to deliver negative than a positive alpha in monthly returns. To ascertain if these alphas differ significantly from the returns offered by the benchmark portfolios, we turn to the heteroskedastic-adjusted *t*-statistics on the 355 alpha estimates.

Only 27 funds' alphas are significant at the 5% level and all of these funds have negative alpha estimates.[13] 328 alpha estimates (and all of the positive abnormal performance) are not statistically different from zero, meaning the returns from these funds are very similar to the returns from the diversified portfolio. Our results show that there is strong evidence that funds were more likely to underperform the benchmark model than to beat it.

The asymmetry in the fund alpha supports the results from the EW section; the non-significant alpha estimate for EW portfolio manifests itself in the non-significant alpha of many individual fund alphas. The added insight from the Jensen's alpha measure is that statistically significant fund returns correspond entirely to underperformance of the benchmark model.

For an investor contemplating using active mutual funds the results are worrisome; any positive abnormal performance relative to the benchmark model is not discernable from luck whereas there appears to be evidence of statistically significant bad skill. This outcome arises even with the presence of survivorship bias, where the results are likely to be skewed in favor of finding funds that outperform the benchmark. Given previous papers such as Carhart (1997) have found that US mutual funds with poor past performance are more likely to perish or remain poorly performing funds than successful funds are likely to remain successful, the above results arguably represent an upper-bound for managers' true skill.[14]

The results are even worse from a practical perspective because active fund managers charge higher fees than passive alternatives. Given the inability of funds to generate significant outperformance in a manner conducive to the presence of skill in forecasting security prices, there is little evidence from this dataset to support paying higher fees in order to beat the benchmark model. Paying such fees appears to be accepting performance that is mainly very similar to the benchmark model, or significantly underperforms the benchmark.

---

[13] The 5% significance is based on undertaking a two-tail test of a particular fund's alpha being different from zero against the null hypothesis that alpha for the same fund is zero.

[14] This assumes the forces driving the dynamic found in Carhart (1997) apply to Japanese funds, or that the funds excluded from the analysis were more likely to have performed poorly than been successful over the sample period.

6. Treynor and Mazuy Measure for Market Timing

The method we use to test for market timing is based upon the Treynor and Mazuy (1966) paper.[15] In their paper, the following regression was used

$$r_{p,t} = \alpha_p + \beta_p r_{m,t} + \gamma_p r_{m,t}^2 + \varepsilon_{p,t}$$

At time t, $r_{p,t}$ is the excess return on the portfolio, $r_{m,t}$ is the excess return on the market and γ_p measures timing ability. If managers have timing ability, we would expect γ_p to be positive and for the portfolio's return to be a convex function of the market's return. The change in this paper, as performed by Bollen and Busse (2001), is to include the four-factor model in the setup so as to not reward managers for exploiting well-known anomalous returns. Hence, the regression becomes

$$r_{p,t} = \alpha_p + \sum_{i=1}^{4} \beta_{p,i} r_{i,t} + \gamma_p r_{m,t}^2 + \varepsilon_{p,t}$$

Each portfolio, p, describes a particular fund's returns. The betas in the summation term are the four-factors in the benchmark model, with the corresponding returns to these factors being given by the $r_{i,t}$. As in the Jensen's alpha section, the regression for each fund is run over the entire 61 months and corrects for the observed heteroskedasticity in the residuals. Of the 355 gamma estimates, 259 are negative and 96 are positive. Table 3 shows the distribution of these estimates.

Table 3. Distribution of market timing ability for each of the 355 funds between April 2011 and April 2016.

Estimated Gamma γ	Frequency of Gamma	Number of Significant Gamma Estimates
$-8 \leq \gamma$	2	2
$-8 < \gamma \leq -7$	8	7
$-7 < \gamma \leq -6$	5	4
$-6 < \gamma \leq -5$	8	6
$-5 < \gamma \leq -4$	12	7
$-4 < \gamma \leq -3$	11	2
$-3 < \gamma \leq -2$	40	1
$-2 < \gamma \leq -1$	104	2
$-1 < \gamma \leq 0$	69	0
$0 < \gamma \leq 1$	38	0
$1 < \gamma \leq 2$	21	5
$2 < \gamma \leq 3$	34	26
$3 < \gamma \leq 4$	1	1
$\gamma > 4$	2	0
Total	**355**	**63**

The results indicate that 73.2% of the funds exhibit no market timing ability; the remaining 26.8% of funds show some timing ability.[16] The heteroskedastic-adjusted t-statistics show that 31 of the 259 negative gamma estimates are significant whereas 32 of the 96 positive gamma estimates are significant both at the 5% significance level.[17] As a result, any significant market-timing ability is as likely to be good as it is bad.

---

[15] Another approach to market timing is the study of Jagannathan and Korajczyk (1986).

[16] These percentages are calculated as follows (2 + 8 + 5 + 8 + 12 + 11 + 40 + 104 + 69)/355 and (38 + 21 + 34 + 1 + 2)/355 from Table 3.

[17] The two-tail test carried out is based on the null hypothesis that a particular fund's gamma equals zero against the alternative hypothesis that the same fund's gamma differs from zero.

It is perhaps unsurprising that many mutual fund returns exhibit no significant timing ability. Bollen and Busse (2001) point out that mutual fund managers can often be constrained by the investment objectives of the fund and are also restricted in their use of leverage and derivatives by regulators.

In addition, the momentum factor returns being included within the benchmark may be a driving force behind many funds failing to exhibit timing ability. Intuitively, market timing is similar to a momentum strategy because timing involves buying the market portfolio when market returns increase; the momentum factor returns involves buying stocks whose price has increased recently. Given the significance of the momentum factor in explaining the cross-sectional fund returns, it is reasonable to suppose that our method underestimates the true market timing ability of fund managers because much of this ability is encapsulated by the momentum factor returns.

Exacerbating this issue is the use of monthly data in returns. Goetzmann et al. (2000) explain that monthly frequency may fail to capture a manager's timing abilities because timing decisions are likely made on a much more frequent basis.[18] Given that Bollen and Busse (2001) found the power of timing tests to be higher for daily data than monthly data, our results likely provide a lower bound for the number of managers with true timing ability. Such a lower bound in timing ability makes the bad-skill in security selection more worrying; if more managers have timing ability yet the stock picking assessment is robust to data frequency, there is likely to be more fund managers who can identify market trends but fail to select securities to deliver significant positive returns thereafter. To illustrate this point, only 6 of the 33 funds with a positive, significant gamma estimate have a positive Jensen's alpha estimate. The other 27 funds have negative alpha estimates and none of these alphas is significant at the 5% level. This suggests that investors should be wary of paying active managers for market timing ability because this ability is unlikely to be transmitted through higher abnormal returns.

Our results also indicate that negative timing ability is associated with poor stock picking ability; only 10 of the 31 funds that possess significant negative timing ability (referred to as bad skill) have a positive (but not significant) alpha. The remaining 21 funds delivered negative alpha, with only one alpha estimate being significant at the 5% level. Consequently, bad skill appears to be transferable, which suggests investors should have been cautious of thinking active managers could overcome bad market timing ability with good stock picking skill.

One potential explanation for the above dynamic is that some managers were too confident in their ability to select securities based on forecasting prices (even when there is no evidence of significant skill) and as such they were reluctant to change their investments as market trends evolved, resulting in significant negative timing ability. Another possible interpretation is that of Braun et al. (2012) who, in the context of open-end life settlement funds, argue that it is possible that fund managers are only interested in collecting fees and not really in the fund's long-term performance, this is especially the case when fees are paid upfront and there can be penalties for early redemption

7. Testing for Normality

We use the Shapiro and Wilk (1965) test to carry out a test of normality. Whilst alternative tests are available, such as the Anderson and Darling (1952); Lilliefors (1967); and Kolmogorov–Smirnov (Massey 1951) tests, it has been shown via Monte Carlo simulation that the Shapiro–Wilk test has the highest power for a given significance level, see Razali and Wah (2011). Using the following hypotheses:

Hypothesis H0: The distribution is normal.

Hypothesis H1: The distribution is not normal.

---

[18] For those concerned of a frequency issue in the results from the Jensen's alpha section, standard tests relating to selection ability are theoretically robust to observation frequency because the estimate is more a function of sample length. See Goetzmann et al. (2000) for more details.

A regression is carried out of each fund on the benchmark four-factor model over the 61 months, calculating robust residuals in each case. Each funds' residuals are saved and ordered from smallest to largest. The test statistic is given by

$$W = \frac{\left(\sum_{i-1}^{61} a_i x_{(i)}\right)^2}{\sum_{i-1}^{61} (x_i - \overline{x})^2}$$

where x_i is the saved residual for a particular fund in a month and $x_{(i)}$ is the ith order statistic

$$a' = (a_1, \ldots, a_{61}) = m'V^{-1}/(m'V^{-1}V^{-1}m)^{1/2}$$

where $m' = (m_1, \ldots, m_{61})$ is the expected values of the residuals, assuming they have been sampled from a standard normal distribution, and V is the variance–covariance matrix of the saved residuals (i.e., the observed residuals not the expected value of the residuals). As there are 61 months of data for each fund, and Shapiro–Wilk's 1965 test is valid up to 50 observations (months), the Royston (1982) adjustment has been used. This ensures W is equal to the squared correlation coefficient and lies between 0 and 1.

The results from this test are omitted for conciseness but in 33 of the 355 test statistics are critical at the 5% level. Given 9.2% of the funds in the dataset exhibit non-normally distributed residuals, the analysis based on the significance of t-statistics has the potential to be misleading. The attention of this paper now turns to analyzing fund performance without imposing the ex-ante assumption of normality onto our regression.

8. The Fama and French (2010) Bootstrap

In the Fama and French (2010) bootstrap, the objective is to draw inferences from the cross-section of true alpha for active funds. The focus is particularly on whether the cross-section of actual alpha estimates differs sufficiently from the corresponding cross-section of a hypothetical world with no skill. For the purposes of extreme performance relative to the four-factor benchmark, the tails of the distribution are important for determining if managers have skill (or bad skill) when it comes to beating the benchmark model.

To have something with which to compare the bootstrap, for each fund i, the following regression is run over the entire 61 months

$$R_{it} = \alpha_i + b_i(R_{Mt} - R_{ft}) + s_i(SMB_t) + h_i(HML_t) + m_i(MOM_t) + \varepsilon_{it}$$

The components of the regression are identical to those of the 355 Jensen's alpha regressions. The 355 sets of parameter and residual estimates are saved for each fund. The t-statistics on the alphas—referred to as 't-alpha' are also saved to ensure we have a distribution of the "true" precision-adjusted abnormal performance of funds.

To create the hypothetical world without managerial fund skill, each fund's estimated alpha is subtracted from the corresponding fund's observed excess returns in all 61 months. This gives the benchmark-adjusted fund return across all 355 funds and the results are saved in a 61 by 355 matrix (in row 1 will be the adjusted returns for month 1 for fund i in column j). If we add to this matrix four columns corresponding to the four-factor returns for each of the 61 months, it is possible to run the 10,000 simulations.

Each simulation involves sampling (with replacement) 61 months across the 61 months in the dataset. Each fund's benchmark-adjusted returns are then regressed on the corresponding month's four-factor returns; the resampling method ensures that the four-factor returns correspond to the same sequence of months as for the funds. This gives 355 estimates of t-alpha per simulation.

This method of resampling has an important advantage. Indeed, Fama and French (2010) point out that because each simulation run is the same random sample of months for all funds, the simulations capture the cross-correlation of fund returns and its effects on the distribution of *t*-alpha. Since the method also jointly samples fund returns and explanatory returns, any correlated heteroskedasticity of the four-factor returns and disturbances of the benchmark model are captured.[19]

The results from the 10,000 simulation runs are given in Table 4. The actual column represents the distribution of *t*-alpha from the initial 355 regressions in this section. The Sim column gives the average value of *t*-alpha at selected percentiles from across the entire simulations. The Percent < Actual column shows the percent of simulation runs that delivered *t*-alpha less than the actual *t*-alpha for a given percentile.

Table 4 makes grim reading from an active management perspective; evidence of active funds beating the four-factor benchmark model requires a relatively small proportion of simulation runs delivering *t*-alpha more than the actual *t*-alpha. Given only 14% of simulated *t*-alpha are less than actual *t*-alpha at the 99% percentile, there is strong evidence that active managers did exhibit much skill in beating the returns of the benchmark model.

Making this result even more emphatic is some oversampling of fund returns owing to all the funds being alive for 8 or more months (a restriction imposed by Fama and French in their 2010 paper). Hence, the *t*-alpha distribution from the simulation runs tends to have more degrees of freedom and thinner tails than the estimates for actual fund returns; even when we are more likely to find evidence of outperformance by funds compared to a world with no skill, the results contradict this overwhelmingly.

On the negative fund performance, active funds appeared to do significantly better; over half of the simulation runs delivered *t*-alpha less than actual counterparts at the 1% level. Whilst this may be due to active managers being proficient at not underperforming the benchmark model, survivorship bias may play a role in distorting the results; funds with highly significant negative *t*-alpha may be more likely to perish than other funds. Future research may look to apply the bootstrap to a dataset free of survivorship bias.

Our results suggest that funds were relatively successful at not achieving extremely poor results but were also rather bad at delivering extremely good performance. This somewhat supports the lack of significance in the majority of the Jensen's alpha estimates; active management failed resoundingly to deliver significantly positive alpha regardless of whether we impose the normality restriction in the regressions.

Fama and French (2010) provide some further drawbacks to this bootstrap method. First, randomly sampling months means the effects of serial autocorrelation are lost, although following on from the research into autocorrelation in US stock returns by Fama (1965), it is suggested this is a minor issue. Additionally, any effects of time variation of parameter estimates are lost by randomly sampling months. Ercolani et al. (2018) have implemented the bootstrapping methodology whilst allowing for time varying parameters. It is left to future research to extend their analysis of US mutual funds to Japanese Mutual Funds.[20]

---

[19] It is for this reason that the initial regressions in this section did not correct for the heteroscedasticity.

[20] As outlined in the 2018 paper, previous research has found parameter estimates change over five-year cycles. Hence, to extend this analysis, future research may have to extend the returns history of this dataset.

Table 4. Comparison of the actual and simulated *t*-alpha for Japanese Mutual Funds using the *t*-alpha Fama and French (2010) bootstrap method.

Percentage	Simulated	Actual	Percent < Actual
1%	−4.8246	−4.4872	53%
2%	−4.1366	−3.8401	54%
3%	−3.5866	−3.1627	66%
4%	−3.2047	−2.4852	81%
5%	−2.9171	−2.2832	85%
6%	−2.6953	−2.1936	77%
7%	−2.5401	−2.0443	77%
8%	−2.4016	−1.7890	82%
9%	−2.2734	−1.7021	82%
10%	−2.1739	−1.6254	78%
20%	−1.5219	−1.0662	73%
30%	−1.1299	−0.7645	68%
40%	−0.8356	−0.6476	58%
50%	−0.5803	−0.5149	51%
60%	−0.3414	−0.4217	46%
70%	−0.0826	−0.2249	47%
80%	0.2335	−0.0242	37%
90%	0.7003	0.2688	24%
91%	0.7596	0.3235	22%
92%	0.8296	0.3627	21%
93%	0.9023	0.4546	24%
94%	0.9839	0.5468	26%
95%	1.0726	0.6307	24.%
96%	1.2024	0.7389	23%
97%	1.3614	0.8488	21%
98%	1.5688	0.9300	15%
99%	1.8594	1.1557	14%

An alternative issue for further research would be to use Kosowski et al. (2006) bootstrap technique to determine if there are any changes to the inferences; their paper finds more evidence of fund manager skill than Fama and French (2010). The technique involves running independent simulations for each fund, which has the benefit of ensuring the number of months a fund is in the simulation exactly matches the number of months of returns for that fund.

9. Persistence in Relative Skill

The method for identifying the persistence of skill is as follows: for a one year formation period a regression is run of each fund's excess returns on the four-factor model to estimate the particular fund's alpha for a given year. Each fund alpha is then ranked relative to the alphas of other funds for the corresponding year; this yields five distinct rankings, one for each year. The persistence is then calculated as the percentage of the top (bottom) 30 performing funds that remain in the top (bottom) 30 over the following few years. The percentages for each year after the formation period are averaged over the number of available portfolios; a one-year formation period for year 3 would only have two years of data after the formation period. The results are given Table 5.

Table 5 indicates that a poorly performing fund was more likely to remain a poorly performing fund, relative to its peers, than a good performing fund was to remain successful. This is not surprising in light of research into US mutual funds; Carhart (1997) find that the probability of a loser remaining a loser was higher than the probability of a winner remaining a winner. In their study Busse et al. (2014) also find little evidence of persistence in performance covering 27 countries. Our results suggest an asymmetry in the persistence of relative skill, even in the presence of survivorship bias.

Table 5. Persistence in the extreme 60 Japanese mutual funds, according to each fund's alpha over rolling one-year formation periods

	Years after Formation Period			
	Year 1	**Year 2**	**Year 3**	**Year 4**
Percentage that remain in top 30	8.33%	13.33%	10.00%	3.33%
Percentage that remain in bottom 30	18.33%	31.67%	45.00%	23.33%

Combined with the fact the persistence in the top 30 funds is rather small, the results suggest it is difficult for an investor to isolate skilled funds by ranking them based on prior abnormal performance. This fits in with the previous sections; if there had been strong abnormal return persistence, we would have expected it to show up in the significance of alpha or the Fama and French (2010) bootstrap. To further highlight this point, Table 6 shows the reversal in relative abnormal performance; after each formation period, the percentage of top (bottom) 30 funds who become bottom (top) 30 funds is calculated.

Table 6. Reversal in the alpha of the extreme 60 Japanese mutual funds

	Years after Formation Period			
	Year 1	**Year 2**	**Year 3**	**Year 4**
Percentage of top 30 that become bottom 30	16.67%	10.00%	3.33%	16.67%
Percentage of bottom 30 that become top 30	21.67%	6.67%	10.00%	3.33%

Given the large reversal in fortunes for both the top and bottom 30 funds, there appears to be little evidence of skill persistence. One could argue this shows luck plays a significant role in determining fund performance relative to its peers. Alternatively, the relationship could suggest that active managers are poor at adapting to different investment landscapes. Should various investment styles work better at different times, and if there were sufficient differences in investment styles across funds, the reversal in funds' fortunes relative to their peers would be unsurprising.[21] The results could also illustrate that Japanese equity markets are extremely efficient, therefore it is extremely difficult to use active management to beat the benchmark in a significant way. The large degree of luck in generating abnormal returns is then illustrated by the low persistence (and some strong reversal) in extreme alpha rankings.

Alongside inferences of significant absolute benchmark underperformance from previous sections, the asymmetry in relative skill persistence made fund selection a difficult task for investors; skill was relatively scarce and extreme skill was more likely to remain bad than remain good. A large portion of luck therefore accompanied successful selection of ex-post successful funds by investors. In summary, the prevalence and persistence of active mutual funds' skill did not appear sufficient to justify using active management to outperform the benchmark when passive alternatives were available.

10. Conclusions

This paper has analyzed the skill of active managers in 355 Japanese mutual funds. Compared to the four-factor benchmark model, the results indicate that managers had little ability to outperform the benchmark in a significant manner. The average excess return was not statistically different from the benchmark.

[21] It is left to further research to determine if funds fail to adapt their investment approach, such as through risk management, asset class selection and security selection, sufficiently in order to capture the reversal.

Even when managers had market timing ability, few were able to translate this skill into delivering significant positive abnormal returns to investors. These results are robust to the normality assumption since the bootstrap technique showed that only 14% of funds would have been expected to outperform a world with no manager skill. This suggests that any outperformance was largely luck-based rather than indicative of the average manager's investment approach. Allowing for parameter estimates to adjust over time would ensure this statement could be made with more certainty, especially in light of the differences uncovered in US mutual funds by Ercolani et al. (2018) compared to Fama and French (2010). The frequency of the data may be cause for concern in the market timing tests; monthly data will fail to capture managers updating their market exposure more regular than once a month. Hence, our results may have understated market timing ability. This is particularly the case given the returns to market timing may have been subsumed by the momentum factor returns in the benchmark model. However, since Jensen's alpha is robust to data frequency, investors may not concern themselves with this issue if managers with timing ability are unable to deliver statistically significant positive Jensen's alpha. Consequently, investors may want to better understand the skills their fees are rewarding, especially in light of our results indicating the transferability of bad skill between stock picking and market timing.

The results are somewhat contradictory regarding underperformance; the bootstrap method suggests active funds did considerably better at not underperforming than would be expected in a hypothetical no-skill world, whereas Jensen's alpha indicated significant performance is entirely negative. An issue with such a result is survivorship bias in the dataset; it seems likely that many of the funds were similar to the benchmark because it was these funds that were the best placed to survive over an extended period of time.

On the persistence side, our results show that there was an asymmetry in funds' relative success; poorly performing funds were more likely to remain poorly performing funds than successful funds were to remain successful. The persistence levels were also relatively low—under 50% for both types of funds across all years. Our overall results highlight the difficulty, historically, in using active mutual funds in Japan to beat the benchmark model. A possible avenue for future research is the approach employed by Pouliot (2016) using a U-statistic process to construct two statistics each devised to test different hypotheses regarding a one-time change in either the alpha or betas of the regression model. One statistic is devised to test jointly for a change in the intercept or slope while the second tests for a one-time change in slope that is robust to a change in the intercept.

Author Contributions: Conceptualization, K.P. and H.P.; methodology, K.P. and H.P.; formal analysis, K.P. and H.P.; writing—review and editing, K.P. and H.P.

Funding: This research received no external funding.

Acknowledgments: We are extremely grateful to the two referees for their extensive set of comments that led to significant improvements in the paper. The usual caveat applies.

Disclaimer: The views expressed in the paper are attributed entirely to the authors and do not necessarily reflect the views of the places at which they work.

Conflicts of Interest: The authors declare no conflict of interest.

References

Anderson, Theodore W., and Donald A. Darling. 1952. Asymptotic Theory of Certain 'Goodness of Fit' Criteria Based on Stochastic Processes. *Annals of Mathematical Statistics* 23: 193–212. [CrossRef]

Asness, Clifford. 2011. Momentum in Japan: The Exception that Proves the Rule. *Journal of Portfolio Management* 3: 67–75. [CrossRef]

Barras, Laurent, Olivier Scaillet, and Russ Wermers. 2010. False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas. *The Journal of Finance* 65: 179–216. [CrossRef]

Berk, Jonathan B., and Richard C. Green. 2004. Mutual Fund Flows in Rational Markets. *Journal of Political Economy* 112: 1269–95. [CrossRef]

Bollen, Nicolas P. B., and Jeffrey A. Busse. 2001. On the Timing Ability of Mutual Fund Managers. *The Journal of Finance* 56: 1075–94. [CrossRef]

Braun, Alexander, Nadine Gatzert, and Hato Schmeiser. 2012. Performance and Risks of Open-End Life Settlement Funds. *The Journal of Risk and Insurance* 79: 193–230. [CrossRef]

Busse, Jeffrey A., Amit Goyal, and Sunil Wahal. 2014. Investing in a Global World. *Review of Finance* 18: 561–90. [CrossRef]

Cai, Jun, Kalok C. Chan, and Takeshi Yamada. 1997. The Performance of Japanese Mutual Funds. *The Review of Financial Studies* 10: 237–74. [CrossRef]

Carhart, Mark M. 1997. On Persistence in Mutual Fund Performance. *The Journal of Finance* 52: 57–82. [CrossRef]

Chen, Zhiwu, and Peter J. Knez. 1996. Portfolio Performance Measurement: Theory and Applications. *The Review of Financial Studies* 9: 511–55. [CrossRef]

Chevalier, Judith, and Glenn Ellison. 1997. Risk Taking by Mutual Funds as a Response to Incentives. *Journal of Political Economy* 105: 1167–200. [CrossRef]

Cuthbertson, Keith, Dirk Nitzsche, and Niall O'Sullivan. 2016. A Review of Behavioural and Management Effects in Mutual Fund Performance. *International Review of Financial Analysis* 44: 162–76. [CrossRef]

Cuthbertson, Keith, Dirk Nitzsche, and Niall O'Sullivan. 2008. UK Mutual Fund Performance: Skill or Luck? *Journal of Empirical Finance* 15: 613–34. [CrossRef]

Cuthbertson, Keith, Dirk Nitzsche, and Niall O'Sullivan. 2012. False Discoveries in UK Mutual Fund Performance. *European Financial Management* 18: 444–63. [CrossRef]

Dahlquist, Magnus, and Paul Söderlind. 1999. Evaluating Portfolio Performance with Stochastic Discount Factors. *The Journal of Business* 72: 347–83. [CrossRef]

Dybvig, Philip H., and Stephen A. Ross. 1985. The Analytics of Performance Measurement Using a Security Market Line. *The Journal of Finance* 40: 401–16. [CrossRef]

Ercolani, Marco G., William Pouliot, and Joanne S. Ercolani. 2018. Luck Versus Skill Over Time: Time Varying Performance in the Cross-Section of Mutual Fund Returns. *Applied Economics* 50: 3686–701. [CrossRef]

Fama, Eugene F. 1965. The Behavior of Stock-Market Prices. *The Journal of Business* 38: 34–105. [CrossRef]

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

Fama, Eugene F., and Kenneth R. French. 2010. Luck Versus Skill in the Cross-Section of Mutual Fund Returns. *The Journal of Finance* 65: 1915–47. [CrossRef]

Fama, Eugene F., and Kenneth R. French. 2012. Size, Value, and Momentum in International Stock Returns. *Journal of Financial Economics* 105: 457–72. [CrossRef]

Farnsworth, Heber, Wayne E. Ferson, David Jackson, and Steven Todd. 2002. *Performance Evaluation with Stochastic Discount Factors*. NBER Working Paper No. 8791. Cambridge, MA, USA: National Bureau of Economic Research.

Ferson, Wayne E., and Rudi W. Schadt. 1996. Measuring Fund Strategy and Performance in Changing Economic Conditions. *The Journal of Finance* 51: 425–61. [CrossRef]

Goetzmann, William N., Jonathan Ingersoll Jr., and Zoran Ivković. 2000. Monthly Measurement of Daily Timers. *The Journal of Financial and Quantitative Analysis* 35: 257–90. [CrossRef]

Graham, John R., and Campbell R. Harvey. 1996. Market Timing Ability and Volatility Implied in Investment Newsletters' Asset Allocation Recommendations. *Journal of Financial Economics* 42: 397–421. [CrossRef]

Grant, Dwight. 1977. Portfolio Performance and the Cost of Timing Decisions. *The Journal of Finance* 32: 837–846.

Grinblatt, Mark, and Sheridan Titman. 1992. Performance Persistence in Mutual Funds. *The Journal of Finance* 47: 1977–84. [CrossRef]

Henriksson, Roy D. 1984. Market Timing and Mutual Fund Performance: An Empirical Investigation. *Journal of Business* 57: 73–96. [CrossRef]

Henriksson, Roy D., and Robert C. Merton. 1981. On Market Timing and Investment Performance. II. Statistical Procedures for Evaluating Forecasting Skills. *Journal of Business* 54: 513–33. [CrossRef]

Jagannathan, Ravi, and Robert A. Korajczyk. 1986. Assessing the Market Timing Performance of Managed Portfolios. *Journal of Business* 59: 217–35. [CrossRef]

Jensen, Michael C. 1968. The Performance of Mutual Funds in the Period 1945–1964. *The Journal of Finance* 23: 389–416. [CrossRef]

Jensen, Michael C., Fischer Black, and Myron S. Scholes. 1972. *The Capital Asset Pricing Model: Some Empirical Tests*. Studies in the Theory of Capital Markets. Santa Barbara: Praeger Publishers Inc.

Kahneman, Daniel, and Mark W. Riepe. 1998. Aspects of Investor Psychology. *The Journal of Portfolio Management* 24: 52–65. [CrossRef]

Kosowski, Robert, Allan Timmermann, Russ Wermers, and Hal White. 2006. Can Mutual Fund Stars Really Pick Stocks? New Evidence from a Bootstrap Analysis. *The Journal of Finance* 61: 2551–95. [CrossRef]

Lilliefors, Hubert W. 1967. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of American Statistical Association* 62: 399–402. [CrossRef]

Malkiel, Burton G. 1995. Returns from Investing in Equity Mutual Funds 1971 to 1991. *The Journal of Finance* 50: 549–72. [CrossRef]

Massey, Frank J. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association* 46: 68–78. [CrossRef]

Otten, Roger, and Dennis Bams. 2002. European Mutual Fund Performance. *European Financial Management* 8: 75–101. [CrossRef]

Pouliot, William. 2016. Robust Tests for Change in Intercept and Slope in Linear Regression Models with Application to Manager Performance in the Mutual Fund Industry. *Economic Modelling* 58: 523–34. [CrossRef]

Razali, Nornadiah Mohd, and Yap Bee Wah. 2011. Power Comparisons of Shapiro–Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. *Journal of Statistical Modeling and Analytics* 2: 21–33.

Royston, J. Patrick. 1982. An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 31: 115–24. [CrossRef]

Shapiro, Samuel Sanford, and Martin B. Wilk. 1965. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 52: 591–611. [CrossRef]

Sharpe, William F. 1991. The Arithmetic of Active Management. *Financial Analysts Journal* 47: 7–9. [CrossRef]

Treynor, Jack L. 1965. How to Rate Management of Investment Funds. *Harvard Business Review* 43: 63–75.

Treynor, Jack, and Kay Mazuy. 1966. Can Mutual Funds Outguess the Market? *Harvard Business Review* 44: 131–36.