MUTUAL FUND PERFORMANCE IN THE NETHERLANDS: LUCK OR SKILL?

MASTER THESIS IN FINANCIAL ECONOMICS

Author: M. Kool (S4232569)
Supervisor: Dr. S. Füllbrunn
Faculty: Nijmegen School of Management
Date: 14th of August 2017

Radboud University

ABSTRACT

This study applies the cross-section bootstrap method of Kosowski et al. (2006) to a sample of 14 surviving actively managed Dutch open-end equity funds over the 1992-2015 period. Using this method, it is able to identify whether an individual fund's outperformance is the product of 'luck' or 'skill'. This provides an indication to investors whether active fund management is worth the costs. An advantage of the method over parametric methods is that it does not require a fund's idiosyncratic risk to be normally distributed. The results of the study provide evidence of luck for positive outperforming funds in the sample and evidence of misfortune for most negative outperformers. The majority of funds underperforms the benchmark and only the worst funds in the sample provide weak evidence of bad skill. In addition, the worst negative funds persistently underperform the benchmark over four non-overlapping sub-periods. Considering the sample is not free of survival bias and has a small sample size, the true proportion of bad skill is probably much higher. The outcomes indicate investors rather invest in low-cost index-tracker funds than in actively managed Dutch equity funds.

Table of contents

1. Introduction

Imagine you are an unexperienced investor who wins the lottery and subsequently you find a legion of mutual fund managers on your doorstep, claiming they are worth every cent to actively invest your money in the equity market. Is it worth paying them for their stock picking 'skills', or are they just as dependent on 'luck' as any other? This is an important question to answer, since if they are dependent on luck you might as well invest your money in a low-cost passive index-tracker. The mutual fund literature has been focussed on questions like these for decades, but only the largest two players have been studied extensively (US and UK). The focus of this study is the Dutch mutual fund industry, but in order to arrive at the central research question we first have to discuss the US and UK related mutual fund literature.

To simplify this vast amount of mutual fund literature, Cuthbertson, Nitzsche, and O'Sullivan (2008) distinguish between two key issues. The first one is whether mutual funds on average have positive, negative or zero abnormal returns compared to a benchmark factor model. Without going into full detail, a factor model is able to price a fund's excess return by using one or more explanatory factors. The constant alpha (α) in the factor model denotes which part of the return cannot be priced by the factors and is therefore an indicator for abnormal performance. Investors are eager to invest in mutual funds with large positive alphas, as they provide a premium compared to the benchmark. Negative alpha funds are avoided, since they underperform the benchmark and thus provide negative abnormal returns. The second key issue in the mutual fund literature is whether outperformance is persistent over a longer period of time and can be identified *ex-ante*. This is important to investors in the mutual fund industry, as persistent positive abnormal returns provide less risk for the future. The possibility to identify these funds ex-ante would give investors a chance to develop an optimal investment strategy.

The general finding for the US and UK regarding the first key issue is that there is almost no or little positive abnormal fund performance, but wider evidence of negative abnormal fund performance (e.g. Lakonishok et al., 1992; Grinblatt et al., 1995; Malkiel, 1995; Carhart, 1997; Blake & Timmerman, 1998; Chevalier & Ellison, 1999; Wermers, 2000; Baks et al., 2001; Pastor & Stambaugh, 2002; Fama & French, 2010). These studies generate individual fund alphas by running factor models over a fund's historical returns. Their results thus indicate that the vast majority of funds produces a negative alpha, while there are limited or no positive alpha funds. An extremely small number of large positive alpha funds may exist though, and these are precisely the outlier funds investors are determined to find.

However, even if these are found, it is questionable whether historical alphas are a suitable instrument for making future investment decisions. Past successes may not be persistent over time and current 'fund stars' may underperform in the future when their results turn out to be an outcome of luck. The second key issue in the literature is therefore whether outperformance can be predicted ex-ante and to which extent it is persistent over time. Carhart (1997) for example studies US performance persistence by examining whether an observed alpha remains positive or negative over 1- to 3-year periods, by ranking funds into deciles based on their historical alphas and rebalancing them on an annual basis. He finds that positive outperformers are not persistent over time, whereas the worst mutual funds keep underperforming the benchmark. Other studies on the US and UK find that persistence in positive outperformance is shorter than one year or absent, even though numerous evidence points towards long-term persistence in underperformance (e.g. Hendricks et al., 1993; Brown and Goetzmann, 1995; Chevalier and Allison, 1997; Blake and Timmerman, 1998; Allen and Tan, 1999; Fletcher and Forbes, 2002; Bollen and Busse, 2005; Cuthbertson et al., 2008).

Carhart (1997) argues similar to Chevalier and Allison (1997) that this lack of persistent positive abnormal fund performance can be explained by momentum. Funds with large positive alphas are simply the product of luck as the managers by chance have constructed a portfolio of winner stocks. As momentum is only a temporary phenomenon and funds run out of luck, they do not outperform the benchmark longer than one year. Another explanation for the lack of persistent abnormal fund performance is given by the rational competitive model of Berk and Green (2004). The model theorizes that funds with positive abnormal returns cannot be persistent over time as their success results in a large inflow of funds and increasing marginal costs. Even the most skilled funds can therefore not produce persistent positive abnormal returns, as the combination of an increased inflow of fund money and decreasing returns to scale eventually causes a lesser performance.

Predicting positive outperformance ex-ante turns out to be nearly impossible in the previous studies, but this could be well caused by their inability to find a fitting explanatory variable. Instead of using past performance, a more recent study by Amihud and Goyenko (2013) proposes to use the R^2 obtained from a multifactor benchmark model. They argue that a lower R^2 represents a higher level of selectivity (or activity) as it indicates a larger deviation from the common factors. By sorting twenty-five US fund portfolios based on their historical R^2 and alpha, they find that a lower R^2 in one period significantly predicts positive outperformance in the next period. Cremers and Petajisto (2009) also attempt to relate the level of activity to performance, by using the proportion by which a fund's portfolio differs from the benchmark index holding. They find that positive persistent outperformance can be admitted to a larger deviation from the benchmark.

What to make of these findings from an investor point of view? Given that positive abnormal fund performance is in most cases not persistent over time, it is a poor call to make investment decisions solely on the basis of historical alphas. An investor has to additionally consider influences like the level of selectivity and performance persistence related conditions. Heuer, Merkle, and Weber (2016) argue however that investors keep chasing positive alphas as they do not take into account the influence of volatility and the chance that a positive alpha does not necessarily indicate skill, but can be the product of sampling variation. Jordan and Riley (2015) for example find that volatile funds may exhibit attractive returns over a short period, but significantly underperform less volatile funds over a longer period. If investors are not able to take this influence of volatility into account, they might not be aware of the risk they are taking. This becomes especially worrisome in the light of a study by Solomon, Soltes, and Sosyura (2014), who find that mutual funds successfully use window dressing in advertising to exploit investor irrationalities. In addition, Barras, Scaillet, and Wermers (2010) argue that investors are unaware of the fact that the chance of finding a positive significant alpha caused by pure luck increases when the sample size increases, resulting in falsely assigning skill to pure luck funds.

From the previous discussed studies it should have become evident that the concepts of *skill* and *luck* play a central role in the mutual fund literature and are important from an investor point of view. Some argue that skilled managers are non-existent and positive outperformers are merely lucky, whereas others think that truly gifted managers may exist and outperform the market with their stock picking skills, albeit not on a persistent basis. Kosowski, Timmermann, Wermers, and White (2006) have further explored these ideas and developed a cross-section bootstrap method to distinguish between luck and skill in mutual fund alphas. They argue that 'conventional' studies like Carhart's (1997) ignore the non-normal distribution from which individual fund alphas are generated and the cross-sectional non-parametric distribution in which they result. These conventional studies thus make claims on the basis of the normality assumption while the actual distribution is non-parametric, which causes biased *p*-values. The bootstrap method of Kosowski et al. (2006) does not require normality and uses random resampling of residuals from a factor model to simulate pseudo excess-returns. These are in turn used to generate simulated alphas under the null hypothesis of no outperformance. Instead of using a parametric *p*-value which only takes a fund's individual alpha distribution into account, they use a bootstrapped *p*-value that compares an individual fund's alpha to all other funds in the sample, the cross-section of funds. They propose that sampling variation (luck) cannot be the only cause of extreme alphas if the cross-section of simulated alphas generates far fewer extreme values than the empirically observed data: evidence of true stock picking skills. Kosowski et al. (2006) consider a positive outperformer skilled when its empirically observed alpha is positioned within the 5% upper-tail area of the simulated cross-section.

Using this bootstrap method, Kosowski et al. (2006) and Cuthbertson et al. (2008) find evidence for respectively the US and the UK that positive abnormal fund performance for the best mutual equity funds is attributable to stock picking skills, while the overall majority of good performers has positive abnormal returns due to luck. Negative abnormal fund performance for most underperformers is not caused by misfortune, but by truly bad stock picking skills. More importantly, the evidence from the US indicates that skilled positive outperformers have persistent abnormal returns. Negative abnormal returns persist in both the US and the UK. The UK results for the division of skill and luck hold when using the false discovery rate method of Barras et al. (2010), even though the group of true outperformers is smaller (Cuthbertson et al., 2012). Performance persistence for positive alpha funds vanishes and only weak evidence for persistence of negative alphas remains. A similar bootstrap study on the US by Fama and French (2010) finds little evidence of positive skill, but they jointly sample fund returns instead of using independent simulations for each fund. Performance persistence for mutual fund stars is weak or absent and tends to vanish after 1992.

The results from the latter studies imply that there might be an extremely small group of skilled fund managers that are worth the costs of active management. These studies are however restricted to the largest mutual fund markets (US and UK), while other countries have considerable mutual fund industries as well. Investment researcher Morningstar for example classified the Netherlands in 2015 as the mutual fund market with the best experience to European investors based on regulation, taxation, disclosure, expenses and sales and media. The Dutch mutual fund industry was ranked third on a global level, right after the US and South-Korean fund markets, after it greatly improved fund fees and expenses by banning retrocessions. In addition, a more dated study by Ter Horst, Nijman, and de Roon (1998) finds that active funds mainly investing in Dutch equity largely outperform a passive portfolio of indices. They do however not address whether this outperformance is caused by luck or skill, while the previous discussed literature has shown this is important from an investor point of view.

The aim of this study is therefore to apply the cross-section bootstrap method of Kosowski et al. (2006) to the Dutch mutual fund market to determine whether the abnormal performance of individual Dutch mutual funds is simply based on luck or the product of true stock picking skills. By using the cross-section bootstrap method, it is the first study on Dutch mutual fund performance that takes into account non-normalities in the cross-section of mutual fund alphas. The central research question is formulated as follows: *Is the ex-post performance of individual Dutch mutual equity funds in the 1992-2015 period due to luck or skill?* This study's cross-section bootstrap is based on the European Fama-French three-factor model, stretches over the entire 1992-2015 period and takes into account the excess returns of 14 active Dutch open-end mutual equity funds.

The key findings in the sample of this study are that there is only weak evidence of bad skill amongst the worst Dutch mutual funds, but no evidence of skilled positive outperformance. Most funds are in fact unlucky negative outperformers, only few funds indicate positive outperformance caused by luck. In addition, the worst funds keep underperforming the benchmark in all sub-periods, which is an indication for persistence in underperformance. Executing the bootstrap with the Capital Asset Pricing Model (CAPM) instead of the Fama-French three-factor model provides similar results. The outcomes of this study are contradicting the studies of Kosowski et al. (2006) and Cuthbertson et al. (2008), by finding no superior skilled outperformance. Given the sample is small and not free of survival bias, the actual proportion of bad skill in the true 1992-2015 population of Dutch funds is most likely higher. The results imply that if investors are eager to diversify on the equity market, they should rather invest in low-cost index-trackers than in actively managed Dutch equity funds.

The rest of the study proceeds as follows. Section 2 describes the standard factor models in the mutual fund literature and provides a justification for the use of a European benchmark model. Section 3 evaluates the non-normality of mutual funds' idiosyncratic risk and details the bootstrap procedure. Section 4 describes the data used in the study, after which section 5 selects a 'best' factor model and presents the results. These are discussed in section 6, after which section 7 concludes.

2. Factor models

The bootstrap method makes use of an underlying factor model by using the residuals from the model to generate pseudo excess returns and simulated alphas. This section discusses the most important factor models in the mutual fund literature (2.1), their general construction (2.2) and provides a justification for using European factor models (2.3). The specifics of the bootstrap method are described in section 3.

2.1. Brief description of key factor models in the mutual fund literature

Factor models aim to explain variation in returns by using one or more factors as explanatory variable(s). In fact they explain the variation in excess returns, which is the extra return one can earn on the stock market compared to the time-value of money. Over several decennia three models have been extensively discussed and used by economists: the Capital Asset Pricing Model (CAPM), the Fama-French three-factor model and the Carhart four-factor model. Each of the three factor models incorporates alpha (α), a constant that is the difference between a model's theoretical return and the actually observed return. Alpha is therefore perceived as the rate of return by which a stock or mutual fund portfolio outperforms the market and an indicator for manager performance. Investors prefer high positive alphas, since these indicate a higher return then one would expect based on the factors. Alphas smaller than zero indicate underperformance and are therefore avoided, whereas alphas equal to zero imply a similar return as the benchmark and no added value (or damage) from manager performance. Index-trackers are therefore often considered zero-alpha funds. Note that a positive alpha does not necessarily represent a positive return, it just implies that the actually observed return is higher than the model predicted it would be. If a factor model is the right and only predictor for variation in stock returns, alpha should be equal to zero.

The first factor model to discuss is the CAPM as proposed by Sharpe (1964) and Lintner (1965). The model assumes that the only explanation for variation in returns is the premium or discount one receives on the asset beta (β), a coefficient that captures the volatility of a particular asset compared to the entire market. CAPM can therefore be considered a one-factor model. All changes in the risk of a stock are explained by changes in the covariance between a particular stock and the market. A β of 1 implies that the stock's price movement is equal to that of the market, whereas a β smaller than 1 indicates that the stock is less volatile than the market. An investor earns a premium return when taking more risk (volatility) than the market ($\beta>1$) and a discount when taking less risk than the market ($\beta<1$). Studies that have tested CAPM however show that the model suffers from a great number of anomalies and has restricted predictive power (see Bornholt (2013) for a complete discussion).

Fama and French (1992, 1993, 1996) therefore extend the CAPM to a three-factor model that incorporates the original CAPM β, book-to-market ratio and size. The model hereby makes an additional correction for the finding that low book-to-market ratio stocks are consistently outperforming high book-to-market ratio stocks and the finding by Banz (1981) that small firms tend to outperform big firms. The model has been able to successfully explain anomalies of the CAPM and has significant explanatory power, even though enthusiasm has dwindled as the size-effect has vanished over time and a sound causal chain is still missing. In addition, the model is unable to capture momentum return effects as described by Jegadeesh and Titman (1993, 2001). Momentum describes the tendency of winner stocks to keep increasing and the propensity of loser stocks to keep decreasing. Carhart (1997) therefore introduces a four-factor model that adds a momentum factor to the Fama-French three-factor model, making it possible to capture a premium on winner stocks and a discount on loser stocks.

The previous paragraphs briefly summarized the *unconditional* versions of the main factor models, but there are many *conditional* variations. These are left aside in this study and do not result in different outcomes in the studies of Kosowski et al. (2006) and Cuthbertson et al. (2008). It should be clear by now that the selection of a particular model yields some implications for the amount of explained variation. Including more factors does however not necessarily imply a better factor model, as each added factor needs to make a 'meaningful' contribution and requires some theoretical back-up. Most of the literature on US mutual fund performance uses Carhart's four-factor model and shows its explanatory power is significantly better than the Fama-French three-factor model (e.g. Kosowski et al., 2006; Fama and French, 2010). The momentum-factor is often not significant for the UK market, for which reason UK studies make use of the Fama-French three-factor model (Cuthbertson et al., 2008). As it is unclear to this point which model fits the Netherlands best, this study runs both the Fama-French three-factor model and the Carhart four-factor model in order to decide on the criteria of significance, the amount of explained variation (adjusted R^2) and the Schwarz Information Criterion (SIC) which model serves the bootstrap best. SIC generates a score for each model based on the goodness of fit and the added value of increasing the number of parameters to reach this degree. The lowest score represents the desired model as it is thought to be the most plausible *a posteriori* (Cavanaugh and Neath, 1999).

2.2. General construction of the factor models

The unconditional Carhart (1997) four-factor model is formulated in equation *1)*, where $r_{i,t}$ is the excess return for fund *i* in month *t*, composed as the fund return minus the risk-free rate. The alpha for fund *i* is denoted by α_i. $RMRF_t$ is the excess return on a weighted market index, whereas SMB_t, HML_t and

MOM_t are respectively the factor-mimicking portfolios for size, book-to-market ratio and momentum in month *t*. The coefficients represent the weight of a factor on an individual fund's excess returns. The residuals of the model are represented by $\varepsilon_{i,t}$ and T_i is the number of observations on fund i. The *RMRF*-factor represents the intuition from the CAPM, whereas the Fama-French three-factor model can be derived straightforward by dismissing the Carhart momentum factor (*MOM*). $RMRF_t$ is operationalized as R_m-R_f, where R_m represents the weighted market index return and R_f the risk-free rate (usually equal to a T-bill).

$$r_{i,t} = \alpha_i + \beta_i RMRF_t + s_i SMB_t + h_i HML_t + m_i MOM_t + \varepsilon_{i,t} \quad 1)$$

The size factor SMB_t and book-to-market factor HML_t are constructed as follows. The proxy for size is market capitalization and consists of the stock price multiplied by the number of outstanding common stocks. Microcap stocks are often dismissed. The book-to-market (BE/ME) ratio is calculated by dividing the book value of equity (BE) by its market capitalization (ME). The book value of equity is often operationalized as common shareholders' equity, which is calculated by subtracting total liabilities from total assets. All stocks in the market portfolio are first sorted by size, by taking the top 90% as big stocks and the bottom 10% as small stocks. These are then grouped into three categories of BE/ME ratios (*Low*, *Medium* and *High*) by using the 30th and 70th percentiles as breaking points.[1] The result is six portfolios (2x3) of stocks sorted by size and BE/ME (*SH, SM, SL, BH, BM, BL*). The portfolios are reshuffled at the end of each June in year *t*, creating a period of July of year *t* till June of year t_{+1}.The reshuffling of size takes place on the basis of the market capitalization at the end of June in year *t*, but the BE/ME reshuffling at the end of June in year *t* takes place on the basis of the BE/ME from the end of December of year t_{-1}. These yearly sorted six portfolios can then be used to construct the monthly values for *SMB* and *HML* as expressed by equations *2)* and *3):*

2) $$SMB = \frac{1}{3}(SH + SM + SL) - \frac{1}{3}(BH + BM + BL)$$

3) $$HML = \frac{1}{2}(SH + BH) - \frac{1}{2}(SL + BL)$$

4) $$MOM = \frac{1}{2}(SH + BH) - \frac{1}{2}(SL + BL)$$

[1] Breaking points for all factors can differ per model, depending on how well a market is capitalized and diversified. Fama and French (2012) use the given breaking points for regional models like the European factor model.

Equation *2)* expresses the equal-weight average return on the small stock portfolios minus the average return on the big stock portfolios, whereas equation *3)* expresses the average return on the high BE/ME portfolios minus the average of the low BE/ME portfolios. The momentum factor *MOM* is in this study constructed similar to Fama and French (2010). The advantage of this approach over Carhart's (1997) operationalization of momentum (*PR1YR*) is that Fama and French's (2010) method corrects for size related effects. It is comparable to the construction of *HML* by first sorting the stocks again by size, but now the second step involves sorting by winners (*H*) and losers (*L*). This is executed by taking the 30% top performers (*H*) and the 30% worst performers (*L*), based on the previous 11-month cumulative return. Portfolios formed at the end of month t_{-1} are sorted on the average of the returns from month t_{-12} to month t_{-2}.[2] These are then separated in three momentum categories by using the 30th and 70th percentiles as breaking points, resulting in a high (*H*), low (*L*) and medium (*M*) momentum category per size category. Note that the sorting of the size categories takes place on the basis of the market capitalization in the 6th previous month. The result of the complete sorting procedure is six portfolios sorted by size and lagged momentum (2x3).[3] *MOM* for each month *t* is then constructed by subtracting the loser portfolios from the winner portfolios, as expressed in equation *4)*.

2.3. European benchmark

This study uses the European stock market as a basis for the development of the factors, similar to Otten and Bams (2002). The alphas of the model can thus be interpreted as how well Dutch mutual funds perform compared to the European market. Even though Moerman (2005) finds that domestic factor models still outperform the European factor model, he notes that these factors are merging as a consequence of increasing market integration, especially for European countries with a relatively high number of listed stocks like the Netherlands. This increasing correlation of European stock returns is also found by Bekaert, Hodrick and Zhang (2009). Fama and French (2012) add to this that a European factor model explains domestic stock returns sufficiently well when using a three-factor model, while the four-factor model suffers from general problems. An advantage of the European factor model is that it overcomes potential problems arising from too few stocks in the Dutch market to create well-diversified portfolios. Even though Dutch market capitalization is relatively large on a European level, it is fairly small compared to the US or other well-developed markets. Using the European market as a benchmark therefore results in more reliable and more diversified factors, but its model's explanatory power might be lower than is the case in domestic models.

[2] Fama and French (2010) mention it is conventional in the momentum literature to leave out month t_{-1} when constructing the *11*-month return for month *t*.

[3] The formula for *MOM* might look similar to the formula for *HML*, but *H* and *L* represent fundamentally different variables in each case.

3. The cross-section bootstrap method

The previous section discussed the factor models and their construction. This section describes the reasons why a bootstrap method is preferred over conventional methods (3.1) and the specifics of the bootstrap procedure (3.2).

3.1. Non-normality of mutual funds

The advantage of the cross-section bootstrap method over conventional methods is that it does not assume that that a fund's idiosyncratic risk has a known parametric distribution. It allows to even include and assess the fund alphas that are based on an extremely non-normal distribution and positioned in the outer tails of the cross-section distribution (Kosowski et al., 2006). To comprehend the meaning of these sentences, one has to understand its concepts. The cross-section here simply refers to all funds in a sample, whereas idiosyncratic risk represents a fund's specific risk. Without going into the statistical details, a normal distribution is characterized by that its values are centred in a symmetrical way around the mean of the distribution in a bell-curved shape. The frequency by which values deviate from the mean decreases the more standard deviations they are located from the centre of the distribution. If this is not the case, the distribution is considered non-normal. The mean and median are no longer necessarily equal and skewness can cause fat tails, implying a much higher number of extreme values than is possible under the standard normal distribution. Most 'conventional' statistical methods however require the residuals of a model to approach the standard normal distribution. The results of conventional statistical methods can therefore be biased if the distribution is non-normal (non-parametric) instead of normal (parametric). Politis and Romano (1994) for example mention that methods based on the normality assumption generate unreliable *p*-values when the actual distribution is non-parametric. These *p*-values are important in the mutual fund literature, as they are used to indicate whether a fund's outperformance is statistically different from zero. Bootstrapping does not require normality and can therefore help to create more reliable *p*-values when a distribution is non-normal.

To provide a justification for the use of the cross-section bootstrap method, Kosowski et al. (2006) mention four reasons why non-normality is likely to occur in the residuals of the alphas of individual mutual funds. The first reason is that active fund managers invest severely in a restricted number of stocks, which makes their portfolio return sensitive to movements in individual assets. While the standard normal distribution assumes that equally-weighting non-normally distributed stocks in the portfolio approaches normality, this equal-weighting is in fact absent when investing heavily in a limited number of stocks. Second, individual stocks show different levels of serial

correlation in returns over time, which causes skewness in the residuals. Third, the returns from the market-benchmark may be non-normally distributed, causing co-skewness between individual stock returns and the benchmark. The fourth and final argument is that there is the possibility of heterogeneous risk-taking across time. Kosowski et al. (2006) add to this that even if all individual fund alphas follow a normal distribution, this does not imply the cross-section of mutual fund alphas is normally distributed. In fact, it is likely that the cross-section of alphas is still non-normal in that case due to cross-sectional correlations in the residuals and heterogeneous risk taking across funds. Given these arguments for non-normally distributed alphas, Kosowski et al. (2006) propose the use of the cross-section bootstrap method as it does not require the normality assumption. The bootstrap method can much better approach the true standard probability of properly rejecting a null hypothesis than conventional statistical methods (Horowitz, 2003).

3.2. The bootstrap process

Cross-section bootstrapping is able to reveal for each individual fund whether its abnormal performance is based on sampling variation or skill. It does so by comparing an empirically observed individual alpha to a cross section of simulated alphas, that is a set of simulated alphas from all other funds in the sample. These alphas are simulated by using the residuals and vectors from an estimated factor model (the original alpha from this estimated model is saved and considered the empirically observed or real world alpha). The residuals are randomly resampled and used together with the saved vectors to create pseudo-returns, under the null of no outperformance ($\alpha_i = 0$). These are in turn regressed on the original factor model to create a simulated alpha, which is a product of luck. The entire process is repeated a thousand times for each fund (B=1000), resulting in a distribution of simulated alphas per fund called the luck distribution. Using a more schematic description, it is possible to describe the bootstrap process in six steps:

I) ***Model estimation***: estimate the model of equilibrium returns with a factor model for each individual fund *i*. Equation *1)* and discussed variations can be used for this purpose.

II) ***Residual resampling***: save the factor-vectors and the residuals from the model estimation. Draw a random sample with replacement of length T_i for each fund *i* from the residuals $\varepsilon_{i,t}$. This creates a pseudo-time series of resampled residuals.

III) **Generate pseudo excess returns**: maintain the factor vectors in chronological order. Use the resampled residuals and saved vectors to simulate a pseudo excess returns series for fund *i,* under the null hypothesis that alpha (α_i) is zero. See equation *5)* with regards to the Carhart four-factor model.

IV) **Generating alpha:** regress the pseudo excess returns series on the original factor model. This creates a simulated alpha

V) **Repetition**: steps *II-IV* are repeated a 1000 times for each fund *i,* creating a distribution of simulated alphas that is only existing because of sampling variation (luck).

VI) **Sorting and generating p-values**: the simulated α_i's (or $t_{\alpha i}$'s) of each individual fund are then sorted from highest to lowest, resulting in a sorted luck distribution per individual fund and a cross-sectional luck distribution across all funds. This allows to generate bootstrapped *p*-values that compare an individual empirically observed alpha to the cross-section of simulated alphas.

5) $$\hat{r}_{i,t} = \hat{\beta}_i RMRF_t + \hat{s}_i SMB_t + \hat{h}_i HML_t + \hat{m}_i MOM_t + \hat{\varepsilon}_{i,t}$$

The sorting process allows to compare the position of an individual empirically observed alpha to the cross-section of simulated alphas. This procedure thus takes into account the luck distributions of all funds in the sample, instead of an alpha's individual distribution. The best fund's empirically observed α_{max} can now be compared to all funds by taking the cross-sectional distribution of all simulated α_{max}*'s,* $f(\alpha_{max})$. This cross-sectional $f(\alpha_{max})$ represents the distribution of each highest alpha simulation per individual fund. A sample with 500 funds would for example result in 5000 simulations and a $f(\alpha_{max})$ of 500 simulations, representing each fund's highest alpha simulation. Cuthbertson et al. (2008) give an example of how to decide whether an empirically observed α_{max} is due to luck or skill. A fund's alpha can be assigned to be a product of skill with 95% confidence when the empirically observed α_{max} exceeds the 5% upper tail cut-off point in the distribution of $f(\alpha_{max})$. Put differently, the null hypothesis that α_{max} is caused by sampling variation is rejected with 95% confidence when it is located above the 95th percentile of $f(\alpha_{max})$ and interpreted as evidence that there must be skill. Of course this can be applied to any individual empirically observed positive alpha, by checking whether that alpha exceeds the 5% upper tail cut-off point within the cross-sectional distribution of funds. This corresponds to a bootstrapped *p*-value lower than 0.05. The previous example is graphically illustrated by figure 1 panel A. Panel A presents a possible cross-sectional distribution of positive simulated $\alpha_{max's}$**.** The reference line represents the 95th percentile cut-off point. If the actually observed alpha has a value that is located to the right of the reference line and thus positioned in the upper 5%-area, it is at least more extreme than 95% of all simulated $\alpha_{max's.}$ In the bootstrap method of Kosowski et al.

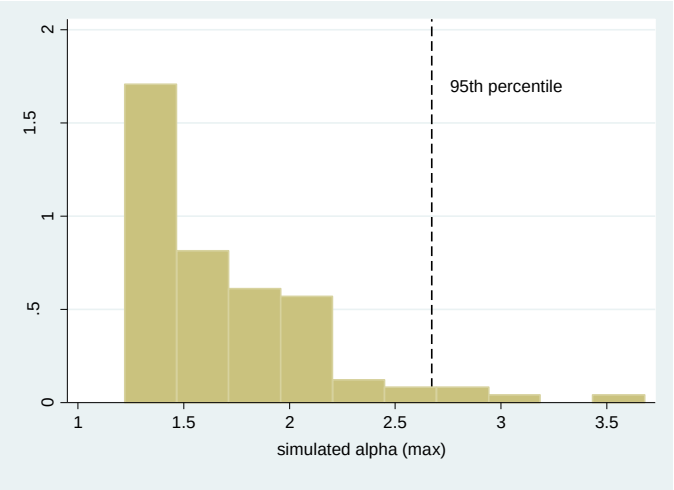
(2006) this is interpreted as skill, as it is unlikely that such an extreme value has been caused solely by luck. Likewise, negative outperformance is compared to the cross-sectional distribution of simulated $\alpha_{min's}$ (see figure 1 panel B). If an empirically observed fund alpha is located in the area beneath the 5th percentile represented by the reference line, its negative outperformance is attributable to bad skill instead of misfortune. Put differently, if the empirically observed negative alpha is located in the lower 5%-area, it performs worse than at least 95% of all simulated $\alpha_{min's,}$ which is interpreted as evidence of bad skill.

Similar to Cuthbertson et al. (2008) this study uses one-sided tests. For positive outperformers this implies the null hypothesis H_0 that the empirically observed alpha is smaller than or equal to zero (H_A that alpha is larger than zero). For negative outperformers this implies the H_0 that the empirically observed alpha is larger than or equal to zero (H_A that alpha is smaller than zero). The bootstrapped *p*-value for positive alphas represents the percentage of cases in which the empirically observed alpha is lower than the cross-section of simulated alphas. For negative alphas this is simply the reverse. A bootstrapped *p*-value of 0.005 in the positive alpha case would for instance indicate that only 0.5% of the corresponding simulated cross-section exceeds the empirically observed alpha. For negative alpha's this would indicate that only 0.5% of the simulations generates worse (more negative) values than the empirically observed alpha. Kosowski et al. (2006) attribute a fund's alpha to skill if the bootstrapped *p*-value is lower than 0.05 (5%), meaning the actually observed alpha is extremer than 95% of the simulated alphas in the corresponding cross-section. During the analysis this study makes use of a *mild* bootstrapped *p*-value and a *strict* bootstrapped *p*-value. *Strict* and *mild* refer to which level in the cross-section of simulations the empirically observed data is compared to. Given this study incorporates 14 funds, the *mild* bootstrapped *p*-value compares all empirically observed negative (positive) funds to the worst (best) 14 simulations in the cross-section, similar to Kosowski et al. (2006) and Cuthbertson et al. (2008). The *strict* bootstrapped *p*-value compares each ranked empirical alpha to its own ranking place in the cross-section. The worst (best) empirical fund is compared to the worst (best) 14 simulations in the cross-section, the 2nd worst (best) fund to the 2nd worst (best) 14 simulations etc. The *strict* bootstrapped *p*-value makes it therefore slightly more difficult to be (un)fortunate and easier to be (un)skilled.

Even though the previous examples were based on alpha, this study mostly interprets on the basis of sorting *t*-alpha. Brown, Goetzmann, and Ross (1992) propose to use the *t*-statistic of alpha instead of α, as it has superior statistical characteristics and helps to further alleviate a possible survival bias. Kosowski et al. (2006) add to this that the cross-sectional distribution of the *t*-statistic has more desirable statistical properties than the alpha, as alpha does not correct for time-varying risk-taking and the heterogeneous fund volatilities in which it results. Since *t*-alpha divides alpha by its standard

deviation (volatility), the cross-section of *t*-alphas is better able to approach the standard normal distribution than the cross-section of alphas. The rest of the study therefore mostly analyses and sorts on the basis of t_α, similar to Kosowski et al. (2006) and Cuthbertson et al. (2008).

Panel A: positive outperformance simulated cross-section



Panel B: negative outperformance simulated cross-section

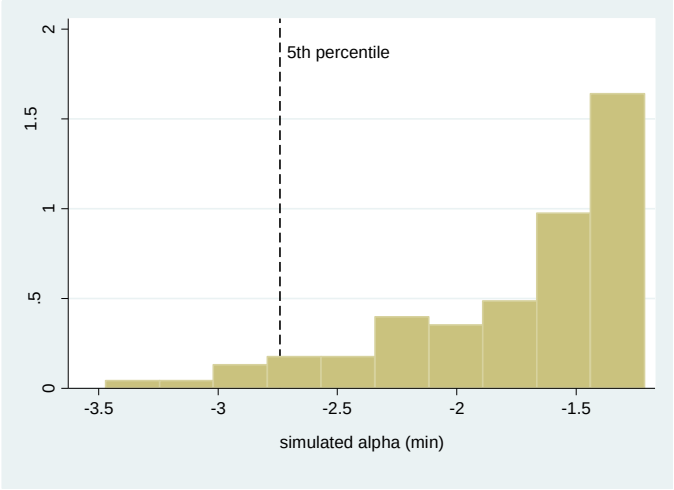


Figure 1: Cross-sectional distribution of simulated alpha's and cut-off points

Panel A reports a possible cross-sectional distribution of simulated maximum alpha's $f(\alpha_{max})$ and the 5% upper-tail cut-off point represented by the reference line. Panel B reports a possible cross-sectional distribution of simulated minimum alpha's $f(\alpha_{min})$ and the 5% lower-tail cut-off point.

Unlike Kosowski et al. (2006) and Cuthbertson et al. (2008), the data population of this study does not consist of hundreds of mutual funds (see section 4). The result is that one will be producing bootstrapped *p*-values on the basis of 14 funds as the entire cross-section, which makes it much harder to be (un)skilled instead of (un)fortunate. As mentioned, Kosowski et al. (2006) use a cut-off bootstrapped *p*-value of 0.05, but for a sample of 14 funds this would imply that there is only skill when the empirically observed alpha (*t*-alpha) is more extreme than all simulations. It is not unlikely that one simulation in the cross-section is always more extreme than an empirically observed value, as it is bootstrapped from the same data point. A lower cut-off point (bootstrapped *p*-value < 0.143) will therefore be taken into account in addition to the 0.05 cut-off point. A bootstrapped *p*-value of 0.143 in a sample of 14 funds indicates that 2 simulated alphas (*t*-alphas) in the corresponding cross-section are extremer than the empirically observed alpha (*t*-alpha). The use of a 0.143 bootstrapped *p*-value cut-off point in addition to the 0.05 cut-off point is to some degree defendable. However, its evidence is weaker than the narrower cut-off point handled by Kosowski et al. (2006), especially when it considers a *strict* bootstrapped *p*-value. The strongest evidence of skill would be to find a *mild* bootstrapped *p*-value that is smaller than 0.05.

4. Data

The data for the European factors are retrieved from Kenneth French's website and are based on the combined markets of Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom (Fama and French, 2012).[4] The complete fund dataset consists of over 2000 open-end alive and dead mutual funds and is retrieved from Datastream's fund market list. Using the same criteria as Cuthbertson et al. (2008), only open-end mutual funds that invest primarily in domestic equity (at least 50%) in the 1992-2015 period are incorporated, resulting in a sample of 26 funds.[5] Open-end fund shares can only be bought and sold back by the investor to the mutual fund at the daily net asset value (NAV). Three index tracker funds are removed from this sample by using Morningstar's online fund assessments, as their returns are clearly not a product of stock picking abilities (Cuthbertson et al. (2008) do this as well). One fund is removed as it turns out to be closed-end instead of open-end. Both dead and alive funds are included to overcome the survival basis described by Elton, Gruber, and Blake (1996) and Carhart (1997). Survival bias is the tendency for failed mutual funds to be excluded from the sample, resulting in an overrepresentation of funds that perform well. This overrepresentation leads towards an overestimation of positive outperformance compared to the entire population of funds. A possible reason for survival bias is that poor performing funds leave the market as money from investors flows to better performing funds over time (Chevalier and Ellison, 1999). As long as a fund has existed for 36 months or more ($T_i \geq 36$) it is included in the sample of this study, the same criterion Cuthbertson et al. (2008) use. In this way there is a balance between a sufficient number of data points to overcome noise problems and a possible survival bias. In the case of duplicate funds with similar returns, the fund with the smallest amount of observations is dropped. The same accounts for mergers and dead funds that are double listed in Datastream. Eight funds are dropped based on these criteria.[6] Taking all inclusion criteria into account, there remain 14 mutual funds with 2362 observations. These are presented in table 1, with T_i as the number of monthly observations. The returns for the remaining 14 funds are extracted from Thomson Reuters Datastream. This return does take into account ongoing annual fees like manager costs and other operating expenses, but is gross of other costs like extra fees, taxes and loads (transaction costs). All values are based on US dollars, as the factors provided by Fama and French (2012) are calculated with US dollars and US T-Bills.

[4] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html#Developed

[5] Cuthbertson et al. (2008) use an 80% boundary, while this study only has access to Thomson Reuters' Lipper codes. These simply deviate between funds with 50% or more of their activities in a particular country. It is however assumable on the basis of publicly available fund reports that many of the selected mutual funds invest more than 80% of their portfolio in Dutch equity, most of them even 100%.

[6] Teslin Capital Management's two funds are both incorporated as they have independent returns.

Surprisingly there are no offshore funds in the remaining sample, indicating that the fund market for Dutch equity is either *1)* largely run by domestic funds, *2)* by funds that only invest a small part of their portfolio in Dutch equity and a larger part in other equities or *3)* missing data in Datastream and/or Thomson Reuters' Lipper codes. Another concerning element is that there is data for all funds at the end of 2015, while there is only data for one fund at the start of 1992. Given these two findings, the data in Datastream's market list or Lipper's fund ratings may not be free of survival bias. Ter Horst et al. (1998) support this idea as they report tens of Dutch mutual equity funds by the end of 1996. Checking with Morningstar's online fund searcher indicates that this sample represents the market of surviving funds, which happens to be the entire Dutch equity fund market in 2015. The 2015 Dutch mutual fund market has only got a small number of players and these are typically large in size as they are the product of mergers and takeovers. Unfortunately this study has to proceed with a sample that is small and not free of survival bias, which makes it complex to extend the results to the entire fund market in the 1992-2015 period and could have an influence on the generation of the bootstrapped *p*-values (see the discussion in sub-section 6.1).

Table 1: Case selection of Dutch mutual equity funds

Fund name	T_i	***Enters dataset***
Achmea	*183*	*2000*
Add Value Fund NV	*106*	*2007*
Allianz	*146*	*2004*
BNP Paribas Netherlands	*190*	*2000*
De Goudse Nederlandse Aandelen	*133*	*2004*
Delta Lloyd Deelnemingenfonds	*256*	*1994*
Generali Aandelenfonds Nederland	*167*	*2002*
Kempen Orange Fund NV	*288*	*1992*
Nationale Nederlanden	*158*	*2002*
Nederlands Aandelenfonds	*115*	*2006*
Robeco Hollands Bezit	*260*	*1994*
Teslin Capital Management BV Darlin	*61*	*2010*
Teslin Capital Management BV Todlin	*61*	*2010*
Zwitserleven Aandelenfonds	*238*	*1996*

5. Results

Sub-section 5.1 generates the fund α's and *t*-alpha's with both the Carhart four-factor and the Fama-French three-factor model and selects a 'best' model. Sub-section 5.2 continues by executing the bootstrap technique on the residuals of the Fama-French three-factor model for the entire 1992-2015 period. Sub-section 5.3 analyses four five-year non-overlapping sub-periods, after which the bootstrap is performed under the CAPM as a robustness check in sub-section 5.4. The results and limitations of this study are discussed in section 6.

5.1. Factor model selection

First one has to determine which factor model serves the bootstrap best, by running both the Carhart four-factor model and the Fama-French three-factor model. Table 2 displays the general results of the Carhart four-factor regressions for all funds in the sample (the cross-section).

Table 2: Carhart four-factor regression for all funds

	Mean	Std. dev.	Min	Max
α (% p.m.)	*-.2609*	*.2609*	*-.7577*	*.2822*
R_m-R_f	*1.0773*	*.0455*	*.9615*	*1.1333*
SMB	*.3575*	*.3552*	*-.1174*	*.9771*
HML	*.0037*	*.1321*	*-.2251*	*.2544*
MOM	*-.0217*	*.0984*	*-.2311*	*.0957*
Adj. R^2	*0.7941*		*.5020*	*0.9094*

* The reported standard errors are Newey-West (1987) heteroscedasticity- and autocorrelation adjusted standard errors.

The mean adjusted R^2 of 79.41% is quite strong and in general funds generate negative alphas that are not statistically different from zero (using a *t*-test at 5% significance). The R_m-R_f coefficient (β), representing the CAPM intuition, is extremely significant. *SMB, HML* and *MOM* are all insignificant, but this is the cause of large differences between individual funds. Table 3 therefore examines the Carhart four-factor model at the individual fund level. To check for autocorrelation the model was tested with Breusch-Godfrey LM tests with lags up to six months, which control whether the variance in error terms is constant (Breusch & Pagan, 1979). The model tests positive for serial correlation (unreported).

Table 3: Carhart four-factor regressions for each individual fund

	Constant		***R_m-R_f***		***HML***		***SMB***		***MOM***		
	α (% p.m.)	*t-statistic*	*coeff.*	*t-statistic*	*coeff.*	*t-statistic*	*coeff.*	*t-statistic*	*coeff.*	*t-statistic*	*adj. R^2*
Achmea	*-.404*	*-2.701*	*1.089*	*25.149*	*-.108*	*-1.237*	*.123*	*1.194*	*.009*	*0.134*	*0.8325*
Add Value	*.282*	*1.007*	*1.021*	*11.738*	*.076*	*0.516*	*.810*	*4.865*	*-.062*	*-0.578*	*0.7905*
Allianz	*-.758*	*-3.609*	*1.052*	*13.017*	*.100*	*0.583*	*.503*	*3.036*	*.079*	*1.344*	*0.7778*
BNP Paribas	*-.569*	*-3.892*	*1.082*	*23.511*	*-.003*	*-0.039*	*.201*	*2.007*	*.054*	*0.974*	*0.8638*
Delta Lloyd	*-.266*	*-0.813*	*0.962*	*14.577*	*.254*	*1.888*	*.558*	*3.170*	*-.176*	*-1.787*	*0.5020*
Generali	*-.280*	*-2.126*	*1.113*	*21.462*	*-.160*	*-1.348*	*.012*	*0.088*	*.013*	*0.279*	*0.8604*
De Goudse	*-.356*	*-1.746*	*1.077*	*14.409*	*.058*	*0.354*	*.382*	*2.033*	*.096*	*1.311*	*0.7621*
Kempen Orange	*-.021*	*-0.114*	*1.057*	*32.987*	*.171*	*2.289*	*.744*	*10.249*	*.039*	*0.736*	*0.7737*
Nederlandse Aandelenfonds	*-.116*	*-0.721*	*1.133*	*23.713*	*-.016*	*-0.127*	*.051*	*0.393*	*.058*	*1.047*	*0.9094*
Nationale Nederlanden	*-.193*	*-1.372*	*1.118*	*23.448*	*-.116*	*-0.762*	*.102*	*0.681*	*-.032*	*-0.522*	*0.8315*
Robeco	*-.263*	*-1.982*	*1.074*	*26.264*	*.052*	*0.744*	*-.028*	*-0.335*	*-.053*	*-1.027*	*0.8229*
Teslin Darlin	*-.074*	*-0.203*	*1.082*	*12.770*	*-.085*	*-0.429*	*.977*	*3.970*	*-.231*	*-1.867*	*0.7234*
Teslin Todlin	*-.286*	*-1.069*	*1.093*	*13.661*	*-.225*	*-1.141*	*.686*	*4.111*	*-.127*	*-1.293*	*0.7824*
Zwitserleven	*-.347*	*-2.973*	*1.130*	*32.559*	*.052*	*0.808*	*-.117*	*-1.604*	*.031*	*0.912*	*0.8845*

* The t-statistics of this table are computed with Newey-West (1987) adjusted standard errors. Each row represents an individual fund, whereas each two columns report the load of a factor on a fund's return (coeff.) and the corresponding *t*-statistic.

As a solution the model is executed similar to Kosowski et al. (2006) and Cuthbertson et al. (2008) with Newey-West (1987) heteroscedasticity- and autocorrelation adjusted standard errors. At the individual fund level one can notice that Delta Lloyd is an outlier in terms of adjusted R^2. The R_m-R_f coefficient is extremely significant and shows a positive direction. The results for the other factors are mixed. For the majority of funds *SMB* has got a positive and significant coefficient. There seems to be little added value from incorporating *HML* in the regression, as the coefficient switches between a positive and a negative effect and barely shows to be statistically significant. The same accounts for momentum factor *MOM*. With regards to non-normality, ten out of fourteen funds exhibit a non-normal distribution in the residuals of their alphas, which would justify the use of the bootstrap method over conventional methods operating on the normality assumption.

A surprising element of table 3 is that all fund alphas are either negative or not statistically different from zero (based on their parametric *p*-value), which is a remarkable result given the sample is not free of survival bias. No fund is able to outperform the (European) market, while the expectation was that there would be an overrepresentation of positive alpha funds. This result could be explained by Berk and Green's (2004) rational competitive model. Their model theorizes that even if funds positively outperform for a while, an increasing fund inflow and rising marginal costs cause the fund to underperform over a longer period.

To make a comparison between the main factor models we now turn to the Fama-French three-factor model. Table 4 reports the general results of the three-factor model for all funds in the sample. With a mean adjusted R^2 of 79.35% the model explains slightly less variation than the four-factor model.

Table 4: Fama-French three-factor regression for all funds

	Mean	**Std. dev.**	**Min**	**Max**
α (% p.m.)	*-.2821*	*0.2322*	*-0.6839*	*0.2400*
R_m-R_f	*1.0791*	*0.0377*	*1.0052*	*1.1252*
SMB	*.3582*	*0.3561*	*-0.1151*	*0.9785*
HML	*0.0143*	*0.1319*	*-0.1636*	*0.3214*
Adj. R^2	*0.7935*		*0.4952*	*0.9094*

* The reported standard errors are Newey-West (1987) heteroscedasticity- and autocorrelation adjusted standard errors.

Table 5: Fama-French three-factor regressions for each individual fund

	Constant		***R_m-R_f***		***HML***		***SMB***		
Fund name	*α (% p.m.)*	*t-statistic*	*coeff.*	*t-statistic*	*coeff.*	*t-statistic*	*coeff.*	*t-statistic*	*adj. R^2*
Achmea	*-.394*	*-2.897*	*1.085*	*24.475*	*-.109*	*-1.221*	*.125*	*1.191*	*0.8335*
Add Value	*.240*	*0.814*	*1.032*	*14.292*	*.122*	*0.717*	*.821*	*5.251*	*0.7916*
Allianz	*-.684*	*-3.475*	*1.041*	*13.277*	*.046*	*0.280*	*.504*	*3.063*	*0.7778*
BNP Paribas	*-.517*	*-3.840*	*1.064*	*23.873*	*-.015*	*-0.218*	*.217*	*2.229*	*0.8634*
Delta Lloyd	*-.486*	*-1.296*	*1.005*	*17.265*	*.321*	*2.322*	*.546*	*3.024*	*0.4952*
Generali	*-.266*	*-2.047*	*1.109*	*21.743*	*-.164*	*-1.379*	*.015*	*0.108*	*0.8612*
De Goudse	*-.268*	*-1.363*	*1.063*	*14.477*	*-.009*	*-0.056*	*.379*	*2.017*	*0.7618*
Kempen Orange	*.026*	*0.144*	*1.049*	*38.958*	*.156*	*1.872*	*.747*	*10.260*	*0.7739*
Nederlandse Aandelenfonds	*-.073*	*-0.450*	*1.123*	*22.579*	*-.058*	*-0.498*	*.044*	*0.337*	*0.9094*
Nationale Nederlanden	*-.224*	*-1.524*	*1.125*	*22.064*	*-.100*	*-0.686*	*.099*	*0.649*	*0.8323*
Robeco	*-.328*	*-2.529*	*1.087*	*23.175*	*.072*	*1.046*	*-.032*	*-0.359*	*0.8224*
Teslin Darlin	*-.272*	*-0.700*	*1.099*	*13.700*	*.049*	*0.262*	*.978*	*3.888*	*0.7190*
Teslin Todlin	*-.395*	*-1.342*	*1.102*	*15.856*	*-.151*	*-0.746*	*.687*	*4.406*	*0.7829*
Zwitserleven	*-.310*	*-2.757*	*1.123*	*30.493*	*.040*	*0.650*	*-.115*	*-1.562*	*0.8846*

* The t-statistics of this table are computed with Newey-West (1987) adjusted standard errors. Each row represents an individual fund, whereas each two columns report the load of a factor on a fund's return (coeff.) and the corresponding *t*-statistic.

To check for serial correlation the model is tested with Breusch-Godfrey LM tests with lags up to six months. Similar to the Carhart four-factor model, the three-factor model tests positive for serial correlation. The model is therefore executed by using Newey-West (1987) adjusted standard errors. Again the distributions of ten out of fourteen individual fund alphas turn out to be non-normal, providing a foundation for the use of the bootstrap method. The average fund generates a negative alpha that is not statistically different from zero. Similar to the four-factor model, the coefficient for R_m-R_f is extremely significant. On a general level the factors for size and book-to-market ratio are not statistically different from zero, however, large differences can be found at the individual fund level. Table 5 therefore examines the Fama-French three-factor model at the individual fund level.

At the individual fund level one can notice that the three-factor model does a reasonable job for all funds except for Delta Lloyd in terms of adjusted R^2. Whereas the coefficients for R_m-R_f are positive and extremely significant, the coefficients for *SMB* and *HML* give mixed results. For the majority of funds *SMB* has got a positive and significant coefficient, whereas *HML* only shows to be positive and significant for two funds.

From the discussion on the models it becomes clear that the Fama-French three-factor model is the 'best' factor model for studying Dutch fund performance when using a European benchmark, especially as the coefficient for momentum the Carhart four-factor model is not statistically different from zero. The Schwarz Information Criterion (SIC) further acknowledges this with a minimum score for the Fama-French model (848.19 vs. 852.05). The main bootstrap will therefore be executed by using the vectors and residuals from the Fama-French three-factor model.

5.2. Main bootstrap analysis

The bootstrap has generated 1000 simulated alphas (α) and *t*-statistics (t_α) for each of the 14 funds in the sample, a total of 14,000 simulations. These are summarized in table A1 and A2 of the appendix. The sorting process is for example able to show the 14 best t_α*'s* (or α*'s*) and the 14 worst t_α's (α*'s*) in the simulated cross-section. The results from the Fama-French three-factor model presented in table 5 are sorted from best t_α (or α) to worst t_α (α) to rank each individual fund. Now it is possible to compare the worst empirically observed t_α to the cross-sectional distribution of all worst t_α's generated by sampling variation. As explained in section 3.2, this study mainly interprets on the basis of sorting t_α**,** but funds sorted by α are also reported to demonstrate the difference in inference**.** Scaling alpha by standard error makes a necessary correction for short-lived funds and funds that take a higher level of risk, as their standard error is likely to be larger (Kosowski et al., 2006).

Table 6: The cross-section of mutual funds 1992-2015

	Fund position												
	Positive α		***Negative α***										
	Best	*2nd*	*12th*	*11th*	*10th*	*9th*	*8th*	*7th*	*6th*	*5th*	*4th*	*3rd*	*2nd*	*Worst*
Panel A: α-sorted													
Actual α (% p.m)	*0.240*	*0.026*	*-0.073*	*-0.224*	*-0.266*	*-0.268*	*-0.272*	*-0.310*	*-0.328*	*-0.394*	*-0.395*	*-0.486*	*-0.517*	*-0.684*
Bootstrapped p-value (strict)	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*0.929*	*0.929*	*0.714*	*0.786*	*0.500*
Bootstrapped p-value (mild)	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*0.857*	*0.786*	*0.500*
Parametric p-value	*0.418*	*0.886*	*0.653*	*0.130*	*0.042*	*0.175*	*0.487*	*0.006*	*0.012*	*0.004*	*0.185*	*0.196*	*0.000*	*0.001*
Panel B: t_α-sorted	***Positive t_α***		***Negative t_α***											
Actual t_α	*0.814*	*0.144*	*-0.450*	*-0.700*	*-1.296*	*-1.342*	*-1.363*	*-1.524*	*-2.047*	*-2.529*	*-2.757*	*-2.897*	*-3.475*	*-3.840*
Bootstrapped p-value (strict)	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*0.929*	*0.714*	*0.214*	*0.214*	*0.071*	*0.071*
Bootstrapped p-value (mild)	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*0.929*	*0.857*	*0.857*	*0.714*	*0.071*
Parametric p-value	*0.418*	*0.886*	*0.653*	*0.487*	*0.196*	*0.185*	*0.175*	*0.130*	*0.042*	*0.012*	*0.006*	*0.004*	*0.001*	*0.000*

* Table 6 shows the position of funds within the cross-section of simulations, either sorted by α (panel A) or by t_α (panel b). The bootstrapped *p*-value indicates what percentage of the simulations from the cross-sectional luck distribution is above the actually observed value in the case of positive outperformance and what percentage of the simulations from the cross-sectional luck distribution is below the actually observed value in the case of negative outperformance. A p-value of 1.000 in the case of positive outperformance for example indicates that all simulated α's (or t_α's) within the cross-sectional distribution are higher than the empirically observed data, thus a clear indication for luck. A p-value of 0.000 in the case of negative outperformance indicates that all simulated funds within the cross-sectional distribution perform better (less negative α's or t_α's) than the empirically observed data, which translates to bad skill. The parametric p-value simply resembles the position of the *t*-value from its own estimated alpha. Strict and mild refer to which group in the simulated data the empirically observed data is compared to. The strict bootstrapped p-value compares for example the worst empirical fund to the worst 14 simulations, the 2nd worst fund to the 2nd worst 14 simulations etc. The mild bootstrapped *p*-value compares all negative funds to the worst 14 simulations (making it slightly easier to be unfortunate instead of unskilled).

The results of the cross-section analysis for the 1992 - 2015 period are presented in table 6. Panel A shows the results of the bootstrap method for each fund sorted by its original Fama-French alpha. Panel B reports the results for each fund sorted by $t_{\alpha,}$ obtained from the original Fama-French regression (see table 5). Each column represents the results for an individual fund. The thick line between columns 2 (2nd) and 3 (12th) indicates that the first two columns are positive outperformers, whereas columns 3-14 report the results for negative outperformers. The first row reports the actual α as a percentage per month, indicating the return that cannot be explained by the other factors in the model. The second row reports the *strict* bootstrapped *p*-value. The *strict* bootstrapped *p*-value compares the value of an empirically observed alpha (or t-alpha) to its corresponding ranked distribution in the cross-section. The worst (best) empirical fund is for example compared to the worst (best) 14 simulations in the cross-section, the 2nd worst (best) fund to the 2nd worst (best) 14 simulations, the 3rd worst (best) fund to the 3rd worst (best) 14 simulations etc. It could therefore also be classified as the *ranked* bootstrapped *p*-value. The fourth row reports the *mild* bootstrapped *p*-value. Remember this is the same bootstrapped *p*-value used by Kosowski et al. (2006), as it compares the position of all real world positive (negative) funds to the best (worst) 14 bootstrap simulations in the cross-section. It is called *mild* because it is easier with this *p*-value to be unfortunate than is the case with the *strict* bootstrapped *p*-value. A bootstrapped *p*-value of 1.000 means that all simulations in the corresponding cross-section of funds are extremer than that individual fund's real world alpha (*t*-alpha in the case of panel B) and thus a clear indication of luck. A bootstrapped *p*-value of 0.000 indicates that the empirically observed alpha (*t*-alpha) is extremer than all simulations in the corresponding cross-section, thus indicating skill. The parametric *p*-value simply resembles the *p*-value from the original Fama-French regression in table 5. It is used in the *conventional* mutual fund literature to determine whether outperformance is statistically different from zero, but is biased as it is based on the normality assumption. The previous statistics are also reported in panel B, but now the funds are sorted by *t*-alpha instead of alpha. The first row of panel B therefore reports *t*-alpha instead of alpha.

There are clear differences in ranking when using the *t*-statistic of alpha instead of alpha.[7] BNP Paribas is for example the worst fund when ranking by t-alpha, whereas Allianz is the worst fund when ranking by alpha. The bootstrapped *p*-values of the two positive funds (columns 1 and 2) in panel B

[7] Ranking the empirically observed data by α gives the following order for positive α's: Add Value Fund NV (best), Kempen Orange Fund NV (2nd best). The same ranking of empirically observed negative α's results in: Nederlands Aandelenfonds, Nationale Nederlanden, Generali Aandelenfonds Nederland, De Goudse Nederlandse Aandelen, Teslin Capital Darlin, Zwitserleven Aandelenfonds, Robeco Hollands Bezit, Achmea, Teslin Capital Todlin, Delta Lloyd Deelnemingenfonds, BNP Paribas Netherlands (2nd worst), Allianz (worst). Ranking the empirically observed data by t_α gives the following order for positive t_α's: Add Value Fund NV (best), Kempen Orange Fund NV (2nd best). The same ranking of empirically observed negative t_α's results in: Nederlands Aandelenfonds, Teslin Capital Darlin, Delta Lloyd Deelnemingenfonds, Teslin Capital Todlin, De Goudse Nederlandse Aandelen, Nationale Nederlanden, Generali Aandelenfonds Nederland, Robeco Hollands Bezit, Zwitserleven Aandelenfonds, Achmea, Allianz (2nd worst), BNP Paribas Netherlands (worst).

clearly show their positive outperformance is due to luck (both *mild* and *strict*). Their bootstrapped *p*-values of 1.000 indicate that all simulations in the corresponding cross-sectional luck distribution generated extremer (higher) values than they did. As they are not able to beat any simulation from the luck distribution, their performance is attributed to luck. What is interesting about the negative outperformers (columns 3-14) is that most perform badly due to misfortune in both panel A and B. Their bootstrapped *p*-values are much higher than any of two cut-off points. In Kosowski et al. (2006) and Cuthbertson et al. (2008) on the contrary, almost all negative outperformers perform badly as a consequence of their own bad skill (see the discussion in sub-section 6.1). Only the worst two funds reported in columns 13 and 14 (*strict* bootstrapped *p*-values of 0.071) show evidence of bad skill when using a bootstrapped *p*-value of 0.143 as cut-off point to reject the null that outperformance is caused by luck. When using the stricter cut-off point ($p<0.05$) of Kosowski et al. (2006), their negative outperformance is attributable to bad luck. When using a *mild* bootstrapped *p*-value instead of a *strict* bootstrapped *p*-value, only one bad skill fund remains when using a 0.143 *p*-value cut-off point. The chosen cut-off point seems to be quite influential on whether a fund is unskilled or misfortunate. Therefore there is only weak evidence that the worst funds in the sample underperform the benchmark as a consequence of their own bad skill. The alphas corresponding to the worst two *t*-alphas in panel B are -0.517% p.m. and -0.684% p.m. This signifies that they underperform the benchmark, albeit not extremely strong. The worst funds in Kosowski et al. (2006) for example generate much lower corresponding alphas (-3.6% p.m.).

Note the difference in inference between the bootstrapped *p*-values and the parametric *p*-values for the 3^{rd}-6^{th} worst performing funds in panel B. Conventional studies that operate on the normality assumption would have assigned these funds' negative outperformance to bad skill ($p<0.05$), whereas both types of bootstrapped *p*-values indicate bad luck, regardless of the cut-off point. The difference between this study's findings and Kosowski et al. (2006) and Cuthbertson et al. (2008) is that the latter two find a small number of skilled positive outperformers in the extreme positive tail of the distribution. They do however draw conclusions from a much larger sample of funds, so their positive tail is much richer in number and data than this study (see sub-section 5.5).

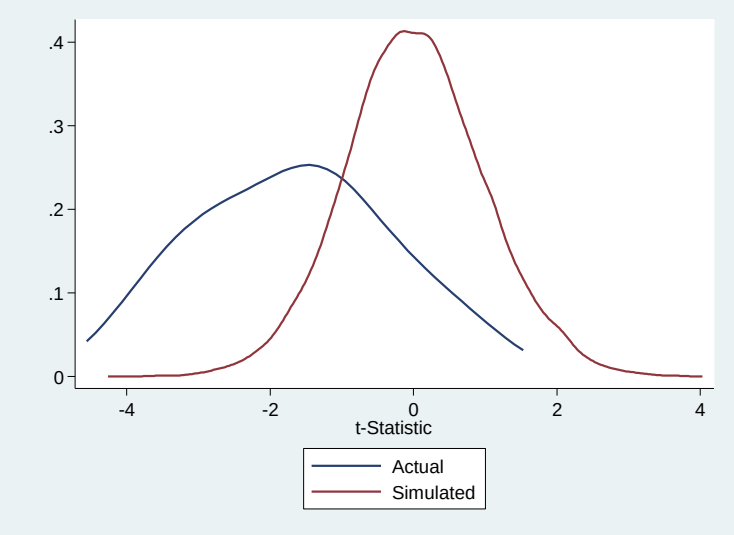


Figure 2: Actual distribution vs. bootstrapped distribution. *The figure plots the kernel density estimate of the probability density function (PDF). 'Actual' (blue) indicates the distribution of the t-alpha from the original Fama-French regressions, whereas 'simulated' (red) plots the distribution of all simulations from the bootstrap.*

Figure 2 plots the kernel density estimate of the probability density function (PDF) for both the bootstrapped data (14,000 simulations) and the empirical data (14 funds) with regards to the *t*-alpha. The figure allows to graphically compare the distribution of the empirically observed data to the distribution of the bootstrap simulations. The x-axis represents all possible values of the *t*-statistic, whereas the y-axis provides a value to calculate the probability that *t*-alpha lies in a small interval around the corresponding *t*-alpha on the x-axis. This figure allows for an alternative interpretation of the bootstrap results, as it is now able to compare how many funds one would expect by sampling variation to have a specific *t*-alpha to the actual number of funds with this *t*-alpha in the real world (Cuthbertson et al., 2008). The difference between the simulated distribution and the real world (actual) distribution indicates that one would expect less *t*-alphas with strong negative values than actually observed in the real world data. This can therefore be interpreted as some evidence of bad skill for the worst funds. The PDF indicates that the empirically observed positive tail of t_α's is explained rather well by sampling variation, as it falls within the bootstrapped luck distribution. By luck alone one would have expected more real world funds to positively outperform than currently is the case. This does not account for the negative tail of the empirically observed distribution, as it is positioned far to the left of the bootstrapped distribution. The PDF therefore provides some additional evidence that there is bad skill amongst the worst funds, but mind differences are amplified as 'actual' is only based on a total number of 14 funds. The reason bad skill is more obvious in figure 2 than in the results from table 6 is that the latter's bootstrapped *p*-values are generated on the basis of a distribution *within* the distribution of simulations, so it really is an alternative way to examine the results.

5.3. Sub-period analysis

The analysis on the entire 1992-2015 period has shown there is only evidence of skill amongst the worst performers, most funds are simply lucky of unfortunate. The absence of positive skilled outperformers could be explained by Berk and Green's (2004) causal mechanism of increasing marginal costs, but this idea not been tested yet. If their theory on an inflow of money and increasing marginal costs for positive outperformers is true, it could be the case that there are skilled positive outperformers over smaller periods of time that do not show up when bootstrapping the entire 1992-2015 period. To further test this idea this sub-section executes the bootstrap on four non-overlapping five-year sub-periods. To be included a fund must at least have 36 observations in the sub-period, otherwise there might be too many noise-related problems. Table 7 reports the number of funds per period and the Schwarz Information Criterion (SIC) scores for both the Carhart four-factor and the Fama-French three-factor model. The Fama-French three-factor model has the minimum SIC score in each sub-period, indicating it is the superior model for all sub-period bootstraps. This sub-section only reports *t*-statistic sorted bootstrap results, as the *t*-statistic has superior statistical properties over alpha. As the number of funds becomes extremely small, especially in the first two sub-periods, the results from this sub-section with regards to luck and skill have to be handled with care (see the discussion in sub-section 6.1).

Table 7: Sub-period model description

Period	***Carhart 4-factor SIC score***	***Fame-French 3-factor SIC score***	***Funds included (T≥36)***
January 1995 - December 1999	*318.71*	*315.76*	*4*
January 2000 - December 2004	*294.73*	*293.23*	*6*
January 2005 – December 2009	*300.49*	*297.46*	*11*
January 2010 – December 2014	*287.64*	*284.65*	*14*

Table 8 reports the bootstrap results for the January 1995 - December 1999 period. Panel A reports four levels from the cross-section of bootstrap simulations, out of 1000 possible levels. Each column represents one level from the sorted cross-sectional distribution. The first column for example reports

the highest simulated *t*-alphas generated by executing the bootstrap procedure on 4 funds. Columns 2-4 report the corresponding cross-sectional distribution for the three worst *t*-alphas. Panel B reports the inference statistics similar to panel B in table 6, except now the second row reports the alpha that corresponds to the sorted *t*-alpha in the first row. The two panels expose that having an even smaller sample of funds has a large impact on the generation of the bootstrapped *p*-value. The bootstrapped *p*-value in such a small sample only compares the empirically observed value to four simulations, including a simulation from its own luck distribution. The conclusion for all funds in the sample is simply that their outperformance is caused by (mis)fortune, as no real world *t*-alpha is able to beat the luck distribution (bootstrapped *p*-values of 1.000 in all cases). There are large difference between the bootstrapped *p*-values and the parametric *p*-values, for the worst fund in the sample this makes the difference between being unskilled or unlucky, but it is doubtful whether the bootstrap is more reliable than the parametric *p*-values in such a small sample (see discussion in sub-section 6.1).

Table 8: Skill vs. luck January 1995 - December 1999

Panel A: Bootstrap simulations

Positive t_α		*Negative t_α*	
Best	***3rd***	***2nd***	***Worst***
6.154	*-3.494*	*-4.080*	*-4.715*
4.856	*-3.362*	*-3.777*	*-3.930*
4.298	*-3.337*	*-3.434*	*-3.721*
3.638	*-1.980*	*-2.047*	*-2.112*

Panel B: Inference

	Positive t_α		*Negative t_α*	
Fund	***Best***	***3rd***	***2nd***	***Worst***
Actual t_α	*0.055*	*-0.033*	*-0.528*	*-2.023*
Corresponding α	*0.013*	*-0.016*	*-0.158*	*-2.022*
Strict p-value	*1.000*	*1.000*	*1.000*	*1.000*
Mild p-value	*1.000*	*1.000*	*1.000*	*1.000*
Parametric p-value	*0.956*	*0.974*	*0.600*	*0.048*

** Panel A reports the bootstrap generated t_α's and sorts them from high to low, leaving the individual luck distribution intact. Panel B presents the results of the inference procedure sorted by t_α. The underlying factor model is the Fama-French three-factor model and reported t-statistics are based on Newey-West errors. The parametric p-value simply resembles the position of the t-value from its own estimated alpha. Strict and mild refer to which group in the simulated data the empirically observed data is compared to. The strict bootstrapped p-value compares a real world t-alpha to its corresponding ranked distribution in the cross-section of funds. The mild bootstrapped p-value compares all negative funds to the worst 4 simulations in the cross-section.*

Table 9 reports the bootstrap results for the January 2000 – December 2004 period, similar to table 8. Panel A reports the bootstrap simulations for the best simulations and the 5 worst simulations in the cross-section, separated by the thick line between columns 1 and 2. The reason only these levels are reported is that there are 5 negative outperformers and 1 positive outperformer in the sample. Panel B reports the sorted *t*-alpha and its corresponding alpha and *p*-values. The first column represents the only positive *t*-alpha from the original Fama-French regression, columns 2-6 the negative *t*-alphas. The results are similar to the 1995-1999 period, as no outperformance can be admitted to genuine stock picking skills (bootstrapped *p*-values are all larger than 0.05 or 0.143). Most funds underperform the benchmark because of bad luck. The four worst funds would have been classified as bad skilled funds when using the parametric *p*-value, whereas they are unlucky when using the bootstrapped *p*-value. See the discussion in sub-section 6.1 on the reliability of bootstrapped *p*-values in small samples.

Table 9: Skill vs. luck January 2000 – December 2004

Panel A: Bootstrap simulations

Positive t_α	***Negative*** t_α				
Best	***5th***	***4th***	***3rd***	***2nd***	***Worst***
5.919	*-3.683*	*-3.715*	*-3.998*	*-4.887*	*-7.238*
5.514	*-3.522*	*-3.700*	*-3.887*	*-4.389*	*-4.953*
4.835	*-3.269*	*-3.311*	*-3.800*	*-4.142*	*-4.328*
3.959	*-3.236*	*-3.246*	*-3.592*	*-3.950*	*-4.159*
3.332	*-3.021*	*-3.159*	*-3.321*	*-3.635*	*-3.689*

Panel B: Inference

	Positive t_α	*Negative* t_α				
Fund	***Best***	***5th***	***4th***	***3rd***	***2nd***	***Worst***
Actual t_α	*0.411*	*-0.962*	*-2.694*	*-2.798*	*-3.636*	*-3.914*
Corresponding α	*0.334*	*-0.477*	*-0.788*	*-0.927*	*-0.990*	*-1.067*
Strict p-value	*1.000*	*1.000*	*1.000*	*1.000*	*0.667*	*0.667*
Mild p-value	*1.000*	*1.000*	*1.000*	*1.000*	*0.833*	*0.667*
Parametric p-value	*0.682*	*0.340*	*0.009*	*0.007*	*0.001*	*0.000*

** Panel A reports the bootstrap generated t_α's and sorts them from highest to lowest, leaving a fund's individual luck distribution intact. Panel B presents the results of the inference procedure sorted by t_α. The underlying factor model is the Fama-French three-factor model and reported t-statistics are based on Newey-West (1987) errors. The parametric p-value simply resembles the position of the t-value from its own estimated alpha. Strict and mild refer to which group in the simulated data the empirically observed data is compared to. The strict bootstrapped p-value compares for example the worst empirical fund to the worst 6 simulations, the 2nd worst fund to the 2nd worst 6 simulations etc. The mild bootstrapped p-value compares all negative funds to the worst 6 simulations.*

Table 10: Skill vs. luck January 2005 – December 2009 & January 2010 – December 2014

Panel A: January 2005 – December 2009

	Positive t_α						*Negative t_α*				
Fund	***Best***	***2nd***	***3rd***	***4th***	***5th***	***6th***	***5th***	***4th***	***3rd***	***2nd***	***Worst***
Actual t_α	*1.835*	*0.948*	*0.535*	*0.478*	*0.431*	*0.302*	*-0.434*	*-0.555*	*-0.925*	*-1.321*	*-1.555*
Corresponding α	*0.425*	*0.213*	*0.134*	*0.138*	*0.105*	*0.064*	*-0.111*	*-0.132*	*-0.225*	*-0.352*	*-0.352*
Bootstrapped p-value (strict)	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*
Bootstrapped p-value (mild)	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*
Parametric p-value	*0.072*	*0.347*	*0.595*	*0.635*	*0.668*	*0.763*	*0.666*	*0.581*	*0.359*	*0.192*	*0.126*

Panel B: January 2010 – December 2014

Fund	***14th***	***13th***	***12th***	***11th***	***10th***	***9th***	***8th***	***7th***	***6th***	***5th***	***4th***	***3rd***	***2nd***	***Worst***
Actual t_α	*-0.297*	*-0.516*	*-0.607*	*-0.914*	*-0.993*	*-1.136*	*-1.429*	*-1.437*	*-1.441*	*-1.820*	*-2.099*	*-2.244*	*-2.279*	*-6.358*
Corresponding α	*-0.119*	*-0.104*	*-0.081*	*-0.345*	*-0.176*	*-0.324*	*-0.442*	*-0.425*	*-0.279*	*-0.348*	*-0.528*	*-0.423*	*-0.395*	*-1.096*
Bootstrapped p-value (strict)	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*0.000*
Bootstrapped p-value (mild)	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*0.000*
Parametric p-value	*0.768*	*0.608*	*0.546*	*0.366*	*0.325*	*0.261*	*0.160*	*0.156*	*0.155*	*0.074*	*0.040*	*0.029*	*0.027*	*0.000*

** Panel A reports the inference procedure for the 2005-2009 period and is based on the bootstrap simulations reported in table A3 of the appendix. Panel B reports the inference procedure for the 2010-2014 period and is based on the bootstrap simulations reported in table A4 of the appendix. The underlying factor model is the Fama-French three-factor model and reported t-statistics are based on Newey-West errors. The parametric p-value simply resembles the position of the t-value from its own estimated alpha. Strict and mild refer to which group in the simulated data the empirically observed data is compared to. The strict bootstrapped p-value compares the t-*alpha to its corresponding ranked cross-sectional distribution*. The mild bootstrapped p-value compares all negative funds to the worst generated t-alphas from the cross-section of funds.*

Table 10 reports the bootstrap results for the January 2005 – December 2009 period in panel A and the bootstrap results for the January 2010 – December 2014 period in panel B (see tables A3 and A4 in the appendix for the corresponding simulations). The composition of the panels is similar to panel B in tables 8 and 9. The first rows in both panels of table 10 report the actually observed *t*-alpha, sorted from highest to lowest. The second row reports the alpha that corresponds to the *t*-alpha in the first row. Rows 3-5 report the different *p*-values. Each column displays the bootstrap results for an individual fund. In panel A columns 1-6 report the results for positive *t*-alpha funds, whereas columns 7-11 do so for negative *t*-alphas (the thick line between the 6th and 7th column serves as a boundary). In panel B one can note this boundary is missing, as the original Fama-French regressions have only generated 14 negative *t*-alphas for this period.

Panel A presents more positive outperformance than observed in any other period, but no fund is able to produce a higher t_α than the cross-sectional luck distribution (both for *mild* and *strict* bootstrapped *p*-values). Their outperformance is therefore clearly attributable to luck. Likewise, all five negative outperformers produce *t*-alphas that are higher (less negative) than the cross-sectional luck distribution, indicating all outperformance is due to misfortune instead of bad skill. All funds in the 2010-2014 period have produced negative t_α's, of which some have substantial corresponding negative alphas. The entire population of funds except for the worst fund has negatively outperformed because of misfortune.

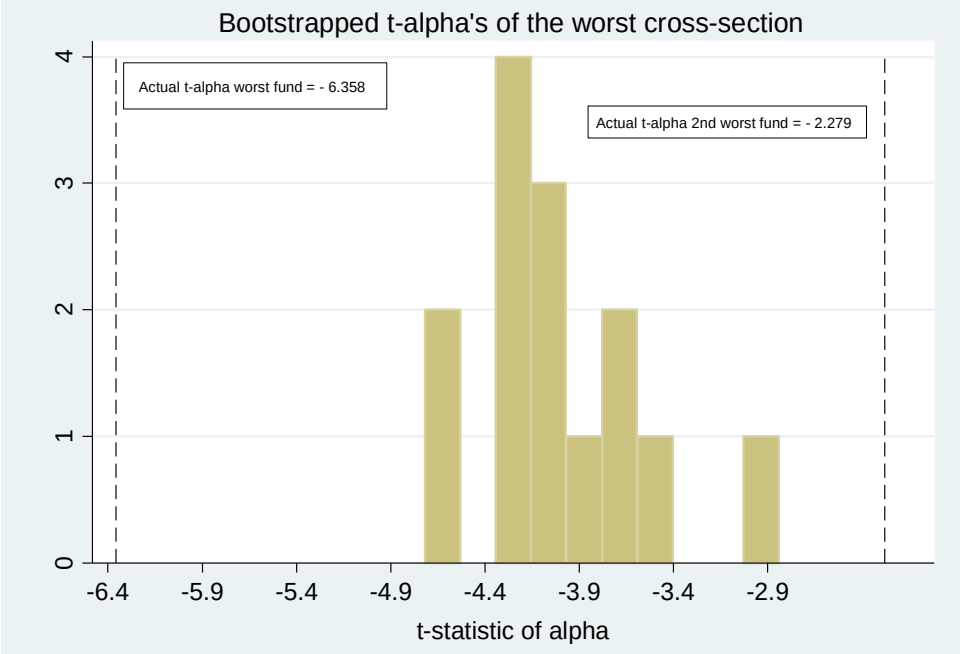


Figure 3: Histogram of worst cross-sectional bootstrap t_α simulations 2010-2014

The result of the worst fund, Delta Lloyd, is remarkable. Not only does this fund have an extremely large empirically observed negative t_α (-6.358), even within its own luck distribution no simulation is more extreme than the actually observed t_α. With a bootstrapped *p*-value of 0.000 (either *mild* or *strict*) the fund is even attributed with bad skill when using the 0.05 bootstrapped *p*-value cut-off point of Kosowski et al. (2006). This is strong evidence of bad skill. To further illustrate this, the histogram of figure 3 plots the worst simulated cross-sectional distribution of *t*-alphas based on appendix table A4, column 14. The bars report the values of the worst simulated *t*-alphas from each individual fund's luck distribution by frequency of appearance in the cross-section. The actual t_α's of the worst two funds are represented by the reference lines. The reference line on the left corresponds to Delta Lloyd's actual t_α (-6.358), the reference line on the right to the 2nd worst fund (t_α=-2.279). Delta Lloyd's actual t_α is clearly positioned much lower than all the cross-sectional bootstrap simulations represented by the bars, resulting in a bootstrapped *p*-value of 0.000 (both *mild* and *strict*). On the other hand, the 2nd worst fund has a higher *t*-statistic than any of the simulations in the cross-section, which results in a *mild* bootstrapped *p*-value of 1.000. This *p*-value indicates that all simulations in the cross-section are extremer (lower) than the actually observed *t*-alpha. Note that in order to calculate the *strict* bootstrapped *p*-value of the 2nd worst fund, one would have to consider its position in the cross-sectional distribution of 2nd worst simulated *t*-alphas (reported in appendix table A4, column 13).

With regards to performance persistence, there are three funds that negatively outperform the benchmark three sub-periods in a row: BNP Paribas, Achmea and Allianz.[8] These are exactly the funds belonging to the worst group of funds in the 1992-2015 period analysis. It is interesting however that none of these funds provide evidence of bad skill in the sub-periods, while two of them (BNP Paribas and Allianz) provide weak evidence of bad skill in the full period analysis. Delta Lloyd provides strong evidence of bad skill in one sub-period (2010-2014), while its cross-sectional ranking in the complete 1992-2015 period is not so dreadful and caused by misfortune. Different cross-sections and time-periods are likely the cause of this (see the discussion in 6.2).

5.4. CAPM bootstrap

The model selection procedure of sub-section 5.1 picked the Fama-French three-factor model over the Carhart four-factor model, as momentum factor *MOM* turned out to be insignificant and the four-factor model SIC score turned out to be inferior. So far this study has not taken into account the Capital Asset Pricing Model (CAPM), as the three- and four-factor model have been central to the mutual fund literature. It is however interesting from a robustness point of view to observe whether there emerge substantial differences when using another model than the Fama-French three-factor model. Kosowski

[8] The ranking of funds by name is unreported in the tables, but is available upon request from the author.

et al. (2006) for example compare many different conditional and unconditional factor models, but find no critical differences. The CAPM factor R_m-R_f turned out to be extremely significant in section 5.1., but the SIC score of the CAPM (*852.65*) is still inferior to the Fama-French three-factor model (*848.19*) and the Carhart 4-factor model (*852.05*).

Table 11 therefore reports the results of the bootstrap over the 1992-2015 period, but now with the CAPM as underlying factor model. Panel A reports the summary statistics of the original CAPM regressions. The results of the Fama-French three-factor (FF3) model are reported as well in parentheses to easily compare the two models. The first row reports the statistics for the CAPM alpha expressed in percentage per month, the second row similarly expresses the FF3 alpha in parentheses. The third row reports the statistics for the R_m-R_f coefficient from the CAPM, the fourth row does so for the FF3 model. Likewise, row 5 and 6 express the statistics of the adjusted R^2 for both models. Columns 1-4 report the values for mean, standard deviation and the extremes. Panel B reports the bootstrap inference similar to previous tables. Each column reports the statistics for an individual fund and the first row reports the empirically observed cross-section of sorted *t*-alphas from highest (first column) to lowest (14th column). The thick line between columns 2 and 3 indicates that the first two columns indicate positive outperformers, whereas columns 3-14 report negative outperformers. The second row reports the sorted *t*-alphas of the FF3 model in parentheses. Rows 3 and 4 report the alpha that corresponds to the reported *t*-alpha for respectively the CAPM and the FF3 model. Rows 5-7 report the possible set of *p*-values for the CAPM. The corresponding set of bootstrap simulations is reported in table A5 of the appendix.

Panel A reports only minor differences. The Fama-French three-factor model tends to generate slightly extremer values than the CAPM for all statistics except the maximum adjusted R^2 and the standard deviation of the R_m-R_f coefficient. The CAPM for example generates a mean negative α that is 5.34 basis points higher than the three-factor model. It is however striking how much variation is explained by CAPM alone. Panel B reports the CAPM bootstrap inference and reveals some small differences with the Fama-French three-factor model (see table 6, panel B). The number of positive and negative funds remains the same, just as their cross-sectional ranking. Most funds in the CAPM bootstrap underperform because of misfortune, some underperform as a consequence of their own bad skill. No skilled positive outperformers emerges since all bootstrapped *p*-values (1.000) reported in columns 1 and 2 are far above any cut-off point. A difference between the CAPM inference and the Fama-French three-factor inference can be found in the negative tail of the distribution. If one considers the *strict* bootstrapped *p*-value of the 3rd worst firm in the CAPM this fund negatively outperforms because of bad skill, whereas under the Fama-French three-factor model this fund is unfortunate. When using the mild *p*-value however this fund is also unfortunate under CAPM. The conclusions with regards to luck and skill are similar for the 2nd worst fund, even though the

bootstrapped *p*-values are lower in the CAPM bootstrap. The worst fund in the CAPM is unfortunate both under the *strict* and the *mild* bootstrapped *p*-value, as the bootstrapped *p*-value is higher than the 0.05 cut-off point of Kosowski et al. (2006) and equal to the additional 0.143 cut-off point (must be lower than 0.143 to indicate bad skill). Remember that in the Fama-French analysis this fund was only badly skilled when considering a 0.143 cut-off point. Taking these small differences and the mutual fund literature into account, there seems little reason to believe that different models will yield completely different outcomes.

Table 11: CAPM bootstrap 1992-2015

Panel A: CAPM summary statistics

	Mean	Std. dev.	Min	Max
α (% p.m.)	*-.2287*	*0.2243*	*-0.6118*	*0.2255*
(FF3)	*(-.2821)*	*(0.2322)*	*(-0.6839)*	*(0.2400)*
R_m-R_f	*1.0583*	*0.0428*	*0.9884*	*1.1364*
(FF3)	*(1.0791)*	*(0.0377)*	*(1.0052)*	*(1.1252)*
Adj. R^2	*0.7718*		*0.4615*	*0.9106*
(FF3)	*(0.7935)*		*(0.4952)*	*(0.9094)*

Panel B: The cross-section of funds 1992-2015 CAPM

	Positive t_α		*Negative t_α*											
Fund	***Best***	***2nd***	***12th***	***11th***	***10th***	***9th***	***8th***	***7th***	***6th***	***5th***	***4th***	***3rd***	***2nd***	***Worst***
Actual t_α	*0.676*	*0.593*	*-0.300*	*-0.338*	*-0.657*	*-0.924*	*-1.097*	*-1.346*	*-2.145*	*-2.282*	*-2.536*	*-2.998*	*-3.117*	*-3.544*
	(0.814)	*(0.144)*	*(-0.450)*	*(-0.700)*	*(-1.296)*	*(-1.342)*	*(-1.363)*	*(-1.524)*	*(-2.047)*	*(-2.529)*	*(-2.757)*	*(-2.897)*	*(-3.475)*	*(-3.840)*
Corresponding α	*0.226*	*0.113*	*-0.117*	*-0.054*	*-0.187*	*-0.355*	*-0.214*	*-0.194*	*-0.274*	*-0.311*	*-0.309*	*-0.612*	*-0.417*	*-0.496*
	(0.024)	*(0.025)*	*(-0.073)*	*(-0.272)*	*(-0.485)*	*(-0.395)*	*(-0.268)*	*(-0.224)*	*(-0.266)*	*(-0.328)*	*(-0.310)*	*(-0.394)*	*(-0.684)*	*(-0.517)*
Bootstrapped p-value (strict)	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*0.929*	*0.929*	*0.643*	*0.000*	*0.000*	*0.143*
Bootstrapped p-value (mild)	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*0.929*	*0.714*	*0.500*	*0.143*
Parametric p-value	*0.500*	*0.554*	*0.765*	*0.736*	*0.514*	*0.356*	*0.275*	*0.180*	*0.033*	*0.023*	*0.012*	*0.003*	*0.002*	*0.000*

** Panel A reports the summary statistics of the CAPM regressions over the 1992-2015 period with Newey-West adjusted standard errors. Panel B reports the corresponding bootstrap inference and is based on the bootstrap simulations reported in table A5 of the appendix. The parametric p-value simply resembles the position of the t-value from its own estimated alpha. Strict and mild refer to which group in the simulated data the empirically observed data is compared to. The strict bootstrapped p-value compares for example the worst empirical fund to the worst 14 simulations in the cross-section, the 2nd worst fund to the 2nd worst 14 simulations etc. The mild bootstrapped p-value compares all negative funds to the worst 14 simulations in the cross-section. The results of the Fama-French three-factor (FF3) regressions are reported in parentheses. All t-*alphas are *based on* Newey-West (1987) heteroscedasticity- and autocorrelation adjusted standard errors.

6. Discussion

The central question to this study is whether the ex-post performance of Dutch mutual equity funds in the 1992-2015 period is due to luck or skill. The key finding is that there is substantial evidence of (mis)fortune for most funds in the sample and weak evidence of bad skill amongst some of the worst funds. These worst funds keep underperforming the benchmark in the sub-period analysis. No evidence of positive stock picking emerges.

6.1. Small sample and survival bias

With only 14 funds in the sample, and even less in most sub-periods, the bootstrapped *p*-values are generated on the basis of a small cross-sectional comparison group. Kosowski et al. (2006) mention the bootstrap becomes more important in smaller samples, as these result in a larger deviation from the standard normality function than larger samples. The bootstrapped *p*-values deviate stronger from their corresponding parametric *p*-values, making it more difficult to be (un)skilled than to be (un)lucky. At first sight this seems to make the weak evidence of bad skill slightly stronger, however, their 'smaller' samples are still much larger than this study's small sample. It is possible therefore that the sample is too small for correct bootstrap inference, especially in some of the sub-periods. The effect of having an extremely small sample in this specific bootstrap method is to the knowledge of the author ungrounded territory in the literature. Statistical studies on bootstrap methods concentrate on the confidence intervals of the entire simulated distribution, while the bootstrap method of Kosowski et al. (2006) uses the cross-section distribution of best or worst simulations for each individual fund (a distribution within the distribution of simulations). Chernick (2008) notes with regards to bootstrap methods in general, that even though very small sample sizes can be problematic, sample sizes as small as 14 can give 'surprisingly' good results. He mentions that the main problem with small samples is "*that with only a few values to select from, the bootstrap sample will underrepresent the true variability since observations are frequently repeated and bootstrap samples, themselves, can repeat*" (p. 173). As this study requires at least 36 observations to be included, the chance of reproducing a resample is extremely small. The simulated distributions confirm this. It remains a guess, but based on Chernick (2008) it could well be that the severity of small sample size related problems is influenced by the extent to which the sample is representative of the true population of funds. The presence of survival bias could be more problematic if this if the case, especially as Kosowski et al. (2006) point out that short-lived funds are likely to generate more extreme alphas. Increasing the number of repetitions in the bootstrap from B=1,000 to B=10,000 could help in closer approaching the true variability, but whether this is enough remains doubtful. Future research could study the severity of this problem, by

comparing results from larger samples and smaller samples when the true population of funds is known. When it comes to the sub-periods with the smallest samples, it might be better to use parametric *p*-values than bootstrapped *p*-values, even if this ignores the non-normality from which an individual fund's alpha is generated. The bootstrap correction in such a small sample might be worse than the unreliable parametric *p*-value.

Even though Kosowski et al. (2006) and Cuthbertson et al. (2008) indicate that survival bias has a minimal impact on their final bootstrapping results, this might not be the case with this study's survival bias due to the small number of funds in the sample. Incorporating these dead funds could well increase the chance of surviving funds at the negative end of the performance scale to be unfortunate instead of unskilled, due to an increased proportion of strong negative outperformance in the cross-section of funds (a fatter negative tail). Most dead funds are in fact liquidated as a consequence of their own bad performance (Elton et al., 1996). Given the evidence of bad skill for the worst funds in this study's sample is weak, these could be unfortunate within the true population of funds. A dataset free of survival bias probably leads to more similar proportions of bad skill as found in the studies of Kosowski et al. (2006) and Cuthbertson et al. (2008).

With regards to the positive end of the performance scale, it could be that a very small number of non-surviving funds was actually skilled, but merged or taken over by other mutual funds because of their performance (Cuthbertson et al., 2008). Given the Dutch mutual fund market only has a small number of domestic equity funds, which are mostly large players because of mergers and takeovers, it cannot be excluded that there are skilled positive outperformers among the 'dead' funds not incorporated in this study's sample. There is little reason to believe that the positive outperformers in this study's sample would become skilled instead of lucky when the survival bias is lifted, considering their parametric *p*-values are already quite large (1.000). Taking the previous discussion on sample size and survival bias into account, future research on the division of skill and luck in the Netherlands could definitely improve with a completer dataset.

6.2. Performance persistence

The analysis of the five-year sub-periods provides no evidence of positive stock picking skills and supports the idea that most funds are in fact unfortunate or lucky, even though these results have to be handled with care considering the previous discussion on small sample size and survival bias. Only the worst fund in the left tail of the 2010-2014 period cross-sectional distribution demonstrates truly bad skill. Positive short-term persistent outperformance might however surface when evaluating funds several times a year. Bollen and Busse (2005) for example observe short-lived superior abnormal returns when using quarterly measurement periods. Berk and Green's (2004) causal mechanism of

fund money inflow, increasing marginal costs and decreasing performance with regards to positive skilled outperformers can therefore neither be verified nor rejected by the sub-period analysis. Their correlational claim that no fund is able to positively outperform over a longer period of time seems supported by the sub-period analysis.

The differences in bad skill between the main analysis and the sub-period analysis could be explained by the positioning of the sub-periods. The positions of the four five-year non-overlapping sub-periods are chosen arbitrarily and a change in the starting and ending point of a period could have some impact on the results. In addition, some years that are incorporated in the whole period analysis are not included in the sub-period analysis. Another possible reason could be the circumstance that the sub-periods have an even smaller number of funds than the full period analysis. As mentioned in sub-section 6.1 this could have a large impact on the inference procedure and the production of the bootstrapped *p*-values, as the cross-sectional comparison group becomes particularly small.

Other studies like Cuthbertson et al. (2008) in addition use Carhart's (1997) overlapping persistence measure in their analyses, which rolls over past t_α's to find links in outperformance between different periods. Using this method, Kosowski et al. (2006) and Cuthbertson et al. (2008) find evidence for persistence in underperformance for the worst funds, but persistence in positive outperformance is only found for a small group of positive skilled funds in the US. Carhart's (1997) persistence measure could indeed reveal short-term persistence in positive outperformance and provide stronger evidence of persistent underperformance for the worst Dutch mutual funds. But even if there is historical short-term persistent positive outperformance, this would mean little to today's investors as they are (or should be) interested in future long-term persistent positive outperformance that is identifiable *ex-ante*.

6.3. Investor perspective of the Dutch mutual fund industry

The results of this study, in addition to the existing mutual fund literature, still provide enough reasons for investors to assume that actively managed Dutch funds are almost certainly not worth the costs. As most funds are simply (un)lucky, their historical alphas provide no certainty for the future. Investing in actively managed Dutch equity funds on the basis of historical alphas is therefore risky and even if positive skilled outperformance exists *ex-post*, it might be difficult to identify these *ex-ante*, especially in the case of unexperienced investors. Investing in a passive benchmark like an index-tracker fund has in general got less costs and probably provides higher returns than investing in actively managed Dutch equity funds. And what about other factors like transaction costs (loads), different kinds of incidental fees and taxes? From an investor perspective this means even more bad news, as this study uses the return that is gross of such costs. A study by Blake, Caulfield, Ioannidis, and Tonks (2017) illustrates

this point by using cross-section bootstrapping to provide evidence that no UK mutual fund manager is able to beat the luck distribution when using net returns (includes management fees, gross of taxes) instead of gross returns (net of ongoing operating costs, gross of manager fees and taxes). Some funds are able to positively outperform when gross of manager fees, but given no outperformance remains when including these indicates that any positive outperformance does not end on the bank account of the investor, but flows right into the pockets of the fund managers.

One could wonder based on the results of this study and the mutual fund literature why the global active mutual fund industry is still large and existing anyway, given most evidence points towards (persistent) underperformance or zero abnormal returns. Bailey, Kumar, and Ng (2011) give a motivation by showing that *unsophisticated* investors suffer from behavioural biases like framing, narrowing and overconfidence. *Sophisticated* investors better avail bad mutual funds by holding a large proportion of funds over time, avoiding high-expense funds and allocating more of their money in index funds. The biases of *unsophisticated* investors cause them to time their fund purchases badly, trade them frequently as a consequence of chasing behaviour and allocate more of their money in high-expense actively managed funds than in low-cost index funds. It is therefore important to keep informing investors about the risks and costs of active fund management.

6.4. Other points of interest

The inability of the Carhart four-factor model to produce a significant momentum coefficient could be explained by the composition of the benchmark in this study. As the benchmark is based on the markets of several European countries and the momentum effect possibly still too domestic, it might be difficult to translate regional momentum effects to domestic returns. This could be achieved by generating and using factors based solely on a Dutch benchmark, but it is questionable whether a Dutch market portfolio has sufficient stocks to be considered well-diversified enough. The prospect of increasing market integration provides a better opportunity. Given the result from the main analysis is robust under CAPM, and the finding in Kosowski et al. (2006) and Cuthbertson et al. (2008) that model selection has a negligible impact on the division of skill and luck, it seems unlikely that a significant Carhart four-factor model will produce completely different results.

Another point of interest is the discussion in Kosowski et al. (2006) on other sensitivities and alternative bootstraps. One of them is the time-series behaviour of factor models, as the bootstrap results assume that residuals are identically and independently distributed (IID), even though Hall (1992) mentions the basic bootstrap procedure offers natural robustness in case IID is violated. Kosowski et al. (2006) therefore run a robustness test in which they randomly vary the block length of the resampling process up to 10 months instead of 1 month, which allows for dependence in return

residuals over time. The resulting bootstrapped *p*-value Kosowski et al. (2006) present remains nearly identical to the basic bootstrap procedure with block lengths of 1 month. The same accounts for robustness tests in which 1) both the factor returns and the residuals are resampled, 2) residuals are randomly replaced across funds, 3) the number of required observations is varied and 4) a robustness test in which they control for omitted factors. The latter robustness test checks whether an omitted and thus non-priced factor accounts for a fund's outperformance instead of manager performance. Kosowski et al. (2006) mention that this is only likely under the condition that a fund is relatively short-lived and exists in the period this omitted factor has an effect on the fund's return. As the sample of this study is not free of survivor bias and has a substantial number of observations per fund, there seems little reason to assume the results are severely influenced by omitted factors. The other robustness tests could be of added value to this study, but no large differences are expected.

Besides these robustness tests, Fama and French (2010) still have additional remarks on the method of Kosowski et al. (2006). Whereas Kosowski et al. (2006) randomly resample individual fund residuals like this study, Fama and French (2010) argue to jointly sample fund and explanatory returns, taking into account the monthly cross-sectional distribution of the residuals. In this way possible effects of cross-sectional correlated movements in the residuals and returns are not ignored. The bootstrap method of Kosowski et al. (2006) is vulnerable to 'pooling over time' in the presence of altering bearish and bullish periods, which results in narrower confidence intervals compared to Fama and French's (2010) method (Blake et al., 2017). Fama and French (2010) therefore argue that the results of Kosowski et al. (2006) are biased towards skilled positive outperformance. This actually also points out a bias towards bad skill, but the bias seems stronger for positive outperformance (Fama and French, 2010, p. 1940). Cheng and Yan (2017) further confirm this in a Monte Carlo study in which they find that in the presence of cross-sectional dependence in fund returns, the cross-sectional bootstrap method of Kosowski et al. (2006) overestimates the proportion of skilled outperformers. Using Newey-West standard errors might even exacerbate this overestimation (Petersen, 2009). As the sample on the Dutch fund market does not find any skilled positive outperformance, the bias is not concerning to the outcomes of this study, but future research should take this into account. The result that the worst funds in this study's sample indicate weak evidence of bad skill could possibly be altered by using Fama and French's (2010) bootstrapping method, but this does not change this study's recommendations. Blake et al. (2017) further compare the two cross-section bootstrap methods under equal conditions (same market, inclusion criteria and time period) and find evidence of (mis)fortune and bad skill with both, but only the method of Kosowski et al. (2006) provides evidence of good stock picking skills. These superior returns are however completely extracted by managers through fees, so from an investor point of view it might not matter which of the two bootstrap methods is used.

7. Conclusion

Using the cross-section bootstrap method of Kosowski et al. (2006), this study aims to answer the question whether the *ex-post* performance of Dutch mutual equity funds in the 1992-2015 period is due to luck or skill. The study consists of a sample of 14 actively managed Dutch open-end domestic equity funds, which represents the entire active population of funds in 2015. The cross-section bootstrap method is preferred over parametric methods used in the conventional mutual fund literature, as it does not require a fund's idiosyncratic risk to be normally distributed. This is an essential element of the bootstrap method, as there appear many non-normalities in the cross-section of mutual fund alphas and the individual fund alpha distributions. The main analysis is executed with the European Fama-French three-factor model, but running the bootstrap with the CAPM provides similar results. Due to a survival bias in the dataset, the study has not been able to incorporate any dead funds.

The results of this study provide evidence of luck for all positive outperforming funds and evidence of misfortune for most negative outperformers. Only the worst funds in the sample provide some evidence of bad skill and persistently underperform the benchmark over four non-overlapping five-year sub-periods. Most funds generate negative abnormal returns. Contrary to the US and UK results of Kosowski et al. (2006) and Cuthbertson et al. (2008), the outcomes provide no evidence of genuine positive stock picking skills. However, it is likely that the survival bias and the small sample size have an influence on these results. The actual proportion of funds with bad skill is most likely higher in the true population. Future research should therefore consider a larger sample that is free of survival bias. The answer to the central question is nevertheless that the ex-post performance of Dutch funds in the 1992-2015 period is almost completely attributable to (mis)fortune and bad skill. Investors are therefore advised to invest in low-cost index-trackers, instead of actively managed Dutch equity funds. The results recommend that supervisory bodies like The Dutch Authority for the Financial Markets (AFM) keep informing (unexperienced) investors about the risks and costs of active fund management.

Besides these recommendations, the results provide plenty of leads for future research. This study is for instance limited in the sense that it is descriptive in nature. It does not explain the causes of the division of luck and skill in the Dutch mutual fund industry. The results of Amihud and Goyenko (2013) and Cremers and Petajisto (2009) provide reason to believe skill is related to the level of selectivity, but whether this serves the Dutch case is unknown. Combining their methods with cross-section bootstrapping could provide more reliable results. Statistical research on the limits of the cross-section bootstrap might further help in studying luck and skill in countries with smaller fund industries.

References

Allen, D. E., & Tan, M. L. (1999). A test of the persistence in the performance of UK managed funds. *Journal of Business Finance and Accounting*, *26*(5–6), 559–593.

Amihud, Y., & Goyenko, R. (2013). Mutual Fund's R^2 as predictor of performance. *Review of Financial Studies*, *26*(3), 667–694.

Bailey, W., Kumar, A., & Ng, D. (2011). Behavioral biases of mutual fund investors. *Journal of Financial Economics*, *102*(1), 1–27.

Baks, K. P., Metrick, A., & Wachter, J. (2001). Should investors avoid all actively managed mutual funds? A study in Bayesian performance evaluation. *The Journal of Finance*, *56*(1), 45–85.

Banz, R. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, *9*(1), 3-18.

Barras, L., Scaillet, O., & Wermers, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The Journal of Finance*, *65*(1), 179–216.

Bekaert, G., Hodrick, R., & Zhang, X. (2009). International Stock Return Co-movements. *The Journal of Finance, 64*(6), 2591-2626.

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

Blake, D., & Timmermann, A. (1998). Mutual Fund Performance: Evidence from the UK. *European Finance Review*, *2*(1), 57–77.

Blake, D., Caulfield, T., Ioannidis, C. & Tonks, I. (2017). New Evidence on Mutual Fund Performance: A Comparison of Alternative Bootstrap Methods. *Journal of Financial and Quantitative Analysis, 52*(3), 1279-1299.

Bollen, N. P. B., & Busse, J. A. (2005). Short-term persistence in mutual fund performance. *Review of Financial Studies*, *18*(2), 569–597.

Bornholt, G. (2013). The Failure of the Capital Asset Pricing Model (CAPM): An Update and Discussion. *Abacus*, *49*(SUPPL.1), 36–43.

Breusch, T. S., & Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, *47*(5), 1287–1294.

Brown, S., Goetzmann, W., & Ross, S. (1992). Survivorship bias in performance studies. *Review of Financial Studies*, *5*(4), 553–580.

Brown, S., & Goetzmann, W. (1995). Performance Persistence. *The Journal of Finance*, *50*(2), 679-698.

Cavanaugh, J. E., & Neath, A. A. (1999). Generalizing the Derivation of the Schwarz Information Criterion. *Communications in Statistics - Theory and Methods*, *28*(1), 49–66.

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

Chen, H., Jegadeesh, N., & Wermers, R. (2000). The Value of Active Mutual Fund Management: An Examination of the Stockholdings and Trades of Fund Managers. *The Journal of Financial and Quantitative Analysis*, *35*(3), 343-368.

Cheng, T., & Yan, C. (2017). Evaluating the size of the bootstrap method for fund performance evaluation. *Economics Letters*, *156*, 36–41.

Chernick, M. R. (2008). *Bootstrap methods: a guide for practitioners and researchers* (2nd ed.). Hoboken, NJ: Wiley.

Chevalier, J., & Ellison, G. (1997). Risk Taking by Mutual Funds as a Response to Incentives. *Journal of Political Economy*, *105*(6), 1167–1200.

Chevalier, J., & Ellison, G. (1999). Are Some Mutual Fund Managers Better Than Others? Cross-Sectional Patterns in Behavior and Performance. *The Journal of Finance*, *54*(3), 875–899.

Cremers, K. J. M., & Petajisto, A. (2009). How Active Is Your Fund Manager? A New Measure That Predicts Performance. *Review of Financial Studies*, *22*(9), 3329–3365.

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

Cuthbertson, K., Nitzsche, D., & O'Sullivan, N. (2010). Mutual Fund Performance: Measurement and Evidence. *Financial Markets, Institutions & Instruments*, *19*(2), 95–187.

Cuthbertson, K., Nitzsche, D., & O'Sullivan, N. (2012). False Discoveries in UK Mutual Fund Performance. *European Financial Management*, *18*(3), 444–463.

Elton, E. J., Gruber, M. J., & Blake, C. R. (1996). Survivorship Bias and Mutual Fund Performance. *Review of Financial Studies*, *9*(4), 1097–1120.

Fama, E. F., & French, K. R. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance*, *47*(2), 427-465.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bond. *Journal of Financial Economics, 33*(3), 3–56.

Fama, E. F. and French, K. R. (1996). Multifactor Explanations of Asset Pricing Anomalies. *The Journal of Finance, 51*(1), 55–84.

Fama, E. F., & French, K. R. (2010). Luck versus Skill in the Cross-section of Mutual Fund Returns. *The Journal of Finance*, *65*(5), 1915–1947.

Fama, E. F., & French, K. R. (2012). Size, value, and momentum in international stock returns. *Journal of Financial Economics*, *105*(3), 457–472.

Fletcher, J., & Forbes, D. (2002). An exploration of the persistence of UK unit trust performance. *Journal of Empirical Finance*, *9*(5), 475–493.

Grinblatt, B. M., Titman, S., & Wermers, R. (1995). Momentum Investment Strategies, Portfolio Performance, and Herding: A Study of Mutual Fund Behavior. *American Economic Review*, *85*(5), 1088–1105.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer.

Heuer, J., Merkle, C., & Weber, M. (2016). Fooled by Randomness: Investor Perception of Fund Manager Skill. *Review of Finance*, *21*(2), 605-635.

Hendricks, D., Patel, J., & Zeckhauser, R. (1993). Hot Hands in Mutual Funds: Short Run Persistence of Performance, 1974-1988. *The Journal of Finance*, *48*(1), 93–130.

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

Jegadeesh, N., & Titman, S. (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance, 48*(1), 65–91.

Jegadeesh, N., & Titman, S. (2001). Profitability of momentum strategies: An evaluation of alternative explanations. *The Journal of Finance, 56*(2), 699–720.

Jordan, B. D., & Riley, T. B. (2015). Volatility and mutual fund manager skill. *Journal of Financial Economics*, *118*(2), 289–298.

Kosowski, R., Timmermann, A., Wermers, R., & White, H. (2006). Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis. *The Journal of Finance*, *61*(6), 2551–2595.

Lakonishok, Shleifer, A., Vishny, R., & Hart, O. (1992). The structure and performance of the money management industry. *Brookings Papers on Economic Activity: Microeconomics,* 339-391.

Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *The Journal of Finance, 20*(4), 587–615.

Malkiel, B. G. (1995). Returns from Investing in Equity Mutual Funds 1971 to 1991. *The Journal of Finance*, *50*(2), 549–572.

Moerman, G. A. (2005). *How Domestic is the Fama and French Three-Factor Model? An Application to the Euro Area* (ERIM Report Series Reference No. ERS-2005-035-F&A). Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=738363

Newey, W. K., & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, *55*(3), 703-708.

Otten, R., & Bams, D. (2002). European Mutual Fund Performance. *European Financial Management*, *8*(1), 75–101.

Pastor, L., & Stambaugh, R. F. (2002). Mutual fund performance and seemingly unrelated assets. *Journal of Financial Economics*, *63*(3), 315–349.

Petersen, M. A. (2009). Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. *The Review of Financial Studies*, *22*(1), 435-480.

Politis, D. N., & Romano, J. P. (1994). The Stationary Bootstrap. *Journal of the American Statistical Association*, *89*(428), 1303–1313.

Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance, 19*(3), 425–442.

Solomon, D. H., Soltes, E., & Sosyura, D. (2014). Winners in the spotlight: Media coverage of fund holdings as a driver of flows. *Journal of Financial Economics*, *113*(1), 53–72.

Ter Horst, J., Nijman, T., & de Roon, F. (1998). *Style analysis and performance evaluation of Dutch mutual funds* (CentER discussion paper 9850). Retrieved from https://www.researchgate.net/profile/Frans_De_Roon/publication/4783286_Style_analysis_and_performance_evaluation_of_Dutch_mutual_funds/links/02e7e5182ac76678de000000.pdf

Wermers, R. (2000). Mutual fund performance: An empirical decomposition into stock-picking talent, style, transactions costs, and expenses. *The Journal of Finance*, *55*(4), 1655–1695.

Appendix – Supplementary tables

Table A1: Bootstrap simulations 1992-2015 period for α

Positive α		***Negative α***											
Best	**2nd**	**12th**	**11th**	**10th**	**9th**	**8th**	**7th**	**6th**	**5th**	**4th**	**3rd**	**2nd**	**Worst**
1.501	*1.228*	*-0.963*	*-0.978*	*-0.980*	*-0.992*	*-0.994*	*-1.037*	*-1.079*	*-1.081*	*-1.107*	*-1.203*	*-1.203*	*-1.400*
1.127	*0.987*	*-0.814*	*-0.819*	*-0.843*	*-0.888*	*-0.889*	*-0.900*	*-0.973*	*-0.992*	*-1.029*	*-1.031*	*-1.097*	*-1.153*
0.953	*0.913*	*-0.761*	*-0.795*	*-0.817*	*-0.828*	*-0.843*	*-0.854*	*-0.908*	*-0.933*	*-0.969*	*-1.029*	*-1.094*	*-1.132*
0.932	*0.871*	*-0.725*	*-0.728*	*-0.758*	*-0.824*	*-0.840*	*-0.844*	*-0.872*	*-0.872*	*-0.891*	*-0.935*	*-1.093*	*-1.099*
0.924	*0.824*	*-0.647*	*-0.665*	*-0.669*	*-0.676*	*-0.683*	*-0.690*	*-0.743*	*-0.759*	*-0.765*	*-0.778*	*-0.808*	*-1.042*
0.912	*0.815*	*-0.612*	*-0.614*	*-0.659*	*-0.662*	*-0.679*	*-0.686*	*-0.698*	*-0.707*	*-0.728*	*-0.732*	*-0.745*	*-0.820*
0.861	*0.620*	*-0.491*	*-0.492*	*-0.501*	*-0.511*	*-0.512*	*-0.520*	*-0.535*	*-0.564*	*-0.585*	*-0.607*	*-0.675*	*-0.759*
0.659	*0.615*	*-0.479*	*-0.480*	*-0.482*	*-0.493*	*-0.496*	*-0.496*	*-0.496*	*-0.519*	*-0.526*	*-0.560*	*-0.572*	*-0.651*
0.630	*0.601*	*-0.435*	*-0.442*	*-0.464*	*-0.469*	*-0.491*	*-0.492*	*-0.496*	*-0.501*	*-0.506*	*-0.515*	*-0.572*	*-0.602*
0.546	*0.529*	*-0.433*	*-0.435*	*-0.445*	*-0.450*	*-0.451*	*-0.452*	*-0.453*	*-0.458*	*-0.475*	*-0.507*	*-0.554*	*-0.586*
0.535	*0.526*	*-0.372*	*-0.378*	*-0.381*	*-0.386*	*-0.393*	*-0.419*	*-0.441*	*-0.441*	*-0.450*	*-0.482*	*-0.533*	*-0.574*
0.512	*0.490*	*-0.367*	*-0.377*	*-0.380*	*-0.384*	*-0.386*	*-0.397*	*-0.410*	*-0.418*	*-0.444*	*-0.463*	*-0.483*	*-0.502*
0.489	*0.469*	*-0.340*	*-0.351*	*-0.366*	*-0.382*	*-0.386*	*-0.386*	*-0.398*	*-0.412*	*-0.426*	*-0.446*	*-0.462*	*-0.472*
0.405	*0.390*	*-0.324*	*-0.333*	*-0.333*	*-0.342*	*-0.372*	*-0.373*	*-0.374*	*-0.378*	*-0.391*	*-0.417*	*-0.418*	*-0.452*

* The table represents the result of the bootstrap simulations with regards to the alphas in the 1992-2015 period and is based on the Fama-French three-factor model. The column for *Worst* for example represents the lowest simulated α's from the 1000 bootstrap iterations for each individual fund (the cross-section). Since there are two positive alpha funds in the Fama-French regressions and twelve negative alpha funds, this table only reports the two best columns for positive alphas and the worst twelve for negative alphas.

Table A2: Bootstrap simulations 1992-2015 period for t_α

Positive t_α		***Negative t_α***											
Best,	***2nd***	***12th***	***11th***	***10th***	***9th***	***8th***	***7th***	***6th***	***5th***	***4th***	***3rd***	***2nd***	***Worst***
4.028	*3.545*	*-2.441*	*-2.525*	*-2.563*	*-2.690*	*-2.748*	*-2.781*	*-2.809*	*-2.875*	*-2.901*	*-3.102*	*-3.536*	*-4.259*
3.709	*3.371*	*-2.431*	*-2.497*	*-2.497*	*-2.573*	*-2.691*	*-2.743*	*-2.746*	*-2.820*	*-2.878*	*-3.009*	*-3.180*	*-3.672*
3.685	*3.347*	*-2.354*	*-2.393*	*-2.414*	*-2.449*	*-2.463*	*-2.563*	*-2.679*	*-2.731*	*-2.848*	*-2.953*	*-3.138*	*-3.663*
3.679	*3.296*	*-2.297*	*-2.369*	*-2.401*	*-2.404*	*-2.454*	*-2.558*	*-2.618*	*-2.707*	*-2.745*	*-2.894*	*-3.118*	*-3.487*
3.612	*3.234*	*-2.297*	*-2.337*	*-2.382*	*-2.396*	*-2.452*	*-2.466*	*-2.586*	*-2.669*	*-2.719*	*-2.877*	*-3.024*	*-3.470*
3.421	*3.171*	*-2.267*	*-2.331*	*-2.339*	*-2.385*	*-2.437*	*-2.466*	*-2.518*	*-2.626*	*-2.671*	*-2.799*	*-2.947*	*-3.462*
3.383	*3.063*	*-2.212*	*-2.267*	*-2.322*	*-2.377*	*-2.433*	*-2.464*	*-2.512*	*-2.618*	*-2.639*	*-2.695*	*-2.894*	*-3.198*
3.242	*3.002*	*-2.172*	*-2.178*	*-2.222*	*-2.234*	*-2.429*	*-2.458*	*-2.482*	*-2.558*	*-2.624*	*-2.678*	*-2.817*	*-3.197*
3.127	*2.930*	*-2.150*	*-2.158*	*-2.189*	*-2.225*	*-2.427*	*-2.453*	*-2.479*	*-2.506*	*-2.601*	*-2.667*	*-2.791*	*-3.128*
3.059	*2.849*	*-2.136*	*-2.143*	*-2.187*	*-2.220*	*-2.289*	*-2.290*	*-2.338*	*-2.422*	*-2.512*	*-2.620*	*-2.718*	*-3.049*
3.016	*2.664*	*-2.099*	*-2.126*	*-2.179*	*-2.217*	*-2.244*	*-2.278*	*-2.324*	*-2.371*	*-2.491*	*-2.564*	*-2.710*	*-2.993*
2.976	*2.664*	*-2.099*	*-2.114*	*-2.145*	*-2.210*	*-2.232*	*-2.260*	*-2.293*	*-2.345*	*-2.468*	*-2.523*	*-2.704*	*-2.978*
2.745	*2.626*	*-2.013*	*-2.045*	*-2.052*	*-2.193*	*-2.200*	*-2.232*	*-2.283*	*-2.342*	*-2.376*	*-2.455*	*-2.673*	*-2.721*
2.694	*2.558*	*-1.782*	*-1.784*	*-1.855*	*-1.860*	*-1.867*	*-1.872*	*-1.883*	*-1.991*	*-2.054*	*-2.189*	*-2.354*	*-2.390*

* The table represents the result of the bootstrap simulations with regards to the alpha t-statistic in the 1992-2015 period and is based on the Fama-French three-factor model. The column for *Worst* for example represents the lowest simulated t_α's from the 1000 bootstrap iterations on all the funds (the cross-section). Since there are two positive t_α funds in the Fama-French regressions and twelve negative t_α funds, this table only reports the two best columns and the worst twelve columns.

Table A3: Bootstrap simulations t_α January 2005 – December 2009

Positive t_α						***Negative t_α***				
Best	***2nd***	***3rd***	***4th***	***5th***	***6th***	***5th***	***4th***	***3rd***	***2nd***	***Worst***
4.218	*3.572*	*3.500*	*3.256*	*3.170*	*3.119*	*-3.429*	*-3.777*	*-4.030*	*-4.341*	*-5.225*
4.035	*3.549*	*3.425*	*3.131*	*3.129*	*3.064*	*-3.142*	*-3.380*	*-3.593*	*-3.877*	*-4.919*
3.917	*3.431*	*3.306*	*3.124*	*3.085*	*3.060*	*-3.130*	*-3.340*	*-3.549*	*-3.720*	*-4.813*
3.682	*3.402*	*3.294*	*3.120*	*2.845*	*2.825*	*-3.027*	*-3.159*	*-3.357*	*-3.698*	*-4.068*
3.620	*3.398*	*3.266*	*2.897*	*2.789*	*2.737*	*-3.026*	*-3.150*	*-3.325*	*-3.419*	*-3.842*
3.570	*3.386*	*3.215*	*2.836*	*2.685*	*2.579*	*-2.971*	*-3.090*	*-3.116*	*-3.360*	*-3.788*
3.347	*3.012*	*2.974*	*2.808*	*2.627*	*2.578*	*-2.925*	*-3.005*	*-3.048*	*-3.330*	*-3.699*
3.338	*2.900*	*2.786*	*2.779*	*2.606*	*2.574*	*-2.901*	*-2.905*	*-3.013*	*-3.326*	*-3.624*
3.125	*2.869*	*2.732*	*2.660*	*2.596*	*2.534*	*-2.831*	*-2.899*	*-2.983*	*-3.159*	*-3.458*
3.012	*2.835*	*2.667*	*2.644*	*2.595*	*2.524*	*-2.780*	*-2.791*	*-2.916*	*-3.114*	*-3.268*
2.892	*2.826*	*2.654*	*2.637*	*2.580*	*2.441*	*-2.576*	*-2.665*	*-2.709*	*-2.763*	*-3.019*

* The table represents the result of the bootstrap simulations with regards to the alpha t-statistic in the 2005-2009 period and is based on the Fama-French three-factor model. The column for *Worst* for example represents the lowest simulated t_α of each individual fund in the cross-section out of 1000 bootstrap iterations. Since there are six positive t_α funds in the Newey-West regressions and five negative t_α funds, I only report the six best columns and the worst five columns.

Table A4: Bootstrap simulations t_α January 2010 – December 2014

14th	*13th*	*12th*	*11th*	*10th*	*9th*	*8th*	*7th*	*6th*	*5th*	*4th*	*3rd*	*2nd*	*Worst*
-2.646	*-2.679*	*-2.694*	*-2.846*	*-2.871*	*-2.897*	*-3.075*	*-3.102*	*-3.236*	*-3.443*	*-3.644*	*-3.816*	*-3.913*	*-4.720*
-2.578	*-2.662*	*-2.688*	*-2.765*	*-2.782*	*-2.877*	*-2.928*	*-2.973*	*-3.214*	*-3.305*	*-3.556*	*-3.627*	*-3.910*	*-4.641*
-2.569	*-2.600*	*-2.651*	*-2.699*	*-2.759*	*-2.853*	*-2.892*	*-2.970*	*-2.993*	*-3.283*	*-3.465*	*-3.612*	*-3.876*	*-4.316*
-2.537	*-2.538*	*-2.581*	*-2.655*	*-2.700*	*-2.718*	*-2.888*	*-2.929*	*-2.969*	*-3.248*	*-3.301*	*-3.592*	*-3.859*	*-4.290*
-2.524	*-2.535*	*-2.572*	*-2.597*	*-2.657*	*-2.687*	*-2.775*	*-2.839*	*-2.884*	*-3.076*	*-3.250*	*-3.271*	*-3.675*	*-4.289*
-2.485	*-2.532*	*-2.564*	*-2.581*	*-2.613*	*-2.643*	*-2.760*	*-2.810*	*-2.873*	*-3.039*	*-3.233*	*-3.258*	*-3.628*	*-4.194*
-2.460	*-2.491*	*-2.542*	*-2.543*	*-2.580*	*-2.623*	*-2.659*	*-2.769*	*-2.788*	*-2.995*	*-3.132*	*-3.233*	*-3.513*	*-4.095*
-2.390	*-2.403*	*-2.432*	*-2.458*	*-2.545*	*-2.608*	*-2.629*	*-2.766*	*-2.787*	*-2.969*	*-3.125*	*-3.216*	*-3.351*	*-4.075*
-2.348	*-2.393*	*-2.427*	*-2.453*	*-2.493*	*-2.554*	*-2.561*	*-2.756*	*-2.776*	*-2.920*	*-3.118*	*-3.178*	*-3.311*	*-3.974*
-2.311	*-2.351*	*-2.400*	*-2.403*	*-2.417*	*-2.426*	*-2.555*	*-2.738*	*-2.772*	*-2.807*	*-3.086*	*-3.176*	*-3.293*	*-3.875*
-2.236	*-2.255*	*-2.354*	*-2.355*	*-2.382*	*-2.420*	*-2.509*	*-2.634*	*-2.662*	*-2.806*	*-2.985*	*-3.131*	*-3.281*	*-3.722*
-2.228	*-2.247*	*-2.310*	*-2.323*	*-2.324*	*-2.411*	*-2.460*	*-2.623*	*-2.649*	*-2.749*	*-2.969*	*-3.105*	*-3.184*	*-3.660*
-2.212	*-2.244*	*-2.309*	*-2.314*	*-2.314*	*-2.384*	*-2.448*	*-2.453*	*-2.471*	*-2.590*	*-2.613*	*-2.811*	*-2.991*	*-3.411*
-1.981	*-2.009*	*-2.058*	*-2.091*	*-2.204*	*-2.240*	*-2.296*	*-2.303*	*-2.403*	*-2.490*	*-2.608*	*-2.610*	*-2.814*	*-2.840*

* The table represents the results of the bootstrap simulations with regards to the alpha *t*-statistic in the 2010-2014 period and is based on the Fama-French three-factor model. The column for *Worst* for example represents the lowest simulated t_α's from the 1000 bootstrap iterations on all the funds (the cross-section). Since there are fourteen negative t_α funds in the Newey-West regressions, this table only report the fourteen worst columns.

Table A5: CAPM bootstrap simulations 1992-2015 t_α

Positive t_α		***Negative t_α***											
Best,	***2nd***	***12th***	***11th***	***10th***	***9th***	***8th***	***7th***	***6th***	***5th***	***4th***	***3rd***	***2nd***	***Worst***
4.099	*3.619*	*-2.423*	*-2.438*	*-2.463*	*-2.494*	*-2.624*	*-2.680*	*-2.754*	*-2.820*	*-2.874*	*-2.956*	*-3.042*	*-4.461*
3.599	*3.420*	*-2.378*	*-2.398*	*-2.412*	*-2.473*	*-2.554*	*-2.607*	*-2.720*	*-2.727*	*-2.748*	*-2.911*	*-3.025*	*-3.581*
3.547	*3.305*	*-2.320*	*-2.370*	*-2.409*	*-2.470*	*-2.484*	*-2.534*	*-2.664*	*-2.725*	*-2.728*	*-2.880*	*-3.000*	*-3.416*
3.535	*3.304*	*-2.293*	*-2.364*	*-2.395*	*-2.404*	*-2.429*	*-2.492*	*-2.618*	*-2.644*	*-2.664*	*-2.831*	*-2.999*	*-3.364*
3.511	*3.285*	*-2.238*	*-2.336*	*-2.351*	*-2.369*	*-2.385*	*-2.480*	*-2.541*	*-2.581*	*-2.646*	*-2.812*	*-2.982*	*-3.302*
3.415	*3.161*	*-2.237*	*-2.282*	*-2.301*	*-2.359*	*-2.361*	*-2.411*	*-2.476*	*-2.577*	*-2.626*	*-2.811*	*-2.960*	*-3.236*
3.352	*2.902*	*-2.228*	*-2.252*	*-2.289*	*-2.334*	*-2.348*	*-2.389*	*-2.455*	*-2.465*	*-2.524*	*-2.677*	*-2.889*	*-3.156*
3.327	*2.877*	*-2.168*	*-2.212*	*-2.287*	*-2.307*	*-2.343*	*-2.359*	*-2.422*	*-2.440*	*-2.507*	*-2.569*	*-2.862*	*-3.088*
3.225	*2.788*	*-2.147*	*-2.185*	*-2.265*	*-2.290*	*-2.341*	*-2.346*	*-2.382*	*-2.435*	*-2.498*	*-2.546*	*-2.831*	*-3.061*
2.914	*2.729*	*-2.143*	*-2.167*	*-2.228*	*-2.288*	*-2.329*	*-2.333*	*-2.362*	*-2.405*	*-2.460*	*-2.545*	*-2.808*	*-3.015*
2.760	*2.713*	*-2.127*	*-2.160*	*-2.217*	*-2.264*	*-2.271*	*-2.303*	*-2.338*	*-2.390*	*-2.438*	*-2.532*	*-2.594*	*-2.974*
2.759	*2.701*	*-2.072*	*-2.154*	*-2.202*	*-2.214*	*-2.243*	*-2.286*	*-2.329*	*-2.343*	*-2.425*	*-2.466*	*-2.565*	*-2.923*
2.733	*2.683*	*-2.025*	*-2.101*	*-2.140*	*-2.162*	*-2.173*	*-2.173*	*-2.298*	*-2.331*	*-2.365*	*-2.423*	*-2.446*	*-2.605*
2.728	*2.487*	*-1.803*	*-1.819*	*-1.826*	*-1.830*	*-1.844*	*-1.882*	*-1.889*	*-1.979*	*-2.050*	*-2.186*	*-2.345*	*-2.477*

* The table represents the results of the bootstrap simulations with regards to the alpha *t*-statistic in the 1992-2015 period and is based on the CAPM. The column for *Worst* for example represents the lowest simulated t_α's from the 1000 bootstrap iterations in the cross-section. Since there are two positive t_α funds in the Newey-West regressions and twelve negative t_α funds, this table only report the two best columns and the worst twelve columns.