Capital Allocation and the Market for Mutual Funds: Inspecting the Mechanism*

Jules H. van Binsbergen
University of Pennsylvania and NBER

Jeong Ho (John) Kim
Emory University

Soohun Kim
Korea Advanced Institute of Science and Technology

July 26, 2021

Abstract

We exploit heterogeneity in decreasing returns to scale parameters across funds to analyze their effects on capital allocation decisions in the mutual fund market. We find strong evidence that steeper decreasing returns to scale attenuate flow sensitivity to performance, which has a large effect on equilibrium fund sizes. Our results are consistent with a rational model for active management. We argue that an important fraction of cross-sectional variation in fund sizes is due to investors rationally anticipating the effects of scale on return performance.

*We are grateful for comments from our discussants, Jun Kyung Auh, Laurent Barras, Jaewon Choi, and Yi Li. We are also grateful for comments from Nikolai Roussanov, Robert Stambaugh, and Youchang Wu, conference participants at the American Finance Association Annual Meeting, the China International Risk Forum, the Econometric Society World Congress, the Fixed Income and Financial Institutions Conference, the Frontiers of Factor Investing Conference, and the Northern Finance Association Meetings, and seminar participants at Boston Fed, Emory University, and Georgia Tech.

1 Introduction

An important question in financial economics is whether investors efficiently allocate capital across financial assets. Under the standard neoclassical assumptions, investors compete with each other for positive present value opportunities, and by doing so, remove them in equilibrium. In the case of mutual funds, the literature has argued that decreasing returns to scale (DRS) play a key role in equilibrating the mutual fund market (Berk and Green (2004)). Because the percentage fee that mutual funds charge changes infrequently, the bulk of the equilibration process operates through the size (or Assets Under Management (AUM)) of the fund. When good news about a mutual fund arrives, rational Bayesian updating will lead investors to view the fund as a positive Net Present Value (NPV) buying opportunity at its current size. In response, flows will go to that fund. As the fund grows, the manager of the fund finds it increasingly harder to put the new inflows to good use, leading to a deterioration of the performance of the fund. The flows into the fund will stop when the fund is no longer a positive NPV investment, and the fund's abnormal return to investors has reverted back to zero.

In this paper we investigate this equilibrating mechanism more closely. In particular, if the above-mentioned equilibration process is at work, we should expect to find that the degree of decreasing returns to scale (DRS) can have implications for the flow sensitivity to performance (FSP). While there is much evidence that an active fund's ability to outperform its benchmark declines as its size increases,[1] there is surprisingly little empirical work devoted to whether investors account for the adverse effects of fund scale in making their capital allocation decisions.

We address this important question by formally deriving and empirically testing what a rational model for active management implies about the relation between returns to scale and flow sensitivity to performance. Using a theory model similar to that of Berk and

[1] See, for example, Chen et al. (2004), Yan (2008), Ferreira et al. (2013), and Zhu (2018).

Green (2004), we show that steeper decreasing returns to scale attenuate flow sensitivity to performance. In the model, investors rationally interpret high performance as evidence of the manager's superior skill, so good performance results in an inflow of funds. More relevant to our hypothesis, the magnitude of the capital response is primarily driven by the extent of decreasing returns to scale. As a fund's returns decrease in scale more steeply, the positive net alpha is competed away with a smaller amount of capital inflows, making flows less sensitive to performance.

To test this theoretical insight, one needs a source of heterogeneity in decreasing returns to scale. One also needs to observe investor reactions to this heterogeneity. Indeed, we demonstrate that there is a substantial amount of heterogeneity in DRS across individual funds,[2] with correspondingly heterogeneous flow sensitivity to performance across funds. Our approach can be interpreted as inferring how the subjective size-performance relation, perceived by investors in real time, is incorporated into the flow-performance relation going forward. We find that a steeper decreasing returns to scale parameter predicts a lower sensitivity of flows to performance, consistent with the main prediction of our model.

One of the challenges in estimating the effect of decreasing returns to scale on flow sensitivity to performance is the estimation error in fund-specific DRS. As a result, the point estimates of the DRS-FSP relation using DRS estimates from simple fund-by-fund regressions are likely to suffer from an errors-in-variables bias. To gauge the severity of attenuation bias, we first adjust these simple estimates of the DRS-FSP relation for the errors-in-variable bias, assuming that the errors are of the classical type. As expected, they tend to be too small in magnitude. In turn, we address the attenuation bias associated with estimating the DRS-FSP relation by using the instrumental variables (IV) approach. In particular, we instrument for a fund's DRS estimate by using two fund-level variables: (i) the fund's exposure to the momentum factor and (ii) the fund's % inflow over the prior 1 year.

[2]Barras, Gagliardini, and Scaillet (2021) also provide empirical evidence that not only skill but also scalability vary substantially across funds.

We find that the degree of DRS is weaker for higher-momentum funds and funds that have experienced investor inflows.[3] Importantly, we show that while the statistical significance of the DRS-FSP relation is unaffected by using the IV approach, the IV estimates become substantially more negative, and their magnitudes are similar to those implied by the classical measurement error assumption, suggesting that the IV approach indeed has alleviated the errors-in-variables problem.[4]

Next we turn to the economic significance of our estimates. In particular, we assess how equilibrium fund size is affected by the cross-sectional variation in decreasing returns to scale parameters. This exercise does require model assumptions. We calibrate a rational model in the spirit of Berk and Green (2004). After simulating data in which investors know the DRS can vary by fund, we check how much of the simulated size can be explained by counterfactual fund sizes computed under the assumption that the investors believe the DRS is the same for all funds. We find that at least 34% of the cross-sectional variance of fund sizes can be related to cross-sectional variation in decreasing returns to scale parameters. More importantly, although we do not target the DRS-FSP relation in our calibration, our model produces DRS-FSP relation estimates that are quantitatively very similar to those obtained from the data. Thus, it appears that the magnitude of the DRS-FSP relation estimates from the data is consistent with what the model predicts. This result suggests that the model does a good job of capturing capital allocation patterns in the data.

Beyond implications for fund flows, steeper decreasing returns to scale have implications for fund size in equilibrium. In the model, fund size in equilibrium is proportional to the

[3] To the extent that hypotheses of decreasing returns to scale have been motivated by liquidity constraints, returns should be decreasing less steeply for funds experiencing investor inflows, which are likely to be cash-rich and thus more liquid funds. On the other hand, funds with higher momentum exposures are likely to hold winner (which tend to be liquid) rather than loser stocks (which tend to be illiquid) and thus more liquid portfolios, facing milder decreasing returns to scale.

[4] In the appendix, we explore yet another way to alleviate the errors-in-variables problem by relating the heterogeneity in decreasing returns to scale to a broad set of fund characteristics. Again, we show that while the statistical significance of the DRS-FSP relation is unaffected by using characteristic-based DRS, the point estimates become substantially more negative, and their magnitudes are comparable to those implied by the classical measurement error assumption and the IV approach.

ratio of perceived skill over diseconomies of scale, which predicts that, all else equal, the decreasing returns to scale parameter should be lower for larger funds. This prediction is confirmed in our empirical analysis. Moreover, if investors update their beliefs about skill as in the model, their perception of optimal size ought to converge to true optimal size as funds grow older. Consistent with this argument, we find that estimates for the optimal size largely explain capital allocation across older funds in the data. We measure (log) optimal size of a fund by the average ratio of the fund's net alpha (adjusted for returns to scale) to the fund's individual DRS parameter. We show that older fund's size continues to be significantly related to our measure of its optimal size even when we control for an alternative measure of optimal size that assumes fund scale has the same effect on performance for all funds. Again, investors seem to account not only for average decreasing returns to scale, but also for the heterogeneity of decreasing returns to scale across funds.

Taken together, our results demonstrate that investors do account for the adverse effects of fund scale in making their capital allocation decisions, and that the rational expectations equilibrium does a reasonable job of approximating the observed equilibrium in the mutual fund market. In contrast, the previous literature has often deemed mutual fund investors as naive return chasers because fund flows respond to past performance even though performance is not persistent.[5] Furthermore, many papers in the mutual fund literature have documented that mutual fund returns show little evidence of outperformance.[6] While these findings led many researchers to question the rationality of mutual fund investors, Berk and Green (2004) argue that they are consistent with a model of how competition between rational investors determines the net alpha in equilibrium. We contribute to this debate by presenting findings that are hard to reconcile with anything other than the existence of rational fund flows.

[5] See Chevalier and Ellison (1997) and Sirri and Tufano (1998), among others.

[6] See Jensen (1968), Malkiel (1995), Gruber (1996), Fama and French (2010), and Del Guercio and Reuter (2013), among others.

2 Definitions and hypotheses

To formally derive our hypothesis, we use the notation and setup presented in Berk and van Binsbergen (2016). Let q_{it} denote assets under management (AUM) of fund i at time t and let a_i denote a parameter that describes the skill of the manager of fund i. At time t, investors use the time t information set I_t to update their beliefs on a_i resulting in the distribution function $g_t(a_i)$ implying that the expectation of a_i at time t is:

$$\theta_{it} \equiv E[a_i | I_t] = \int a_i g_t(a_i) \, da_i. \tag{1}$$

We assume throughout that $g_t(\cdot)$ is not a degenerate distribution function. Let R^n_{it} denote the return in excess of the risk free rate earned by investors in fund i at time t. This return can be split up into the excess return of the manager's *benchmark*, R^B_{it}, and a deviation from the benchmark ε_{it}:

$$R^n_{it} = R^B_{it} + \varepsilon_{it}. \tag{2}$$

Note that q_{it}, R^n_{it} and R^B_{it} are elements of I_t. Let $\alpha_{it}(q)$ denote investors' subjective expectation of $\varepsilon_{i,t+1}$ when investing in fund i that has size q between time t and $t+1$, and let it be equal to:

$$\alpha_{it}(q) = \theta_{it} - h_i(q), \tag{3}$$

where $h_i(q)$ is a strictly increasing function of q that captures the decreasing returns to scale the manager faces, which can vary by fund. In equilibrium, the size of the fund q_{it} adjusts to ensure that there are no positive net present value investment opportunities so $\alpha_{it}(q_{it}) = 0$ and

$$\theta_{it} = h_i(q_{it}). \tag{4}$$

At time $t+1$, the investor observes the manager's return outperformance, ε_{it+1}, which is a signal that is informative about a_i. The conditional distribution function of $\varepsilon_{i,t+1}$ at time t,

$f_t(\varepsilon_{it+1})$, satisfies the following condition in equilibrium:

$$E[\varepsilon_{it+1} | I_t] = \int \varepsilon_{it+1} f_t(\varepsilon_{it+1}) \, d\varepsilon_{it+1} = \alpha_{it}(q_{it}) = 0. \tag{5}$$

In other words, the manager's return outperformance can be expressed as follows:

$$\begin{aligned} \varepsilon_{it+1} &= a_i - h_i(q_{it}) + \epsilon_{it+1} \\ &= s_{it+1} - h_i(q_{it}), \end{aligned}$$

where $s_{it+1} = a_i + \epsilon_{it+1}$. Our hypothesis relies on the insight that good news, that is, high s_{it}, implies good news about a_i and bad news, low s_{it}, implies bad news about a_i. The following lemma shows that this condition holds generally. That is, θ_{it} is a strictly increasing function of s_{it}.

Lemma 1 *If the likelihood ratio* $f_t(s_{it+1} | a_i) / f_t(s_{it+1} | a_i^c)$ *is monotone in* s_{it+1}*, increasing if* $a_i > a_i^c$ *and decreasing otherwise,*

$$\frac{\partial \theta_{it+1}}{\partial s_{it+1}} > 0. \tag{6}$$

Proof. *See Milgrom (1981).* ■

In addition, we assume that the costs that manager i faces in expanding the fund's scale is given by:

$$h_i(q) = b_i h(q), \tag{7}$$

where $b_i > 0$ is a parameter that captures the cross-sectional variation in the fund's returns to scale technology and $h(q)$ is a strictly increasing function of q, which essentially determines the form of decreasing returns to scale technology that is common across all funds. Using (7) to rewrite (4) now gives

$$q_{it} = h^{-1}\left(\frac{\theta_{it}}{b_i}\right). \tag{8}$$

The following lemma shows how the size of the fund q_{it} depends on the information in s_{it} or the parameter b_i.

Lemma 2

$$\frac{\partial q_{it}}{\partial s_{it}} = \frac{1}{b_i h'(q_{it})} \frac{\partial \theta_{it}}{\partial s_{it}} \tag{9}$$

and

$$\frac{\partial q_{it}}{\partial b_i} = -\frac{h(q_{it})}{b_i h'(q_{it})}. \tag{10}$$

Proof. *See appendix.* ■

Next, let the flow of capital into mutual fund i at time t be denoted by F_{it}, that is,

$$F_{it+1} \equiv \log(q_{it+1}/q_{it}).$$

Differentiating this expression with respect to s_{it+1},

$$\frac{\partial F_{it+1}}{\partial s_{it+1}} = \frac{1}{q_{it+1}} \frac{dq_{it+1}}{ds_{it+1}} = \frac{1}{q_{it+1}} \frac{1}{b_i h'(q_{it+1})} \frac{\partial \theta_{it+1}}{\partial s_{it+1}} > 0,$$

where the second equality follows from (9) and the inequality follows from Lemma 1, so good (bad) performance results in an inflow (outflow) of funds. This result is one of the important insights from Berk and Green (2004).

Given the importance of returns to scale technology in determining the size of a fund, a natural question to ask is, what is the implication of steeper decreasing returns to scale for the flow-performance relation? We answer this question by computing the derivative of the flow-performance sensitivity with respect to b_i:

$$\begin{aligned}
\frac{\partial}{\partial b_i}\left(\frac{\partial F_{it+1}}{\partial s_{it+1}}\right) &= \frac{\partial}{\partial b_i}\left(\frac{1}{q_{it+1}}\frac{1}{b_i h'\left(q_{it+1}\right)}\right)\frac{\partial \theta_{it+1}}{\partial s_{it+1}} \\
&= -\frac{q_{it+1}h'\left(q_{it+1}\right)+\frac{\partial q_{it+1}}{\partial b_i}\left(b_i h'\left(q_{it+1}\right)+q_{it+1}b_i h''\left(q_{it+1}\right)\right)}{q_{it+1}^2\left(b_i h'\left(q_{it+1}\right)\right)^2}\frac{\partial \theta_{it+1}}{\partial s_{it+1}} \\
&= -\frac{q_{it+1}h'\left(q_{it+1}\right)-h\left(q_{it+1}\right)\left(1+\frac{q_{it+1}h''\left(q_{it+1}\right)}{h'\left(q_{it+1}\right)}\right)}{q_{it+1}^2\left(b_i h'\left(q_{it+1}\right)\right)^2}\frac{\partial \theta_{it+1}}{\partial s_{it+1}}, \qquad (11)
\end{aligned}$$

where the first equality uses the fact that $\frac{\partial}{\partial b_i}\left(\frac{\partial \theta_{it+1}}{\partial s_{it+1}}\right)=0$, since θ_{it+1} is solely a function of the history of realized signals and is not a function of b_i, and the last equality invokes expression (10). What (11) combined with Lemma 1 tells us is that steeper decreasing returns to scale must lead to a smaller flow of funds response to performance if and only if

$$q_{it+1}h'\left(q_{it+1}\right)-h\left(q_{it+1}\right)\left(1+\frac{q_{it+1}h''\left(q_{it+1}\right)}{h'\left(q_{it+1}\right)}\right)>0. \qquad (12)$$

Unfortunately, the left hand side of equation (12) is not easy to sign without further assumptions. To assess whether this condition holds, we rely on the second-order approximation to the decreasing returns to scale technology:

$$h\left(q\right)\simeq h_0+h_1\log\left(q\right)+h_2\log\left(q\right)^2, \qquad (13)$$

where h_i for $i=\{0,1,2\}$ are the coefficients in the second-order approximation. This approximation nests exactly specifying the technology as logarithmic, most commonly considered in empirical studies, if we set $h_1>0$ and $h_0=h_2=0$. Going forward, we set $h_0=0$. This assumption is without loss of generality, because we can rewrite the skill parameter as $a_i'=a_i-b_ih_0$, which, in turn, renders $h_0'=0$. The following proposition shows that, under approximation (13), condition (12) holds generally. That is, steeper decreasing returns to scale lead to a weaker flow response to performance. We take this as our main hypothesis that we will take to the data.

Proposition 3 *Under approximation (13), the derivative of the flow-performance sensitivity with respect to the decreasing returns to scale parameter is negative, that is,*

$$\frac{\partial}{\partial b_i}\left(\frac{\partial F_{it+1}}{\partial s_{it+1}}\right) < 0.$$

Proof. *See appendix.* ■

3 Data

Our data come from CRSP and Morningstar. We require that funds appear in both the CRSP and Morningstar databases, which allows us to validate data accuracy across the two. We merge CRSP and Morningstar based on funds' tickers, CUSIPs, and names. We then compare assets and returns across the two sources in an effort to check the accuracy of each match following Berk and van Binsbergen (2015) and Pástor, Stambaugh, and Taylor (2015). We refer the readers to the data appendices of those papers for the details. Our mutual fund data set contains 3,066 actively managed domestic equity-only mutual funds in the United States between 1985 and 2014.

We use Morningstar Category to categorize funds into nine groups corresponding to Morningstar's 3×3 stylebox (large value, mid-cap growth, etc.). We also use keywords in the Primary Prospectus Benchmark variable in Morningstar to exclude bond funds, international funds, target funds, real estate funds, sector funds, and other non-equity funds. We drop funds identified by CRSP or Morningstar as index funds, in addition to funds whose name contains "index." We also drop any fund observations before the fund's (inflation-adjusted) AUM reaches $5 million.

We now define the key variables used in our empirical analysis: fund performance, fund size, and fund flows. Summary statistics are in Table 1.

3.1 Fund Performance

We take two approaches to measuring fund performance. First, we use the standard risk-based approach. The recent literature finds that investors use the CAPM in making their capital allocation decisions (Berk and van Binsbergen (2016)), and hence we adopt the CAPM. In this case the risk adjustment R_{it}^{CAPM} is given by:

$$R_{it}^{\text{CAPM}} = \beta_{it}\text{MKT}_t,$$

where MKT_t is the realized excess return on the market portfolio and β_{it} is the market beta of fund i. We estimate β_{it} by regressing the fund's excess return to investors onto the market portfolio over the sixty months prior to month t. Because we need historical data of sufficient length to produce reliable beta estimates, we require a fund to have at least two years of track record to estimate the fund's betas from the rolling window regressions.

Second, we follow Berk and van Binsbergen (2015) by taking the set of passively managed index funds offered by Vanguard as the alternative investment opportunity set.[7] We then define the Vanguard benchmark as the closest portfolio in that set to the mutual fund. Let R_t^j denote the excess return earned by investors in the j'th Vanguard index fund at time t. Then the Vanguard benchmark return for fund i is given by:

$$R_{it}^{\text{Vanguard}} = \sum_{j=1}^{n(t)} \beta_i^j R_t^j,$$

where $n(t)$ is the total number of index funds offered by Vanguard at time t and β_i^j is obtained from the appropriate linear projection of active mutual fund i onto the set of Vanguard index funds. As pointed out by Berk and van Binsbergen (2015), by using Vanguard funds as the benchmark, we ensure that this alternative investment opportunity set was marketed and tradable at the time. Again, we require a minimum of 24 months of data to estimate β_i^j's

[7] See Table 1 of that paper for the list of Vanguard Index Funds used to calculate the Vanguard benchmark.

necessary for defining the Vanguard benchmark for fund i.

Our measures of fund performance are then $\widehat{\alpha}_{it}^{\text{CAPM}}$ and $\widehat{\alpha}_{it}^{\text{Vanguard}}$, the realized return for the fund in month t less R_{it}^{CAPM} and R_{it}^{Vanguard}. The average of $\widehat{\alpha}_{it}^{\text{CAPM}}$ is +1.3 bp per month, whereas the average $\widehat{\alpha}_{it}^{\text{Vanguard}}$ is −1.5 bp per month.

3.2 Fund Size and Flows

We adjust all AUM numbers by inflation by expressing all numbers in January 1, 2000 dollars. Adjusting AUM by inflation reflects the notion that the fund's real (rather than nominal) size is relevant for capturing decreasing returns to scale in active management. That is, lagged real AUM corresponds to q_{it-1} in the model from Section 2. There is considerable dispersion in real AUM: the inner-quartile range is from \$45 million to \$622 million, while the 99th percentile is orders of magnitude larger at \$15 billion.

Fund flows are measured in two different ways. First, as in the model, we define fund flow F as the logarithmic change in real AUM, that is, the percentage change in fund size. Alternatively, we calculate flows for fund i in month t as:

$$F_{it} = \frac{AUM_{it} - AUM_{it-1}\left(1 + R_{it}\right)}{AUM_{it-1}\left(1 + R_{it}\right)},$$

where AUM_{it} is the nominal AUM of fund i at the end of month t, and R_{it} is the total return of fund i in month t.[8] Under this more traditional definition of F, flows represent the percentage change in new assets. The flow of fund data contain some implausible outliers, so we winsorize each of the two flow variables at its 1st and 99th percentiles. Mean percentage changes (per month) in fund size and in new assets are 0.8% and 0.5%, respectively.

[8]Note that we use $AUM_{it-1}\left(1 + R_{it}\right)$ in the denominator rather than AUM_{it-1}, which is typically used in much of the existing literature on fund flows. Unfortunately, this definition distorts the flow for very large negative returns, as shown by Berk and Green (2004): for example, liquidition of a fund, i.e., $AUM_{it} = 0$, implies a flow of $-\left(1 + R_{it}\right)$. Our measure of the flow of funds is equal to, and correctly so, −1 in this case. Regardless, our findings are unaffected by using the more common definition of the flow.

4 Method

Our analysis relies on a theoretical link between decreasing returns to scale and flow sensitivity to returns. We discuss how we estimate each part in the following sections.

4.1 Fund-Specific Decreasing Returns to Scale (DRS)

Empirically, we assume that the net alpha that manager i generates by actively managing money is given by:

$$\alpha_{it} = a_i - b_i \log(q_{it-1}) + \epsilon_{it}, \tag{14}$$

where a_i is the fund fixed effect, b_i captures the size effect, which can vary by fund, and q_{it-1} is the dollar size of the fund.[9]

The simple regression model in equation (14) corresponds to the model in Section 2. This model further assumes the form of the fund's decreasing returns to scale technology is logarithmic, which is often used to empirically analyze the nature of returns to scale due to severe skewness in dollar fund size.

We depart from much of the literature by allowing for heterogeneity in the size-performance relation across funds. Indeed, the effect of scale on a fund's performance is unlikely to be constant across funds. For example, a fund's returns should be decreasing in scale more steeply for those that have to invest in small and illiquid stocks.

Given that it is not clear a priori why and how the size-performance relation depends on which fund characteristics, we prefer to estimate fund-specific a_i and b_i parameters in our main analysis. For each fund i at time t, we run the time-series regression of $\alpha_{i\tau}$ on

[9] To the extent that fee changes are significant, it is possible that our results going forward might be sensitive to whether we use the net alpha or the gross alpha in equation (14). We report the former set of results but find that the latter results lead to the same conclusions. In fact, our results are stronger in the unreported results using the gross alpha in equation (14). This robustness is consistent with the evidence in the existing literature: fee changes are rare, so they are unlikely to play an important role in equilibrating the mutual fund market.

$\log(q_{i\tau-1})$ using sixty months of its data before time t.[10] Estimating b_i fund-by-fund leads to imprecise estimates especially for funds with short track records, so we require at least three years of data to estimate fund-specific returns to scale of a mutual fund.

The estimate of b_i, $\widehat{b}_{it}^m$, is obtained from (14) using sixty months of the data for fund i prior to time t, where the alpha can be estimated under model $m \in \{\text{CAPM}, \text{Vanguard}\}$. Intuitively, these estimates represent, for investors who use model m in making capital allocation decisions, their perception of the effect of size on performance for fund i at time t based on information prior to time t.

Panel A of Figure 1 shows how the cross-sectional distribution of $\widehat{b}_{it}$ using the CAPM alpha varies over time. For each month in 1991 through 2014, the figure plots the average as well as the percentiles of the estimated fund-specific b parameters across all funds operating in that month. The plot shows considerable heterogeneity in decreasing returns to scale across funds. For example, the interquartile range is more than 3 times larger than the estimates' cross-sectional median in a typical month; in fact, this ratio can be almost as large as 22 in some months. We find that, for the average fund, one percent increase in fund size is typically associated with a sizeable decrease in fund performance of about 0.8 basis points (bp) per month. This evidence suggests that the subjective size-performance relation, perceived by investors in real time, provides identifying variation in the extent of decreasing returns to scale.

Panel B of Figure 1 shows the time evolution of $\widehat{b}_{it}$ when we take Vanguard index funds as the alternative investment opportunity set. Similar to our main measure in Panel A, the alternative measure exhibits a clear heterogeneity in diseconomies of scale across funds, though these estimates typically indicate milder decreasing returns to scale.

[10] This is consistent with Pástor, Stambaugh, and Taylor (2015) and Zhu (2018), who also run simple OLS regressions when examining returns to scale fund by fund.

4.2 Fund-Specific Flow Sensitivity to Performance (FSP)

We estimate the fund-specific flow sensitivities to past performance by estimating the following regression fund by fund:

$$F_{it} = c_i + \gamma_i P_{it-1} + \upsilon_{it}, \tag{15}$$

where P_{it-1} is annual alpha for the year leading to month $t-1$, computed by compounding the monthly alphas as follows:

$$P_{it-1} = \prod_{s=t-12}^{t-1} \left(1 + R_{is}^{n} - R_{is}^{B}\right) - 1.$$

This regression is consistent with empirical evidence that investors do not respond immediately. For example, Berk and van Binsbergen (2016) and Barber, Huang, and Odean (2016) show that flows respond to recent returns, as well as distant returns. Parameter $\gamma_i > 0$ captures the positive time-series relation between performance and fund flows, which can vary by fund.

At time t, we calculate the fund's flow sensitivity to performance by estimating (15) using its data over the subsequent 5 years. For fund i, let $\widehat{FSP}_{it}^{m}$ be the estimated flow-performance regression coefficient of that model, where the performance can be estimated under model $m \in \{\text{CAPM}, \text{Vanguard}\}$. To avoid using imprecise estimates, we require these coefficient estimates to be obtained from at least three years of data. For the average fund, we observe that an increase of 1% in the monthly CAPM alpha is associated with an increase of 1.3% in monthly flows next month.

Figures 2 and 3 display the evolution of the distributions of $\widehat{FSP}_{it}$ by plotting the average as well as the percentiles of the estimated flow sensitivities to performance at each point of time. In Figure 2, we estimate the FSP's using the change in fund size to capture flows; in Figure 3, we estimate the FSP's using the change in new assets to define F. Panel A in each figure shows the distribution using the CAPM alpha, and Panel B shows the distribution when net alpha is computed using Vanguard index funds as benchmark portfolios. Note that the results are very similar across the two figures, manifesting considerable heterogeneities in the flow-performance relation across funds. More importantly, these figures show that while the average $\widehat{FSP}_{it}$ for both versions of the flow variable do not exhibit any obvious trend, they are certainly time varying. As the red dashed lines in the figures make clear, the distributions remain roughly the same over our sample period, conditional on the median.

5 Results

5.1 DRS and Flow Sensitivity to Performance

To investigate whether fund-specific decreasing returns to scale parameters are related to capital allocation decisions, we run panel regressions of fund i's flow sensitivity to performance going forward in month t, $\widehat{FSP}_{it}$, on the fund's returns to scale estimated as of the previous month-end, $\widehat{b}_{it}$. We test the null hypothesis that the slope on $\widehat{b}_{it}$ is zero.[11] We report the results in Tables 2 and 3.[12] In Panel A, we report the results using the change in fund size to capture flows; in Panel B, we examine their robustness using the change in new assets to define F. The first two columns in each panel use the CAPM as the benchmark, while the last two columns use Vanguard index funds as the benchmark.

We show results based on raw estimates in Table 2. We focus on variation coming from the market equilibrating mechanism beyond differences in sensitivity across funds and over time by including month and fund fixed effects. The fund fixed effects absorb the cross-sectional variation in flow/performance sensitivity that is due to differences in investor clientele across

[11]Surely, not only the independent variable, but the dependent variable are measured imprecisely. The measurement error in $\widehat{b}_{it}$ will bias the OLS estimator toward zero. While the measurement error in $\widehat{FSP}_{it}$ will not induce bias in the OLS coefficients, it will render their variance larger. For now, we do not worry, as the errors-in-variables problem will work against us from finding a statistically significant relation that the model predicts.

[12]Tables 2 and 3 report the double clustered (by fund and time) standard errors.

funds, while the time fixed effects soak up any variation in flow/performance sensitivity due to investor attention allocation over time. Indeed, there is evidence of clientele differences because some investors tend to update faster than others,[13] and Figures 2 and 3 show how the average as well as the median of flow-performance dynamics vary considerably over time.[14]

In the odd columns, we only include month and fund fixed effects. The results in Panel A are consistent with the main prediction of our model: the estimated coefficients on $\widehat{b}_{it}$ are negative and highly significant, with t-statistics of -5.6 in column 1 and -4.9 in column 3. These findings are unaffected by including a host of controls in the even columns, where we add proxies for participation costs, as considered by Huang, Wei, and Yan (2007),[15] as well as performance volatility and fund age.[16] The slopes on $\widehat{b}_{it}$ remain negative and highly significant, with t-statistics of -6.4 in column 2 and -4.8 in column 4, and their magnitudes are larger compared to odd columns where controls are excluded.

In Panel B, the same conclusions continue to hold when we consider $\widehat{FSP}_{it}$ estimated using the more traditional definition of F: the percentage change in new assets. Just like in Panel A, the coefficients on $\widehat{b}_{it}$ are significantly negative, and they increase in magnitude when we include a host of controls. While these negative slope coefficients are smaller in magnitude than their counterparts in the first panel, we suggest that they are also more affected by the attenuation bias, resulting from the measurement error in $\widehat{b}_{it}$, when we address the estimation error in scale effects in the next section.

Table 3 repeats this exercise with percentile ranks in each month based on $\widehat{b}_{it}$ and $\widehat{FSP}_{it}$. In this case, we do not use month fixed effects, as percentile ranks already soak up any time variation in the flow-performance relation. In each column, the estimated coefficient on $\widehat{b}_{it}$

[13] See Berk and Tonks (2007).

[14] Ferreira et al. (2012) discuss the role of economic, financial, and mutual fund industry development in determining the flow-performance relation across countries. Note that our fixed-effect approach already controls for these factors, since time fixed effects subsume any potential time-series variation in FSP due to different stages of development in the US.

[15] Specifically, we use marketing expenses, star family affiliation, family size, as well as fund size, to proxy for the variation in investors' information costs across funds.

[16] Huang, Wei, and Yan (2012) find that the flow-performance sensitivity is weaker for funds with more volatile past performance and longer track records.

is significantly negative at the 1% confidence level. Again, the addition of other potential determinants of the flow-performance relationship makes the slope coefficients on $\widehat{b}_{it}$ even more negative (compare columns 1 and 3 against 2 and 4 in each panel, respectively).

To summarize, we find a strong negative relation between decreasing returns to scale and flow sensitivity to performance. This relation, which is statistically significant, is consistent with the presence of investors rationally accounting for the adverse effects of fund scale in making their capital allocation decisions. Unfortunately, these coefficient values are not yet easily interpretable in economic terms. In Section 5.1.4, we propose a way of assessing the economic magnitude of these estimated coefficients by computing counterfactual fund sizes.

5.1.1 DRS-FSP Relation Under the Classical Measurement Error Assumption

We have estimated fund-specific b_i parameters based on a rolling estimation window. As noted earlier, estimating b_i fund by fund leads to imprecise estimates especially for funds with short track records. To gauge the severity of attenuation bias, we adjust the estimated coefficients on $\widehat{b}_{it}$ in Table 2 for the errors-in-variable (EIV) problem, assuming that the errors are of the classical type: they are purely random, have mean zero, and are uncorrelated with the regressors, including the actual b_i, and with the regression errors. Using the standard errors of $\widehat{b}_{it}$ to estimate the variance of measurement error in b_i, we can calculate the EIV-adjusted coefficients, reported in the last row of each panel.

As expected, the simple estimates of the DRS-FSP relation tend to be too small in magnitude. For example, when the DRS-FSP relation is esimated based on the CAPM to measure fund performance and on the change in fund size to capture flows controlling for other potential determinants of the flow-performance relationship (i.e., column 2 in Panel A), the coefficient becomes substantially more negative with the EIV adjustment: -7.80, compared to -1.17 without this adjustment. Bias is even more severe for estimates based on the Vanguard benchmark than those based on the CAPM. When the DRS-FSP relation is estimated using the Vanguard benchmark, the EIV-adjusted coefficients are 7 to 10 times larger than their simple-estimate counterparts (see the last two columns of Table 2). Of course, these results are only true if the errors are indeed of the classical type, but they illustrate that our estimates of the DRS-FSP relation are likely to be severely biased toward zero.[17]

5.1.2 DRS-FSP Relation Under the Instrumental Variables Approach

We address the attenuation bias associated with estimating the DRS-FSP relation by using the instrumental variables (IV) approach. In particular, we instrument for $\widehat{b}_{it}$ by using two fund-level variables: (i) $\widehat{\beta}_{it}^{mom}$, fund i's momentum exposure over the sixty months up to month $t-1$, and (ii) "Past inflow," the fund's % inflow over the prior 1 year. Note that we can implement our IV estimator via two-stage least squares. We first regress $\widehat{b}_{it}$ on "Past inflow" and $\widehat{\beta}_{it}^{mom}$, and then we regress $\widehat{FSP}_{it}$ on the fitted values from the first-stage regression. Both regressions include fund and month fixed effects, as well as the controls in Table 2. Table 4 shows the results from the IV procedure. These regressions are most comparable to our estimates in columns 2 and 4 of Table 2.

To be valid instruments for $\widehat{b}_{it}$, "Past inflow" and $\widehat{\beta}_{it}^{mom}$ must satisfy the relevance and exclusion conditions. The relevance condition requires that $\widehat{b}_{it}$ be significantly related to "Past inflow" and $\widehat{\beta}_{it}^{mom}$ in the first-stage regression. Why do we expect these characteristics to affect how scale impacts performance? Hypotheses of decreasing returns to scale have been motivated by liquidity constraints. Lower liquidity of a fund's assets is likely to make the fund's returns decrease in scale more steeply because it faces larger total price impact costs. This logic suggests that returns should be decreasing less steeply for funds experiencing investor inflows, which are likely to be cash-rich and thus more liquid funds. Higher-momentum funds might also exhibit milder decreasing returns to scale. The reason is that the momentum factor goes long on winners (which tend to be liquid) and short on

[17] Barras, Gagliardini, and Scaillet (2021) make a closely related observation that the error-in-variable bias can have significant impact on the cross-sectional distribution of scale coefficient b_i.

losers (which tend to be illiquid).[18] Funds with higher momentum exposures are likely to hold winner rather than loser stocks and thus more liquid portfolios, facing milder decreasing returns to scale. Indeed, $\widehat{b}_{it}$ is significantly negatively related to both "Past inflow" and $\widehat{\beta}_{it}^{mom}$, regardless of whether it is estimated using the CAPM or the Vanguard benchmark. We also report the first-stage F-statistics on the excluded instruments, which are well above 10.

The exclusion condition requires that the instruments should be unrelated to the innovation in the dependent variable, i.e., the future flow-performance sensitivity for each fund. This condition is likely to hold as well, since two factors fully determine the magnitude of capital response to performance in a rational model — the degree of decreasing returns to scale, and the prior and posterior beliefs about managerial skill. This means that "Past inflow" and $\widehat{\beta}_{it}^{mom}$ are unlikely to have information helpful in predicting a fund's flow-performance relation above and beyond their implications for the fund's DRS technology. Since we have a single regressor that is measured imprecisely and two instruments, it is possible to carry out the overidentifying restrictions test, whose p-values are reported in Table 4. This procedure is unable to reject the null that all the instruments are exogenous.

Consistent with the presence of severe attenuation bias, the IV estimates of the relation between $\widehat{FSP}_{it}$ and $\widehat{b}_{it}$ are substantially more negative than those in columns 2 and 4 in Table 2, and the relation continues to be highly statistically significant, with t-statistics that are more negative than -4 in Table 4, except in column 2 of Panel A, where it is -2.22.[19] The IV estimates are 6 times larger than the OLS estimates using the change in fund size to capture flows, and their magnitudes are similar to those implied by the classical measurement error assumption. Using the change in new assets to define F, the IV estimates are even more negative than those implied by the classical measurement error assumption.

[18] Avramov, Cheng, and Hameed (2016) provide empirical evidence supporting this fact.

[19] Table 4 reports the double clustered (by fund and time) standard errors.

5.1.3 DRS-FSP Relation Using the Characteristic Component of DRS

In Appendix B, we explore which fund characteristics are correlated with the observed heterogeneity in returns to scale. Based on this analysis, we obtain an economically interpretable component of $\widehat{b}_i$ based on fund characteristics, using which we re-estimate the DRS-FSP relation. The characteristic-based approach taken there exploits additional fund characteristics that are relevant for identifying variation in DRS, compared to the IV approach in the previous section. While we do not test whether this method actually increases the accuracy of the DRS estimates, the prior evidence of fund-level DRS depending on fund characteristics suggests that it *is* likely to deliver a more accurate measure of b_i and thus is yet another reasonable way to mitigate the errors-in-variable problem. Indeed, when we conduct the analysis using the characteristic component of DRS, the estimates of the DRS-FSP relation become substantially more negative than in Table 2 and they are comparable in magnitude to those implied by the classical measurement error assumption and the IV approach.

5.1.4 Simulated DRS-FSP Relation

In this section, we use our model to ask how much capital is allocated the way it is because of these differences in decreasing returns to scale. Specifically, we compute counterfactual fund sizes by assuming the investors believe a priori that returns are decreasing in scale at the same (average) rate for all funds.

Two factors fully determine the magnitude of capital response to performance in a rational model — the degree of decreasing returns to scale, and the prior and posterior beliefs about managerial skill. This means that, for a given value of b in equation (14), the prior uncertainty about a, σ_0, can be inferred from the flow-performance relation, as long as investors update their posteriors with the history of returns as Bayesians.

We simulate benchmark-adjusted fund returns from equation (14). It is straightforward to show that the mean of investors' posteriors will satisfy the following recursion:

$$\theta_{it} = \theta_{it-1} + \frac{\sigma_{i0}^2}{\sigma^2 + t\sigma_{i0}^2}\alpha_{it},$$

where θ_{i0} is the mean of the initial prior. Using (8), we compute fund size as follows:

$$q_{it} = \exp\left(\frac{\theta_{it}}{b_i}\right).$$

We begin by tying down the model parameters that can be set directly. Following Berk and Green (2004), we set $\text{Std}(\varepsilon) = 20\%$ per year, or 5.77% per month. Investors' prior on a fund's ability is that a_i is normally distributed with mean θ_0 and standard deviation σ_0. Since investors are assumed to have rational expectations, this is also the distribution from which we draw each fund's skill. We shall also assume that funds shut down the first time $\theta_{it} < \overline{\theta}$, where we set $\overline{\theta} = 0$.[20] These parameter values are summarized in Panel A of Table 5. It is straightforward to see that the only remaining parameters that we need to set for simulating data are b, θ_0 and σ_0.

The empirical distribution of b is approximated by a generalized Pareto (GP) distribution, from which we draw b randomly. In that case, assuming that θ_0 is independent of b gives rise to distributions of fund size considerably more disperse than in our actual sample. Specifically, the simulated fund sizes tend to be too big for funds whose returns decrease in scale more gradually, while the simulated fund sizes tend to be too small for those that exhibit steeper decreasing returns to scale. In turn, we model the prior mean as a linear function of b, $\theta_0(b)$. Our approach is to fit the parameters governing this function such that the simulated mean and standard deviation of log fund size essentially match the empirical benchmark values of 5.12 and 1.86, respectively.[21]

[20] Intuitively, managers incur fixed costs of operation each period. These costs can be, for example, overhead, back-office expenses, and the opportunity cost of the manager's time. Managers will optimally choose to exit when they cannot cover their fixed costs.

[21] Note that there generally exist multiple ways prior mean as a function of b for which the simulated mean and standard deviation of log fund size can match the empirical benchmark values. To pick a single

We set the prior uncertainty (σ_0) to match the average flow-performance relation in the data. To this end, we construct $2,500$ samples of simulated panel data for $10,000$ funds over 100 months. We simulate a given sample by first drawing each fund's DRS b_i randomly from a GP distribution consistent with the distribution of fund-specific b estimates, while drawing the fund's skill a_i from a normal distribution with mean $\theta_0(b_i)$ and standard deviation σ_0. Next, we draw the random values of ε_{it}, building up the panel data of r_{it} and q_{it}. For each fund in the sample, we run the following regression using data for just that fund to estimate its flow-performance sensitivity:

$$\log(q_{it}/q_{it-1}) = c_i + \gamma_i r_{it} + \upsilon_{it}.$$

Given all other parameters, we set σ_0 so that the mean of the average $\widehat{\gamma}_i$ across simulated samples matches the average $\widehat{FSP}_{it}$ in our actual sample. Panel B of Table 5 shows the value of σ_0 that resulted from this process. It also contains the values of the parameters governing the GP distribution of DRS and those of the parameters governing the prior mean that we use in our simulation analysis. The last three columns of Panel B report all the moments that we target in our calibration, as well as their values in both the actual and simulated data. Note that the simulated moments in the model closely match the target moments.

Recall from Table 4 that steeper decreasing returns to scale imply less flow sensitivity to performance. For example, as shown in column 1 of Panel A, the DRS-FSP relation is -7.249. To assess the economic magnitude of such estimates, we estimate the DRS-FSP relation in each of the simulated samples. Panel A of Table 6 shows summary statistics of these estimates across simulated samples from the calibrated model. Perhaps surprisingly, the DRS-FSP relation estimates in the calibrated model tend to be slightly smaller in magnitude compared to column 1 of Panel A of Table 4. But importantly, the empirical DRS-FSP

function, we impose the additional constraint that the simulated mean of log fund size is decreasing in b. This constraint is motivated by empirical evidence presented later in Section 5.2: steeper decreasing returns to scale shrink fund size.

relation estimate lies comfortably within the 95% confidence interval for simulated DRS-FSP relation estimates and vice versa. Thus, it appears that the magnitude of the DRS-FSP relation estimates from the data is consistent with what the model predicts. This result suggests that the calibrated model does a good job of capturing capital allocation patterns in the data.

To quantitatively assess the role of heterogeneity in returns to scale in capital allocation, we must construct a counterfactual. We construct the counterfactual by assuming investors learn about skill based on distorted beliefs that the fund exhibits average decreasing returns to scale. Specifically, the counterfactual investors assume that $b_i = 0.0105$ for all funds.[22] Then, updating investors' beliefs with the history of its returns under the counterfactual assumption, we compute what the size of the fund would have been.

Again, we construct $2,500$ samples of simulated panel data for $10,000$ funds over 100 months. To simulate a given sample, we first draw each fund's DRS b_i randomly from a GP distribution consistent with the distribution of fund-specific b estimates, while we draw the fund's skill a_i from a normal distribution with mean $\theta_0(b_i)$ and standard deviation σ_0. Next, we draw the random values of ε_{it}, building up the panel data of r_{it} and q_{it}. For every i and t, we compute the fund's size under the counterfactual, q_{it}^C, as detailed above. Finally, for each sample, we calculate the R^2 from a regression of $\log(q_{it})$ on $\log(q_{it}^C)$ to check the goodness of fit by the counterfactual. Here, 1 minus the R^2 can be interpreted as the fraction of capital allocation explained by individual heterogeneity in decreasing returns to scale.

We report the results from counterfactual simulations in Panel B of Table 6. The first two rows show summary statistics of the coefficient estimates from the regression of $\log(q_{it})$ on $\log\left(q_{it}^C\right)$ across simulated samples; the last row shows summary statistics of the R^2 from this regression across simulated samples.

The counterfactually computed fund sizes explain about 66% of the variation of simu-

[22] Note that the counterfactual investors have rational expectations about the skill level, i.e., they know the true $\theta_0(b_i)$ for each fund's DRS parameter b_i. So this assumption ensures that differences between the simulated outcomes and the counterfactual sizes are not driven by distorted beliefs about the skill level.

lated fund sizes. While counterfactual sizes are positively related to actual sizes, they are considerably larger than actual sizes and their distributions are substantially tighter than those of actual sizes. On one hand, since the size of a fund is inversely proportional to DRS and, in turn, strictly convex in DRS, Jensen's inequality implies that the counterfactual investors naturally overestimate the average fund size. On the other hand, since differences in DRS across funds is a major source of the cross-sectional variation in fund size, these counterfactual investors naturally underestimate the true dispersion in fund size. Thus, the counterfactuals ignoring heterogeneity in DRS are very different than the actual size. In this sense, we can interpret 1 minus the R^2 as a lower bound on the role of heterogeneity in returns to scale on capital allocation: at least 34% of the cross-sectional variance of fund sizes can be related to cross-sectional variation in decreasing returns to scale parameters, which is economically significant.

To summarize, Table 6 shows that a significant fraction of equilibrium capital allocation can be plausibly be explained by investor response to differences in decreasing returns to scale. Not only are fund sizes in the data quantitatively consistent with what our simple model predicts they should be, the magnitude of empirical DRS-FSP relation estimates are consistent with what our simple model would predict.

5.2 DRS and Fund Size in Equilibrium

While the main implication of our model is that steeper decreasing returns to scale attenuate flow sensitivity to performance, another immediate implication is that steeper decreasing returns to scale shrink fund size. Recall that fund size in equilibrium is proportional to the ratio of perceived skill over diseconomies of scale (see equation (8)). This implies that large funds either earn a high gross alpha on the first dollar and/or implement strategies that are highly scalable. We now investigate the importance of the latter effect, while explicitly controlling for the former effect, as well as for fund style and fund age. Table 7 presents the results of this exercise.

To control for the effect of perceived skill, we first form quintile groups sorted on fund fixed effects estimated as of month $t-1$, $\widehat{a}_{it}$. Then, within each $\widehat{a}_{it}$ quintile, we sort funds into five groups based on fund-specific returns to scale estimated as of month $t-1$, $\widehat{b}_{it}$. Both $\widehat{a}_{it}$ and $\widehat{b}_{it}$ are computed from estimating (14) using sixty months of the data for fund i prior to time t. We conduct double sorts of funds belonging to the same Morningstar category and to the same age category.[23] After forming the 5×5 $\widehat{a}_{it}$ and $\widehat{b}_{it}$ groups, we average fund sizes, as measured by log real AUM in month t, of each $\widehat{b}_{it}$ quintile over the five $\widehat{a}_{it}$ groups. This characteristic control procedure creates a set of quintile $\widehat{b}_{it}$ groups with similar levels of perceived skill, and with near-identical distributions of fund style and fund age. Thus, these quintile $\widehat{b}_{it}$ groups control for differences in skill, as well as fund style and fund age.

Panel A of Table 7 reports average fund sizes of the 25 $\widehat{a}_{it} \times \widehat{b}_{it}$ groups using the CAPM as the benchmark. The column labeled "Average" reports the average month-end fund sizes of the $\widehat{b}_{it}$ quintiles, controlling for $\widehat{a}_{it}$, fund style and fund age. The row labeled "High-low" reports the differences in average sizes between the first and fifth quintile $\widehat{b}_{it}$ groups within each $\widehat{a}_{it}$ quintile.[24] The difference in average sizes in the bottom right entry of Panel A indicates that the sizes of funds that are perceived to face steepest decreasing returns to scale tend to be 80% smaller than those of funds that are perceived to be relatively immune to the adverse scale effects. This difference has a robust t-statistic around -20. Hence, steeper decreasing returns to scale shrink fund size, consistent with the above prediction of our model. Importantly, this effect is not only statistically but also economically significant. The patterns within each $\widehat{a}_{it}$ quintile moving from low $\widehat{b}_{it}$ to high $\widehat{b}_{it}$ funds (reading down each column) are very similar, except for the first $\widehat{a}_{it}$ quintile, in which the sizes of funds with steeper perceived DRS are larger than those of funds with milder perceived DRS. But this difference is both economically and statistically insignificant.[25] Panel B of Table 7 repeats

[23] Specifically, we assign funds to one of three samples based on fund age: $[0, 5]$, $(5, 10]$, and > 10 years.

[24] To adjust for the strong persistence in fund size, we report standard errors of these differences in average fund sizes between quintile portfolio 5 (high $\widehat{b}_{it}$) and quintile portfolio 1 (low $\widehat{b}_{it}$) using 60 Newey-West lags.

[25] The reason for this pattern is as follows. In the first $\widehat{a}_{it}$ quintile, we find that the funds' estimates of both a_i and b_i are typically negative. To the extent that all funds must face decreasing returns to scale in

the same exercise as Panel A, except we use Vanguard index funds as the benchmark. We find the same qualitative patterns.

In summary, steeper decreasing returns to scale shrink fund size in the data. This relation, which is not only statistically but also economically significant, is consistent with the presence of investors rationally accounting for the adverse effects of fund scale in making their capital allocation decisions. Less important but also noteworthy is that the sizes of funds with higher perceived skill tend to be larger than those of funds with lower perceived skill (reading from left to right in each panel), again consistent with our model.

5.2.1 DRS and Optimal Fund Size

Thus far, we have used heterogeneity in decreasing returns to scale across funds and over time to test whether investors respond to the adverse effects of fund scale in making their capital allocation decisions. If investors update their beliefs about skill as in the model, their perception of optimal size ought to converge to true optimal size over a fund's lifetime. This implies that the sizes of older funds should be more closely related to their optimal sizes based on the model than those of younger funds. In this section, we test this prediction and find empirical support for it.

We have estimated fund-specific b_i parameters based on a rolling estimation window. As noted earlier, estimating b_i fund by fund leads to imprecise estimates. In particular, about 28% of the funds in our sample end up with negative $\widehat{b}_i$. While this is not an issue when focusing only on the *relative* steepness of the DRS technology as in the rest of the paper, it is a problem for computing the optimal fund size, which requires that $b_i > 0$ since, theoretically, all mutual funds must face decreasing returns to scale in equilibrium. A straightforward way to deal with this econometric shortcoming is to "shrink" the OLS estimates toward their prior mean, i.e., the average fund-level DRS parameter in our sample, denoted by b^{RD2}, which we

equilibrium, these funds' perceived DRS are likely to be close to zero as well as close to each other. Hence, the cross-sectional variation in $\widehat{b}_{it}$ is mostly driven by random noise in this $\widehat{a}_{it}$ quintile, so fund sizes do not meaningfully differ based on differences in $\widehat{b}_{it}$ within this sample.

estimate using the recursive demeaning procedure of Zhu (2018).[26] Measuring performance using the CAPM, $\widehat{b}^{RD2}$ is statistically significant, indicating that an 1% increase in fund size is associated with a decrease in the fund's CAPM alpha of 0.42 bp per month.[27] All of the resulting fund-specific DRS values, denoted by $\widehat{b}_i^{shrunk}$, are positive. Then, the skill parameter a for fund i can be estimated as:

$$\widehat{a}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} \left(\widehat{\alpha}_{it} + \widehat{b}_{it}^{shrunk} \log \left(q_{it-1} \right) \right),$$

where $\widehat{\alpha}_{it}$ is the net alpha, and T_i is the number of observations for fund i. We employ the average value of the ratios $\widehat{a}_i / \widehat{b}_{it}^{shrunk}$ over a fund's lifetime to get an estimate for the optimal fund size, $\log(\widehat{q}_i^*)$.[28] However, this measure of optimal fund size can be different than what investors believe to be optimal (ex-post) if they do not account for individual heterogeneity in decreasing returns to scale. In this case, the valid proxy for the optimal size, perceived by investors, would be the optimal size estimated assuming that the fund size has the same effect on performance across all funds. We can then estimate the parameter a for fund i as:

$$\widehat{a}_i^{RD2} = \frac{1}{T_i} \sum_{t=1}^{T_i} \left(\widehat{\alpha}_{it} + \widehat{b}^{RD2} \log \left(q_{it-1} \right) \right).$$

The alternative measure of optimal fund size $\log\left(\widehat{q}_i^{*RD2}\right)$ is calculated as $\widehat{a}_i^{RD2} / \widehat{b}^{RD2}$.

To test the above prediction, we examine how the relation between log real AUM and our measure of optimal fund size depends on fund age. Specifically, we assign funds to one of three samples based on fund age: $[0, 5]$, $(5, 10]$, and > 10 years. In each age sample, we

[26] Pástor, Stambaugh, and Taylor (2015) analyze the nature of returns to scale by developing a recursive demeaning procedure. They find coefficients indicative of decreasing returns to scale both at the fund level and at the industry level, though only the latter is statistically significant. Zhu (2018) improves upon the empirical strategy in PST (by using more recent fund sizes as the instrument) and establishes strong evidence of fund-level diseconomies of scale.

[27] Using Vanguard index funds as benchmarks, the coefficient estimate is again statistically significant, indicating a decrease in fund performance of 0.0013% per month for an 1% increase in fund size.

[28] Note that the optimal fund size here is the size at which the benchmark-adjusted net return is expected to be zero. This is different than, but related to, the optimal amount the manager chooses to actively manage (Berk and van Binsbergen (2015)).

run panel regressions of fund i's log real AUM in month t on the fund's log optimal fund size estimate, $\log(\widehat{q}_i^*)$. We report the results in the first three columns of Table 8.[29] In Panel A, we report the results using the CAPM as the benchmark; in Panel B, we use Vanguard index funds as the benchmark.

Across all three age groups, the estimated coefficients on $\log(\widehat{q}_i^*)$ are positive and highly statistically significant. More importantly, the coefficient values increase over a typical fund's lifetime, indicating that this positive relation between the fund's size and its optimal size is stronger for older funds. As the fund ages, investors learn about its optimal size, implying that the equilibrium size is closer to this optimal size measure. In addition, the R^2 of the regressions confirm this insight. The R^2 in the > 10 age sample is the highest and the R^2 decreases monotonically as we move to the samples of ages $(5, 10]$ and $[0, 5]$ funds.

Next, we split our measure of optimal fund size into two components: (i) that explained by $\log\left(\widehat{q}_i^{*RD2}\right)$, and (ii) that unexplained by $\log\left(\widehat{q}_i^{*RD2}\right)$. In columns 4 through 6, we run the multiple regression of $\log(q_{it})$ on both components of $\log(\widehat{q}_i^*)$ in all three age-sorted samples. We find that the slope on the component of $\log(\widehat{q}_i^*)$ explained by $\log\left(\widehat{q}_i^{*RD2}\right)$ is positive and significant across all age-sorted samples. The slope on the component of $\log(\widehat{q}_i^*)$ unexplained by $\log\left(\widehat{q}_i^{*RD2}\right)$ is positive and significant in the > 10 sample, but it is negative in samples of younger funds.

The significantly positive coefficient on the component of $\log(\widehat{q}_i^*)$ unexplained by $\log\left(\widehat{q}_i^{*RD2}\right)$ in the multiple regression using the > 10 sample, whose R^2 remains about the same as in column 3, reveals that investors *do* account for the fund heterogeneity in decreasing returns to scale when allocating their capital to older funds. On the other hand, negative coefficients on the component of $\log(\widehat{q}_i^*)$ unexplained by $\log\left(\widehat{q}_i^{*RD2}\right)$ in the multiple regressions using samples of younger funds suggest that investors allocate their money to younger funds based on the simpe version of optimal size. Our results offer the following narrative. Investors want to account for heterogeneity in decreasing returns to scale, but need to learn about

[29] Table 8 reports the double clustered (by fund and time) standard errors.

fund-specific values. Given that such fund-specific information is not yet available for young funds, investors use the sample-wide b instead. In particular, the investors only use the $\widehat{q}_i^*$ estimate in making their capital allocation decisions when a fund grows old enough such that the remaining Bayesian uncertainty on fund-specific b is relatively modest. Thus, it appears that investors in the data might be learning not only about skill but also about decreasing returns to scale.[30] We leave the task of examining the capital allocation implications of learning about fund heterogeneity in decreasing returns to scale technology for future research.

In short, the estimates of optimal size largely explains capital allocation to older funds. Both measures of optimal fund size matter, which is consistent with our narrative that investors account for not only the presence of decreasing returns to scale, but the heterogeneity of decreasing returns to scale across funds.

6 Conclusion

The main contribution of this paper is to provide and verify predictions unique to a rational model for active management: the role of decreasing returns to scale in equilibrating the market for mutual funds. Not only do we find that steeper decreasing returns to scale attenuate flow sensitivity to performance, we also find that differences in decreasing returns to scale across funds are quantitatively important for explaining capital allocation in the market for mutual funds. Interestingly, the magnitude of empirical DRS-FSP relation estimates are consistent with what our simple model predicts. This result suggests that the rational model for active management does a good job of capturing capital allocation patterns in the data. Overall, our results support that, as a group, investors in the mutual fund market are *not* naive.

[30] For formal models that relate capital allocation to learning about returns to scale, see Pastor and Stambaugh (2012) and Kim (2017).

Appendix

A Proofs

A.1 Proof of Lemma 2

First, note that ε_{it} does not contain information about managerial ability that is not already contained in s_{it}. Because rescaling the fund's returns to scale technology (i.e., changing the parameter b_i) does not change the signal s_{it}, we can conclude that

$$\frac{\partial \theta_{it}}{\partial b_i} = 0. \tag{16}$$

Now differentiating (8) with respect to s_{it}, using the Inverse Function Theorem, and using the fact that these signals are independent of b_i (i.e., $\partial b_i / \partial s_{it} = 0$), gives

$$\frac{\partial q_{it}}{\partial s_{it}} = \frac{1}{h'(q_{it})} \frac{\partial (\theta_{it}/b_i)}{\partial s_{it}} = \frac{1}{b_i h'(q_{it})} \frac{\partial \theta_{it}}{\partial s_{it}},$$

Similarly, differentiate (8) with respect to b_i, use the Inverse Function Theorem, and use (16) to substitute for $\partial \theta_{it}/\partial b_i$ in this expression. This gives (10).

A.2 Proof of Proposition 3

Under approximation (13), the left-hand side of (12) is then given by:

$$\begin{aligned}
& h_1 + 2h_2 \log(q_{it+1}) - \left(h_1 \log(q_{it+1}) + h_2 \log(q_{it+1})^2\right) \left(1 + \frac{\frac{2h_2 - (h_1 + 2h_2 \log(q_{it+1}))}{q_{it+1}}}{\frac{h_1 + 2h_2 \log(q_{it+1})}{q_{it+1}}}\right) \\
= \; & h_1 + 2h_2 \log(q_{it+1}) - \left(h_1 \log(q_{it+1}) + h_2 \log(q_{it+1})^2\right) \frac{2h_2}{h_1 + 2h_2 \log(q_{it+1})} \\
= \; & \frac{(h_1 + 2h_2 \log(q_{it+1}))^2 - 2\left(h_1 \log(q_{it+1}) + h_2 \log(q_{it+1})^2\right) h_2}{h_1 + 2h_2 \log(q_{it+1})} \\
= \; & \frac{(h_1 + h_2 \log(q_{it+1}))^2 + h_2^2 \log(q_{it+1})^2}{h_1 + 2h_2 \log(q_{it+1})}.
\end{aligned}$$

The numerator of this expression is the sum of two squares, so it is positive. Note that the denominator can be rewritten as the product of q_{it+1} and $h'(q_{it+1})$ under the given approximation. Recall that $h(q)$ is a strictly increasing function of q, reflecting the fact that all mutual funds must face decreasing returns to scale in equilibrium. Requiring that, under the approximation, $h'(q_{it+1}) > 0$ is also ensured, this means that the denominator is positive as well. It then follows immediately that condition (12) holds, which completes the proof.

B Determinants of Fund-Level Returns to Scale

In this appendix, we investigate which fund characteristics are correlated with the observed heterogeneity in returns to scale. We explore a number of characteristics that seem relevant a priori (also from the previous literature) for heterogeneity in returns to scale: the number of managers, volatility, expense ratios, marketing expenses, an international exposure indicator, turnover, and log fund size. In analyzing the dependence of returns to scale on fund characteristics, we control for the contributions of fund flows as well as the loadings on the four Fama-French-Carhart (FFC) factors to capture fund style and risk.[31]

The first characteristic, $NMgr$, is the number of managers managing the fund. About 59% of our funds are multi-manager funds. The second characteristic, $Std(Alpha)$, is the standard deviation of a fund's alphas over the prior 1 year. The next two characteristics we examine are the fund's expense ratios and marketing expenses. The fifth characteristic, $1(IntExp)$, is an indicator for funds with a high degree of international exposure, defined as follows. We test the null hypothesis that the coefficients on three Vanguard international index funds are 0.[32] For any given fund, the international exposure dummy is equal to one if we reject the null hypothesis at the 5% confidence level. Although we focus our attention on domestic funds, about 24% of them are highly exposed to international shocks. The sixth

[31] We estimate these risk exposures by regressing the fund's return on the four FFC factors over the prior sixty months.

[32] Recall that we use a set of eleven Vanguard index funds to calculate the Vanguard benchmark. Three of these index funds are international: European Stock Index, Pacific Stock Index, and Small-Cap Value Index.

characteristic is the fund's average annual turnover (from CRSP).[33] Median turnover is 64% per year. We also examine whether the fund's log real AUM matters for its DRS technology.

We examine how these characteristics affect the impact of a fund's scale on its performance by running panel regressions of fund i's DRS parameter using only its observations prior to month t, $\widehat{b}_{it}$, on the fund's characteristics at the end of the previous month. Table 9 shows the estimation results.[34] Panel A reports the results using the CAPM as the benchmark; Panel B uses Vanguard index funds as the benchmark.

In both panels, we find significant relations between $\widehat{b}$ and three characteristics: the number of managers, volatility, and expense ratios (see the first three columns of Table 9). On the other hand, we find a statistically insignificant relation between returns to scale and marketing expenses (column 4) of mixed signs. While the relation between returns to scale and international exposure (column 5) is consistently negative, it is insignificant. Finally, we find that the slope on turnover is positive (column 6) and the slope on fund size is negative (column 7). These results are statistically significant for the CAPM, but they are insignificant using the Vanguard benchmark.

When all seven fund characteristics are added at the same time, the estimated slopes on the number of managers, volatility, and expense ratios are robust, indicating steeper decreasing returns to scale for sole-manager funds, higher-volatility funds, and funds charging higher expense ratios. Marketing expenses now enter with a consistently significantly negative slope, indicating that decreasing returns to scale are less pronounced for funds with higher marketing expenses. The relation between returns to scale and international exposure remains statistically insignificantly negative. Finally, the slopes on turnover and on fund size remain positive and negative, respectively, but they are now insignificant regardless of how one defines the benchmark. Therefore, we focus on the results when the four statistically significant fund characteristics are added at the same time (see column 8 of Table 9).

[33] We winsorize turnover at the 1st and 99th percentiles.

[34] Standard errors of these regressions are two-way clustered by fund and time.

While we leave the task of deriving these relations between fund characteristics and diseconomies of scale in an equilibrium model for future research, these results are consistent with the following interpretations. The division of labor within a fund might alleviate the negative impact of size on performance, so it is the fund's assets under management on a per-manager basis that matters for capturing decreasing returns to scale. If so, a multi-manager fund would be able to deploy capital more easily and, consequently, exhibit milder decreasing returns to scale. Pástor, Stambaugh, and Taylor (2015) offer a narrative for why higher-volatility funds might also exhibit steeper decreasing returns to scale, while steeper decreasing returns to scale for funds charging higher expense ratios are consistent with the model of Stambaugh (2020). Finally, gradual decreasing returns to scale for funds with high-marketing expenses are consistent with funds marketing to attract more flows only if they can manage the performance erosion associated with growing fund size.

B.1 DRS-FSP Relation Using the Characteristic Component of DRS

We have estimated fund-specific b_i parameters based on a rolling estimation window. As noted earlier, estimating b_i fund by fund leads to imprecise estimates especially for funds with short track records. Instead of using the coefficient estimates $\widehat{b}_i$ as before, we use the estimates from column 8 of Table 9 to obtain an economically interpretable component of $\widehat{b}_i$ based on fund characteristics. This implementation choice assumes that all the funds with the same fund characteristics share the same b value. While ignoring variation might potentially lead to inaccuracy in quantifying fund-specific b, this method actually seems to increase the accuracy of the b_i estimate by dramatically reducing estimation errors. While about 28% of the funds in our sample end up with negative $\widehat{b}_i$, 2% of their predicted values based on fund characteristics, denoted by $\widehat{b}_i^{Char}$, are negative. These results seem sensible since, theoretically, all mutual funds must face decreasing returns to scale in equilibrium.

As another way to assess the robustness of our results regarding the effect of returns to scale on capital flows, we replace $\widehat{b}_i$ by $\widehat{b}_i^{Char}$ and rerun the regressions in Table 2, whose results are tabulated in Table 10. When we rerun our analysis in Table 2 with characteristic-based DRS, we obtain similar and even stronger results indicating that steeper decreasing returns to scale attenuates flow sensitivity. Table 10 shows that $\widehat{b}_i^{Char}$ has significantly negative slopes throughout, but the coefficients' estimated values become substantially more negative than in Table 2: the estimated coefficients based on the CAPM are more than 6 times larger (compare the first two columns of Tables 2 and 10).

To summarize, when we conduct the analysis using cleaner measures of decreasing returns to scale, our conclusions on the effects of decreasing returns to scale on capital allocation only become stronger. These estimates of the DRS-FSP relation are comparable in magnitude to those implied by the classical measurement error assumption and the IV approach.

References

[1] AVRAMOV, D., S. CHENG, AND A. HAMEED (2016): "Time-Varying Liquidity and Momentum Profits," *Journal of Financial and Quantitative Analysis*, 51(6), 1897–1923.

[2] BARBER, B. M., X. HUANG, AND T. ODEAN (2016): "Which Factors Matter to Investors? Evidence from Mutual Fund Flows," *Review of Financial Studies*, 29(10), 2600–2642.

[3] BARRAS, L., P. GAGLIARDINI, AND O. SCAILLET (2021): "Skill, Scale, and Value Creation in the Mutual Fund Industry," *Journal of Finance*, Forthcoming.

[4] BERK, J. B., AND R. C. GREEN (2004): "Mutual Fund Flows and Performance in Rational Markets," *Journal of Political Economy*, 112(6), 1269–1295.

[5] BERK, J. B., AND I. TONKS (2007): "Return Persistence and Fund Flows in the Worst Performing Mutual Funds," Working Paper, 13042, NBER.

[6] BERK, J. B., AND J. H. VAN BINSBERGEN (2015): "Measuring Skill in the Mutual Fund Industry," *Journal of Financial Economics*, 118(1), 1–20.

[7] BERK, J. B., AND J. H. VAN BINSBERGEN (2016): "Assessing Asset Pricing Models using Revealed Preference," *Journal of Financial Economics*, 119(1), 1–23.

[8] CARHART, M. M. (1997): "On Persistence in Mutual Fund Performance," *Journal of Finance*, 52(1), 57–82.

[9] CHEN, J., H. HONG, M. HONG, AND J. D. KUBIK (2004): "Does Fund Size Erode Mutual Fund Performance? The Role of Liquidity and Organization," *American Economic Review*, 94(5), 1276–1302.

[10] CHEVALIER, J., AND G. ELLISON (1997): "Risk Taking by Mutual Funds as a Response to Incentives," *Journal of Political Economy*, 105(6), 1167–1200.

[11] DEL GUERCIO, D., AND J. REUTER (2013): "Mutual Fund Performance and the Incentive to Generate Alpha," *Journal of Finance*, 69(4), 1673–1704.

[12] FAMA, E. F., AND K. R. FRENCH (2010): "Luck versus Skill in the Cross-Section of Mutual Fund Returns," *Journal of Finance*, 65(5), 1915–1947.

[13] FERREIRA, M. A., A. KESWANI, A. F. MIGUEL, AND S. B. RAMOS (2012): "The Flow-Performance Relationship Around the World," *Journal of Banking & Finance*, 36(6), 1759–1780.

[14] FERREIRA, M. A., A. KESWANI, A. F. MIGUEL, AND S. B. RAMOS (2013): "The Determinants of Mutual Fund Performance: A Cross-Country Study," *Review of Finance*, 17(2), 483–525.

[15] FRANZONI, F., AND M. C. SCHMALZ (2017): "Fund Flows and Market States," *Review of Financial Studies*, 30(8), 2621–2673.

[16] GRUBER, M. J. (1996): "Another Puzzle: The Growth in Actively Managed Mutual Funds," *Journal of Finance*, 51(3), 783–810.

[17] HUANG, J., K. D. WEI, AND H. YAN (2007): "Participation Costs and the Sensitivity of Fund Flows to Past Performance," *Journal of Finance*, 62(3), 1273–1311.

[18] HUANG, J., K. D. WEI, AND H. YAN (2012): "Investor Learning and Mutual Fund Flows," Working Paper, University of Texas at Austin.

[19] JENSEN, M. C. (1968): "The Performance of Mutual Funds in the Period 1945–1964," *Journal of Finance*, 23(2), 389–416.

[20] KIM, J. H. (2017): "Why Has Active Asset Management Grown?" Working Paper, Emory University.

[21] MALKIEL, B. G. (1995): "Returns from Investing in Equity Mutual Funds 1971 to 1991," *Journal of Finance*, 50(2), 549–572.

[22] MILGROM, P. R. (1981): "Good News and Bad News: Representation Theorems and Applications," *Bell Journal of Economics*, 12(2), 380–391.

[23] PÁSTOR, L'., AND R. F. STAMBAUGH (2012): "On the Size of the Active Management Industry," *Journal of Political Economy*, 120(4), 740–781.

[24] PÁSTOR, L'., R. F. STAMBAUGH, AND L. A. TAYLOR (2015): "Scale and Skill in Active Management," *Journal of Financial Economics*, 116(1), 23–45.

[25] SIRRI, E. R., AND P. TUFANO (1998): "Costly Search and Mutual Fund Flows," *Journal of Finance*, 53(5), 1589–1622.

[26] STAMBAUGH, R. F. (2020): "Skill and Profit in Active Management," Working Paper, 26027, NBER.

[27] ZHU, M. (2018): "Informative Fund Size, Managerial Skill, and Investor Rationality," *Journal of Financial Economics*, 130(1), 114–134.

Table 1: Summary Statistics

This table shows summary statistics for our sample of active equity mutual funds from 1985–2014. The unit of observation is the fund/month. All returns are in units of fraction per month. Net return is the return received by investors. Net alpha equals net return minus the return on benchmark portfolio, calculated using the CAPM or using a set of Vanguard index funds. Fund size is the fund's total AUM aggregated across share classes, adjusted by inflation. The numbers are reported in Y2000 $ millions per month. The first version of Flows is the logarithmic change in real AUM (i.e., the percentage change in fund size); the second version of Flows is the monthly change in the fund's net assets not attributable to its return gains or losses. $\widehat{\beta}^{\cdot}_{it}$ are fund i's estimated risk exposures from the regression of the fund's return on the four FFC factors over the sixty months prior to month t. Past inflow (outflow) is equal to the fund's fractional net flow over the prior 1 year if it were positive (negative) and zero otherwise. # of managers is the number of managers managing the fund in a given month. Volatility is the standard deviation of a fund's alphas, calculated over the prior 1 year. All expenses are in units of fraction per year. Marketing expenses is a fund's total fee ratio, defined as the annual expense ratio plus one-seventh of the up-front load fees. Fund age is the number of years since the fund's first offer date (from CRSP or, if missing, from Morningstar). $\widehat{b}_{it}$ is the fund's decreasing returns to scale estimated as of the previous month-end. $\widehat{FSP}_{it}$ is the fund's flow sensitivity to performance going forward.

Panel A: Fund-Level Variables

	# of obs.	Mean	Stdev.	Percentiles 25%	50%	75%
Net return	412,943	0.0077	0.0497	−0.0194	0.0123	0.0385
Net alpha (CAPM Risk Adj.)	346,885	0.0001	0.0209	−0.0105	−0.0002	0.0103
Net alpha (Vanguard BM)	408,313	−0.0001	0.0153	−0.0083	−0.0002	0.0079
Fund size (in 2000 $millions)	410,052	1012	4062	45	165	622
Flows (v.1)	410,048	0.0078	0.0734	−0.0274	0.0077	0.0439
Flows (v.2)	410,048	0.0049	0.0520	−0.0139	−0.0020	0.0142
$\widehat{\beta}^{mkt}_{it}$	346,885	0.9602	0.1462	0.8815	0.9661	1.0444
$\widehat{\beta}^{smb}_{it}$	346,885	0.2080	0.3377	−0.0660	0.1229	0.4646
$\widehat{\beta}^{hml}_{it}$	346,885	−0.0148	0.3158	−0.2310	−0.0096	0.1945
$\widehat{\beta}^{mom}_{it}$	346,885	0.0130	0.1384	−0.0625	0.0051	0.0771
Past inflow	375,999	0.2505	0.6547	0	0	0.1840
Past outflow	375,999	−0.0940	0.1323	−0.1482	−0.0313	0
# of managers	397,741	2.38	2.13	1	2	3
Volatility (CAPM Risk Adj.)	317,350	0.0188	0.0115	0.0105	0.0157	0.0238
Volatility (Vanguard BM)	377,572	0.0140	0.0081	0.0085	0.0121	0.0174
Expense ratio	410,068	0.0124	0.0043	0.0096	0.0119	0.0148
Marketing expenses	245,941	0.0183	0.0061	0.0148	0.0196	0.0218
Fund age	412,249	13.20	13.27	4.50	9.08	16.50

	Panel B: Estimated DRS and FSP					
				Percentiles		
	# of obs.	Mean	Stdev.	25%	50%	75%
$\widehat{b}$ (CAPM Risk Adj.)	$264,879$	0.0082	0.0166	-0.0004	0.0051	0.0136
$\widehat{b}$ (Vanguard BM)	$315,066$	0.0045	0.0112	-0.0008	0.0028	0.0082
$\widehat{FSP}$ (CAPM Risk Adj., v.1)	$283,339$	0.0559	0.2715	-0.0822	0.0426	0.1873
$\widehat{FSP}$ (Vanguard BM, v.1)	$312,487$	0.0971	0.3852	-0.0942	0.0897	0.2879
$\widehat{FSP}$ (CAPM Risk Adj., v.2)	$283,339$	0.1048	0.1893	0.0143	0.0764	0.1697
$\widehat{FSP}$ (Vanguard BM, v.2)	$312,487$	0.1443	0.2848	0.0142	0.1057	0.2441

Table 2: Relation Between DRS and Flow Sensitivity to Performance

The dependent variable in each regression model is $\widehat{FSP}_{it}$, the fund's flow sensitivity to performance going forward, where flow is defined as the % change in fund size in Panel A and as the % change in new assets in Panel B. $\widehat{b}_{it}$ is the fund's decreasing returns to scale estimated as of the previous month-end. The first two columns in each panel use the CAPM as the benchmark, while the last two columns use Vanguard index funds as the benchmark. In the odd columns, we only include month and fund fixed effects; in the even columns, we add proxies for participation costs, as well as performance volatility and fund age. Standard errors, two-way clustered by fund and by month, are in parentheses. We report the EIV-adjusted coefficients in the last row of each panel.

	Panel A:	**Flow as % Change in Fund Size**		
		$\widehat{FSP}_{it}$		
$\widehat{b}_{it}$	-0.826	-1.172	-1.289	-1.486
	(0.148)	(0.184)	(0.264)	(0.310)
Fund FE & Month FE	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
Observations	182,676	114,623	221,743	136,365
Performance relative to	CAPM	CAPM	Vanguard BM	Vanguard BM
EIV Adj. Coefficient	-4.866	-7.802	-9.321	-14.146

	Panel B:	**Flow as % Change in New Assets**		
		$\widehat{FSP}_{it}$		
$\widehat{b}_{it}$	-0.279	-0.428	-0.375	-0.576
	(0.090)	(0.124)	(0.167)	(0.179)
Fund FE & Month FE	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
Observations	182,676	114,623	221,743	136,365
Performance relative to	CAPM	CAPM	Vanguard BM	Vanguard BM
EIV Adj. Coefficient	-1.642	-2.850	-2.712	-5.488

Table 3: Relation Between DRS and FSP Based on Their Percentile Ranks
This table is the same as Table 2 but, instead of using the coefficient estimates $\widehat{b}_{it}$ and $\widehat{FSP}_{it}$, uses their percentile ranks in each month.

	Panel A:	**Flow as % Change in Fund Size**		
		Pctl. rank based on $\widehat{FSP}_{it}$		
Pctl. rank based on $\widehat{b}_{it}$	-0.1023 (0.0116)	-0.1148 (0.0134)	-0.0836 (0.0096)	-0.0906 (0.0117)
Fund FE	Yes	Yes	Yes	Yes
Month FE	No	No	No	No
Controls	No	Yes	No	Yes
Observations	182,676	114,623	221,743	136,365
Performance relative to	CAPM	CAPM	Vanguard BM	Vanguard BM

	Panel B:	**Flow as % Change in New Assets**		
		Pctl. rank based on $\widehat{FSP}_{it}$		
Pctl. rank based on $\widehat{b}_{it}$	-0.0765 (0.0099)	-0.0826 (0.0122)	-0.0572 (0.0086)	-0.0657 (0.0108)
Fund FE	Yes	Yes	Yes	Yes
Month FE	No	No	No	No
Controls	No	Yes	No	Yes
Observations	182,676	114,623	221,743	136,365
Performance relative to	CAPM	CAPM	Vanguard BM	Vanguard BM

Table 4: DRS-FSP Relation Under the Instrumental Variables Approach

The dependent variable in each regression model is $\widehat{FSP}_{it}$, the fund's flow sensitivity to performance going forward, where flow is defined as the % change in fund size in Panel A and as the % change in new assets in Panel B. $\widehat{b}_{it}$ is the fund's decreasing returns to scale estimated as of the previous month-end. Column 1 in each panel uses the CAPM as the benchmark, while column 2 uses Vanguard index funds as the benchmark. We instrument for $\widehat{b}_{it}$ using (i) $\widehat{\beta}_{it}^{mom}$, the fund's momentum exposure and (ii) "Past inflow," the fund's % inflow over the prior 1 year. Specifically, we first regress $\widehat{b}_{it}$ on these two fund-level variables, and then we regress $\widehat{FSP}_{it}$ on the fitted values from the first-stage regression. Both regressions include fund and month fixed effects, as well as the controls in Table 2. Standard errors, two-way clustered by fund and by month, are in parentheses.

	Panel A: Flow as % Change in Fund Size	
	First-Stage Regressions	
	$\widehat{b}_{it}$	
% Inflow Over the Past 1 Year	-0.00328	-0.00143
	(0.000352)	(0.000172)
Momentum Exposure	-0.0157	-0.00742
	(0.00291)	(0.00215)
F-stat of excluded instruments	56.7	42.5
	Second-Stage Regressions	
	$\widehat{FSP}_{it}$	
Predicted $\widehat{b}_{it}$	-7.249	-8.875
	(1.638)	(3.994)
p-val of overidentification test	0.131	0.961
Fund FE	Yes	Yes
Month FE	Yes	Yes
Controls	Yes	Yes
Observations	114,612	136,354
Performance relative to	CAPM	Vanguard BM
Compare to Table 2	Panel A, Column 2	Panel A, Column 4

	Panel B: Flow as % Change in New Assets	
	First-Stage Regressions	
	$\widehat{b}_{it}$	
% Inflow Over the Past 1 Year	-0.00328	-0.00143
	(0.000352)	(0.000172)
Momentum Exposure	-0.0157	-0.00742
	(0.00291)	(0.00215)
F-stat of excluded instruments	56.7	42.5
	Second-Stage Regressions	
	$\widehat{FSP}_{it}$	
Predicted $\widehat{b}_{it}$	-6.100	-13.01
	(1.271)	(3.053)
p-val of overidentification test	0.919	0.972
Fund FE	Yes	Yes
Month FE	Yes	Yes
Controls	Yes	Yes
Observations	114,612	136,354
Performance relative to	CAPM	Vanguard BM
Compare to Table 2	Panel B, Column 2	Panel B, Column 4

Table 5: Parameter Values Used for Simulations

Panel A summarizes the model parameters that we set directly and their parameter values. The empirical distribution of b is approximated by a generalized Pareto (GP) distribution, from which we draw b randomly. In that case, assuming that θ_0 is independent of b gives rise to distributions of fund size considerably more disperse than in our actual sample. Therefore, we model the prior mean as a linear function of b, $\theta_0(b) = \theta_{0a} + \theta_{0b} b$. Our approach is to fit the parameters θ_{0a} and θ_{0b} by essentially matching the simulated mean and standard deviation of log fund size to their empirical counterparts. We set the prior uncertainty (σ_0) to match the average flow-performance relation in the data. Panel B shows the value of σ_0 that resulted from this process. It also contains the values of the parameters governing the GP distribution of DRS and those of the parameters governing the prior mean that we use in our simulation analysis. The last three columns of Panel B report all the moments that we target in our calibration, as well as their values in both the actual and simulated data.

Panel A: Parameters Set Directly

Variable	Symbol	Value
Return standard deviation	σ	5.77%
Exit mean	$\overline{\theta}$	0%

Panel B: Calibrated Parameters

Variable	Symbol	Value	Target	Emp. Value	Sim. Value
$f(b_i \mid \mu_b, \sigma_b, \xi_b) = \left(\frac{1}{\sigma_b}\right)\left(1 + \xi_b \frac{(b_i - \mu_b)}{\sigma_b}\right)^{-1-\frac{1}{\xi_b}}$					
Threshold (location)	μ_b	0.0000	Mean of b_i	0.0092	0.0105
Scale	σ_b	0.0137	Std dev of b_i	0.0100	0.0083
Tail index (shape)	ξ_b	-0.3011	Skewness of b_i	0.9268	0.9296
$\theta_0(b_i) = \theta_{0a} + \theta_{0b} b_i$					
Prior mean for CRS funds	θ_{0a}	0.06%	Mean of $\log(q_{it})$	5.12	5.12
Prior mean slope on DRS	θ_{0b}	4.821	Std dev of $\log(q_{it})$	1.86	1.84
Prior uncertainty	σ_0	0.06%	Mean of $\widehat{\gamma}_i$	0.543	0.543

Table 6: Simulated DRS-FSP Relation

We construct 2,500 samples of simulated panel data for 10,000 funds over 100 months. We simulate a given sample by first drawing each fund's DRS b_i randomly from a GP distribution consistent with the distribution of fund-specific b estimates, while drawing the fund's skill a_i from a normal distribution with mean $\theta_0(b_i)$ and standard deviation σ_0. Next, we draw the random values of ε_{it}, building up the panel data of r_{it} and q_{it}. For each fund in the sample, we run the following regression using data for just that fund to estimate its FSP:

$$\log\left(q_{it}/q_{it-1}\right) = c_i + \gamma_i r_{it} + v_{it}.$$

We then estimate the DRS-FSP relation in each of the simulated samples. Panel A shows summary statistics of these estimates across simulated samples from the calibrated model. To quantitatively assess the role of heterogeneity in returns to scale in capital allocation, we construct a counterfactual by assuming investors learn about skill based on distorted beliefs that $b_i = 0.0105$ for all funds. Then, updating investors' beliefs with the history of its returns under the counterfactual assumption, we compute the fund's size under the counterfactual, q_{it}^C, for every i and t. For each sample, we calculate the R^2 from a regression of $\log(q_{it})$ on $\log(q_{it}^C)$ to check the goodness of fit by the counterfactual. We report the results from counterfactual simulations in Panel B. The first two rows show summary statistics of the coefficient estimates from the regression of $\log(q_{it})$ on $\log\left(q_{it}^C\right)$ across simulated samples; the last row shows summary statistics of the R^2 from this regression across simulated samples.

Panel A: Simulated DRS-FSP Relation

$\widehat{\gamma}_i = k + \lambda b_i + u_i$

			Percentiles						
	Mean	Stdev.	1%	5%	10%	50%	90%	95%	99%
$\widehat{\lambda}$	−6.79	0.489	−8.00	−7.62	−7.43	−6.77	−6.16	−6.02	−5.70
Data	−7.249								

Panel B: Capital Allocation Explained by Counterfactual

$\log(q_{it}) = \pi_0 + \pi_1 \log\left(q_{it}^C\right) + v_{it}$

			Percentiles						
	Mean	Stdev.	1%	5%	10%	50%	90%	95%	99%
$\widehat{\pi}_0$	−622	21.9	−661	−653	−647	−625	−593	−582	−561
$\widehat{\pi}_1$	129	4.49	116	120	123	129	134	135	137
R^2	0.656	0.023	0.593	0.615	0.626	0.659	0.682	0.688	0.697

Table 7: Relation Between DRS and Fund Size

We first form quintile groups sorted on fund fixed effects estimated as of month $t-1$, $\widehat{a}_{it}$. Then, within each $\widehat{a}_{it}$ quintile, we sort funds into five groups based on $\widehat{b}_{it}$. Both $\widehat{a}_{it}$ and $\widehat{b}_{it}$ are computed from estimating (14) using 60 months of the data for fund i prior to time t. We conduct double sorts of funds belonging to the same Morningstar category and to the same age category. After forming the 5×5 $\widehat{a}_{it}$ and $\widehat{b}_{it}$ groups, we average fund sizes, as measured by log real AUM in month t, of each $\widehat{b}_{it}$ quintile over the five $\widehat{a}_{it}$ groups. Panel A reports average fund sizes of the 25 $\widehat{a}_{it} \times \widehat{b}_{it}$ groups using the CAPM as the benchmark; Panel B repeats the same exercise, except we use Vanguard index funds as the benchmark. The column labeled "Average" reports the average month-end fund sizes of the $\widehat{b}_{it}$ quintiles, controlling for $\widehat{a}_{it}$, fund style and fund age. The row labeled "High-low" reports the differences in average sizes between the first and fifth quintile $\widehat{b}_{it}$ groups within each $\widehat{a}_{it}$ quintile. We report standard errors of these differences between quintile 5 (high $\widehat{b}_{it}$) and quintile 1 (low $\widehat{b}_{it}$) using 60 Newey-West lags.

Panel A: CAPM Risk Measure

	$\widehat{a}_{it}$ Quintiles					
Group	1 Low	2	3	4	5	Average
1 Low $\widehat{b}_{it}$	5.454	5.684	6.889	7.379	7.462	6.547
2	5.677	5.977	6.403	6.593	6.644	6.248
3	5.808	5.931	5.955	6.080	6.232	5.996
4	5.939	5.717	5.398	5.543	6.022	5.725
5 High $\widehat{b}_{it}$	5.637	4.755	4.326	4.466	5.527	4.947
High-low	0.183	−0.929	−2.563	−2.913	−1.935	−1.600
	(0.134)	(0.167)	(0.161)	(0.179)	(0.132)	(0.080)

Panel B: Vanguard Benchmark

	$\widehat{a}_{it}$ Quintiles					
Group	1 Low	2	3	4	5	Average
1 Low $\widehat{b}_{it}$	5.381	5.369	6.721	7.186	7.232	6.353
2	5.482	5.727	6.248	6.400	6.370	6.032
3	5.554	5.786	5.806	5.863	5.939	5.782
4	5.748	5.610	5.348	5.346	5.650	5.542
5 High $\widehat{b}_{it}$	5.575	4.683	4.248	4.258	5.205	4.797
High-low	0.194	−0.685	−2.473	−2.928	−2.027	−1.555
	(0.132)	(0.151)	(0.159)	(0.183)	(0.146)	(0.070)

Table 8: Relation Between Optimal Size and Fund Size

The dependent variable in each regression model is the fund's log real AUM in \$ millions (base year is 2000). Our measure of a fund's optimal size, $\log(\widehat{q}_i^*)$, is equal to as the average ratio of the fund's net alpha (adjusted for returns to scale) to the fund's individual DRS parameter; the alternative measure of a fund's optimal size, $\log\left(\widehat{q}_i^{*RD2}\right)$, is calculated assuming fund scale has the same effect on performance for all funds. We assign funds to one of three samples based on fund age: $[0,5]$, $(5,10]$, and > 10 years. In Panel A, we report the results using the CAPM as the benchmark; in Panel B, we use Vanguard index funds as the benchmark. Columns 1–3 show the results from running panel regressions of a fund's log real AUM on the fund's log optimal fund size estimate in each age sample. Columns 4 through 6 show the results from running the multiple regression of $\log(q_{it})$ on two components of $\log(\widehat{q}_i^*)$—that explained by $\log\left(\widehat{q}_i^{*RD2}\right)$ and that unexplained by $\log\left(\widehat{q}_i^{*RD2}\right)$—in all three age-sorted samples. The double clustered (by fund and time) standard errors are in parentheses.

	Panel A: CAPM Risk Measure					
	Dependent Variable: Log Real AUM					
$\log(\widehat{q}_i^*)$	0.334	0.611	0.882			
	(0.019)	(0.022)	(0.022)			
explained by $\log\left(\widehat{q}_i^{*RD2}\right)$				0.459	0.741	0.940
				(0.014)	(0.012)	(0.008)
unexplained by $\log\left(\widehat{q}_i^{*RD2}\right)$				−0.376	−0.185	0.647
				(0.040)	(0.036)	(0.056)
R^2	0.20	0.55	0.81	0.34	0.67	0.83
Observations	64,752	105,347	196,364	64,752	105,347	196,364
Fund ages	$[0,5]$ yr.	$(5,10]$ yr.	> 10 yr.	$[0,5]$ yr.	$(5,10]$ yr.	> 10 yr.

	Panel B: Vanguard Benchmark					
	Dependent Variable: Log Real AUM					
$\log(\widehat{q}_i^*)$	0.161	0.389	0.659			
	(0.014)	(0.011)	(0.024)			
explained by $\log\left(\widehat{q}_i^{*RD2}\right)$				0.273	0.473	0.695
				(0.010)	(0.010)	(0.012)
unexplained by $\log\left(\widehat{q}_i^{*RD2}\right)$				−0.068	−0.051	0.382
				(0.019)	(0.026)	(0.047)
R^2	0.14	0.41	0.68	0.25	0.51	0.71
Observations	75,990	110,319	196,598	75,990	110,319	196,598
Fund ages	$[0,5]$ yr.	$(5,10]$ yr.	> 10 yr.	$[0,5]$ yr.	$(5,10]$ yr.	> 10 yr.

Table 9: Determinants of Decreasing Returns to Scale

The dependent variable in all regressions is $\widehat{b}_{it}$, the fund's adverse scale effect computed using only its data prior to the month. $NMgr$ is the number of managers managing the fund. $Std\,(Alpha)$ is the standard deviation of monthly outperformance for each fund over the prior 1 year. $ExpRatio$ and $MktgExp$ are the fund's expense ratios and marketing expenses, respectively. $1\,(IntExp)$ is an indicator for funds with a high degree of international exposure, defined as follows. We test the null hypothesis that the coefficients on three Vanguard international index funds are 0. For any given fund, the international exposure dummy is equal to one if we reject the null hypothesis at the 5% confidence level. $Turnover$ is the fund's average annual turnover (from CRSP). Finally, $\log(q_{it-1})$ is the log of the fund's lagged real AUM. All regressions control for fund flows ("Past inflow" and "Past outflow") as well as fund style and risk, as proxied by the loadings on the four Fama-French-Carhart (FFC) factors $(\widehat{\beta}^{\cdot}_{it})$. Panel A reports the results using the CAPM as the benchmark; Panel B uses Vanguard index funds as the benchmark. Standard errors, two-way clustered by fund and time, are in parentheses.

Panel A: CAPM Risk Measure

	Dependent Variable: $\widehat{b}_{it}$							
$NMgr$	−0.000436							−0.000326
	(0.000075)							(0.000101)
$Std\,(Alpha)$		0.2764						0.2542
		(0.0234)						(0.0273)
$ExpRatio$			0.2701					0.2564
			(0.0524)					(0.0787)
$MktgExp$				0.0041				−0.1391
				(0.0488)				(0.0533)
$1\,(IntExp)$					−0.000243			
					(0.000410)			
$Turnover$						0.000835		
						(0.000297)		
$\log(q_{it-1})$							−0.000234	
							(0.000127)	
Flows Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Style & Risk Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	247,977	252,328	252,328	154,976	252,328	247,187	252,328	151,996

Panel B: Vanguard Benchmark

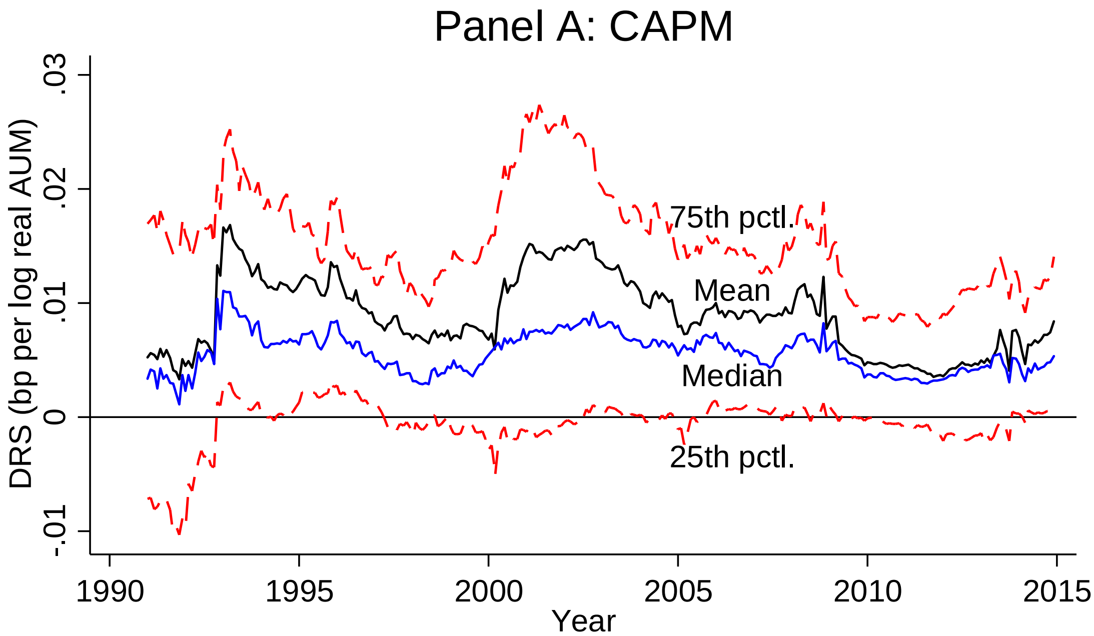
	Dependent Variable: $\widehat{b}_{it}$							
$NMgr$	−0.000267							−0.000252
	(0.000043)							(0.000068)
$Std\,(Alpha)$		0.1610						0.1625
		(0.0187)						(0.0233)
$ExpRatio$			0.1434					0.1315
			(0.0330)					(0.0510)
$MktgExp$				−0.0509				−0.1340
				(0.0367)				(0.0408)
$1\,(IntExp)$					−0.000002			
					(0.000266)			
$Turnover$						0.000109		
						(0.000175)		
$\log\left(q_{it-1}\right)$							−0.000059	
							(0.000077)	
Flows Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Style & Risk Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	294, 808	300, 843	300, 843	180, 891	300, 843	294, 582	300, 843	176, 745

Table 10: DRS-FSP Relation Using the Characteristic Component of DRS

This table is the same as Table 2 but, instead of using the coefficient estimate $\widehat{b}_{it}$ as before, uses its characteristic component estimated from the specification in column 8 of Table 9, denoted by $\widehat{b}_{it}^{Char}$.

	Panel A:	**Flow as % Change in Fund Size**		
		$\widehat{FSP}_{it}$		
$\widehat{b}_{it}^{Char}$	-5.020	-7.666	-3.646	-6.273
	(0.971)	(1.203)	(2.165)	(2.717)
Fund FE & Month FE	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
Observations	112,964	112,060	134,106	132,791
Performance relative to	CAPM	CAPM	Vanguard BM	Vanguard BM

	Panel B:	**Flow as % Change in New Assets**		
		$\widehat{FSP}_{it}$		
$\widehat{b}_{it}^{Char}$	-2.656	-3.984	-2.800	-6.920
	(0.663)	(0.826)	(1.396)	(1.691)
Fund FE & Month FE	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
Observations	112,964	112,060	134,106	132,791
Performance relative to	CAPM	CAPM	Vanguard BM	Vanguard BM



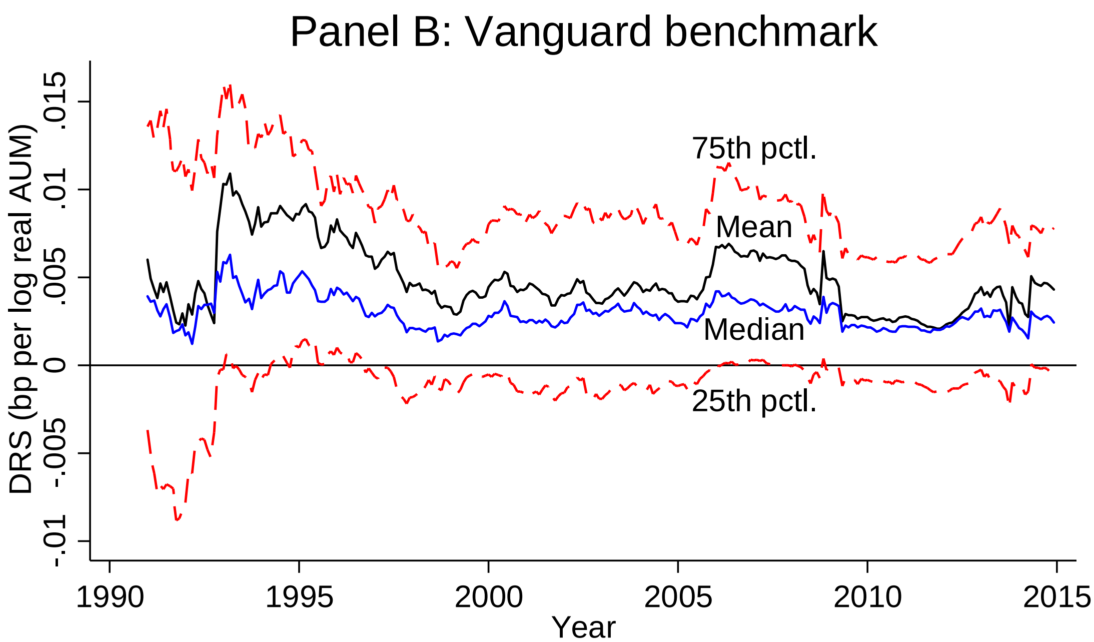


Figure 1: **Distribution of individual decreasing returns to scale (DRS) parameters over time:** The figure plots each month's mean and percentiles of estimated size effect on performance across all funds operating during that month. Panel A estimates DRS when we calculate outperformance relative to the CAPM. Panel B estimates DRS when we calculate outperformance relative to the Vanguard benchmark.

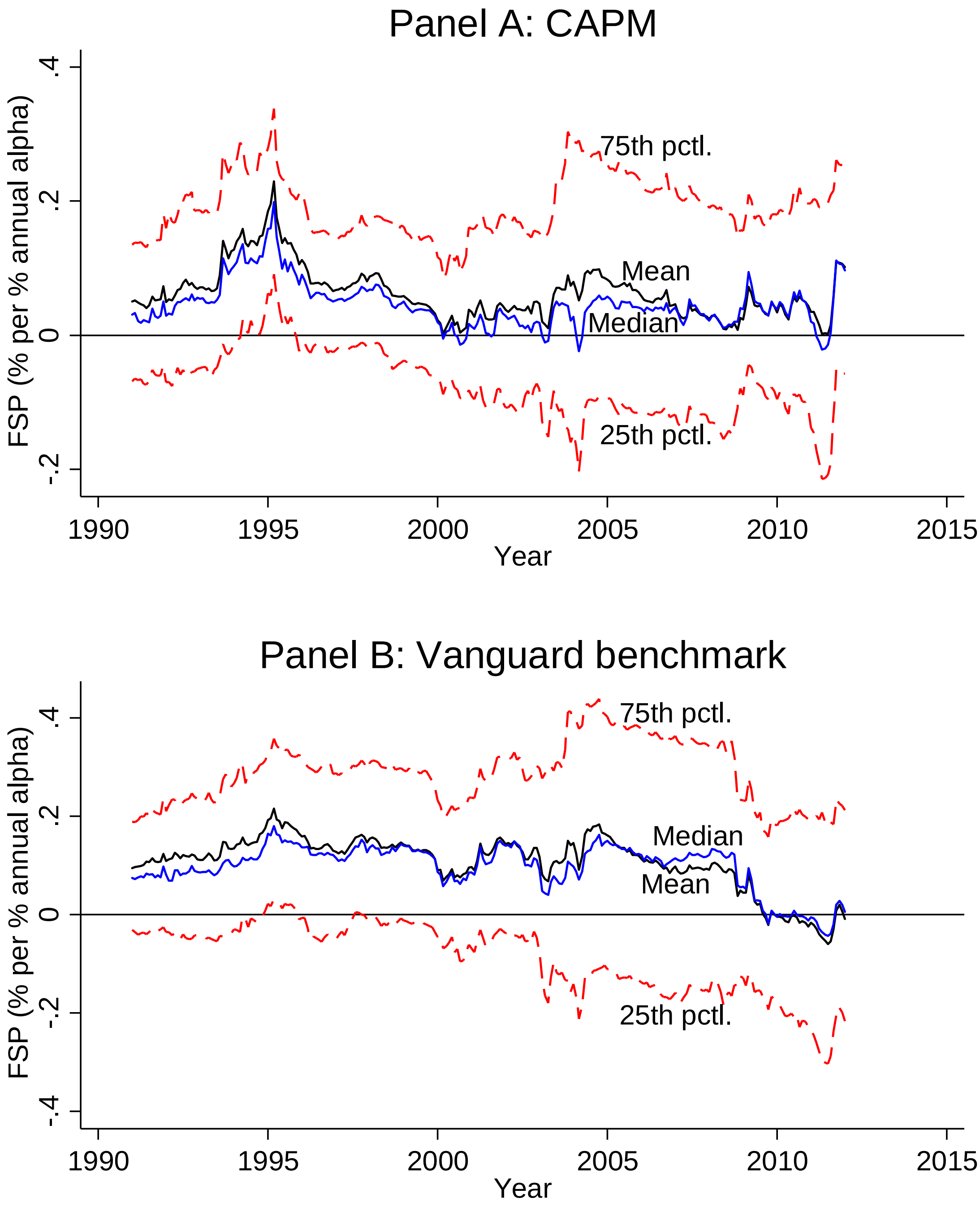
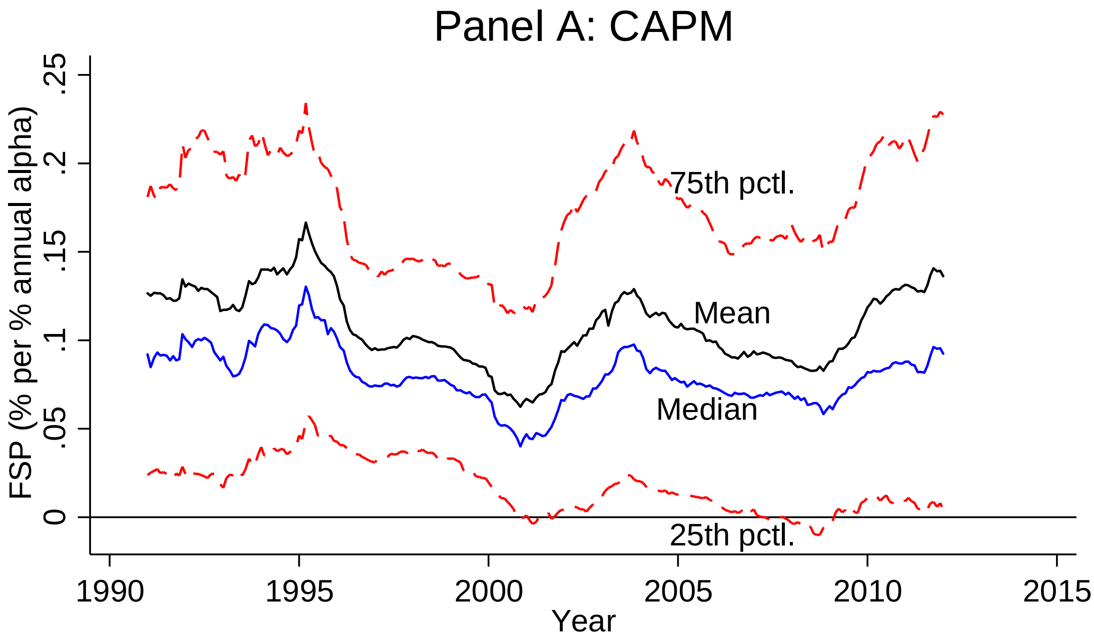


Figure 2: **Distribution of individual flow sensitivity to performance (FSP, v.1) over time:** The figure plots each month's mean and percentiles of estimated % change in real AUM due to performance across all funds operating during that month. Panel A estimates FSP when we calculate outperformance relative to the CAPM. Panel B estimates FSP when we calculate outperformance relative to the Vanguard benchmark.



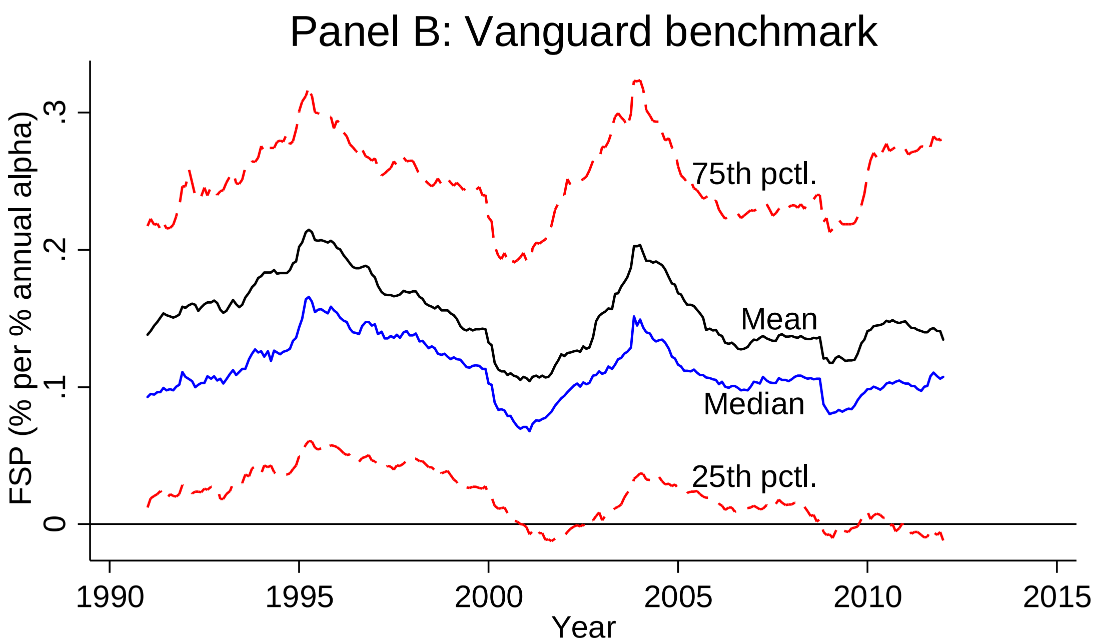


Figure 3: **Distribution of individual flow sensitivity to performance (FSP, v.2) over time:** The figure plots each month's mean and percentiles of estimated % change in new assets due to performance across all funds operating during that month. Panel A estimates FSP when we calculate outperformance relative to the CAPM. Panel B estimates FSP when we calculate outperformance relative to the Vanguard benchmark.