Performance Persistence of Dutch Pension Funds

Xiaohong Huang · Ronald J. Mahieu

Abstract   This paper studies the investment performance of pension funds with a focus on their ability in implementing their intended investment strategy. We use a sample of Dutch industry-wide pension funds, which are obliged by law to report their investment performance according to the so-called *z-score*. The *z-score* is a risk-adjusted performance measure with benchmark settings predefined by Dutch law. We find that pension funds as a group cannot beat their self-selected benchmarks consistently. Applying a cross-sectional portfolio approach we find evidence that the largest pension funds outperform the smallest funds.

Keywords   Pension fund investment · Performance evaluation · *z-score*

JEL Classification   G23 · C12

X. Huang (✉)
School of Management and Governance, University of Twente, and Netspar, Drienerlolaan 5,
7522 NB, Enschede, The Netherlands
e-mail: x.huang@utwente.nl

R. J. Mahieu
Tilburg School of Economics and Management, Tilburg University, and Netspar,
Warandelaan 2, 5000 LE, Tilburg, The Netherlands
e-mail: r.j.mahieu@uvt.nl

1 Introduction

The aggregated market value of Dutch pension fund portfolios is large. At the end of year 2008 the total asset size of Dutch pension funds was €605 billion. This number is 1.2 times larger than the value of the GDP, which was €488.5 billion in 2008. In contrast, the assets managed by collective investment schemes such as mutual funds and hedge funds, are only about €84 billion.[1] Most of these pension assets are associated to industry-related pension funds (€409 billion), which manage the pension savings for the majority of Dutch employees. The sheer size of this category of pension funds and their significant role in providing retirement income warrants a careful investigation of the performance of their investment portfolios.

In the Netherlands a mandatory industry-wide pension fund is a multi-sponsor pension plan that provides pension services to all employees of the companies affiliated to a particular industry.[2] Employees of these companies are obliged to participate in these schemes. The mandatory feature of these plans leads to a legal requirement that pension funds should report their investment performance in terms of a so-called *z-score*, a risk-adjusted measure of their investment returns. The *z-score* methodology is fixed by Dutch law (Vrijstellingsbesluit Wet Bpf 2000). Most importantly, the prescribed risk adjustment is not consistent with *ex post* risk exposure. If a fund fails a performance test based on the *z-score*, it loses its mandatory status. Individual participating companies can then leave the fund and either join another pension fund or establish their own fund.

In this paper, we use a unique data set of *z-score* observations to provide a cross-sectional and longitudinal description of the investment performance of Dutch mandatory industry-wide pension funds. Our study adds to the current literature on pension fund performance. It provides another piece of evidence that pension funds do not add value in implementing investment strategies with respect to the indicated benchmarks. Our study also shows the variation in performance across funds of different sizes, revealing that the biggest fund group persistently outperform the smallest fund group. This finding corroborates the ongoing consolidation in the pension fund sector.

Pension fund performance is often measured by the total investment return of the fund portfolio, which is in general determined by the strategic asset allocation and the implementation of this allocation. The strategic allocation is typically set by the trustees with the help of consultants and investment advisors. The implementation of the strategic portfolio is delegated to internal or external asset managers with different specializations.[3] Our paper focuses on the quality of the implementation, and this is measured by the fund's overall portfolio performance in excess of an *a priori* agreed-upon benchmark portfolio.

A pension fund portfolio typically consists of various asset classes. The study on investment performance can be performed at both the individual asset class level as

---

[1] According to statistics on the website of the Dutch central bank (DNB): www.dnb.nl.

[2] According to statistics from the Dutch central bank (DNB), in 2010 about 88% of the industry-wide funds provide defined benefit schemes and the rest are defined contribution and mixed schemes.

[3] A new trend is that an external investment firm acts as a fiduciary asset manager, who structures and monitors the total investment process from strategic asset allocation to individual asset manager selection.

well as at the level of the overall fund portfolio. One stream of the literature is evaluating the performance on an asset class level, such as Lakonishok et al. (1992), Coggin et al. (1993), Busse et al. (2006), Tonks (2005) and Bauer et al. (2007). The other stream of the literature, which is relatively small, is related to performance evaluation at the overall fund level. Brinson et al. (1986) and Ippolito and Turner (1987) are the pioneers in this line of research. A most recent paper is Blake et al. (1999) on the UK pension funds. Our paper fits in this second stream of literature and addresses the question whether pension funds outperform their benchmarks using the *z-score* framework. In that way our study adds to the limited, but growing, literature on pension fund investment performance. The lack of empirical studies on pension fund investment performance up to now may be related to the fact that there is little detailed information available on the asset allocations and on the returns of individual components of the pension fund investment portfolios.[4] The Dutch sample used in our study overcomes this problem, and provides a risk-adjusted measurement that accounts for fund- and period- specific asset allocations and returns.

The rest of the paper is organized as follows. In Sect. 2 we describe previous studies on pension fund investment performance. Section 3 provides some background on the investment processes at Dutch pension funds, and introduces the *z-score* in detail. Section 4 describes the associated data. The results are discussed in Sect. 5. Section 6 concludes.

2 Literature Review

In this part we provide an overview of the most important papers on pension fund performance evaluation at the overall fund level. Brinson et al. (1986) and Ippolito and Turner (1987) are the pioneers in this line of research. A more recent paper is Blake et al. (1999) on the UK pension funds.

Performance evaluation requires the use of appropriate benchmarks. In general there are two ways to construct benchmark portfolios. One way is to use risk factors, like equity/bond market index returns, or the returns on specifically designed portfolios as in the Fama and French methodology. Subsequently, the loading on these factors are found by regressing the returns on the pension fund portfolio on these risk factor returns. The benchmark portfolio is then composed by the estimated loadings and the returns on the risk factors. This is the approach taken in Coggin et al. (1993), Tonks (2005), and Busse et al. (2006). The other way is to use explicit information on the asset allocation holdings of a pension fund portfolio, as Brinson et al. (1986), Ippolito and Turner (1987), and Blake et al. (1999). In that case the benchmark portfolio is calculated by multiplying the directly measured weights with the associated returns on the individual asset classes.

There are two parameters in constructing a holding-based benchmark portfolio, the holdings/allocations and the index returns to respective asset classes. The benchmark

---

[4] There is some information on asset holdings of pension funds and these data are used to test the determinants of investment policies empirically, such as Alestalo and Puttonen (2006), Gerber and Weber (2007) and Bikker and de Dreu (2009). The lack of returns on individual asset classes, however, limits these authors from more detailed evaluations on investment performance.

portfolio used in both Brinson et al. (1986) and Ippolito and Turner (1987) considers only three broad asset classes: stocks, bonds and cash. However, nowadays the investment opportunities used by pension funds range over many more asset categories, as was recognized by Blake et al. (1999). Though considering more asset classes, Blake et al. (1999) apply the same index for a given asset class for all funds. The data set in our empirical analysis holds more detailed information on different asset categories. In addition, the index for each asset category differs across funds, which allows us to compute excess returns more precisely.

Both Brinson et al. (1986) and Blake et al. (1999) document underperformance with respect to the benchmark portfolio. Ippolito and Turner (1987) find mixed results depending on the choice of benchmark, but show that larger pension funds outperformed smaller funds substantially. These findings motivate the formulation of similar hypotheses for the Dutch case in the *z-score* environment.

3 Investment Process and the *z-score*

The investment process of a pension fund starts with an Asset Liability Management (ALM) study, which results in an investment policy represented by a strategic asset allocation. From the strategic asset allocation trustees define an annual investment plan, which specifies allocations for detailed asset categories. Then trustees assign mandates for each asset category to a selected group of asset managers. These managers can be either in-house or external, one or multiple, with passive or active style. This paper looks at the quality of implementing the annual investment plan.

The success of implementation is measured by the differences between the actual returns and the returns attainable from strict adherence to the annual investment plan. A benchmark portfolio needs to be defined that represents the annual investment plan. This benchmark portfolio is a hypothetical portfolio, which is "structurally identical to the investment strategy without whatever active management takes place" as defined in Logue and Rader (1998) (p. 168) or a "passive mix with the same style" as in Sharpe (1992). Our performance measure defines such a benchmark portfolio. See the example in Table 1. The benchmark portfolio has a twofold purpose. First, the index for each component portfolio is used by trustees to evaluate the performance of individual asset managers for a particular asset class. Second, the overall return from the benchmark portfolio serves as a return target. In our study we use the benchmark portfolio for its second purpose to evaluate the quality of investment implementations by asset managers.

The remainder of this section introduces the performance measure we use in this paper. Since 1998 every Dutch mandatory industry-wide pension fund must compute a so-called *z-score* to reflect their investment performance. The methodology of the *z-score* is set by Dutch law (Vrijstellingsbesluit Wet Bpf 2000). The *z-score* is the difference between the actual return and the return on a predefined benchmark portfolio, net of expenses, and normalized by the riskiness of the portfolio, as in the following equation:

Table 1 An example of a benchmark portfolio

Assets	Weight (%)	Range (%)	Index
Fixed income	*75*	*65–85*	
Governments	70	60–80	Citigroup gov bond index
Corporates	15	10–20	Citigroup non-EGBI EMU index
Private loans	15	10–20	Customized private loan index
Equity	*15*	*5–25*	
Europe	40	30–50	MSCI Europe
USA	20	10–30	MSCI North America
Pacific	15	5–25	MSCI Pacific
EM global	25	15–35	MSCI EM global
Real estate	*5*	*0–10*	
Residential	50	25–75	ROZ-IPD residential
Shops	50	25–75	ROZ-IPD Retail
Alternatives	*5*	*0–10*	
Commodities	50	0–100	DJ-AIG Commodity Index
Hedge fund	50	0–100	Euro 7-day Libid

This is a reproduction of a benchmark portfolio. It specifies the weighting and the indices used for different investment styles. The range specifies the bound within which an active asset manager must control the allocation
Source 2006 annual report of the Agriculture and Food Supply Pension Fund (www.company.info)

$$z_{i,t} = \frac{(R_{p,i,t} - c_{p,i,t}) - (R_{b,i,t} - c_{b,i,t})}{E_{i,t}}$$

where $R_{p,i,t}$ and $c_{p,i,t}$ are the gross investment return and internal investment cost of pension fund i at time t respectively. The internal investment cost also includes the fees paid to the external asset managers and investment related custodian and administrative cost. $R_{b,i,t}$ is the fund i's benchmark portfolio return using market indices in the respective asset categories at time t. See Table 1 for an example. $c_{b,i,t}$ is the associated investment cost of the benchmark portfolio which depends on the percentage of equity investment in the portfolio.[5] The benchmark portfolio is determined by trustees at the beginning of each year and fixed for 1 year. Specifically, the weights and the index for various asset classes in the benchmark portfolio are defined *a priori*. In addition the index should represent the asset class, be investable and objectively measurable.[6] The benchmark return represents the return that an individual investor can obtain if he invests in the benchmark portfolio, and the difference between the realized return and the benchmark return reflects the excess return that a pension fund can earn by selecting the right internal or external asset managers. The pre-selected benchmark portfolio also excludes the possibility of manipulation in calculating the *z-score*.

---

[5] This cost is presented in Vrijstellingsbesluit Wet Bpf (2000), and range from 0.10 to 0.22%. It varies to the equity proportion of the pension portfolio.

[6] See Article 5.3 in Vrijstellingsbesluit Wet Bpf (2000).

To enable the comparison across pension funds with different investment strategies, the excess returns are scaled by the riskiness of the asset mix in the benchmark portfolio ($E_{i,t}$). The asset mix for this purpose contains two major categories: equity and fixed income (including cash). The riskiness is measured by the variance of the benchmark portfolio return. According to Vrijstellingsbesluit Wet Bpf (2000), the riskiness of equity and fixed income investment is set at 2.6 and 0.6%.[7] Nederlands Pensioen and Beleggingsnieuws (December 1, 2005) reports that these values are based on the standard deviations of the realized excess returns on these asset categories. The standard deviations are calculated by WM Company on the population of Dutch pension funds over the period of 1992–1996. The risk parameters are kept fixed over subsequent years, and are used in *z-score* calculations up to this day. The reported *z-score* is audited by external accountants.

The way the *z-score* is constructed reveals that it is not a measure to evaluate the effectiveness of the investment plan, but rather a measure of the quality of implementing the investment plan. The benchmark used in the calculation reflects the *ex ante* investment plan for a particular fund for the upcoming year. Therefore the *z-score* accurately shows the fund's ability in beating its own benchmarks. A positive (negative) *z-score* means that the fund has successfully implemented (failed to implement) its investment plan. This success (failure) can be attributed to a fund's skill in selecting and monitoring its asset managers. A high (low) *z-score* reflects the relatively good (poor) ability of the fund in executing its investment plan.

The underlying arguments for the *z-score* are related to creating a standardized risk measure that can be used by a regulator to judge whether a fund's investment performance is sufficient. The statistical test, called *performance test*, is used to support this decision. In Vrijstellingsbesluit Wet Bpf (2000) it is stated that if a fund falls to the lowest 10% percentile, measured over a period of 5 years, its performance is regarded as insufficient. Based on the central limit theorem, the test statistic is calculated as $P_{5 \text{ year}} = (\sum_{t=1}^{5} Z_{i,t})/\sqrt{5}$, and is assumed to be asymptotically normally distributed. The critical value of the test is -1.28, which corresponds to a confidence level of 90% for a standardized normal distribution. If the test statistic is less than -1.28, the industry-wide pension fund fails the performance test. The consequence of failing the performance test is that the members of the industry fund have the right to leave the fund, i.e. they have the option to join another pension fund or establish their own.

We use the *z-score* to examine the quality of investment implementation by pension funds over time and over funds (cross-sectionally). We realize that there are some serious concerns related to the *z-score* performance test. A first concern is that the benchmark portfolio is a static benchmark, in which the weighting of different asset styles is fixed for 1 year. As a result, the intertemporal changes in the investment plan during the year cannot be captured by the benchmark portfolio used in the *z-score* calculation. But, investment plan changes do change the return of the actual benchmark portfolio. This can invalidate a fair evaluation of the implementation quality, because part of the deviations is due to the change of the benchmark portfolio and has nothing to do with the implementation ability of the selected asset managers.

---

[7] For example, if a fund has an asset mix of 60% equity and 40% fixed income, then $E_{i,t} = 0.6 * 2.6\% + 0.4 * 0.6\% = 1.8\%$.

We believe the concern of a static benchmark portfolio is more of a conceptual problem rather than a practical one due to the following practical observations. Firstly, fixed weighting is a general rule, but the benchmark portfolio is allowed to be changed once when there is a considerable change in the liability structure or the old investment plan is obviously no longer appropriate for the fund.[8] Secondly, changing the investment plan during the year is more of a practice per January 1, 2007 when the regulation on financial assessment is implemented, which requires the investment plan to match the market value of liabilities. Thus during our sample period 1998–2006 we do not expect material changes in the investment plan during the year.

A second concern is the risk adjustment in the denominator of the *z-score*, where the riskiness only considers equity and fixed income investments, which are kept fixed over time. The ignorance of, for example, real estate investments might lead to improper risk adjustment for certain pension funds. However, the *z-score* makes the risk adjustment at two levels. In addition to the risk adjustment in the denominator, the benchmark portfolio in the numerator adjusts the risk in, for example real estate, by integrating the excess returns in the equity asset class (see Table 1). Insofar as the standard deviations of the *excess* returns on real estate are not that different from the standard deviations on equity *excess* returns the bias in the risk measure is limited.

Another concern is the moral hazard issue arising from the fixed values of 2.6 and 0.6 as the riskiness of the equity and fixed income investments, where pension funds can take advantage in calculating their scores. If a fund takes more risks than what is assumed in the benchmark, its *z-score* can be inflated. This can cause a problem if we find outperformance for the average pension fund because the outperformance can be driven by higher risk exposures in the pension industry than their benchmark portfolios would indicate. On the other hand it would not be an issue if no outperformance is detected. Cross sectionally, if a fund takes a higher excessive risk over its own benchmark than another fund, this risk-taking strategy will give this fund an advantage in the *z-score* ranking. However, a higher *z-score* does not necessarily mean that the fund takes higher excessive risk over the benchmark, because there are other factors that could explain the higher performance. We realize that excessive risk exposures constitute a limitation of using *z-scores* as a measure for explaining cross sectional outperformance.

4 Data

We use the annual report information from publications of the Dutch industry-wide pension fund association.[9] In addition we obtain data from *companyinfo* which collects and composes aggregate financial information of companies and organizations

---

[8] See Article 5.4 in Vrijstellingsbesluit Wet Bpf (2000). As of November 1, 2007 funds are allowed to adjust their norm portfolio twice a year. In addition, some funds, such as Pension fund "*Vervoer*", use a floating benchmark moving with portfolio development.

[9] In Dutch this association is called the *Vereniging van Bedrijfstakpensioenfondsen* (VB). See their website at www.vb.nl.

Table 2 Descriptive statistics of pension fund size

	Invested assets (billion euro)	Total participants	Log (total participants)
Minimum	0.04	2,579	3.41
Maximum	208.9	2,657,000	6.42
Mean	7.2	211,630	4.70
SD	2.96	479,130	0.69

Descriptive statistics of fund sizes measured, respectively by total invested assets, total number of participants, and the logarithm value of the total participants

including pension funds.[10] We merged and verified data from both sources. When there is a discrepancy between the *z-scores* from the two sources, we used the *z-score* reported in a fund's annual report.

Our sample runs from 1998 through 2006 and covers the entire population of mandatory industry-wide pension funds.[11] We do not use the data for 2007 and later years, because in 2007 pension funds are required to value their liabilities at market prices. The new regulation causes many funds to change their investment plans or benchmark portfolios during the year. This change in the investment plan itself is often not accounted for in the benchmark portfolio when the *z-score* is computed. As a result the *z-score* can be an inaccurate measure of excess returns over the benchmark portfolio after January 1, 2007.

Over the sample period, the number of funds varies between 59 and 65 for a number of reasons. Some funds either entered the industry category or became mandatory after 1998, two funds merged, and two funds were sold to insurance companies. In the end, we have a balanced sample of 57 funds reporting *z-scores*. No funds have become non-compulsory or cease during our sample period and thus our sample does not suffer from survivorship bias.

Since there is no considerable change in the relative sizes of the pension funds in our sample, we use the number of total participants in 2006 as a proxy for the fund size. We also use the value of total invested assets in 2006 as an alternative measure, and find all results maintain. This is also confirmed by the correlation coefficient of 0.88 between the two measures. The size data is obtained from the 2006 annual reports of all pension funds and shown in Table 2. The smallest fund in the sample has an amount of 2,579 participants, and the largest fund has 2.657 million participants. In terms of invested assets, the smallest fund is €40.9 million, the largest is €208.9 billion. Both measures reflect a large size spread among Dutch pension funds. Figure 1 shows that the distribution of total participants is skewed with a long right tail. Many funds are relatively small, with a small number of large funds with more than 1 million participants.

---

[10] See their website at http://www.company.info.

[11] According to DNB 2007 statistics there are 71 mandatory industry-wide pension funds including 7 pre-pension funds which provide pensions for early retirement. Only mandatory funds are required to report *z-scores*.
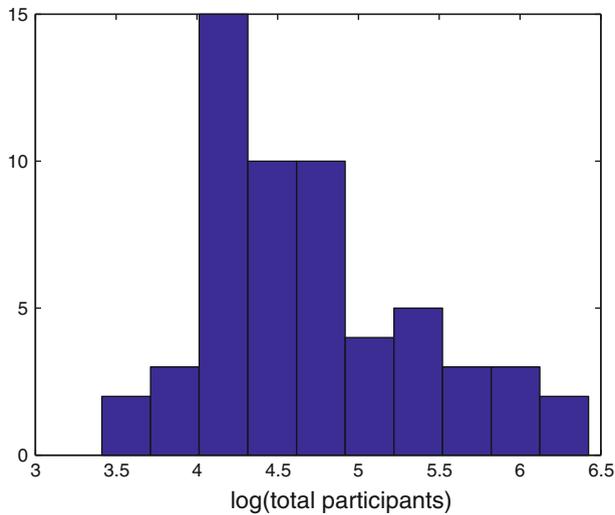
Fig. 1 Size histogram of 57 pension funds in 2006. This figure draws the histogram of 57 pension funds based on the amount of their total participants in 2006

5 Empirical Results

The *z-score* is based on the fund-specific benchmark portfolio and reflects a fund's ability in beating its own *ex ante* benchmarks. Descriptive statistics in Table 3 show that throughout the sample period the average *z-score* varies around 0 suggesting that the *z-score* measures out/under-performance.

We perform a *t*-test to examine whether the average *z-scores* are different from zero. During the period of 1998 through 2000 and the period of 2005 and 2006, the average *z-score* is positive at 5% significance level, while in 2002 and 2004 the average

Table 3 Descriptive statistics of the *z-scores*

	1998	1999	2000	2001	2002	2003	2004	2005	2006	Pooled
Mean	0.26**	0.27**	0.29***	0.08	−0.89***	0.14	−0.39***	0.30***	0.30***	0.03
Med.	0.14	0.19	0.28	−0.08	−1.00	0.04	−0.39	0.25	0.14	0.02
Max.	2.25	3.43	3.44	3.84	0.80	1.74	1.34	2.30	2.27	3.84
Min.	−3.07	−1.22	−1.59	−2.25	−2.91	−1.14	−1.79	−0.87	−0.58	−3.07
Std.	0.89	0.92	0.81	0.90	0.80	0.56	0.56	0.62	0.56	0.84
Skew	−0.38	0.84	0.59	0.98	−0.26	0.67	0.16	1.15	1.03	0.22
Kurt	5.65	4.20	6.17	6.98	3.31	4.23	3.79	4.85	4.41	5.29
Obs.	59	59	60	61	62	63	65	64	62	555
t-stat	2.21	2.25	2.75	0.69	−8.83	1.98	−5.67	3.90	4.26	0.97

Descriptive statistics for the *z-scores* of Dutch industry-wide pension funds over the period of 1998–2006. (*), (**), (***) indicate significance at the 10%, 5% and 1% levels, respectively. *t*-Statistics on the bottom line indicate whether the mean *z-score* for each year and for the pooled sample is different from 0

z-score is negative. This seems to suggest that the performance of pension funds as measured by the *z-score* fluctuates considerably over time. Given the construction of the *z-score* it is not clear however, whether this performance can be attributed to true skill or incomplete risk adjustment. The pooled average *z-score* is not significantly different from 0, which indicates that over a long period pension funds as a group do not outperform their benchmarks significantly. This result is in line with the limited selection and timing abilities of asset managers by pension fund trustees documented in Goyal and Wahal (2008).

5.1 Performance Persistence

The descriptive statistics show that the *average* pension fund is not able to beat its benchmark over time. In this section we focus on the performance persistence of the pension funds in our sample. In mutual fund research most studies indicate that there is no performance persistence.[12] As discussed in Berk and Green (2004), within a rational market framework, this is due to the free movement of competitive capital. In the pension fund industry, however, mandates stay with one asset manager often more than 2 years. There is no competitive supply of capital to pension asset managers, which may lead to some performance persistence. Below we present a number of ways to explore whether performance persistence can be detected in our sample using the *z-score* as a performance measure. In these tests we use a balanced sample of 57 funds with a complete set of reported *z-scores* in all 9 years.

Like many other financial data sets in asset pricing our data set of *z-scores* is characterized by cross-sectional correlation among the error terms when regressing current performance on past performances. We apply the two-step regression procedure of Fama and MacBeth (1973) in order to correct for cross-sectional error correlation in a panel setting. Table 3 indicates that in a certain year when the *z-score* of one fund is unusably high, the *z-score* of another fund is also likely to be high, i.e. there are positive correlations among *z-scores*. Therefore a pooled times-series cross-section regression is not suitable in analyzing our data. See also Cochrane (2001) and Petersen (2009).

We first run cross-sectional regressions of the current *z-scores* on the past *z-scores* on a yearly basis as in

$$z_{i,t} = a_t + b_t z_{i,t-1} + \epsilon_{i,t}, \quad i = 1, \ldots, 57, \quad t = 1999, \ldots, 2006.$$

over the period 1999–2006. Using standard OLS we obtain a time series of the slope coefficient estimates $(\hat{b}_t)$ for 8 years. Then we perform a *t*-test on the average estimated coefficient, shown in Panel A of Table 4. This number $(\overline{\hat{b}_t} = 0.07)$ indicates

---

[12] Among others see Gruber (1996), Carhart (1997), Bollen and Busse (2001). Some recent studies though point out short-run persistence when using daily and monthly returns and certain performance measures such as Bollen and Busse (2005) and Huij and Verbeek (2007).

Table 4 Persistence tests based on regression and ranking

Year	1999	2000	2001	2002	2003	2004	2005	2006
Panel A: regression								
$\hat{b}_t$	−0.39	−0.17	−0.10	0.27	−0.13	0.28	0.27	0.50
$\overline{\hat{b}}_t$				0.07	(0.62)			

Year	1998–1999	1999–2000	2000–2001	2001–2002	2002–2003	2003–2004	2004–2005	2005–2006
Panel B: ranking								
$\rho_{t,t-1}$	−0.19	−0.28	−0.02	0.22	−0.29	0.15	0.15	0.44
$\overline{\rho}_{t,t-1}$				0.02	(0.24)			

Panel A reports the slope coefficients ($\hat{b}_t$) from the cross-sectional Fama–MacBeth regression $z_{i,t} = a_t + b_t z_{i,t-1} + \epsilon_{i,t}$ for each year $t = 1999$–2006. $\overline{\hat{b}}_t$ is the averaged value over time of the $\hat{b}_t$ coefficients. *Panel B* reports the Spearman rank correlation coefficient over time. t-statistics are within parentheses

that past *z-scores* are positively related to current *z-scores*, but not in a statistically significant way. We conclude that pension funds as a group do not show performance persistence.

We also performed a Spearman rank correlation test for persistence, which does not require a distributional assumption. Each year we rank the funds based on their *z-scores*. The Spearman rank correlation coefficient for two consecutive years is then computed as

$$\rho_{t,t-1} = 1 - \frac{6 \sum_{i=1}^{N} d_{i,t,t-1}^2}{N(N^2 - 1)},$$

where $\sum_{i=1}^{N} d_{i,t,t-1}^2$ is the sum of squared differences of ranks over two consecutive years for all funds. $N = 57$ is the number of funds/ranks in our sample. For our 9-year sample, we obtain a time series of correlation coefficients for 8 years. As in the previous regression test, we apply a t-test using the average and the standard deviation of the time series, shown in Panel B of Table 4. We find again that the average coefficient ($\overline{\rho}_{t,t-1} = 0.02$) is not significantly different from zero, corroborating our earlier result using the Fama–MacBeth method.

The results from the Fama–MacBeth regressions may be influenced by outliers. One way to deal with this is to construct fund groups based on their past performance as in Fama and French (1992). Subsequently, we can analyze the performance on these sorted fund groups. Specifically, every year 3 fund groups are formed based on their *z-scores* in the previous year. Then for each individual group the current average *z-score* is computed. Repeating this for each year, we obtain a times series of average *z-scores* for the 3 fund groups. These are reported in panel A of Table 5. If performance is persistent, the best-performing fund group should provide the best performance in the subsequent years again. However, our results show that in some years the best performing group from the past year provides the worst performance in this year. For example, the best performing fund group in 1998 becomes the worst in 1999, having an average *z-score* of −0.10. The paired sample t-tests reported in Panel B show that

Table 5 Persistence test based on fund groups

Performance rank ($t-1$)	z-scores of 3 fund groups								
	1999	2000		2001	2002	2003	2004	2005	2006
Panel A									
Top (past best)	−0.10	0.17		0.05	−0.73	0.08	−0.28	0.51	0.80
Mid	0.56	0.20		0.09	−1.02	−0.02	−0.32	0.32	0.00
Bottom (past worst)	0.36	0.46		−0.09	−0.96	0.32	−0.42	0.14	0.06

	Paired sample *t*-tests		
	Mean of paired difference	SD	*t*-stat
Panel B			
Top–Mid	0.08	0.40	0.59
Mid–Bottom	−0.01	0.21	−0.08
Top–Bottom	0.08	0.39	0.56

Panel A reports the *z-score* in each year of a fund group formed on their previous year's *z-scores*. *Panel B* reports the paired sample *t*-test for mean differences. With a degrees of freedom equal to 7, critical values of 10% (*), 5% (**), 1% (***) significance level are 1.42, 1.90, and 3, respectively

Table 6 Migration statistics

Groups based on past performance	Fund groups based on current performance		
	Top	Mid	Bottom
Top	40% (0.95)	29% (−1.27)	31% (−0.43)
Mid	30% (−1.00)	28% (−1.20)	42% (2.02**)
Bottom	30% (−0.46)	43% (1.96**)	27% (−1.68*)
Total	1	1	1

This table reports fund migrations among groups sorted on performance. Every year funds are assigned to top, mid and bottom groups respectively according to their *z-scores* in that year. The column shows the composition of the current group that comes from the past top, mid or bottom group respectively. In parentheses are the *t*-statistics testing whether the percentage is equal to $33\frac{1}{3}$% for a sample of 57 funds. With a degrees of freedom equal to 7, critical values of 10% (*), 5% (**), 1% (***) significance level are 1.42, 1.90, and 3 respectively

none of the test statistics is statistically different from zero. This again confirms that there is no persistence in pension fund performance over time.[13]

In order to better understand the non-persistence results so far, we look further into the composition of the fund groups over time by applying the ideas from Fama and French (2007). These authors investigate how individual firms migrate from one portfolio to another over time and study its contribution to the cross-section returns. Each column in Table 6 reports the percentages of funds in the current group that originated from the previous year's top, mid and bottom fund groups, respectively.

---

[13] We also performed the analysis based on 5 fund groups (available upon request), but the results do not change.

We find that funds are not "sticky" to their group, and move considerably among the top, mid and bottom groups. For example, of the current top group, 30% are funds that were in the previous year's bottom group, and another 30% come from the mid group of the previous year. Of the current bottom group, 31 and 42% are the funds from the past top and the past mid group respectively. We test the hypothesis of random migration of funds among the three groups. The null hypothesis is that the migration probabilities are all equal to 1/3. The test statistics show that for six out of the nine possible migrations we cannot reject the hypothesis at a 5% significance level. Among the other three migrations, either past poor performing funds tend to move up in the next year, or past mediocre funds tend to move downward in the next year. Our inability to reject the null hypothesis of random migration indicates that performance persistence does not exist in our sample. In an unreported table we also examine the contribution of migrated funds to the *z-scores*.[14] We find in many situations a large part of the *z-score* of a bottom group is contributed by the funds that used to be in the top group, while the top group obtains a large chunk of its *z-score* from the funds that used to in the bottom group. Such dramatic changes of performance attribution between years reflect that past performance does not tell us much about future performance. In sum, the migration analysis underlines the lack of performance persistence that we found earlier.

5.2 Performance and Fund Size

The analysis so far shows that the Dutch industry-wide pension funds as a whole do not show any out- or under-performance with respect to their benchmarks. This, however, does not exclude the situation where a subgroup may outperform another subgroup. Thus it is interesting to investigate the cross-sectional difference among funds. Ambachtsheer et al. (1998) examine 80 US and Canadian pension funds for the period 1993–1996 and find that large fund size is an important driver for high investment performance. Reasons are that a large size brings economies of scale in operating cost and enables funds to support a full-time professional management team. Following these arguments we test whether fund size might be a differentiating factor in performance for our sample.

In order to measure the effect of size on the *z-score* of a fund we perform two cross-sectional regressions. The first regression takes the average *z-score* of a fund over time as the dependent variable and the second set of regressions performs this regression on an annual basis. Size is measured by the logarithm of the total number of participants in a fund in 2006, including active and inactive participants and retirees (see also Fig. 1). Panel A in Table 7 shows that size indeed matters. Size alone explains 27% of the variation in averaged fund *z-scores*. Moreover, the coefficient is positive, i.e. larger funds have a higher *z-score* than smaller funds on average. This result is consistent with earlier findings in the literature on pension fund investment performance. Goyal and Wahal (2008) study the decision of hiring and firing asset managers in US pension funds. They find that fund size can explain the post-hiring excess returns, and suggest that large size allows pension fund sponsors to develop

---

[14] This table is available upon request.

Table 7 Pension fund performance and size

Variable			Coef.	*t*-statistic					
Panel A: Regressing on time-average z-scores									
Log(number of participants)			0.18	4.86***					
Adjusted R-squared			0.27						
	1998	1999	2000	2001	2002	2003	2004	2005	2006
Panel B: regressing on annual z-scores each year									
beta	0.02	0.16	0.19	−0.05	0.49	−0.03	0.27	0.29	0.32
t-stat	0.11	0.88	1.19	−0.36	3.49***	−0.32	2.56***	2.49***	3.09***
R-squared	0.00	0.01	0.03	0.00	0.18	0.00	0.11	0.10	0.15

Panel A regresses the time-average z-score on the fund's size. Panel B regresses the z-score on fund size on a yearly basis. In both cases, the fund size is measured by the logarithm of a fund's total participants in 2006 including active and inactive participants and retirees. (***) indicates a significant level of 1%

Table 8 Average *z-score* of size groups over time

	3 size groups								
Year	1998	1999	2000	2001	2002	2003	2004	2005	2006
Panel A									
1 (largest)	0.38	0.35	0.49	0.16	−0.43	−0.09	−0.20	0.53	0.45
2	0.25	0.11	0.37	−0.32	−1.04	0.33	−0.39	0.26	0.31
3 (smallest)	0.22	0.36	−0.02	0.20	−1.24	0.14	−0.43	0.18	0.10
					Paired sample *t*-test				
Panel B									
Mean of paired difference					0.19				
t-stat					2.03**				

Panel A reports the equally-weighted z-score of each size group. Since relative fund size does not change over time, the three size groups are based on the number of total participants in 2006. Panel B reports the paired sample t-test for the z-score difference between the largest size group and the smallest size group. With a degrees of freedom equal to 8, critical values of 10%, 5%, 1% significance level are 1.40, 1.86, and 2.90, respectively. (*),(**),(***) indicate a significant level at 10%, 5% and 1%

expertise in selecting asset managers. Bauer et al. (2007) study the mandate size of delegated portfolios in pension funds. They find size is not a factor driving the benchmark adjusted net return, but size does bring economies of scale in reducing costs of external managers. Both these reasonings support our findings on size, but we cannot distinguish which of these explanations would apply to our findings.

We also performed regressions of the *z-score* on the fund's size on a yearly basis. Panel B in Table 7 shows that only in the years 2002, and 2004 till 2006 size has some explanatory power. In order to get more information on why this might occur we again construct a number of size-based fund groups, and examine their performance over time.

Table 8 reports the average *z-scores* of three size groups (Panel A). Each group contains the same number of funds (19 funds for each group). There is a clear difference

in the *z-scores* between the largest and the smallest size group. This is confirmed by the paired sample *t*-tests in Panel B. From Panel A it can be seen that there are non-monotonous performance patterns among groups in different years. For example, in the year 2003, the middle-sized group outperforms the other groups. And in 1999 and 2001 the smaller funds had the best performance.[15]

Although performance is not monotonously increasing with size, we do see (panel B) that the largest size group outperforms the smallest group *on average*. To relieve the concern over the power of the *t*-test in this relatively small sample, we also perform a Wilcoxon signed rank test, which is more robust for small samples. Again, this test indicates that largest funds outperform smallest ones. Our results are consistent with the findings on US and Canadian pension funds in Bauer et al. (2007) that size is an important factor explaining pension fund investment performance.[16]

The drivers behind the size effect cannot be investigated in this paper due to limitations of our data. The existing literature provides several explanations such as negotiation power in lowering costs, reputation effect, better monitoring of asset managers, or more expertise in selecting superior asset managers. Bikker et al. (2009) show that larger pension funds invest more in risky assets than smaller funds, which implies that larger funds earn higher expected returns than smaller funds. In our paper it is impossible to investigate this explanation due to the lack of data on asset allocations in both the benchmark (*ex ante*) as well as the realized (*ex post*) portfolio. As a result, we cannot check whether the prescribed risk correction in the *z-score* appropriately corrects for the risk that pension funds take. However, as the *ex ante* benchmark constructions and the subsequent *z-score* calculations are audited by external supervisors, this provides some confidence that the reported *z-scores* do reflect some important and valuable information about the performance of pension funds.

6 Concluding Remarks

One of the main tasks of pension fund trustees is to design an investment strategy that is consistent with the short and long term goals of the fund, and to implement this strategy effectively. This paper focuses on the investment implementation capabilities of pension fund trustees of Dutch sectorial pension funds. We investigate the added value of pension funds in delegating and monitoring their investment activities. For this purpose we use the *z-score* that Dutch mandatory industry-wide pension funds are obliged to publish in order to show their net-of-fees investment performance relative

---

[15] As shown in Fig. 1, the size distribution is skewed with a tail to the right, suggesting that most funds are small except a few extremely large funds. We adapted the construction of size groups by removing some outliers or forming unbalanced size groups such as categories based on total assets with unequal number of funds, to reflect this size asymmetry. We find that our results and conclusions do not change. These tables are available upon request.

[16] We also investigated whether there exists outperformance in the largest size group and underperformance in the smallest size group. We split our sample into subgroups in terms of size as measured by the amount of total participants. We redo the tests on average *z-scores*, the Fama–MacBeth regressions and the Spearman rank correlation tests for all the subgroups. The results show that there is no out- or under-performance and no performance persistence for all size subgroups. This indicates that our earlier result on the size effect is not affected by specific performance results in the subgroups. Results are available upon request.

to an *a priori* self-selected benchmark. The risk correction of the *z-score* is based on the *ex ante* benchmark composition, but the standard deviations on the excess returns are fixed by law. The scores intend to reflect the implementation quality of the strategic asset allocation.

Compared with retail investors, pension funds are more resourceful in carrying out an investment strategy. They can receive extensive help from advisors and consultants, gain valuable information before making the decisions, and can establish desired procedures to monitor the investment process. We would expect that pension fund trustees are able to select and recruit a superior group of internal and/or external asset managers and establish effective investment management procedures to encourage their asset managers to beat the pre-agreed benchmarks. The inconvenience of moving a large amount of pension assets across different asset managers or asset categories may also predict some type of performance persistence.

We have studied and reported *z-scores* on a comprehensive and unique data set of industry-wide pension funds in the Netherlands. We find that pension funds do not outperform their benchmarks consistently over time. In addition to annual performance tests, we also included a test to check whether funds showed performance persistence by analyzing whether funds moved (migrated) from one performance group to another. This migration test showed that the null hypothesis of random movement of funds from one group to another could not be rejected, thereby suggesting that any performance persistence is absent in our sample. All these tests imply that pension funds on average do not add additional investment value in implementing their investment plans. This conclusion also holds when we sort the pension funds into three equally-sized groups based on the total number of associated participants (active, inactive and retirees).

However, we do find that the largest funds perform significantly better than smaller funds when measured over the whole sample 1999–2006. This might be attributed to factors like economies of scale in costs, expertise in asset manager selection, or effective monitoring of asset managers. However, more detailed data on the composition of pension fund asset portfolios is needed to substantiate the validity of these arguments. Nevertheless, our results are consistent with the empirical trend of smaller pension funds merging with, or being acquired by, bigger funds in order to improve their investment performance.

All our results are based on the *z-score* as a performance measure. However, the construction of the *z-score* raises some doubts on using this measure as a proper investment performance measure. Especially the risk correction based on only two investment categories (equity and fixed income) is rather naive. The academic literature has proposed a large number of performance measures that could improve the risk correction. Also, the rather static implementation of the *z-score* at a yearly basis should be reassessed as well. We realize that our results may be influenced by these aspects, but we also stress that industry-wide pension funds in the Netherlands need to report their investment performance in the same manner. As the *z-score* is prescribed by Dutch law, and failing the performance test may have severe consequences for individual pension funds, it seems strange that the shortcomings of the *z-score* methodology have not attracted much more attention.

The credit crisis in 2008 has hit the Dutch pension fund severely and many funds became underfunded. Some funds even had to cut nominal benefits. This has raised

more attention to manage mismatch risk between liabilities and assets rather than financial asset performance alone. Liability-driven investment is being considered more and more. This leads to investment performance being more and more evaluated against liabilities rather than against financial benchmarks. No doubt the *z-score* initiated in the late 1990s has served the purpose of measuring the investment performance across pension funds for the period between late 1990s and early 2000s. However, with changes in the demographic structure, the regulatory and accounting environment, a more advanced performance measurement measure is needed. In this light, and also including the methodological issues raised above, we expect that the use of the *z-score* in evaluating the pension fund investment performance will lose its attractiveness.

References

Alestalo, N., & Puttonen, V. (2006). Asset allocation in finnish pension funds. *Journal of Pension Economics and Finance, 5*, 27–44.

Ambachtsheer, K., Capelle, R., & Scheibelhut, T. (1998). Improving pension fund performance. *Financial Analysts Journal, 54*(6), 15–20.

Bauer, R., Frehen, R., Lum, H., & Otten, R. (2007). *The performance of US pension funds: New insights into the agency costs debate*. Working Paper, Maastricht University.

Berk, J. B., & Green, R. C. (2004). Mutual fund flows and performance in rational markets'. *Journal of Political Economy 112*(6), 1269–1295.

Bikker, J. A., Broeders, D., Hollanders, D., & Ponds, E. (2009). Pension funds' asset allocation and participant age: A test of the life-cycle model. *Journal of Risk and Insurance* (forthcoming).

Bikker, J. A., & de Dreu, J. (2009). Pension fund efficiency: The impact of scale, governance and plan design. *Journal of Pension Economics and Finance, 8*(1), 63–89.

Blake, D., Lehmann, B. N., & Timmermann, A. (1999). Asset allocation dynamics and pension fund performance. *Journal of Business, 72*(4), 429–461.

Bollen, N. P. B., & Busse, J. A. (2001). On the timing ability of mutual fund managers. *Journal of Finance, 56*, 1075–1094.

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

Brinson, G. P., Hood, L. R., & Beebower, G. L. (1986). Determinants of portfolio performance. *Financial Analysts Journal, 43*, 39–44.

Busse, J. A., Goyal, A., & Wahal, S. (2006). Performance persistence in institutional investment management. *EFA 2006 Zurich meetings paper*.

Carhart, M. (1997). On persistence in mutual fund performance. *Journal of Finance, 52*, 57–82.

Cochrane, J. (2001). *Asset pricing*. Princeton: Princeton University Press.

Coggin, D. T., Fabozzi, F. J., & Rahman, S. (1993). The investment performance of U.S. equity pension fund managers: An empirical investigation. *Journal of Finance, 48*(3), 1039–1055.

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy, 81*, 607–636.

Fama, E. F., & French, K. R. (2007). Migration. *Financial Analysts Journal, 63*(3), 48–58.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance, 47*(2), 427–465.

Gerber, D. S., & Weber, R. (2007). Demography and investment behaviour of pension funds; evidence for Switzerland. *Journal of Pension Economics and Finance, 6*, 313–337.

Goyal, A., & Wahal, S. (2008). The selection and termination of investment management firms by plan sponsors. *Journal of Finance, 63*(4), 1805–1847.

Gruber, M. J. (1996). Another puzzle: The growth in actively managed mutual funds. *Journal of Finance, 51*, 783–810.

Huij, J. J., & Verbeek, M. J. C. M. (2007). Cross-sectional learning and short-run persistence in mutual fund performance. *Journal of Banking and Finance, 31*(3), 973–997.

Ippolito, R. A., & Turner, J. A. (1987). Turnover, fees and pension plan performance. *Financial Analysts Journal, 43*(6), 16–26.

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

Logue, D. E., & Rader, J. S. (1998). *Managing pension plans: A comprehensive guide to improving plan performance*. Boston, MA: Harvard Business School Press.

Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *Review of Financial Studies, 22*(1), 435–480.

Sharpe, W. F. (1992). Asset allocation: Management style and performance measurement. *Journal of Portfolio Management, 18*(2), 7–19.

Tonks, I. (2005). Performance persistence of pension-fund managers. *Journal of Business, 78*, 1917–1942.

Vrijstellingsbesluit Wet Bpf (2000). Besluit van December 21, 2000.