Informative Fund Size, Managerial Skill and Investor Rationality

Min Zhu *

Abstract

This paper considers the nature of returns to scale in active management following Pástor, Stambaugh and Taylor (2015) who fail to establish diseconomies of scale at the fund level. Using an enhanced empirical strategy, we find a significant negative impact of fund size on performance. This empirical evidence indicates that fund alpha and fund size are not independent entities. Consequently, skill, rather than being measured by the fund alpha, should be measured by the value that a fund extracts from capital markets. We also show that there is little support for the prevailing belief that as a group, mutual fund investors are naive.

Key words: Mutual funds; managerial skill; diseconomies of scale; investor rationality.
JEL classification: G11; G23; J24.

*Min Zhu is at the Business School of the Queensland University of Technology. I thank John Chen, Adam Clements, Phil Dybvig, Stan Hurn, John Polichronis, Steve Thiele, Thijs van der Heijden, Feng Zhao, participants at the FIRN conference, and seminar participants at the Queensland University of Technology, the Southwestern University of Finance and Economics and the Central University of Finance and Economics. I also thank Shuang Chen for superb research assistance.

1 Introduction

The traditional analytic framework of skill in active management builds on the assumption of constant returns to scale. That is to say, skillwise, a fund with $1 million of assets under management is indistinguishable from a fund with $1 billion of assets under management as long as they both achieve the same risk-adjusted return. Under this framework, skill and scale are independent, and fund size is regarded as uninformative and randomly paired with funds. If the nature of the returns to scale is not constant, fund size then is informative, and unobserved skill is reflected in two observable measures, return and size. Consequently, the traditional framework that studies managerial skill and ignores size fails to fully utilize the available information, which may have biased previous studies against finding differential ability in active management.

Despite the important theoretical and practical implications, the answer to the nature of returns to scale in active management is inconclusive. A number of studies, including Chen, Hong, Huang, and Kubik (2004) and Yan (2008), document a negative relationship between fund size and performance. These studies link diminishing returns to scale to liquidity constraints faced by funds. However, Elton, Gruber, and Blake (2012) report no relationship between size and performance which they attribute to the effect of the diseconomies of returns to scale being offset by the reduction in the expense ratio as a fund increases in size. Ferreira, Keswani, Miguel, and Ramos (2013) carry out an international study and show diseconomies of scale for US funds but not for non-US funds. These studies quantify scale effects based on the ordinary least squares (OLS) approach that directly regresses fund returns on lagged fund sizes. A concern with this approach is that the validity of the model is based on the assumption that fund size is uninformative and randomly distributed among funds, the very hypothesis to be tested.

Recognizing that managerial skill is an omitted variable that is correlated with fund size and performance, Pástor, Stambaugh, and Taylor (2015) utilize fund fixed effects to remove the omitted-variable bias. Further, they use an empirical strategy based on recursive demeaning procedure to examine the size–performance relationship. The procedure runs a panel regression of forward-demeaned returns on the forward-demeaned fund size, while instrumenting for the latter quantity with its backward-demeaned counterpart. Although Pástor, Stambaugh, and Taylor (2015) document that performance decreases as the size of the active mutual fund industry grows, they fail to reject the hypothesis of constant returns to scale at the individual fund level.

In this paper, we revisit the nature of returns to scale in active management following Pástor, Stambaugh, and Taylor (2015). We demonstrate that their empirical strategy suffers an

inherent misspecification resulting from a model restriction which is problematic for the fund size process. As demonstrated by extensive simulation studies, this misspecification increases estimation uncertainty and reduces power in hypothesis tests. In certain parameter combinations which are plausible in reality, it is better to use the biased OLS estimator rather than correcting the omitred-variable bias with Pástor, Stambaugh, and Taylor's (2015) estimator as the benefit of doing so is overshadowed by greatly increased estimating uncertainty. We modify Pástor, Stambaugh, and Taylor's (2015) estimator to make it more suitable for the fund size process. Using this enhanced estimator, we establish an adverse size effect on performance at the fund level with compelling statistical significance. Pástor, Stambaugh, and Taylor (2015) find an average fund-level decreasing returns to scale parameter of -0.220×10^{-6} on dollar fund size with a t-statistic of -0.62, whereas we find an average fund-level decreasing returns to scale parameter of -0.485×10^{-6} with a t-statistic of -2.03. Due to severe skewness in dollar fund size, the logarithm of fund size is often used in quantifying the nature of returns to scale. Applying our enhanced estimator, we report the average negative scale effect of -0.0026 on the logarithm of fund size with a t-statistic of -13.32. Using portfolios sorted on fund size, we further show a substantial amount of individual heterogeneity in decreasing returns to scale.

Decreasing returns to scale at the fund level imply that the fund alpha and the fund size are not two independent entities. Thus, neither provides a complete picture of managerial skill. In a neoclassical world with competitive markets and rational players, Berk and van Binsbergen (2015) prove that the fund alpha, neither net nor gross, measures managerial skill. They demonstrate that the only proper measure in a decreasing returns to scale world is the *value added*, which is the dollar value a fund extracts from capital markets. Using the realized value added, a fund's gross excess return over its benchmark multiplied by the assets under management (AUM), Berk and van Binsbergen (2015) report compelling evidence of long-lasting performance persistence in the active mutual fund industry.

The second main contribution of this paper is to study the relation between the optimal value a fund can create and the value the fund actually delivered. Assuming a loglinear functional form of decreasing returns to scale, we develop an empirical strategy to quantify the maximum value the fund can extract from capital markets. The value a fund actually delivered is measured by Berk and van Binsbergen's (2015) realized value added. In Berk and Green's (2004) and Berk and van Binsbergen's (2015) neoclassical world, both investors and fund managers act rationally to harvest the optimal value. Investors' fund flows ensure all managers have enough capital to extract the maximum value from capital markets, and fund managers index the excess amount if the investors provided surplus capital. In comparing the ideal with the actual, however, we find that the value a typical fund actually added is far short of the

optimum. Overall, 17% of the funds in our sample fail to have enough capital from investors to extract the optimal value. For the funds that have surplus capital, rather than index the excess money, the managers tend to actively manage more than they can handle, running the risk of destroying value. In our sample, 57% of the funds are excessively overfunded in the sense that their expected gross alphas are negative if the funds choose to actively manage all capital provided. Almost all of these excessively overfunded funds fail to index properly, and nearly 90% end up destroying value. It is necessary to bear in mind here that these inferences on whether funds have enough or too much capital depend on the assumption that the relation between size and gross alpha is loglinear. These inferences might change if different functional forms are assumed.

In the third contribution of this paper, we gauge investor rationality by studying the net alpha and its relation to size. Many studies in the mutual fund literature have erroneously used the net alpha to measure managerial skill. As pointed out by Berk and van Binsbergen (2015) and Berk and van Binsbergen (2017), rather than teach us something about managerial skill, net alphas teach us something about the rationality of investors and the competitiveness of financial markets. If the gross alpha is a decreasing function of fund size, the net alpha is also a decreasing function of fund size. This is because the net alpha is the gross alpha minus the fees, and fund fees are quite stable in reality. A positive net alpha implies that investors invest too little in a particular fund. A negative net alpha implies that investors commit too much money to a fund. In contrast to the prevailing belief that mutual fund investors are often naive and irrational in their investment choices, the empirical evidence we provide indicates the existence of sophisticated investors who are capable of rational learning. They can correctly identify positive net present value opportunities and place capital at its most productive.

The rest of the paper is organized as follows. Section 2 provides the theoretical framework for mutual funds. Section 3 contains a description of the sample. Section 4 investigates the fund-level relationship between fund performance and fund size. We start with a detailed discussion of the pros and cons of different econometric estimators and an evaluation of their effectiveness in simulations. This is followed by an empirical examination of the size–performance relationship in our sample. After the fund-level diseconomies of scale are established, we then move in Section 5 to measure skill in active management using value added. In particular, we compare the realized value added with the optimal value added. Section 6 offers insights into investor rationality and the market competitiveness in the active management space, and Section 7 concludes.

2 Theoretical framework for mutual funds

Suppose two investment opportunities are present in the market: an active fund and a passive fund, both benchmarked against the same index. Also suppose that the active fund manager is talented in that she can beat the index after fees. In a neoclassical world with perfectly rational players and no market frictions, the active fund is an investment opportunity with a positive net present value (NPV), and thus attracts investors' capital flows. The capital flows to the active fund would continue to the point at which both funds provide indifferent expected returns to investors. Berk and Green's (2004) rational model of active management states that the net alpha directs investors' capital flows, and a key component in the mechanism to achieve equilibrium is decreasing returns to scale.

Berk and Green's model has several profound implications for mutual fund research. If the net returns to investors are determined in equilibrium by competition between investors, and not by the managers' skill, then the net alpha is never a measure of skill. The net alpha does not teach us anything about managerial ability as claimed by many previous studies. Instead, the net alpha teaches us something about investors. In presence of diseconomies of scale, the gross alpha is a decreasing function of fund size, *i.e.*, $\alpha^g(q)$ with $\alpha^g(q)' < 0$. Recognizing that the net alpha is simply the gross alpha minus the fees and fund fees are quite stable in reality, the net alpha is also a decreasing function of the fund size. Thus, we can gauge the rationality of investors and the competitiveness of capital markets by studying the net alpha and its relation to size. In a decreasing returns to scale world, a positive net alpha indicates that investors have not given enough money to a particular fund; while a negative net alpha suggests that investors have given the fund too much money. In a standard rational expectations world, all net alphas are zero because positive net alphas are competed away by investors, regardless of the fund's skill level.

After ruling out the net alpha as a skill measure, Berk and van Binsbergen (2015; hereafter "BvB") further prove that the gross alpha does not measure skill either. This is because the gross alpha is simply the fees a fund charged in equilibrium. Unless funds choose fees to reflect their skill, the gross alpha does not reflect managerial ability.

In a decreasing returns to scale world, the gross alpha and the fund size are not two independent entities. A proper measure of managerial skill has to augment information from both the fund size and the fund return. BvB's value added measure is such a skill measure. The value added of a fund is defined as the dollar amount of what the fund adds over the benchmark, which is computed as the product of the gross alpha and the fund AUM:

$$V(q) = q\alpha^g(q). \tag{1}$$

The value added at the maximum is given by

$$V^* = \max_q q\alpha^g(q). \tag{2}$$

To ensure a finite solution to the optimization problem above, a necessary (but not sufficient) condition is that $\alpha^g(q)$ decreases with q. Put differently, value added cannot differentiate skill in the absence of decreasing returns to scale at the fund level because in this case V monotonically increases with the fund size, and every fund can add infinite value in theory. We denote the solution to Eq. (2) as q^*, the optimal amount the manager can actively manage.

Under the neoclassical assumptions that managers optimize, markets are competitive and investors are rational, BvB show that the maximum value added can be consistently estimated with a simple measure called the realized value added. To introduce it, we need to put Eq. (1) in context. For the i-th fund, the expected value added between times $t-1$ and t is:

$$V_{it} = q_{it-1}\alpha^g(q_{it-1}) \equiv q_{it-1}\alpha^g_{it},$$

where q_{it-1} is the fund AUM at the end of the previous period, and α^g_{it} is the gross alpha the fund expects to achieve between times $t-1$ and t. Because managers optimize, they would invest at their optimal amount. This effectively means that managers index the excess money when investors provide more capital than the optimal amount q^*_i. As the indexed money earns no alpha, the equilibrium gross alpha is given by:

$$\alpha^g_{it}(q_{it-1}) = \left(\frac{q^*_i}{q_{it-1}}\right)\alpha^g_{it}(q^*_i) + \left(\frac{q_{it-1}-q^*_i}{q_{it-1}}\right)0 = \frac{V^*_i}{q_{it-1}}. \tag{3}$$

Therefore, the product of the gross alpha and the fund AUM is the maximum value the manager can add:

$$V_{it} = q_{it-1}\alpha^g_{it}(q_{it-1}) = V^*_i.$$

An unbiased estimator of V_{it}, and thus V^*_i, is:

$$S_{it} = q_{it-1}r_{it}, \tag{4}$$

where r_{it} is the benchmark-adjusted realized gross return, and $\mathrm{E}[r_{it}] = \alpha^g_{it}$. For a fund that

exists for T_i periods, the estimated skill is the time series average of S_{it}:

$$S_i = \frac{1}{T_i} \sum_{t=1}^{T_i} S_{it}. \tag{5}$$

This is also referred to as the realized value added by BvB which consistently estimates V_i^* under the standard neoclassical assumptions.

As a skill measure, the realized value added (5) reveals the actual dollar amount a fund manager extracted from financial markets. The realized value added is not only intuitive but also robust in the sense that it holds regardless of the functional form of the decreasing returns to scale. BvB provide compelling evidence based on S_i that managerial skill exists and persists.

To summarize, the question of whether and how much mutual fund managers are skilled is an entirely different question from the question of whether investors share in the fruits of the managers' skill. The first question can be answered only by the value added measure, and the second question can be answered with the net alpha measure.

3 Data

The mutual fund data come from Morningstar over the period from January 1995 to December 2014. To avoid survivorship bias, we include live and dead funds. Following the common practice in the mutual fund literature, we restrict the analysis to actively managed domestic equity-only funds in US markets. For this purpose, we include only funds that fall into one of the following nine Morningstar fund categories: large blend (LB), large growth (LG), large value (LV), mid-cap blend (MB), mid-cap growth (MG), mid-cap value (MV), small blend (SB), small growth (SG) and small value (SV).[1] This step effectively excludes bond funds, money market funds, international funds, funds of funds, sector funds, real estate funds, target retirement funds and other non-equity funds. We exclude index funds identified by Morningstar, as well as funds whose name contains "index" or "enhanced-index." This screening process leaves us with 15,766 unique share classes (as identified by SecID) for the sample period we consider.

Many mutual funds offer multiple share classes. These share classes represent claims on the same underlying assets and therefore have the same returns before expenses and loads. The share classes typically differ only in their fee structures (*e.g.*, load versus no load) and/or in their clientele (*e.g.*, institutional versus retail). Therefore, it would be misleading to consider

[1] A mutual fund may experience style drift over time, and correspondingly, its Morningstar category evolves over time. We exclude fund/month observations with the Morningstar category taking values other than the nine listed above.

different share classes of the same fund as separate funds. In this paper, we aggregate all share classes of the same fund. Different share classes of the same funds have a different Morningstar SecID but the same Morningstar FundID. Out of the 15,766 SecIDs, there are 1,617 FundIDs with single SecIDs and 2,915 FundIDs with multiple SecIDs. We manually check the 1,617 FundIDs with a single share class and identify that 156 SecIDs need to be grouped into 55 funds.[2] A fund's AUM is computed by summing the AUM across the fund's different share classes. When aggregating returns, turnovers and expense ratios across share classes, we take the AUM-weighted average across all non-missing values.

We adjust all fund AUM numbers by inflation and express them in January 1, 2014 dollars. A mutual fund enters the sample after its combined AUM across all share classes exceeds $15 million in January 2014 dollars. Once a fund has exceeded this threshold, we keep the fund in the sample no matter what happens to its AUM to avoid sample selection bias. This procedure guards against the incubation bias of Evans (2010). We also drop funds with fewer than two years of data. These screens leave us with a sample of 3,077 mutual funds.

Morningstar assigns each fund a category and designates a benchmark portfolio to each fund category.[3] The Morningstar benchmark does not suffer from cherry picking bias because Morningstar categorizes funds based on their holdings rather than their reported objectives. We follow Pástor, Stambaugh, and Taylor (2015) and calculate the cost of capital by assuming the next best alternative investment opportunity investors have is the Morningstar designated benchmark portfolio. Similar to Pástor, Stambaugh, and Taylor (2015), we also assume that a fund's benchmark beta is equal to one. These two implementation choices are sensible in that first, they effectively avoid handicapping funds by benchmarking them to non-investable strategies, such as the Fama-Fench factor portfolios;[4] and second, they circumvent the need to address the estimation error in mutual fund betas which can be substantial especially for short-lived funds. We construct two performance measures: the benchmark-adjusted gross return and the benchmark-adjusted net return. The benchmark-adjusted gross return is calculated as

[2]We identify share classes using fund names and Morningstar holdings information. For example, "Bartlett Basic Value A," "Bartlett Basic Value C" and "Bartlett Basic Value Y" would be identified as share classes of the same fund.

[3]The Morningstar benchmark portfolios for the nine Morningstar categories are the Russell 1000 Total Return Index for LB, Russell 1000 Growth Total Return Index for LG, Russell 1000 Value Total Return Index for LV, S&P Mid Cap 400 Total Return Index for MB, Russell Mid Cap Growth Total Return Index for MG, Russell Midcap Value Total Return Index for MV, Russell 2000 Total Return Index for SB, Russell 2000 Growth Total Return Index for SG and Russell 2000 Value Total Return Index for SV.

[4]In contrast to the academic practice of using the Fama-French factors as benchmarks in performance evaluation, practitioners typically evaluate an active fund by comparing its performance to a tradable benchmark index. The non-tradable nature of the Fama-French factors disqualifies them as an alternative investment opportunity set faced by a passive investor. Adopting Morningstar benchmark indices ensures that this alternative investment opportunity set is marketed and tradable at the same time. Further, as pointed out by Cremers, Petajisto, and Zitzewitz (2013) and Berk and van Binsbergen (2015), the Fama-French model can produce biased assessments of fund performance.

the difference between a fund's gross return and the fund's benchmark portfolio return, which represents a fund manager's ability to outperform her benchmark. The benchmark-adjusted net return is the difference between a fund's net return and the fund's benchmark return, which captures the return to investors after fees and expenses.

Table 1 lists descriptive statistics of the sample. In the context of the fund's gross return, a fund outperforms its benchmark by 4 bp per month on average, whereas in the context of the net return, a fund underperforms its benchmark by 6 bp per month on average. This is consistent with the negative average net alpha value reported by Pástor, Stambaugh, and Taylor (2015). It is worth noting that we restrict our sample in the same way as Pástor, Stambaugh, and Taylor (2015) to make the results comparable to theirs, but using a longer time sample and larger cross-section as in Berk and van Binsbergen (2015) leads to an average net alpha estimate of zero.[5] The monthly expense ratio is taken as 1/12th of the annual gross expense ratio reported by Morningstar. The average expense ratio is 10 bp per month. Fund size is measured by the AUM at the end of the previous month. The distribution of the AUM is highly skewed to the right with a median value of $261 million and the top range of the AUM in the tens of billions. Fund age is the age of a fund in years at the month end since the fund's inception date. A fund's inception date is taken to be the earliest inception date across the fund's share classes, or if missing, the fund's first offer date in Morningstar. The turnover ratio is defined as the minimum of the aggregate purchases and sales divided by the average annual AUM in percentage. To remove outliers, we winsorize turnover at its 1st and 99th percentiles. The average turnover ratio for our sample is 80.42% per year.

4 Fund-level returns to scale

Denote return for the i-th fund at time t as R_{it}, and the corresponding Morningstar benchmark return as B_{it}. The benchmark-adjusted performance for the i-th fund at time t is

$$r_{it} = R_{it} - B_{it}.$$

Berk and van Binsbergen (2015) make a compelling argument why a tradable index-based adjustment is likely to adjust for fund style and risk more precisely than the commonly used loadings on risk factors. Let x_{it-1} be a proxy for the lagged fund size. In order to describe the relationship between fund size and performance, much of the literature employs a pooled OLS panel regression (see, e.g., Chen, Hong, Huang, and Kubik, 2004, Yan, 2008, Ferreira,

[5]Berk and van Binsbergen (2015) include US equity funds that invest both domestically and aboard. Their whole sample period is from 1977 to 2011.

Keswani, Miguel, and Ramos, 2013):

$$r_{it} = \alpha + \beta x_{it-1} + u_{it}. \tag{6}$$

Two commonly used fund size proxies are the dollar amount of the fund AUM (see, e.g., Pástor, Stambaugh, and Taylor, 2015) and the logarithm of the fund AUM (see, e.g., Chen, Hong, Huang, and Kubik, 2004, Yan, 2008, Elton, Gruber, and Blake, 2012, Ferreira, Keswani, Miguel, and Ramos, 2013). Regression (6) does not take any stand on the relationship between fund size and fund performance. Although the Berk-Green framework implies $\beta < 0$, model (6) also allows for $\beta = 0$, which is constant returns to scale, and $\beta > 0$, which is economies of scale. However, the latter two cases are theoretically unrealistic. A non-negative β implies a fund's investment strategy is infinitely scalable. Put another way, such a fund is a positive NPV investment opportunity regardless of its size. Conversely, an infinitely large fund would become the market and hence a zero gross alpha.

The rest of the section starts with a detailed discussion of the pros and cons of different econometric estimators and an evaluation of their effectiveness in simulations. This is followed by an empirical examination of the size–performance relationship in our sample.

4.1 Methodology

4.1.1 Fixed-effects model and recursive demeaning

Using model (6) to identify the effect of size on performance suffers an endogeneity problem. The model effectively assumes no cross-sectional differences in fund skill; *i.e.*, all funds have the same α, the return generated on the first dollar actively invested. In the presence of heterogeneity in fund skill, a scenario more consistent with reality, the model framework (6) absorbs the skill difference into the error term u_{it}. Denote $u_{it} = s_i + \epsilon_{it}$, where s_i reflects individual fund skill. Let $\boldsymbol{r}$ be the vector of the observed benchmark-adjusted returns, $\boldsymbol{x}$ the vector of regressors, $\boldsymbol{u}$ the error vector and $\boldsymbol{s}$ the vector of individual fund skills. In the case of $\boldsymbol{u} = \boldsymbol{s} + \boldsymbol{\epsilon}$, the asymptotic bias of the OLS estimator $\hat{\beta}$ of model (6) can be formulated as

follows:

$$\begin{aligned} \mathrm{E}(\hat{\beta}_{ols}) - \beta &= \frac{\mathrm{cov}(\boldsymbol{x}, \boldsymbol{r})}{\mathrm{var}(\boldsymbol{x})} - \beta \\ &= \frac{\mathrm{cov}(\boldsymbol{x}, \alpha + \boldsymbol{x}\beta + \boldsymbol{u})}{\mathrm{var}(\boldsymbol{x})} - \beta \\ &= \frac{\mathrm{cov}(\boldsymbol{x}, \boldsymbol{u})}{\mathrm{var}(\boldsymbol{x})} = \frac{\mathrm{cov}(\boldsymbol{x}, \boldsymbol{s})}{\mathrm{var}(\boldsymbol{x})}. \end{aligned} \tag{7}$$

If the unobserved skill component s_i is uncorrelated with the fund size, then s_i affects only the fund performance and not the size effect β. Unfortunately, size and skill are unlikely to be independent. This is because not only do skilled managers foster growth in fund size, but also larger funds can afford to hire better managers. If $\mathrm{cov}(x_i, s_i) \neq 0$, for some but not necessarily all funds, an endogeneity issue arises when putting s_i into the error term, which leads to a biased estimate of β. When the correlation between skill and size is positive, *i.e.*, $\mathrm{cov}(\boldsymbol{x}, \boldsymbol{s}) > 0$, the omitted-variable bias in $\hat{\beta}_{ols}$ is positive.

As pointed out by Pástor, Stambaugh, and Taylor (2015), the omitted-variable bias in regression (6) can be fixed by including a fund fixed effect:

$$r_{it} = \alpha_i + \beta x_{it-1} + \epsilon_{it}. \tag{8}$$

Although the fund fixed effect α_i removes the omitted-variable bias, the effect introduces a finite-sample bias because of the demeaning process used in estimating fixed-effects models.[6] If regressors are strictly exogenous, *i.e.*, they are uncorrelated with past, present and future shocks, then the fixed-effects estimators are free from bias. Alas, this is not the case for the size–performance regression (8). Although the lagged size x_{it-1} and the next period return innovation ϵ_{it} are independent, there is a positive contemporaneous correlation between the fund size and the unexpected return. This positive contemporaneous correlation is due to a mechanical link between the return innovation and the fund size; *i.e.*, the fund size increases automatically after a higher return. Because of this contemporaneous correlation, a demeaning process induces a correlation between the demeaned regressor and the demeaned innovation, which, in turn, causes a finite-sample bias in a fixed-effects estimate of the scale effect β.

To remove the finite sample bias in the fixed-effects estimator, Pástor, Stambaugh, and Taylor (2015) use Moon and Phillips's (2000) recursive demeaning (RD) process. Following

[6]The commonly available fixed-effects estimators involve a demeaning process. The two common estimators available are the within estimator and first-difference estimator. The within estimator demeans the variables by subtracting their full-sample time-series mean. The first-difference estimator demeans the data by differencing the equation (8) across two consecutive time periods.

their notation, we define the recursively forward-demeaned variables for the i-th fund as

$$\bar{r}_{it} = r_{it} - \frac{1}{T_i - t + 1} \sum_{s=t}^{T_i} r_{is},$$
$$\bar{x}_{it-1} = x_{it-1} - \frac{1}{T_i - t + 1} \sum_{s=t}^{T_i} x_{is-1},$$
$$\bar{\epsilon}_{it} = \epsilon_{it} - \frac{1}{T_i - t + 1} \sum_{s=t}^{T_i} \epsilon_{is}.$$

The scale effect β can be estimated from the demeaned model

$$\bar{r}_{it} = \beta \bar{x}_{it-1} + \bar{\epsilon}_{it}, \tag{9}$$

where t spans from 1 to $T_i - 1$. Such a demeaning process sweeps out the fixed effects α_i but introduces a correlation between the demeaned size $\bar{x}_{it-1}$ and the demeaned innovation $\bar{\epsilon}_{it}$; *i.e.*, $\text{Corr}(\bar{x}_{it-1}, \bar{\epsilon}_{it}) \neq 0$. Moon and Phillips (2000) suggest the use of a recursively backward-demeaned regressor as an instrumental variable (IV). We define the recursively backward-demeaned regressor $\underline{x}_{it-1}$, for $t = 2, \cdots, T_i$, as

$$\underline{x}_{it-1} = x_{it-1} - \frac{1}{t-1} \sum_{s=1}^{t-1} x_{is-1}. \tag{10}$$

The variable $\underline{x}_{it-1}$ is qualified as an IV because it is independent of $\bar{\epsilon}_{it}$ but correlated with $\bar{x}_{it-1}$. The IV estimator of the scale effect β is

$$\hat{\beta}_{RD1} = \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i - 1} \bar{x}'_{it-1} \underline{x}_{it-1} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i - 1} \underline{x}'_{it-1} \bar{r}_{it} \right). \tag{11}$$

We call the estimator (11) RD1 which is used by Pástor, Stambaugh, and Taylor (2015)(their Equation (10)) in examining the size–performance relation.

4.1.2 Issues with RD1 and an enhanced estimator

A two-stage least squares interpretation of RD1 is:

$$\bar{x}_{it-1} = \rho \underline{x}_{it-1} + v_{it-1}, \tag{12}$$
$$\bar{r}_{it} = \beta \hat{\bar{x}}_{it-1} + u_{it}, \tag{13}$$

where $\hat{\bar{x}}_{it-1}$ is the fitted value from the first-stage regression (12). That is, we first regress $\bar{x}_{it-1}$ on $\underline{x}_{it-1}$, and then we regress $\bar{r}_{it}$ on the fitted values from the first-stage regression. A zero intercept is imposed on both regressions. The slope estimator from the second-stage regression (13), $\hat{\beta}_{RD1}$, consistently estimates the size effect β. The asymptotic variance of $\hat{\beta}_{RD1}$ is proportional to the inverse of the first stage goodness of fit:

$$\text{var}(\hat{\beta}_{RD1}) \propto \frac{\sigma_u^2}{R_1^2}, \tag{14}$$

where σ_u^2 is the variance of u_{it}, and R_1^2 is the R-squared of the first-stage regression (12). As shown by (14), it is important to achieve a good fit in the first-stage regression. A weak first stage with a small R-squared value is a warning sign. At its best, it leads to a large uncertainty in quantifying the size effect. At its worst, it can cause bias and inconsistency in estimating β.

The imposition of a zero intercept in the second-stage regression (13) is due to the fund fixed effect α_i being removed in the demeaned model (9). Imposing a zero intercept in the first-stage regression (12), however, is not obvious. In fact, it is a risky model choice because a wrongly imposed restriction reduces the fitting and decreases the precision of the RD1 estimator. Below, we demonstrate that this zero intercept restriction in the first-stage regression is problematic for fund size.

Regression theory states that the intercept of the first stage regression (12) is $\text{E}(\bar{x}_{it-1} \mid \underline{x}_{it-1} = 0)$. Restricting it to zero, *i.e.*, $\text{E}(\bar{x}_{it-1}|\underline{x}_{it-1} = 0) = 0$, and plugging in the expressions of $\bar{x}_{it-1}$ and $\underline{x}_{it-1}$, we have $\text{E}\left(x_{it-1} - \frac{1}{T_i-t+1}\sum_{s=t}^{T_i} x_{is-1} \mid x_{it-1} = \frac{1}{t-1}\sum_{s=1}^{t-1} x_{is-1}\right) = 0$, which is equivalent to $\frac{1}{t-1}\sum_{s=1}^{t-1} x_{is-1} = \text{E}\left(\frac{1}{T_i-t+1}\sum_{s=t}^{T_i} x_{is-1} \mid x_{it-1} = \frac{1}{t-1}\sum_{s=1}^{t-1} x_{is-1}\right)$. Taking expectation on both sides of the latter equation leads to $\text{E}\left(\frac{1}{t-1}\sum_{s=1}^{t-1} x_{is-1}\right) = \text{E}\left(\frac{1}{T_i-t+1}\sum_{s=t}^{T_i} x_{is-1}\right)$. The last step is obtained because the law of total expectation $\text{E}(\text{E}(X|Z)) = \text{E}(X)$ holds for any two random variables X and Z. Thus, imposing a zero intercept in the first stage implies that a fund's past average size equals its future average size. It is, however, unrealistic to assume that the size of a typical fund fluctuates around a constant mean over time. Hence, imposing a zero intercept in the first-stage regression (12) would only decrease the goodness of fit of the model which, in turn, reduces the precision of the β estimator in the second stage.

To intuitively illustrate the inappropriateness of such a constraint, we consider four simple hypothetical processes for the log fund AUM (x_t): a) x_t follows a normal distribution $N(5, 1)$, a scenario with the expected fund size being constant ($148 million); b) the first half of the fund size observations follows $N(4, 1)$ and the second half follows $N(5, 1)$, a scenario with the expected fund size being a step function (increasing from $55 million to $148 million); c) $x_t = x_{t-1} + 0.01$, a scenario in which the fund size grows at a constant rate of 1% per month;

and d) $x_t = x_{t-1} + B_t$, a scenario in which the fund size grows with an index B_t. We generate twenty-four fund sizes from each of the four scenarios. The benchmark return B_t is taken as the monthly price returns on the S&P 500 from 2012 to 2013. The choice of the starting value for the fund size in the last two scenarios is not important as the initial size drops out in the demeaning process.

Figure 1 plots the forward-demeaned fund size $\bar{x}_t$ versus the backward-demeaned fund size $\underline{x}_t$ for the four hypothetical fund size processes. The circles represent twenty-two pairs of $(\underline{x}_t, \bar{x}_t)$. The red line depicts the model fitting for the constrained regression $\bar{x}_t = \rho \underline{x}_t + v_t$. The black line shows the regression with an intercept $\bar{x}_t = \psi + \rho \underline{x}_t + v_t$. The red line is forced to cross the origin (the filled diamond) because of the zero intercept constraint. In scenario (a), the fund size fluctuates around a constant mean that satisfies the required condition for imposing a zero intercept. In this ideal case, the black and red lines parallel each other and differ by only a small constant. We know that the true value of the intercept is zero in scenario (a), and a non-zero intercept estimate is due to the finite sample estimation error. This estimation error, however, has little impact on the β estimate in the second stage. In scenario (b) that the expected fund size is a step function, the constrained model starts to depart the data while the black line fits the data well. In scenarios (c) and (d), the restriction greatly constrains the model and results in a nonsensical fit. When the fund size grows at a constant rate, the black line provides a perfect fit, while the red line fails to follow the data altogether. When the fund size grows with a market index, the fitting of the red line is also absurd. Again, the black line delivers a much better fit. As illustrated by these hypothetical examples, the constraint model is inappropriate once the expected fund size is time-varying. With the RD1 approach, it is worrisome that the nonsensical model fit of the constraint regression is carried on into the second stage to quantify the scale effect.

The dynamic properties of the fund size challenge the zero intercept restriction in the first-stage regression of the RD1 approach. The unjustified restriction, in our judgment, can cause model misspecification in the first stage and result in inaccurate and imprecise quantification of the size effect in the second stage. We propose an enhanced recursive demeaning estimator which is more suitable for the size–performance analysis. First, we include an intercept in the first-stage regression in acknowledging that the size of a typical fund is unlikely to fluctuate around a constant mean over time. Second, we use x_{it-1} rather than the backward-demeaned variable $\underline{x}_{it-1}$ as the IV. The qualification of x_{it-1} as an IV lies in the fact that x_{it-1} is obviously correlated with $\bar{x}_{it-1}$ but uncorrelated with $\bar{\epsilon}_{it}$ because it does not contain information after period $t-1$.

The new estimator can be implemented via two-stage least squares:

$$\bar{x}_{it-1} = \psi + \rho x_{it-1} + v_{it-1}, \tag{15}$$

$$\bar{r}_{it} = \beta \bar{x}^*_{it-1} + u_{it}, \tag{16}$$

where $\bar{x}^*_{it-1}$ is the fitted value from the first-stage regression (15). The enhanced estimator is

$$\hat{\beta}_{RD2} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T_i-1} \bar{x}^{*'}_{it-1} \bar{x}^*_{it-1} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T_i-1} \bar{x}^{*'}_{it-1} \bar{r}_{it} \right). \tag{17}$$

We call the estimator (17) RD2.[7]

To perform inference, a robust version of the variance of the recursive demeaning estimators is $\text{var}(\hat{\beta}_{RD}) = (\Omega_{xx})^{-1} \Phi_{ux} (\Omega_{xx})^{-1}$. Let $\hat{\bar{\epsilon}}^1_{it} = \bar{y}_{it} - \hat{\beta}_{RD1} \bar{x}_{it-1}$ and $\hat{\bar{\epsilon}}^2_{it} = \bar{y}_{it} - \hat{\beta}_{RD2} \bar{x}_{it-1}$. For the estimator $\hat{\beta}_{RD1}$, its robust variance is computed by setting $\Omega_{xx} = \sum_{i=1}^{N} \sum_{t=2}^{T_i-1} \bar{x}'_{it-1} \underline{x}_{it-1}$ and $\Phi_{ux} = \sum_{i=1}^{N} \sum_{t=2}^{T_i-1} \sum_{s=2}^{T_i-1} (\underline{x}'_{it-1} \hat{\bar{\epsilon}}^1_{it})(\underline{x}'_{is-1} \hat{\bar{\epsilon}}^1_{is})'$. For $\hat{\beta}_{RD2}$, its robust variance is computed by setting $\Omega_{xx} = \sum_{i=1}^{N} \sum_{t=1}^{T_i-1} \bar{x}^{*'}_{it-1} \bar{x}^*_{it-1}$ and $\Phi_{ux} = \sum_{i=1}^{N} \sum_{t=1}^{T_i-1} \sum_{s=1}^{T_i-1} (\bar{x}^{*'}_{it-1} \hat{\bar{\epsilon}}^2_{it})(\bar{x}^{*'}_{is-1} \hat{\bar{\epsilon}}^2_{is})'$. The t-test and Wald test based on $\text{var}(\hat{\beta}_{RD})$ satisfy the usual property.

4.2 Simulations

In this section, we conduct simulations to evaluate the performance of the different estimators for the size–performance analysis, namely, the simple OLS estimator (OLS), the fixed-effects estimator (FE), RD1 (11) and RD2 (17). In their simulation study, Pástor, Stambaugh, and Taylor (2015) focus on the bias and hypothesis testing performance of the estimators. In addition to these two performance metrics, we also report the standard deviation and the root mean square error (RMSE) of the estimators.[8] An estimator's superiority lies in a careful

[7] Alternatively, the RD2 can also be expressed using x_{it}, $\bar{x}_{it}$ and $\bar{r}_{it}$. For fund i, denote the vectors of its forward-demeaned response, the forward-demeaned regressor and the IV as

$$\bar{\boldsymbol{r}}_i = \begin{pmatrix} \bar{r}_{i1} \\ \bar{r}_{i2} \\ \vdots \\ \bar{r}_{iT_i-1} \end{pmatrix}, \quad \bar{\boldsymbol{x}}_i = \begin{pmatrix} \bar{x}_{i0} \\ \bar{x}_{i1} \\ \vdots \\ \bar{x}_{iT_i-2} \end{pmatrix}, \text{ and } \boldsymbol{z}_i = \begin{pmatrix} 1 & x_{i0} \\ 1 & x_{i1} \\ \vdots & \\ 1 & x_{iT_i-2} \end{pmatrix}.$$

The new proposed estimator is

$$\hat{\beta}_{RD2} = \left(\sum_{i=1}^{N} \bar{\boldsymbol{x}}'_i \boldsymbol{z}_i (\boldsymbol{z}'_i \boldsymbol{z}_i)^{-1} \boldsymbol{z}'_i \bar{\boldsymbol{x}}_i \right)^{-1} \left(\sum_{i=1}^{N} \bar{\boldsymbol{x}}'_i \boldsymbol{z}_i (\boldsymbol{z}'_i \boldsymbol{z}_i)^{-1} \boldsymbol{z}'_i \bar{\boldsymbol{r}}_i \right).$$

[8] The standard deviation of an estimator is calculated as $sd = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (\hat{\beta}_m - \bar{\hat{\beta}})^2}$, where M is the number of simulations, $\hat{\beta}_m$ is the estimate for the mth simulated samples, and $\bar{\hat{\beta}} = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_m$ is the average value of

balance of bias and variance. An unbiased estimator with a large amount of uncertainty can be as useless as a stable estimator with a large bias. Given that bias and standard deviation are two facets of an estimator, the RMSE is a better performance measure than bias or standard deviation alone.

We simulate panel data on funds' returns and sizes. Benchmark-adjusted fund returns are generated from the fixed-effects model (8). Fund sizes are generated from the following process

$$x_{it} - x_{it-1} = \phi + \rho_1 B_{it} + \rho_2 r_{it} + \zeta_{it}, \tag{18}$$

where x_{it} is the natural logarithm of the i-fund's AUM at the end of period t, B_{it} is the fund's benchmark return, and r_{it} is the benchmark-adjusted performance.[9] Parameter ρ_2 captures the contemporaneous correlation between the fund performance and the fund size. We estimate the return and fund size models on our mutual fund data. The size effect β in the return model (8) is estimated to be -0.0026 by the RD2 estimator. More details on the size effect estimation are described in Section 4.3. The corresponding parameter estimates for the size model (18) are $\phi = 0.0019$, $\rho_1 = 1.065$, $\rho_2 = 1.121$ and $std(\zeta) = 0.1$.

Guided by the parameter estimates on our sample, we choose the model parameter values in the simulations as follows:

- Size model (18): the benchmark returns B_{it} are generated from a normal distribution with a mean of 0.009 and a standard deviation of 0.05. These particular parameter choices match the features of the S&P 500 monthly total return series from 1926 to 2014. The other model parameters are $\phi = 0.0019$, $\rho_1 = 1$, $std(z) = 0.1$ and four plausible values of ρ_2: 0.9, 1, 1.1 and 1.2;

- Return model (8): individual fund fixed-effect α_i following a normal distribution $N(0.016, 0.006)$, $std(\epsilon) = 0.026$, and four different values for β: 0, -0.001, -0.003 and -0.01.

The simulation exercise yields insights into the size and power associated with the different estimators. We construct 10,000 samples of simulated panel data for 300 funds over 100 months from the size and return models. All funds start at \$250 million.

all M estimates. The RMSE is calculated as $\sqrt{bias^2 + sd^2}$, where $bias = \bar{\hat{\beta}} - \beta$.

[9] Pástor, Stambaugh, and Taylor (2015) consider a fund size process ignoring benchmark returns,

$$\frac{q_{it}}{q_{it-1}} - 1 = \delta + \gamma r_{it} + v_{it},$$

where q_{it} is the dollar amount of the i-fund's AUM at time t. This is equivalent to modelling the log fund size change as $x_{it} - x_{it-1} = \exp(1 + \delta + \gamma r_{it})$. Our size process fits the actual fund sizes much better than their model with a R-sqauared value of 8% using our model and 0.7% using theirs.

Table 2 presents the simulation results. Panel A summarizes the RMSE of the four estimators across the 10,000 simulated samples. The FE estimator always substantially reduces the RMSE of the simple OLS estimator in all cases. In other words, it pays off to correct the omitted-variable bias in the simple OLS estimator with the FE estimator. However, the RMSE of RD1 is larger than that of the FE estimator in most simulation scenarios. That is to say, it is not worthwhile to correct the finite-sample bias in the FE estimator with RD1. In general, RD1 reduces the RMSE of the OLS estimator, but it is not always the case. For example, when $\beta = -0.003$, a value close to the scale effect estimated on the real data, it is better off to use the biased OLS estimator than RD1 based on the RMSE criterion. In contrast, RD2 offers the smallest RMSE consistently across all simulation scenarios, with the RMSE half of that of the FE estimator.

We next decompose RMSE into bias and standard deviation to understand the driving factor in the RMSE of each estimator. Panel B and C report the bias and standard deviation of the estimators. The omitted-variable bias in the OLS estimator is positive, while the FE estimator is biased downwards. Both RD estimators are essentially unbiased. When it comes to estimation uncertainty, the FE estimator shows the least variation. The two biased-corrected counterparts, RD1 and RD2, eliminate bias at the cost of precision. The RD1 estimates vary the most out of the four estimators considered. Compared with RD1, RD2 yields more precise estimates with standard deviations less than half of those of RD1. The RD1 estimates are particularly variable when $\beta = -0.003$. In this case, the standard deviation of RD1 is five times RD2. Examining Panel B and C, we can see that bias is the main contributor to the RMSE of the OLS and FE estimators, while estimation uncertainty is the key driver in the RMSE of RD1 and RD2. The benefit of correcting bias with RD1 is often hampered by increased estimation uncertainty.

Panel D shows the fraction of simulations in which the null hypothesis ($\beta = 0$) is rejected at the 5% confidence level. The OLS estimator always produces false positives, and the FE estimator also has a distorted size, rejecting the null in 65% to 78% of simulations when the null is actually true. The two RD estimators have the right size, rejecting a true null in 5% of the simulations. However, when it comes to power, the performance of RD1 is unsatisfactory. To a great degree, RD1 lacks sufficient power to reject the null of $\beta = 0$ for plausible magnitudes of the scale effect. For example, when $\beta = -0.003$, RD1 rejects the null about 13% to 17% of the time depending on the value of ρ_2. In contrast, RD2 possesses adequate power to reject the null when the null is false. When $\beta = -0.001$ and $\beta = -0.003$, the power of RD2 is greater than 90%, and its power increases to 100% for $\beta = -0.01$.

In sum, the OLS and the FE estimators are biased which makes it hard to draw convincing

conclusions based on their estimates. The two recursive demeaned estimators are unbiased, but it costs RD1 too much to correct the bias. Quite often, the benefit of reducing bias by RD1 is overshadowed by the increased estimation uncertainty. With its lack of accuracy, RD1 faces a substantial challenge when it comes to inference. In contrast, RD2 has virtually no bias, the right size and adequate power in hypothesis testing.

There are two contributing factors to the superior performance of RD2. The first is the inclusion of an intercept in the first-stage regression, and the second is the stronger instrument by using a more recent fund size measure. We gauge the relative importance of these two factors in term of their contribution in reducing RMSE. Let β_{RD3} be the recursive demeaning estimator obtained by including an intercept in the first-stage regression of $\bar{x}_{it-1}$ on $\underline{x}_{it-1}$. Thus, β_{RD3} removes the unrealistic constraint on fund size, but still uses the backward-demeaned size as the instrument. The reduction in RMSE by moving from the FE estimator to RD2, *i.e.*, $\text{RMSE}(\beta_{FE}) - \text{RMSE}(\beta_{RD2})$, can be decomposed into two terms: the reduction attributed to the inclusion of an intercept, *i.e.*, $\text{RMSE}(\beta_{FE}) - \text{RMSE}(\beta_{RD3})$, and the reduction attributed to the use of the stronger instrument, *i.e.*, $\text{RMSE}(\beta_{RD3}) - \text{RMSE}(\beta_{RD2})$. Across all the simulation scenarios, we find that the inclusion of the intercept contributes on average 98% of the improvement in RMSE and the remaining 2% comes from using the stronger instrument.

We carry out additional simulations to verify the robustness of the results. We consider Pástor, Stambaugh, and Taylor's (2015) size process and the results are presented in Appendix A.1. Like those authors, we find RD1 is bias free but has inadequate power rejecting a false null. RD2, again, dominates RD1 with much smaller RMSEs and sufficient power in hypothesis testing. In unreported simulation studies, we also allow the funds' starting sizes to follow a distribution rather than being fixed, and the results hold.

4.3 Diseconomies of scale

Armed with an understanding of the properties of the four different estimators (the simple OLS, FE and two RD estimators, RD1 and RD2), we are ready to examine the nature of the returns to scale in active management. Table 3 reports the findings of the scale effects for gross and net performance.

In order to compare with Pástor, Stambaugh, and Taylor's (2015) findings, we first consider the linear functional form of decreasing returns to scale, that is, linking the fund performance with the dollar fund AUM. The results are displayed in Panel A of Table 3. The simple OLS estimate is sensitive to the performance measure. The estimated coefficient of the fund size is negative when gross performance is used but positive when net performance is in place. The coefficient estimates obtained using the FE estimator are highly significant and negative

for gross and net performance. As established in Section 4.1 and Section 4.2, the OLS and the FE estimators are biased, and caution should be exercised when interpreting these coefficients. The RD1 estimator yields a negative point estimate. The estimator is, however, unable to reject the null of no relationship with t-statistics of -0.34 and -0.32 on the gross and net performance, respectively. Overall, we find mixed evidence of diminishing returns to scale at the fund level, which is significant and negative under the FE but insignificant and negative under the RD1. These results for the dollar fund size confirm Pástor, Stambaugh, and Taylor's (2015) findings.

We now apply the enhanced estimator RD2. The estimated effect of fund size on performance is negative and statistically significant at the 5% level: -0.485×10^{-6} with a t-statistic of -2.03 on the gross performance and -0.479×10^{-6} with a t-statistic of -1.99 on the net performance. To explore the source driving the difference between the estimates of RD1 and RD2, Table 3 displays the first-stage results. The first-stage intercept is restricted to zero in RD1. When it is allowed to be estimated as in RD2, the estimate is highly significantly different from zero, -292.1 with a t-statistic of -54.15. The model misspecification causes RD1 to have a weak first stage with a small R-squared value of 0.19%, while RD2 achieves a much better first-stage fitting with an R-squared value of 10.04%.

The loglinear functional form of decreasing returns to scale, which links the fund performance with the logarithm of fund AUM, is widely used in the size–performance analysis (see, e.g., Chen, Hong, Huang, and Kubik, 2004, Yan, 2008, Elton, Gruber, and Blake, 2012, Ferreira, Keswani, Miguel, and Ramos, 2013). We present the results using the logarithm of fund AUM in Panel B of Table 3. Three approaches, the OLS, FE and RD2, produce negative point estimates. RD2 gives highly significant estimates, -0.0026 with a t-statistic of -13.32 on the gross performance and a similar outcome on the net performance. The RD1 approach stands out and leads to economies of scale. The estimated size effects are 0.0491 on the gross performance and 0.0466 on the net performance, albeit insignificant. As demonstrated by the simulation studies, a weak first stage is a sign of model misspecification which can cause the RD1 estimates to be exceedingly variable. The goodness of fit of the first stage in RD1 is very poor as indicated by a tiny R-squared value of 0.01%. In contrast, RD2 has a much stronger first stage with an R-squared value of 8.78%. Overall, the evidence stresses the necessity for prudence and caution when making inferences based on the RD1 estimates.

The unbiased and stable estimator RD2 consistently produces significantly negative coefficient estimates regardless of the functional form of the decreasing returns to scale. The average fund-level decreasing returns to scale parameter estimated on the full sample, however, provides an incomplete picture of decreasing returns to scale in the data. It is plausible that

investment strategies of some funds are more scalable than others. That is, there is heterogeneity in the degree of decreasing returns to scale at the fund level. Because the size of a fund is determined endogenously, small funds are likely to differ from large funds in their returns to scale mechanism. We sort the mutual funds based on the decile rankings of their average AUM calculated over the sample period. We consider both linear and loglinear functional forms of decreasing returns to scale: $r_{it} = \alpha_i + \beta q_{it-1} + \epsilon_{it}$ and $r_{it} = \alpha_i + \beta \log(q_{it-1}) + \epsilon_{it}$, where r_{it} is the benchmark-adjusted fund gross performance and q_{it-1} is the lagged dollar fund AUM. In each AUM-sorted portfolio, we estimate β using the panel estimator RD2 under both model specifications and report the estimates as well as their t-statistics in Table 4. The relation between a fund's size and its gross performance is significantly negative across all deciles. It is very interesting to see that the magnitude of β decreases with the fund size regardless of the model specification. Large funds are characterized by their relatively flat decreasing returns to scale technology, while performance of small funds suffers more when fund size increases. The tests based on F-statistics, whose p values are reported in the last row of Table 4, reject the null hypothesis that the β estimates are equal across the ten AUM-sorted fund portfolios.

Table 4 also reveals a substantial amount of heterogeneity in decreasing returns to scale at the fund level. Under the linear functional form of decreasing returns to scale, the coefficient estimates range from -0.192×10^{-6} in the top decile (the group with the largest funds) to -167.578×10^{-6} in the bottom decile. This represents a more than eight-hundred-fold increase in the magnitude of size impact by shifting from the largest fund group to the smallest fund group. The changes in magnitude are less dramatic across the deciles under the loglinear model. The corresponding estimates range from -0.0011 to -0.0030, a three-fold increase in magnitude from the top decile to the bottom decile. To put these magnitudes into some perspective, we consider the impact on fund gross performance when fund sizes are doubled in each decile. Under the linear specification, the average impact of size doubling in a decile is quantified by βq, where β and q are the decreasing returns to scale parameter and the average fund AUM in that decile, respectively. In the loglinear setup, the impact of size doubling is associated with a decrease in expected fund performance of $\beta \log(2)$. As shown by Table 4, the associated average impact of size doubling on performance under the linear model varies from 33.5 bp per month in the 1st decile, 26.5 bp per month in the 5th decile and 7.3 bp per month in the 10th decile. The corresponding impacts under the loglinear model are 21, 11.6 and 7.9 bp per month in the 1st, 5th and 10th deciles, respectively. We see that both models produce a similar negative impact of size doubling on performance in the top decile. However, the size impact quantified by the loglinear model is smaller than that by the linear model in all other deciles. The difference is considerable, it is more than 12 bp per month across five deciles including

the 1st, 5th, 7th, 8th, and the 9th deciles. We use the loglinear specification for inference in the rest of the paper and leave the question of the most suitable functional form of decreasing returns to scale for future research.

In sum, the empirical approaches including OLS, FE and RD1 deliver mixed evidence of decreasing returns to scale at the fund level. Given the various issues associated with these estimators as illustrated in the simulation studies, researchers should exercise caution interpreting these results. Using the enhanced RD2 estimator, we document statistically and economically significant diseconomies of scale at the fund level. We show in Appendix A.2 that the significance of the decreasing returns to scale is unaffected by including controls such as industry size, as well as by adjusting returns using alternative risk models, such as the Fama-Fench three-factor model and the Carhart four-factor model.

5 Value added

In the previous section, we established the decreasing returns to scale at the fund level. BvB prove that in a decreasing returns to scale world, neither the gross alpha nor the net alpha measures managerial skill. Instead, the proper skill measure is value added. Below, we restate their framework with the loglinear functional form of the decreasing returns to scale.

We assume a loglinear relationship between gross alpha and fund size:

$$\alpha^g = a - b\log(q), \tag{19}$$

where q is the dollar amount of the fund AUM. The parameter a is the gross alpha a fund manager earns on the first dollar invested. We have shown in the previous section that $b > 0$ empirically. Intuitively, $b > 0$ is because the manager's investment ideas are in finite supply, and she invests her best ideas first. The parameter b captures how quickly a fund manager runs out of ideas. Under this loglinear functional form of the gross alpha, the optimization problem (2) introduced in Section 2 becomes:

$$V^* = \max_q q(a - b\log(q)). \tag{20}$$

The first-order condition gives the optimal amount the manager can actively manage, q^*:

$$q^* = e^{a/b-1}. \tag{21}$$

Accordingly the maximum value the manager can add is

$$V^* = q^*(a - b\log(q^*)) = be^{a/b-1}. \tag{22}$$

The skill measure (22) tells us the upper bound of the dollar amount a fund manager can extract from financial markets.

The full dynamics of the gross alpha and the value added as a function of the active fund size are depicted in Figure 2. The gross alpha is a nonlinear function of the active fund size which decreases at a decreasing rate with the fund size. The corresponding gross alpha at the optimal fund size q^* is $a/2$. The value added initially increases with the active fund size, reaches the maximum V^* at q^*, and then decreases with the active fund size. The value added is zero when the active fund size equals either zero or $e^{a/b}$. Once the active fund size exceeds $e^{a/b}$, the value added is in negative territory.

As discussed in Section 2, managerial skill for the i-th fund can be potentially quantified by either the realized value added S_i (Eq. 5) or the maximum value added, V_i^*. The realized value added S_i is a model-free measure in the sense that the measure holds regardless of the functional form of the decreasing returns to scale, while V_i^* is model-based. Under the standard neoclassical assumptions and a correctly specified functional form of the decreasing returns to scale, S_i and V_i^* should be consistent with each other.

We estimate fund-specific a and b parameters to compute V_i^*. Estimating b fund by fund leads to imprecise estimates especially for the short-lived funds.[10] To reduce the estimation error, we sort funds into ten portfolios by fund size and estimate b using the panel estimator RD2 in each decile portfolio. This implementation choice assumes that all the funds in a porfolio share the same b value. Contrary to the concern that ignoring within-group variation leads to inaccuracy in quantifying fund-specific b, this method actually increases the accuracy of the b estimate because of the sharp reduction in estimation errors.

Once the fund-specific b estimates are obtained, the parameter a for the i-th fund can be estimated as:

$$\hat{a}_i = \frac{1}{T_i}\sum_{t=1}^{T_i}(r_{it} + \hat{b}_i \log(q_{it-1})), \tag{23}$$

where r_{it} is the benchmark-adjusted gross return, and T_i is the number of observations for the i-th fund in the sample. Table 5 provides summary statistics of the estimates of a within each AUM-sorted porfolio. Recall that a represents the gross alpha earned on the first dollar

[10] One way to estimate the parameters a and b fund by fund is to run a simple OLS regression of fund return on fund size for each individual fund. The estimation error in b is severe. Although in the previous section we established the overall trend of decreasing returns to scale, 29% of the sampled funds end up with negative b estimates using the fund-by-fund OLS regression.

invested in the fund. While the average value of a in each decile increases with fund size, its variation decreases. The smallest funds (1st decile) exhibit the lowest average a but the largest variation in the returns they offer with the first dollar actively invested. The funds in the 9th and 10th deciles have the narrowest cross-sectional distribution of a.

We plug the estimates of a_i and b_i into Eq. (22) to get an estimate for V_i^*, $\hat{V}_i^*$. The realized value added S_i is calculated as the time series average of the product of the lagged fund AUM and its benchmark-adjusted gross return as described in Section 2. For both skill measures, we report two versions of the average value added across funds, the ex ante mean which is the simple average of all individual funds' value added:

$$\bar{S} = \frac{1}{N} S_i, \qquad \text{and} \qquad \bar{V}^* = \frac{1}{N} \hat{V}_i^*, \tag{24}$$

and the ex post mean which is the average value added across funds weighted by the number of periods it appears in the database:

$$\bar{S}_W = \frac{\sum_{i=1}^{N} T_i S_i}{\sum_{i=1}^{N} T_i}, \qquad \text{and} \qquad \bar{V}_W^* = \frac{\sum_{i=1}^{N} T_i \hat{V}_i^*}{\sum_{i=1}^{N} T_i}. \tag{25}$$

Table 6 summarizes the mutual fund skill quantified in the sample. The second column is for the realized value added. The average fund added an economically and statistically significant \$260,000 per month (in Y2014 dollars). The estimate of the mean of the ex ante distribution of talent (a simple cross-sectional average) is much lower at \$10,000 per month and not significantly different from zero. A large variation exists in the distribution of talent across funds. The fund at the 99th percentile cut-off generated close to \$13 million per month, and the median funds lost an average of \$60,000 per month. In total, close to 40% of the funds had positive estimated value added. These figures are consistent with BvB's finding that most funds destroyed value, but because most of the capital is controlled by skilled managers, as a group, the active mutual funds added value. The third column of Table 6 reports the model-based value added measure which is the maximum value a fund could potentially extract from the markets. Again, the ex post mean is greater than the ex ante mean, reflecting the fact that unskilled managers go out of business sooner. The cross-sectional distribution of talent is characterized by a large variation and heavy skewness to the right.

BvB provide compelling evidence that managerial skill exists and persists based on S_i. We focus on why the model-based value added and the realized value added differ, or to put it differently, why so many managers destroy value. Mutual fund managers are among the highest paid cohorts of society. The realized value added seems to indicate that in active management,

the majority of people without a competitive advantage can earn economic rents, a conclusion that is uncomfortable for economists. To shed some light on this puzzle, we have to realize that even if the functional form of the decreasing returns to scale is correctly specified, the realized value added could still be different from the maximum value a manager can add. For these two measures to be consistent, not only are investors rational but also fund managers in that they know their own optimal amounts and operate accordingly. We examine whether this is the case in reality.

Using the insight from Figure 2, we divide the funds into three groups with two critical fund sizes, $q_i^* = e^{a_i/b_i - 1}$ and $q_i^c = e^{a_i/b_i}$: underfunded when $q_i < q_i^*$; moderately overfunded when $q_i \subset (q_i^*, q_i^c]$; and excessively overfunded when $q_i > q_i^c$. The fund size q_i is the average AUM over a fund's existing period in the sample. In each group, we count the number of funds with $S_i \geq 0$ and $S_i < 0$ and report the result in Panel A of Table 7. Overall, 17% of the funds (527 out of 3,077) have AUM less than their optimal amounts, and 25% (781 out of 3,077) have AUM more than their optimal amounts, but still within the positive territory of value added if the managers actively manage all the capital entrusted. There are 57% of the funds (1,769 out of 3,077) excessively overfunded, and the managers would destroy value if they actively manage all the capital provided. With the standard neoclassical assumptions in Berk and Green (2004) and Berk and van Binsbergen (2015), rational managers always invest at their optimal amounts. That is, rational managers borrow if investors provide less capital than their optimal amounts and index the excess money when investors provide more capital than the optimum. If the neoclassical assumptions hold, the distribution of the fund which destroyed value in reality ($S_i < 0$) should not depend on which group the fund belongs. The reality, however, challenges this conclusion. Table 7 shows that the conditional distribution of value-destroying managers differs remarkably across three groups. Only 13% of the funds failed to add value for investors in the underfunded group, about 32% destroyed value in the moderately overfunded group, and this fraction dramatically increased to nearly 90% in the excessively overfunded group. As a result, 83% of the value-destroying funds are from the excessively overfunded group.

To ensure that this finding is not driven by funds with little skill, we filter out the funds whose estimated optimal active size is less than $15 million. This excludes 608 funds with 581 of them in the excessively overfunded group and the remaining 27 in the moderately overfunded group. Panel B of Table 7 lists the updated frequency counts on the filtered data and our findings continue to hold. The excessively overfunded group contributes the most to the pool of the value-destroying funds (77%). The funds in this groups are least likely to create value for investors.

Table 7 suggests that managers do not know their own ability. In particular, they tend to

actively manage assets well above their optimal amounts, running the risk of destroying value. This can be further confirmed by the set of regression results in Table 8. We regress the realized value added S_i on either $\hat{V}_i^*$, the maximum value added, or $\hat{V}_i$, the amount of value added when a manager actively manages all the fund AUM. If the managers had always invested at their optimal amounts, we would see a slope estimate of regressing S_i on $\hat{V}_i^*$ close to one. However, that is not the case. Across the full sample and the three subsamples, namely the underfunded, moderately overfunded and excessively overfunded groups, the slope coefficients of $\hat{V}_i^*$ are all significantly lower than 1. The association between S_i and $\hat{V}_i^*$ even turns negative in the excessively overfunded group as shown by the rank correlation of -0.12 and the negative slope coefficient estimate of -1.51. In contrast, the association between S_i and $\hat{V}_i$ is always positive and stronger stronger than that between S_i and $\hat{V}_i^*$ as evidenced by the larger magnitude of the estimated slope coefficients and the higher R-squared values across the full sample and the three subsamples.

Berk, van Binsbergen, and Liu (2017) also document that managers do not know their own limits. Berk, van Binsbergen, and Liu (2017) find that decisions by mutual fund firms to promote or demote a fund manager in terms of increasing or decreasing her AUM lead to an increase in the manager's value added. Berk, van Binsbergen, and Liu (2017) attribute their finding to mutual fund firms having private information about their managers. While they explain the reason that adding capital creates value is that investors failed to provide managers with enough capital to extract the maximum amount of value from markets, Berk, van Binsbergen, and Liu (2017) are relatively silent on why demotion also creates value. According to the neoclassical assumptions, changing capital should not create value when the manager manages assets above her optimal amount of capital. Our result reveals that the suboptimal behavior in managers in that they fail to index the excess amount might be why reducing capital creates value.

6 Investor rationality

Theoretical models in the mutual fund literature typically assume a significant degree of investor sophistication and learning ability. For instance, Berk and Green (2004) posit that investors can learn about the ability of fund managers based on past performance and allocate money to funds with high expected returns. In equilibrium, the risk-adjusted net returns to investors are zero because positive net alphas are competed away by investors, regardless of the skill level of the fund. The insight from Berk and Green's rational model of money management is that rather than teaching us something about managerial skill, the net alpha teaches us something

about the rationality of investors and the competitiveness of markets. In this section, we infer investor rationality using the net returns. We examine whether investor can allocate their money to where it is most productive both in cross section and over the typical lifetime of a fund.

6.1 Cross-sectional investor rationality

The gross alpha is the net alpha plus the fee charged. Fund fees do not vary much over time. If the gross alpha decreases with the fund size, the net alpha also decreases with the fund size:

$$\alpha^n \quad = \quad a^n - b^n \log(q). \tag{26}$$

We expect that the gross alpha and the net alpha share a similar decreasing rate with size, and the intercept a^n in Eq (26) is smaller than the intercept a in Eq (19) by roughly the fee charged. In a rational market, positive net alphas signal positive NPV investment opportunities for investors. By studying the net alpha and its relation to size, we have a gauge of the rationality of investors and the competitiveness of capital markets.

We adopt the method described in the previous section to estimate parameter a^n and b^n fund by fund, that is, to estimate b^n on ten AUM-sorted portfolios and then back out individual a^n fund by fund. Table 9 lists the estimates of b^n in the size deciles. We obtain the evidence of diseconomies of scale in the net fund performance across all ten portoflios. As expected, the estimated b^n values on the net returns are quite close to those obtained using the gross returns. Further, the fund-specific a^n values are smaller in magnitude compared with those obtained using the gross returns, reflecting the fee charged. The difference between the estimated portfolio means of a and a^n can be used as a proxy for the average fee charged in each decile. The average fee is 16 bp per month in the bottom decile, and decreases to 10 bp per month in the top decile. In general, large funds can better mitigate the adverse impact of an increase in fund size and charge lower fund fees.

The expected net alpha for the i-th fund at time t is given by

$$\hat{\alpha}^n_{it} = \hat{a}^n_i - \hat{b}^n_i \log(q_{it-1}). \tag{27}$$

The associated standard error is given in Appendix A.3. We carry out one-tail significance tests on $\hat{\alpha}^n_{it}$ at the 5% level. If the t-statistic is greater than 1.645, $\hat{\alpha}^n_{it}$ is statistically significantly positive. It is regarded as statistically significantly negative if the t-statistic is less than -1.645. Figure 3 provides the fractions of mutual funds with statistically significantly positive (negative)

net alphas over time. The fractions in Panel A are in terms of the number of funds, and those in Panel B are in terms of the fund AUM.

Regarding the number of funds, the fraction of mutual funds with a significantly positive net alpha drifts from 18% in early 1995 down to around 5% in recent years. A noticeable exception in this downward trend is the spike of the positive net alpha fraction during the Global Financial Crisis (GFC, 2007 to 2009) when the industry size tumbled. Bear in mind that we do not control for luck in the cross-sectional significance tests. That is to say, in a given month, 5% of the funds could show up as significantly positive when their net alphas are actually zero. Overall, investors seem to compete away positive net alpha opportunities, and funds that can deliver alpha have become rare animals in recent years. It is a different story at the other end of the spectrum. The fraction of mutual funds with significantly negative expected net alphas stays above 15% most of the time and fluctuates from as low as 4.6% in 2009 to as high as 25% in 1998. Except for the GFC period, the investment management space appears to be over competitive and investors invest too much money with certain funds. Jointly assessing the fractions of fund number and fund AUM, we learn something about the fund characteristics in each group. As Figure 3 shows, the group of funds with positive net alphas are characterized by small funds as the fraction in terms of the AUM is smaller than the fraction in terms of the fund number. The funds in the negative net alpha group are initially marked by large funds before 2004 but convert to average-sized funds afterward.

Given the dynamic nature of the equilibrium and the noisy learning experience investors have, it is unrealistic to have all the funds perfectly comply with the rational expectations hypothesis all the time. As Figure 3 revealed, roughly one quarter to one third of the funds deviate from the rational expectations hypothesis in the sample period.[11] Do investors effectively learn about a^n and b^n and supply capital to where it is most productive? To answer this question, we examine the capital flows in the different types of funds. Each month, a fund receives a flag based on the statistical test on the net alpha $\hat{\alpha}_{it}^n$ indicating whether it is in equilibrium (insignificant net alpha), offers additional value to investors (significantly positive net alpha) or destroys value (significantly negative net alpha). For each year, we compute the annual net capital flows in each fund, and for each category, we report the fraction of funds with positive net capital inflows as the total number of funds in that category.

Figure 4 depicts the fractions of funds with net capital inflows in each category over the period from 1995 to 2014. The zero net alpha fund group serves as a benchmark. The fraction of funds with inflows in this group reflects the overall trend of capital flows to the active management industry. The active management industry experienced capital inflows in

[11] Because we do not control for luck in the cross-sectional significance tests, our fractions overestimate the non-zero net alpha funds.

the early part of our sample, as indicated by the fraction of funds with inflows in the benchmark group above 0.5 before 2004. The fraction of funds with inflows in the benchmark group has been well below 0.5 since 2006. This is consistent with the data compiled by the Investment Company Institute that actively managed domestic equity mutual funds incurred outflows since 2006 (Investment Company Institute, 2015).

If investors' learning is effective, we would expect the fraction of funds with inflows higher in the positive net alpha group and lower in the negative net alpha group relative to the benchmark trend. Figure 4 confirms our conjecture. The surprising observation is that although the values of a_i^n and b_i^n are unknown to investors, they appear to be able to quickly learn and shift their capital to chase positive NPV opportunities. As shown in Figure 4, funds that are able to deliver alpha experience far greater capital inflows than their less successful peers. As we emphasized above, random chance could well play a role in the way we identify funds. Given that the value-adding group as a proportion of the total funds decreased to around 5% after 2011, the rate of the type I error, it is not surprising that this group has indifferent fractions of fund inflows from the benchmark group in recent years. The negative alpha group consistently experiences a lower fraction of capital inflows than the benchmark group. Our finding implies that the provision of capital by investors to the mutual fund industry is competitive and investors are capable of sophisticated learning.

6.2 Investor rationality over a fund's lifetime

In this section, we investigate investors' capital allocation behavior over the lifetime of a fund. Again, we make use of the fund flag based on the statistical tests on $\hat{\alpha}_{it}^n$: 0 (an insignificant net alpha) indicating that investors have put the right amount of money with the i-th fund at time t; 1 (a significantly positive net alpha) indicating that investors have left money on the table; and -1 (a significantly negative net alpha) indicating that investors have given the fund too much money. We run logit regressions to examine how the probability of the net alpha being positive, negative or insignificant evolves over a fund's lifetime.

Panel A of Table 10 provides the logit regression results. The chance that investors have not given a fund enough money (a positive net alpha) significantly decreases over a fund's lifetime as demonstrated by the large negative coefficient for fund age, -0.189 with a t-statistic of -3.17. Meanwhile, the chance that investors have given the fund too much money (a negative net alpha) increases only slightly with fund age as shown by the small positive coefficient of 0.014 with a t-statistic of 0.33. As a result, the probability that investors have put the right amount of money with a fund (an insignificant net alpha) increases as the fund grows as evidenced by the positive albeit not significant coefficient estimate.

As a quick robustness check, we also provide a model-free way to confirm the result. Each month, we assign funds to four groups based on the fund's age: (0, 3], (3, 6], (6, 10] and > 10 years. We calculate the faction of the three categories of the net alpha in each group each month. Panel B of Table 10 shows the average proportions of the three categories of the net alpha in each age group. Consistent with the findings in Panel A, we observe a clear downward trend in the proportion of the significant positive net alpha and a slight upward trend in the proportions of the negative and insignificant net alpha as the fund age increases. This simple method confirms that investors tend to leave money on the table for younger funds. However, the investors' learning allows them to correct this misallocation over time. We also notice that the money flows are sticky in the negative net alpha funds.

Many studies have shown that fund performance deteriorates over a typical fund's lifetime (see, e.g., Yan, 2008, Ferreira, Keswani, Miguel, and Ramos, 2013, Pástor, Stambaugh, and Taylor, 2015). These studies claim the finding implies that older funds are less skilled than younger ones. Rather than teaching us something about managerial skill, this empirical regularity teaches us something about investors. We showed that although investors invest too little with younger funds, they are able to correct this misallocation over time.

7 Conclusions

This paper addresses several important questions in active investment management. We start with the nature of returns to scale in the active mutual fund management industry which is a question with important theoretical and practical implications. Previous empirical attempts at examining the size–performance relationship are hampered by econometric biases associated with the OLS estimates. This paper builds on the work of Pástor, Stambaugh, and Taylor (2015) which constructs bias-free estimates for analyzing size effects. We address a source of misspecification in their empirical strategy that may have biased them against finding evidence of decreasing returns to scale at the fund level. In contrast to Pástor, Stambaugh, and Taylor (2015), who fail to reject the hypothesis of constant returns to scale at the fund level, we provide strong evidence of decreasing returns to scale at the fund level. Assuming a linear relationship between size and alpha, we report an average fund-level decreasing returns to scale parameter of -0.485×10^{-6} with a t-statistic of -2.03. In case that the relation between size and alpha is loglinear, we find an average fund-level decreasing returns to scale parameter of -0.0026 with a t-statistic of -13.32. Using the size-sorted decile portfolios, we show a large amount of individual heterogeneity in decreasing returns to scale. Both the linear and loglinear models produce significantly negative size impact across all deciles. The loglinear model produces more

stable estimates of decreasing returns to scale than the linear model across deciles. Overall, size impact quantified by the loglinear model is smaller than that by the linear model. Although we adopt the loglinear model in making inference in this paper, the most suitable functional form of decreasing returns to scale is still an open question and we leave it for future research.

Once the fund-level decreasing returns to scale are established, the only proper measure of managerial skill should be value added as proposed by BvB. We compute the dollar amount of the value a fund manager actually delivered to investors and compare it with the maximum value she potentially can create. In a neoclassical world with perfectly rational players, no information asymmetries and no other frictions, these two measures should be consistent with each other. They nevertheless differ greatly in reality. We find widespread suboptimal behavior in managers in that they actively manage more than their optimal amount. This suboptimal behavior is shown to be the major source causing the difference between their potential and what they actually delivered.

BvB clarify the role of the net alpha which does not measure skill; instead, the net alpha measures investor performance. The empirical literature on the behavior of mutual fund investors has suggested that investors are often naive and irrational in their investment choices. Yet theorists have typically assumed investor rationality in constructing models of mutual fund flows and performance. In this paper, we provide the first set of systematic empirical evidence by studying the net alpha and its relation to fund size to infer the rationality of investors and the competitiveness of capital markets. Our finding suggests the existence of sophisticated investors who are capable of rational learning. They correctly shift their capital around to explore positive NPV opportunities. As a result, positive NPV investment opportunities are quickly competed away.

Appendix

A.1 Pástor, Stambaugh, and Taylor's (2015) size process

Pástor, Stambaugh, and Taylor (2015) consider the following size process in their simulation studies:

$$\frac{q_{it}}{q_{it-1}} - 1 = \delta + \gamma r_{it} + v_{it}, \tag{28}$$

where q_{it} is the dollar amount of i-fund's AUM at the end of period t. The parameter γ captures the contemporaneous correlation between the fund performance and the fund size.

In the simulations, we choose $\delta = 0.01$ and $std(v) = 0.054$, which are the OLS estimates of these parameter on our data. The point estimate of γ is 0.61. We consider four different

values of γ: 0.4, 0.6, 0.8 and 1.0. Fund returns are simulated from Model (8). We consider four different values of β: 0, -0.001, -0.003 and -0.01. We simulate individual fund fixed-effect α_i from a normal distribution $N(0.016, 0.006)$. The return innovation follows a normal distribution with a zero mean and a standard deviation of 0.026. We construct 10,000 samples of simulated panel data for 300 funds over 100 months. All funds start at \$250 million.

Table 11 presents the simulation results. Under the size process (28), Pástor, Stambaugh, and Taylor (2015) find that RD1 is essentially unbiased, while the OLS estimator is positively biased, and the FE estimator is negatively biased. When it comes to hypothesis testing, the OLS and the FE estimators have distorted sizes, while RD1 has approximately the right size but lacks the power to reject a false null. Our simulation results closely match Pástor, Stambaugh, and Taylor's (2015) findings. In addition, our results show that RD2 dominates RD1 in all criteria considered. RD2 is unbiased, more stable and of smaller RMSEs. More importantly, RD2 not only has the right size but also possesses adequate power to reject the null when the null is false. For example, when $\beta = -0.001$, RD1 rejects the null about 19% to 40% of the time depending on the value of γ. In contrast, RD2 rejects the null about 60% to 87% of the time, and when $\beta = -0.003$, the power increases to more than 90%.

A.2 Robustness check

We assess whether the diseconomies of scale at the fund level are robust to the inclusion of industry size. Pástor, Stambaugh, and Taylor (2015) find statistically significant evidence of decreasing returns to scale at the mutual fund industry level. That is, as the size of the active mutual fund industry increases, the chance of any given fund outperforming its benchmark declines. They suggest this is because a larger industry intensifies competition among active funds, which, in turn, impedes funds' performance. Similarly to Pástor, Stambaugh, and Taylor (2015), we measure the industry size by adding up the fund AUM across all funds in the sample and then dividing by the stock market capitalization. Endogeneity is not an issue for industry size. We consider a regression of the fund return on industry size and use the standard fixed-effects estimator. The estimated slope coefficient is -0.0341 with a robust t-statistic of -2.26, which confirms the existence of the industry-level decreasing returns to scale reported by Pástor, Stambaugh, and Taylor (2015). We then regress the fund return on the fund size and the industry size. Estimates are obtained via the recursive demeaning procedure in which the forward-demeaned fund size is instrumented but the forward-demeaned industry size is not. The estimated coefficient for the industry size is -0.0291 with a t-statistic of -3.83. The estimated coefficient for the fund size is -0.0025 with a t-statistic of -9.98, which is comparable to the estimated size effect of -0.0026 when the fund size is the only regressor. Thus, the fund-level diminishing returns to scale documented in this study are robust to the

inclusion of the mutual fund industry size.

We check the sensitivity of the results to alternative factor models. Following the convention in the mutual fund literature, we consider measuring performance using abnormal returns adjusting for three factor models, namely, the CAPM, the Fama-French three-factor model and the Carhart four-factor model. The risk factor returns are obtained from Ken French's website. Following Chen, Hong, Huang, and Kubik (2004), we sort the funds at the beginning of each month based on the quintile rankings of their previous-month AUM. The portfolio is equally weighted. We then track these five portfolios for one month and use the entire time series of their monthly gross returns to calculate the loadings to the various risk factors (MKT, SMB, HML, MOM). For each month, a fund inherits the loadings of the quintile portfolio to which the fund belongs. This methods allows a fund to have different loadings when it moves from one size quintile to another during a certain month.

Table 12 reports the factor loadings of the five AUM-sorted fund portfolios using various asset pricing models. The loading on the market portfolio is literally one across all five size-sorted mutual fund portfolios regardless of the asset pricing model. Small funds tend to have higher loadings on SMB and HML. In terms of momentum, all five size portfolios seem to be indifferent on momentum and have negligible loading on momentum. The expected fund return is calculated by using the estimated factor loadings with the realized factor returns (including the return on the risk-free asset), and the risk-adjusted return is calculated as the difference between the realized fund return and the expected fund return. The last row of Table 12 presents the scale effects estimated on abnormal returns by the three alternative asset pricing models. The negative impact of size stays literally the same when shifting from the benchmark portfolio adjustment to the CAPM alpha, -0.0027. The two other risk factors, SMB and HML, shrink the scale effect to -0.0020, but it is still highly statistically significant with a t-statistic of -8.26. Adding the stock momentum does not alter the result. In sum, the negative size–performance relation holds under various asset pricing models.

A.3 Standard error of the net alpha prediction

The relationship between the net alpha and the fund size is $\alpha_{it}^n = a_i^n - b_i^n x_{it-1}$, where x_{it-1} is the logarithm of the fund AUM at time $t-1$. The estimate of b_i^n, $\hat{b}_i^n$, is obtained using the panel estimator RD2. The corresponding standard error is described in Section 4.1.2. The parameter a_i^n is estimated by

$$\hat{a}_i^n = \tilde{r}_i + \hat{b}_i^n \tilde{x}_i,$$

where $\tilde{r}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} r_{it}$ and $\tilde{x}_i = \frac{1}{T_i}\sum_{t=0}^{T_i-1} x_{it}$. Thus, the predicted value of the net alpha is:

$$\begin{aligned}\hat{\alpha}_{it}^n &= \hat{a}_i^n - \hat{b}_i^n x_{it-1} \\ &= \tilde{r}_i + \hat{b}_i^n(\tilde{x}_i - x_{it-1}).\end{aligned}$$

The corresponding variance is:

$$\text{var}(\hat{\alpha}_{it}^n) = \text{var}(\tilde{r}_i) + (\tilde{x}_i - x_{it-1})^2\text{var}(\hat{b}_i^n) + 2\text{cov}\{\tilde{r}_i, \hat{b}_i^n(\tilde{x}_i - x_{it-1})\}. \tag{29}$$

We tackle the three items in the variance expression (29) one by one.

Denote $e_{it} = r_{it} - \hat{\alpha}_{it}^n$, the model predictive error of the net alpha for the i-th fund. The variance of the prediction errors can be estimated as:

$$\hat{\sigma}_i^2 = \frac{1}{T_i - 1}\sum_{t=1}^{T_i} e_{it}^2.$$

The first item in Eq. (29) can be estimated by

$$\text{var}(\tilde{r}_i) = \frac{\hat{\sigma}_i^2}{T_i}. \tag{30}$$

Using the estimator of $\text{var}(\hat{b}_i^n)$ provided in Section 4.1.2, the second item $(\tilde{x}_i - x_{it-1})^2\text{var}(\hat{b}_i^n)$ is straightforward. For the third item, note that

$$\hat{b}_i^n = -\left(\sum_{k=1}^{K}\sum_{t=1}^{T_i-1} \bar{x}_{kt-1}^{*'}\bar{x}_{kt-1}^{*}\right)^{-1}\left(\sum_{k=1}^{K}\sum_{t=1}^{T_i-1} \bar{x}_{kt-1}^{*'}\bar{r}_{kt}\right),$$

where K is the number of funds we use to estimate the decreasing returns to scale parameter b_i^n. For the i-th fund, denote the vector of the benchmark-adjusted net return as $\boldsymbol{r}_i$ and a forward-demeaned transformation matrix M_i as:

$$\boldsymbol{r}_i = \begin{pmatrix} r_{i1} \\ r_{i2} \\ \vdots \\ r_{iT_i} \end{pmatrix} \quad \text{and} \quad M_i = \begin{pmatrix} \frac{T_i-1}{T_i} & -\frac{1}{T_i} & -\frac{1}{T_i} & \cdots & -\frac{1}{T_i} & -\frac{1}{T_i} \\ 0 & \frac{T_i-2}{T_i-1} & -\frac{1}{T_i-1} & \cdots & -\frac{1}{T_i-1} & -\frac{1}{T_i-1} \\ 0 & 0 & \frac{T_i-3}{T_i-2} & \cdots & -\frac{1}{T_i-2} & -\frac{1}{T_i-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{2} & -\frac{1}{2} \end{pmatrix}.$$

The vector of forward-demeaned net returns for the i-th fund is $\bar{\boldsymbol{r}}_i = M_i \boldsymbol{r}_i$. Thus, the estimator $\hat{b}_i^n$ can be rewritten as

$$\hat{b}_i^n = \theta \left(\sum_{k=1}^{K} \bar{x}_{kt-1}^{*'} M_k \boldsymbol{r}_k \right),$$

where $\theta = -\left(\sum_{k=1}^{K} \sum_{t=1}^{T_i - 1} \bar{x}_{kt-1}^{*'} \bar{x}_{kt-1}^{*} \right)^{-1}$ is a constant. Note that $\tilde{r}_i = \frac{1}{T_i} I' \boldsymbol{r}_i$. Assuming that the funds are independent of each other, we have

$$\begin{aligned}
\operatorname{cov}\{\tilde{r}_i, \hat{b}_i^n (\tilde{x}_i - x_{it-1})\} &= \operatorname{cov}\left\{ \frac{1}{T_i} I' \boldsymbol{r}_i, \theta \left(\sum_{k=1}^{K} \bar{x}_{kt-1}^{*'} M_i \boldsymbol{r}_i \right) \right\} \\
&= \operatorname{cov}\left\{ \frac{1}{T_i} I' \boldsymbol{r}_i, \theta \bar{x}_{it-1}^{*'} M_i \boldsymbol{r}_i \right\} \\
&= \operatorname{cov}\left\{ \frac{1}{T_i} I' \boldsymbol{\epsilon}_i, \theta \bar{x}_{it-1}^{*'} M_i \boldsymbol{\epsilon}_i \right\} \\
&= 0. \qquad (31)
\end{aligned}$$

Thus, an estimator for the variance expression (29) is

$$\operatorname{var}(\hat{\alpha}_{it}^n) = \frac{\hat{\sigma}_i^2}{T_i} + (\tilde{x}_i - x_{it-1})^2 \operatorname{var}(\hat{b}_i^n). \qquad (32)$$

References

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

Berk, J. B., van Binsbergen, J. H., 2015. Measuring skill in the mutual fund industry. Journal of Financial Economics 118, 1–20.

Berk, J. B., van Binsbergen, J. H., 2017. Mutual funds in equilibrium. Forthcoming: Annual Review of Financial Economics .

Berk, J. B., van Binsbergen, J. H., Liu, B., 2017. Matching capital and labor. Forthcoming: Journal of Finance .

Chen, J., Hong, H., Huang, M., Kubik, J. D., 2004. Does fund size erode mutual fund performance? the role of liquidity and organization. American Economic Review 94, 1276–1302.

Cremers, M., Petajisto, A., Zitzewitz, E., 2013. Should benchmark indices have alpha? revisiting performance evaluation. Critical Finance Review 2, 1–48.

Elton, E. J., Gruber, M. J., Blake, C. R., 2012. Does mutual fund size matter? the relationship between size and performance. Review of Asset Pricing Studies 2, 31–55.

Evans, R. B., 2010. Mutual fund incubation. Journal of Finance 65, 1581–1611.

Ferreira, M. A., Keswani, A., Miguel, A. F., Ramos, S. B., 2013. The determinants of mutual fund performance: A cross-country study. Review of Finance 2, 483–525.

Investment Company Institute, 2015. 2015 investment company fact book. Investment Company Institute, D.C.

Moon, H. R., Phillips, P. C., 2000. Estimation of autoregressive roots near unity using panel data. Econometric Theory 16, 927–997.

Pástor, Ľ., Stambaugh, R. F., Taylor, L. A., 2015. Scale and skill in active management. Journal of Financial Economics 116, 23–45.

Yan, X. S., 2008. Liquidity, investment style, and the relation between fund size and fund performance. Journal of Financial and Quantitative Analysis 43, 741–767.

Table 1: Descriptive Statistics

	No. fund/ month	mean	sd	Percentiles 1%	25%	50%	75%	99%
Gross ret (%)	410,284	0.85	5.33	-15.24	-1.90	1.36	4.01	13.23
Benchmark-adj gross ret (%)	410,284	0.04	2.68	-8.07	-0.91	0.00	0.93	8.52
Net ret (%)	410,284	0.75	5.33	-15.32	-2.00	1.26	3.91	13.11
Benchmark-adj net ret (%)	410,284	-0.06	2.68	-8.20	-1.01	-0.09	0.83	8.40
Expense ratio (%)	410,284	0.10	0.04	0.02	0.08	0.10	0.12	0.21
AUM ($mil)	410,284	1,493	5679	8	76	261	954	22,001
FundAge	410,284	10.69	9.91	0.33	4.33	8.42	13.92	52.75
Turnover(%)	360,134	80.42	73.56	2.49	33.00	62.00	104.00	355.00

Notes: This table presents descriptive statistics on the fund sample. The sample period is from January 1995 to December 2014. The unit of observation is the fund/month. All returns and expense ratios are monthly figures. Benchmark-adjusted gross return and benchmark-adjusted net return are constructed by subtracting the Morningstar designated benchmark index return from the fund's gross return and net return. *AUM* is the fund's real assets under management in millions of dollars (base year is 2014). *FundAge* is the time in years since the fund's inception date. *Turnover* is defined as the minimum of aggregate purchases and sales divided by the average annual AUM in percentage.

Table 2: Simulation exercise comparing estimators

$\beta(\times 10^3)$	OLS				FE				RD1				RD2			
	$\rho_2 = 0.9$	$\rho_2 = 1$	$\rho_2 = 1.1$	$\rho_2 = 1.2$	$\rho_2 = 0.9$	$\rho_2 = 1$	$\rho_2 = 1.1$	$\rho_2 = 1.2$	$\rho_2 = 0.9$	$\rho_2 = 1$	$\rho_2 = 1.1$	$\rho_2 = 1.2$	$\rho_2 = 0.9$	$\rho_2 = 1$	$\rho_2 = 1.1$	$\rho_2 = 1.2$
Panel A: Root mean square error of $\hat{\beta}$ $(\times 10^3)$																
0	1.31	1.36	1.38	1.40	0.44	0.44	0.45	0.45	0.48	0.44	0.41	0.39	0.23	0.22	0.20	0.20
-1	1.59	1.68	1.75	1.82	0.61	0.63	0.64	0.65	0.70	0.66	0.61	0.58	0.31	0.30	0.28	0.27
-3	2.17	2.36	2.54	2.69	1.07	1.17	1.26	1.36	3.84	3.58	3.77	3.85	0.71	0.74	0.76	0.78
-10	1.83	1.85	1.84	1.84	0.63	0.60	0.57	0.54	1.03	0.85	0.74	0.66	0.33	0.30	0.27	0.25
Panel B: Bias of $\hat{\beta}$ $(\times 10^3)$																
0	1.29	1.33	1.37	1.39	-0.41	-0.41	-0.41	-0.42	0.01	0.00	-0.01	-0.01	0.00	0.01	0.00	-0.00
-1	1.56	1.66	1.74	1.80	-0.57	-0.59	-0.61	-0.62	0.01	0.00	0.00	-0.01	-0.01	-0.00	0.00	-0.00
-3	2.15	2.34	2.52	2.67	-1.03	-1.13	-1.23	-1.33	-0.01	0.02	-0.01	-0.02	-0.00	0.01	-0.02	0.02
-10	1.80	1.82	1.82	1.82	-0.57	-0.55	-0.53	-0.50	-0.00	-0.01	0.02	0.02	0.00	0.00	-0.00	0.00
Panel C: Standard deviation of $\hat{\beta}$ $(\times 10^3)$																
0	0.23	0.22	0.21	0.21	0.17	0.17	0.16	0.15	0.48	0.44	0.41	0.39	0.23	0.22	0.20	0.20
-1	0.27	0.27	0.25	0.25	0.21	0.20	0.20	0.19	0.70	0.66	0.61	0.58	0.31	0.30	0.28	0.27
-3	0.35	0.35	0.34	0.33	0.29	0.30	0.29	0.30	3.84	3.58	3.77	3.85	0.71	0.74	0.76	0.78
-10	0.32	0.30	0.28	0.27	0.24	0.23	0.21	0.20	1.03	0.85	0.74	0.66	0.33	0.30	0.27	0.25
Panel D: Fraction rejecting the null hypothesis $(\beta = 0)$																
0	1.00	1.00	1.00	1.00	0.65	0.70	0.74	0.78	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
-1	0.84	0.91	0.96	0.98	1.00	1.00	1.00	1.00	0.30	0.33	0.37	0.41	0.90	0.92	0.94	0.96
-3	0.92	0.82	0.65	0.49	1.00	1.00	1.00	1.00	0.17	0.16	0.15	0.13	0.98	0.98	0.97	0.96
-10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Notes: This table reports simulation results. The details of the simulation are described in Section 4.2. 10,000 samples of fund size and return pairs are simulated. Simulated returns follow $r_{it} = \alpha_i + \beta x_{it-1} + \epsilon_{it}$, and sizes follow $x_{it} = \phi + x_{it-1} + \rho_1 B_{it} + \rho_2 r_{it} + \zeta_{it}$. In each sample, we estimate β using the OLS, FE, RD1 and RD2. Four performance measures, root mean square error, bias, standard deviation and fraction of samples in which the null of $\beta = 0$ is rejected are reported. These measures are calculated across the 10,000 simulated samples.

Table 3: Relation between size and fund performance

	Gross performance				Net performance			
	OLS	FE	RD1	RD2	OLS	FE	RD1	RD2
Panel A: Dollar fund AUM								
Slope ($\times 10^6$)	-0.010	-0.136	-0.113	-0.485	0.001	-0.132	-0.108	-0.479
	(-0.76)	(-9.05)	(-0.34)	(-2.03)	(0.09)	(-8.84)	(-0.32)	(-1.99)
1st stage intercept	–	–	0	-292.1	–	–	0	–292.1
	–	–	–	(-54.15)	–	–	–	(-54.15)
1st stage R2 (%)	–	–	0.19	10.04	–	–	0.19	10.04
Panel B: Logarithm of fund AUM								
Slope	-0.0001	-0.0016	0.0491	-0.0026	-0.00002	-0.0015	0.0466	-0.0026
	(-1.55)	(-27.78)	(0.58)	(-13.32)	(-0.28)	(-26.64)	(0.58)	(-13.00)
1st stage intercept	–	–	0	-0.8242	–	–	0	-0.8242
	–	–	–	(-191.1)	–	–	–	(-191.1)
1st stage R2 (%)	–	–	0.01	8.78	–	–	0.01	8.78

Notes: This table shows regression parameter estimates of monthly fund return on lagged fund size. Two types of fund return, benchmark-adj gross return (Gross performance) and benchmark-adj net return (Net performance), are used. For the RD estimators, we report two additional statistics associated with the first-stage regression: the intercept estimate and R-squared value. Panel A lists the estimation result using lagged dollar fund AUM, while panel B reports the result for logarithm of lagged fund AUM. The slope estimates in Panel A are multiplied by 10^6 to make them easier to read. Heteroskedasticity-robust t-statistics clustered by style $\times$ month are reported in the parentheses. The robust standard errors are cluster by fund in the RD specifications.

Table 4: Decreasing returns to scale in size deciles.

	Ave. AUM	Dollar fund AUM			Logarithm of fund AUM		
	($mil)	Slope ($\times 10^6$)	t-value	Impact (bp)	Slope	t-value	Impact (bp)
1	20	-167.578	-4.09	-33.5	-0.0030	-2.37	-21.0
2	41	-60.670	-4.80	-24.9	-0.0024	-4.44	-16.9
3	71	-30.070	-4.27	-21.3	-0.0021	-3.24	-14.5
4	112	-15.017	-5.95	-16.8	-0.0018	-3.83	-12.7
5	174	-15.255	-6.45	-26.5	-0.0017	-6.58	-11.6
6	270	-6.959	-6.80	-18.8	-0.0015	-5.71	-10.7
7	432	-6.178	-8.02	-26.7	-0.0015	-7.45	-10.1
8	689	-3.643	-8.48	-25.1	-0.0013	-7.84	-8.8
9	1266	-1.655	-9.79	-21.0	-0.0012	-7.99	-8.3
10	3828	-0.192	-2.13	-7.3	-0.0011	-8.22	-7.9
F-test p-value		0.000			0.000		

Notes: We sort our sample of mutual funds based on the decile rankings of their average AUM. Column two displays the average fund AUM in each decile. We consider two functional forms of decreasing returns to scale, $r_{it} = \alpha_i + \beta q_{it-1} + \epsilon_{it}$ and $r_{it} = \alpha_i + \beta \log(q_{it-1}) + \epsilon_{it}$, where q_{it-1} is the lagged dollar fund AUM. In each AUM sorted portfolio, we estimate β using the panel estimator RD2 for both models and report the associated t-statistics. In each decile, we consider the impact of fund size doubling on fund gross performance.

Table 5: Fund by fund a and b parameters in gross alpha.

	b	a (%)						
		mean	s.e.	1%	25%	50%	75%	99%
1	0.0030	0.66	0.51	-1.02	0.47	0.74	0.91	1.81
2	0.0024	0.75	0.39	-0.58	0.61	0.82	0.95	1.51
3	0.0021	0.79	0.30	-0.32	0.68	0.80	0.93	1.49
4	0.0018	0.79	0.28	-0.09	0.69	0.80	0.93	1.54
5	0.0017	0.83	0.25	0.26	0.70	0.83	0.95	1.58
6	0.0015	0.85	0.21	0.31	0.74	0.85	0.95	1.37
7	0.0015	0.88	0.22	0.25	0.78	0.88	1.00	1.56
8	0.0013	0.87	0.22	0.24	0.77	0.86	0.99	1.42
9	0.0012	0.89	0.17	0.40	0.80	0.88	0.98	1.30
10	0.0011	1.01	0.18	0.63	0.89	0.98	1.14	1.49

Notes: Assume the functional form of decreasing gross alpha to scale to be $\alpha^g = a - b \log(q)$. We sort our sample of mutual funds based on the decile rankings of their average AUM. In each AUM sorted group, an estimate of b is obtained using the panel estimator RD2. The fund by fund estimates of the parameter a are obtained by (23). The summary statistics of the estimated a are presented in each decile group.

Table 6: Value added.

	S_i	$\hat{V}_i^*$
Cross-sectional weighted mean ($\bar{S}_W$ and $\bar{V}_W^*$)	0.26	2.90
Standard error of the mean	0.11	0.38
t-statistic	2.41	7.68
Cross-sectional mean ($\bar{S}$ and $\bar{V}^*$)	0.01	2.18
Standard error of the mean	0.09	0.33
t-statistic	0.08	6.52
1st percentile	-9.93	0.00
5th percentile	-3.15	0.00
10th percentile	-1.44	0.01
50th percentile	-0.06	0.11
90th percentile	0.92	2.00
95th percentile	2.01	6.26
99th percentile	12.77	30.10
Percent with less than zero	61.03%	0
Total number of funds	3077	

Notes: This table reports both realized value added (S_i) and optimal value added ($\hat{V}_i^*$). For each fund, we estimate the average monthly value added, both S_i and $\hat{V}_i^*$. The cross-sectional mean, standard error of mean, t-statistic and percentiles are the statistical properties of this distribution. Percent with less than zero is the fraction of the distribution that has value added estimates less than zero. The cross-sectional weighted mean, standard error of the weighted mean and t-statistic are computed by weighting by the number of periods the fund exists, that is, they are the statistical properties of $\bar{S}_W$ and $\bar{V}_W^*$ defined by Eq. (25). The numbers are reported in Y2014 $ millions per month.

Table 7: Two-way relative frequency table.

	$S_i \geqslant 0$	$S_i < 0$	Total
Panel A: Full sample			
Excessively Overfunded	210	1559	1769
Moderately Overfunded	532	249	781
Underfunded	457	70	527
Total	1199	1878	3077
Panel B: Subset of sample with $q_i^* > 15$			
Excessively Overfunded	184	1004	1188
Moderately Overfunded	524	230	754
Underfunded	457	70	527
Total	1165	1304	2469

Notes: We divide funds into three groups: the underfunded ones with $q_i < q_i^*$, the moderately overfunded ones with $q_i \subset (q_i^*, q_i^c]$, and the excessively overfunded ones with $q_i > q_i^c$. In each group, we count the number of funds with $S_i \geq 0$ and $S_i < 0$. Panel A lists the result on the full mutual fund samples we collect, and Panel B is for the subsample of funds with the estimated optimal active amount greater than 15 millions.

Table 8: Relation among S_i, $\hat{V}_i^*$ and $\hat{V}_i$.

	Full Sample	Underfunded	Moderately Overfunded	Excessively Overfunded
	Panel A: Relation between S_i and $\hat{V}_i^*$			
Intercept	-0.32	-0.76	-0.15	-0.63
	(-3.91)	(-2.55)	(-1.72)	(-8.48)
Slope	0.15	0.24	0.64	-1.51
	(24.41)	(15.99)	(18.92)	(-15.53)
R^2	0.28	0.52	0.36	0.10
Rank cor	0.35	0.51	0.28	-0.12
	Panel B: Relation between S_i and $\hat{V}_i$			
Intercept	-0.11	-0.24	-0.14	-0.28
	(-2.58)	(-1.35)	(-1.07)	(-6.04)
Slope	0.71	0.73	0.91	0.76
	(68.22)	(48.19)	(19.21)	(35.82)
R^2	0.79	0.87	0.37	0.64
Rank cor	0.69	0.56	0.33	0.72

Notes: This table reports the coefficient estimates of regressing the realized value added S_i on $\hat{V}_i^*$, which is the optimal value added, or $\hat{V}_i$, which is the value added when a manager actively manages all the fund AUM. We run regressions on the full sample and three sub-samples, namely, the underfunded funds with $q_i < q_i^*$, the moderately overfunded ones with $q_i \subset (q_i^*, q_i^c]$, and the excessively overfunded ones with $q_i > q_i^c$. Heteroskedasticity-robust t-statistics are reported in the parentheses. The top panel is for the association between S_i and $\hat{V}_i^*$, and the bottom panel is for the association between S_i and $\hat{V}_i$. The last row of each panel displays rank correlation values between the regressor and the response.

Table 9: Fund by fund a^n and b^n parameters in net alpha.

	b^n		a^n (%)						
	est.	t-value	mean	s.e.	1%	25%	50%	75%	99%
1	0.0029	2.51	0.50	0.52	-1.17	0.30	0.58	0.76	1.66
2	0.0024	4.33	0.61	0.38	-0.72	0.49	0.69	0.81	1.37
3	0.0020	3.04	0.66	0.31	-0.49	0.55	0.68	0.80	1.35
4	0.0018	3.71	0.67	0.28	-0.25	0.57	0.68	0.81	1.38
5	0.0016	6.41	0.70	0.25	0.07	0.57	0.70	0.81	1.45
6	0.0015	5.47	0.73	0.21	0.20	0.61	0.73	0.83	1.25
7	0.0014	7.22	0.75	0.22	0.13	0.65	0.75	0.86	1.38
8	0.0012	7.59	0.74	0.21	0.10	0.66	0.75	0.85	1.26
9	0.0011	7.77	0.78	0.17	0.24	0.67	0.77	0.85	1.16
10	0.0011	7.15	0.91	0.18	0.50	0.77	0.88	1.01	1.36

Notes: Assume the functional form of decreasing net alpha to scale to be $\alpha^n = a^n - b^n \log(q)$. We sort our sample of mutual funds based on the decile rankings of their average AUM. In each AUM sorted group, an estimate of b^n is obtained using the panel estimator RD2. The estimated value and the associated heteroskedasticity-robust t-statistics of b^n are reported in the columns two and three. The fund by fund estimates of the parameter a^n are obtained and its summary statistics are presented for each fund group.

Table 10: Relation between net alpha and fund age.

Panel A: Logit Regression

	Positive	Negative	Insignificant
FundAge	-0.189	0.014	0.064
	(-3.17)	(0.33)	(1.46)

Panel B: Average proportion

	Age Group			
	(0, 3]	(3, 6]	(6, 10]	>10
Positive(%)	13.47	7.54	4.86	4.87
Negative(%)	14.38	17.96	17.52	17.58
Insignificant(%)	72.15	74.50	77.62	77.55

Note: This table shows the relation between net alpha and fund age. Panel A presents the results from logit regressions with fund fixed effects. *FundAge* is the number of years since the fund's inception date. *Positive* refers to statistically significantly positive net alpha, *Negative* refers to statistically significantly negative net alpha, and *Insignificant* refers to statistically insignificant net alpha. Panel B lists the average factions of mutual funds with positive/negative/insignificant net alpha in four age-sorted groups.

Table 11: Simulation exercise comparing estimators under Pástor, Stambaugh, and Taylor's (2015) size process

$\beta(\times 10^3)$	OLS				FE				RD1				RD2			
	$\rho_2 = 0.9$	$\rho_2 = 1$	$\rho_2 = 1.1$	$\rho_2 = 1.2$	$\rho_2 = 0.9$	$\rho_2 = 1$	$\rho_2 = 1.1$	$\rho_2 = 1.2$	$\rho_2 = 0.9$	$\rho_2 = 1$	$\rho_2 = 1.1$	$\rho_2 = 1.2$	$\rho_2 = 0.9$	$\rho_2 = 1$	$\rho_2 = 1.1$	$\rho_2 = 1.2$
Panel A: Root mean square error of $\hat{\beta}$ $(\times 10^3)$																
0	2.06	2.31	2.36	2.29	0.64	0.66	0.64	0.61	0.76	0.59	0.49	0.42	0.38	0.31	0.26	0.23
-1	2.46	2.94	3.15	3.22	0.80	0.91	0.96	0.97	0.92	0.77	0.67	0.59	0.45	0.39	0.35	0.32
-3	3.37	4.57	5.36	5.82	1.34	1.91	2.39	2.78	1.80	1.85	1.91	1.91	0.74	0.79	0.85	0.91
-10	3.92	3.61	3.11	2.73	1.65	1.13	0.85	0.71	3.26	1.10	0.73	0.56	0.95	0.49	0.35	0.28
Panel B: Bias of $\hat{\beta}$ $(\times 10^3)$																
0	2.02	2.28	2.34	2.28	-0.57	-0.60	-0.60	-0.58	0.00	-0.01	0.00	0.00	-0.01	0.01	0.00	0.00
-1	2.41	2.91	3.13	3.20	-0.72	-0.86	-0.92	-0.94	-0.01	0.00	0.01	0.00	0.00	-0.01	0.00	0.00
-3	3.32	4.53	5.34	5.81	-1.25	-1.85	-2.33	-2.74	0.02	0.02	0.00	0.03	-0.01	0.00	0.00	0.01
-10	3.85	3.57	3.08	2.71	-1.56	-1.07	-0.80	-0.66	0.02	-0.02	0.00	0.00	0.00	0.00	0.01	0.00
Panel C: Standard deviation of $\hat{\beta}$ $(\times 10^3)$																
0	0.42	0.35	0.31	0.27	0.31	0.26	0.23	0.20	0.76	0.59	0.49	0.42	0.38	0.31	0.26	0.23
-1	0.48	0.42	0.38	0.34	0.35	0.31	0.29	0.26	0.92	0.77	0.67	0.59	0.45	0.39	0.35	0.32
-3	0.62	0.58	0.54	0.48	0.48	0.47	0.47	0.47	1.80	1.85	1.91	1.91	0.74	0.79	0.85	0.91
-10	0.70	0.50	0.39	0.33	0.54	0.38	0.30	0.25	3.26	1.10	0.73	0.56	0.95	0.49	0.35	0.28
Panel D: Fraction rejecting the null hypothesis $(\beta = 0)$																
0	1.00	1.00	1.00	1.00	0.45	0.63	0.76	0.83	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05
-1	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.19	0.26	0.33	0.40	0.60	0.72	0.81	0.87
-3	0.34	0.93	1.00	1.00	1.00	1.00	1.00	1.00	0.39	0.37	0.36	0.36	0.98	0.96	0.94	0.91
-10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.86	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Notes: This table reports simulation results. 10,000 samples of fund size and return pairs are simulated. Fund returns are simulated from the model $r_{it} = \alpha_i + \beta x_{it-1} + \epsilon_{it}$ and fund sizes are from the model $\frac{q_{it}}{q_{it-1}} - 1 = \delta + \gamma r_{it} + v_{it}$. In each sample, we estimate β using the OLS, FE, RD1 and RD2. Four performance measures, root mean square error, bias, standard deviation and fraction of samples in which the null of $\beta = 0$ is rejected are reported. These measures are calculated across the 10,000 simulated samples.

Table 12: Alternative asset pricing models – factor loadings and size-performance relation.

Panel A: Risk factor loadings											
	CAPM		3-Factor model				4-Factor model				
Portfolio	α (%)	MKT	α (%)	MKT	SMB	HML	α (%)	MKT	SMB	HML	MOM
1 (small)	0.14	1.00	0.09	0.97	0.21	0.11	0.09	0.97	0.22	0.11	-0.00
2	0.09	1.01	0.05	0.99	0.22	0.10	0.04	0.99	0.22	0.10	0.01
3	0.07	1.03	0.04	0.99	0.23	0.07	0.03	1.00	0.23	0.08	0.01
4	0.02	1.04	-0.01	1.00	0.21	0.04	-0.02	1.01	0.21	0.05	0.01
5 (large)	0.01	1.02	-0.01	1.00	0.12	0.03	-0.02	1.00	0.12	0.03	0.01
Panel B: Size performance relation											
LOGAUM	-0.0027 (-12.42)		-0.0020 (-8.26)				-0.0020 (-8.26)				

Notes: Panel A of the table reports the loadings of the five equally weighted AUM-sorted fund portfolios on various risk factors. *MKT*, *SMB*, *HML* and *MOM* represent market, size, book-to-maker and momentum factors respectively. Panel B reports size effect estimated under different asset pricing models by the RD2 estimator. Robust t-statistics are reported in the parentheses.

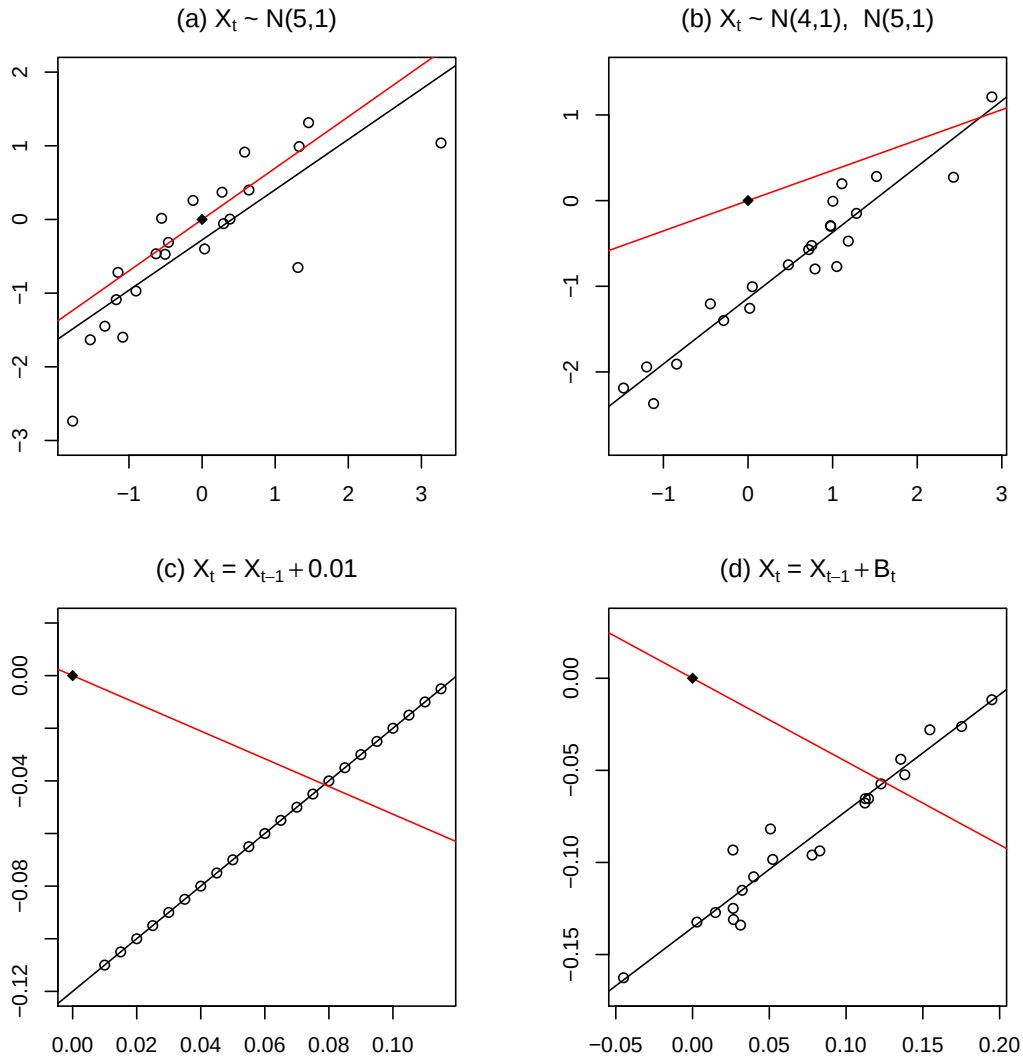

Figure 1: **Relation between forward- and backward-demeaned fund size** This figure plots the forward-demeaned fund size $\bar{x}_t$ (y-axis) versus the backward-demeaned fund size $\underline{x}_t$ (x-axis) of the hypothetical fund size processes. We consider four simple processes for the log fund AUM: a) x_t follows a normal distribution $N(5,1)$, a scenario with a constant mean; b) the first half of the fund size observations follows $N(4,1)$ and the second half follows $N(5,1)$, a scenario with the expected fund size being a step function; c) $x_t = x_{t-1} + 0.01$, a scenario in which the fund size grows at a constant rate of 1% per month; and d) $x_t = x_{t-1} + B_t$, a scenario in which the fund size grows with a market index B_t. We generate twenty-four fund sizes from each of the four scenarios. The benchmark return B_t is taken as the monthly price returns on the S&P 500 from 2012 to 2013. The circles indicate twenty-two pairs of $(\underline{x}_t, \bar{x}_t)$. The black line is the model fitting of $\bar{x}_t = \psi + \rho \underline{x}_t + v_t$, and the red line is the model fitting of $\bar{x}_t = \rho \underline{x}_t + v_t$. The red line is forced to cross the origin (the solid diamond point).

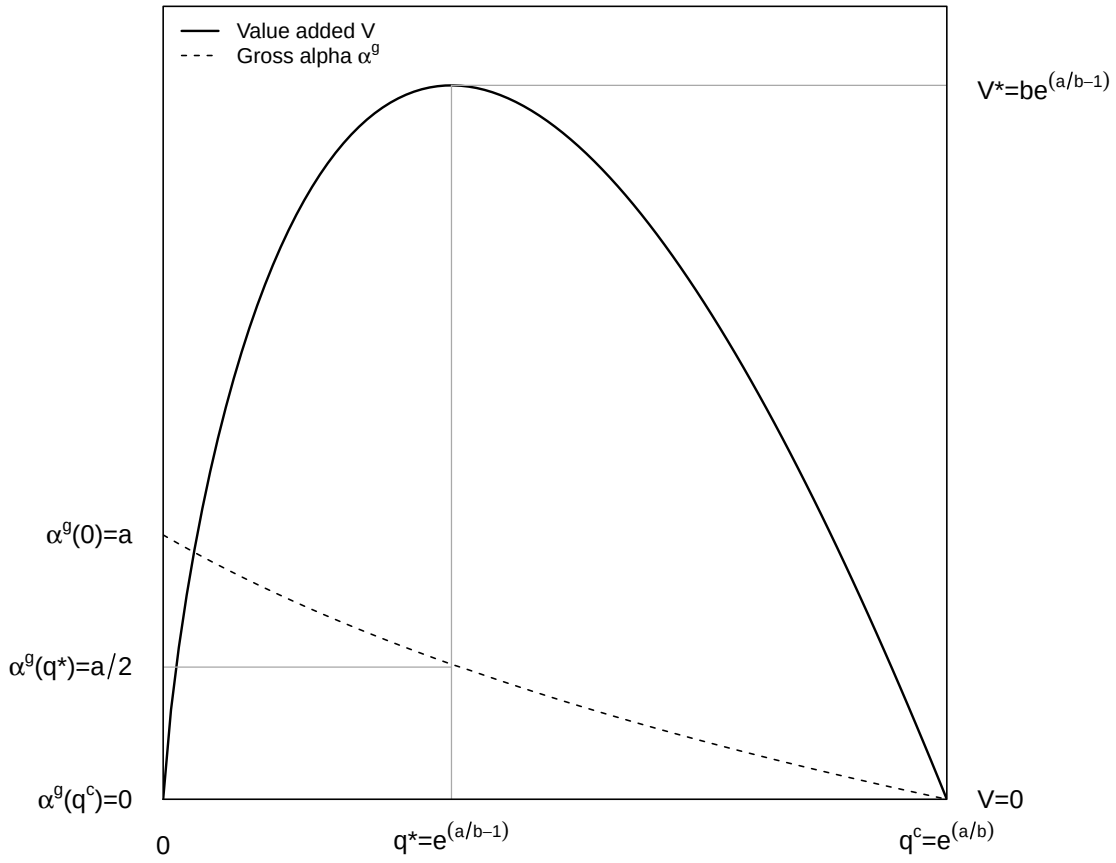

Figure 2: **Value added and gross alpha** This figure displays the value added and the gross alpha as functions of the fund AUM (q). The dashed line plots the gross alpha, $\alpha^g(q) = a - b\log(q)$, and the solid line is the value added, $V(q) = q\alpha^g(q)$. q^* is the optimal amount under active management.

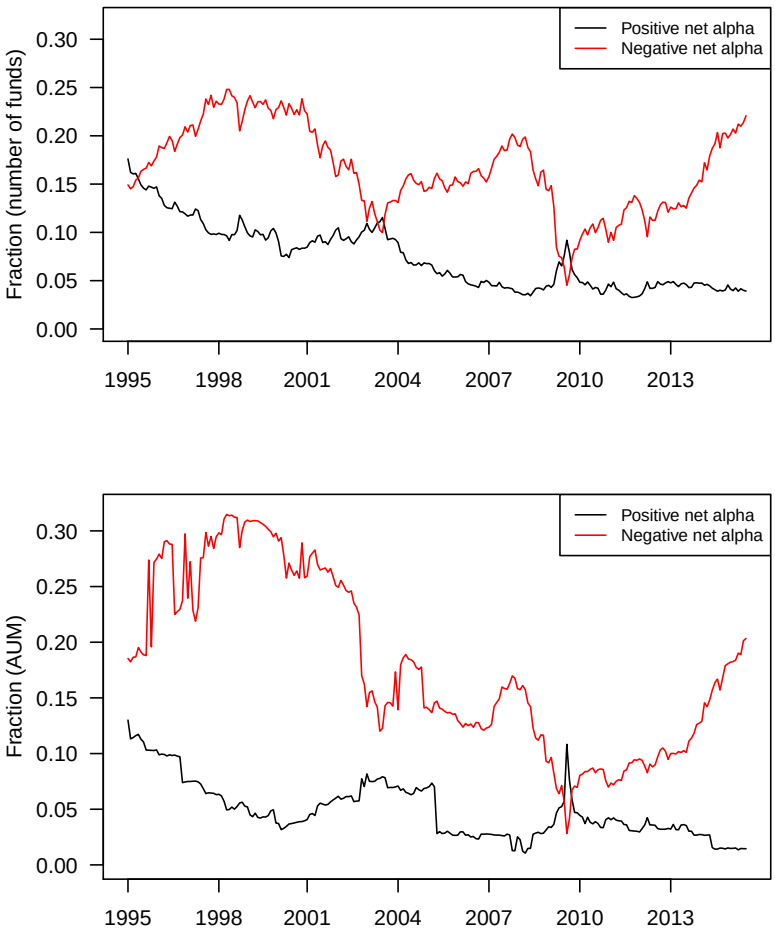

Figure 3: **Fraction of mutual funds with significant net alpha** This graph shows the fraction of mutual funds with a statistically significantly positive net alpha (black line) and the fraction with a statistically significantly negative net alpha (red line). The top panel displays the number of mutual funds with significantly positive (negative) net alpha as a fraction of the total number of existing mutual funds in a given month. The bottom panel displays the assets under management (AUM) of mutual funds with a significantly positive (negative) net alpha as a fraction of the total AUM of all existing mutual funds in a given month.

Figure 4: **Fraction of mutual funds with net capital inflows** This graph displays the number of funds with net capital inflows as a fraction of the total number of funds in three categories, namely, funds with a zero net alpha (green line), funds with a statistically significant positive net alpha (black line) and funds with a statistically significant negative net alpha (red line).