Mutual Fund Flows and Performance in (Imperfectly) Rational Markets?*

Nikolai Roussanov†, Hongxun Ruan‡, and Yanhao Wei §

January 1, 2020

Abstract

Does the observed relationship between mutual fund flows and recent performance represent irrational "return chasing" or rational learning about unobserved fund manager skill? We estimate a structural model of investor beliefs implicit in the fund flows and compare it with the rational Bayesian benchmark that is based on past performance. Our estimates imply that investors are more optimistic about fund manager's average skill than warranted by the historical data. They over-weight recent performance in a manner consistent with models based on the representativeness heuristic, yet respond slowly to changes in these beliefs, consistent with limited attention and/or informational frictions.

Keywords: mutual fund performance, flow-performance relationship, Bayesian learning, representativeness

*We benefited from comments and suggestions by audiences at CKGSB, Hanqing, TAMU and Wharton. Authors are listed in alphabetical order.

†The Wharton School, University of Pennsylvania and NBER

‡Guanghua School of Management, Peking University

§Marshall School of Business, University of Southern California

1 Introduction

Investor flows in and out of mutual funds respond to past fund performance. Conventional wisdom has long held that this flow-performance relationship represents irrational "return chasing" by investors, since past performance does not appear to reliably predict future performance. Berk and Green (2004) (hereafter BG) upended this consensus by pointing out that rational Bayesian learning about unobserved fund manager skill is consistent with a positive flow-performance relationship, while decreasing returns to scale imply a lack of persistence in observed performance. Since both views are *qualitatively* consistent with the empirical evidence, distinguishing between the two is a *quantitative* challenge (e.g., as argued by Cochrane, 2013), which we undertake in this paper.

We proceed in two steps. First, we estimate the optimally filtered dynamics of beliefs of skills about the entire cross-section of U.S. equity mutual funds, together with the shape of returns to scale and the prior beliefs about fund manager's skill, based on their historical *performance* data.[1] In this step, our estimation does not make any assumption on investor behavior driving mutual fund size or flows. As a result, the estimates serve as a rational benchmark that reflects the econometrician's best estimate of manager's skill at a given point in time. We uncover substantial variation in latent manager skill, which is persistent over time but subject to fairly steep decreasing returns to scale.

Second, we show that fund flows respond very weakly to variation in the estimated rational beliefs. At the same time, a measure of misallocation - the difference between actual fund size and the "efficient" fund size implied by the BG model (given the estimated beliefs) strongly predicts subsequent fund performance. Funds that are "too small" relative to their "efficient" fund size subsequently outperform, and vice versa. We proceed to estimate the parameters of the belief process that best fit the observed relationship between past performance and fund flows (assuming standard Bayesian updating of those beliefs). Our estimation reveals that the "average" mutual fund investor is substantially more optimistic about the underlying fund manager skill than is warranted by the historical performance data. Consistent with the "return chasing" view, investor flows appear to respond too strongly to recent performance relative to more distant historical performance, suggesting either mistaken beliefs about mean reversion in manager skill and the relative role of skill and luck in generating performance, or a type of "recency" bias in beliefs, consistent with models based on the representativeness heuristic of Kahnemann and Tversky.[2] At the same time, flows adjust very sluggishly towards these implied beliefs, suggesting

[1] This generalizes the estimation approach we follow in Roussanov et al. (2018).

[2] The representativeness heuristic was introduced by Kahneman and Tversky (1972a), Tversky and Kahneman (1974), Tversky and Kahneman (1983) and can be used as an organizing principle for explaining several related biases in probabilistic decision making that imply over-inference from small samples and excessive reliance on the most recent or salient observations at the expense of prior beliefs or "base rates," as detailed in a survey chapter by Benjamin (2018). Rabin (2002), Rabin and Vayanos (2010), Gennaioli and Shleifer (2010), Bordalo et al. (2017), Bordalo et al. (2018) develop theoretical models based on this principle that apply to various aspects of

a combination of limited attention and information/search frictions might play an important role.

Specifically, the first step of our exercise starts by considering that the BG model relies on fund-level decreasing returns to scale (hereafter DRS) to maintain a non-degenerate cross-sectional distribution of funds.[3] However, existing empirical evidence on this key component of the model is mixed.[4] Early studies quantify the magnitude of DRS by regressing the fund performance on assets under management (hereafter AUM).[5] A potential concern in these studies is that fund size is not randomly assigned. An omitted factor, such as manager's skill, can affect both fund size and fund performance, leading to biases in DRS estimates. Pástor et al. (2015) use fund fixed effects to remove this bias under the assumption of constant skill and find insignificant DRS. Building on BG's framework with a flexible parameterization of DRS, we use Kalman filter to express the conditional posterior of manager's skill at a given point in time as a function of historical fund performance and AUM. This allows us to obtain a consistent estimate of the DRS using maximum-likelihood method (MLE), by matching the model-predicted performance based on these posteriors to the performance data. The estimation does not rely on the cross-sectional correlation between size and performance, which is the source of bias in regression-based estimates. A key difference between our method and Pástor et al. (2015) is that the MLE is efficient, giving us enough power to reject the null that DRS is zero. Assuming the fund size increases by \$100 million, which is about 40% of the median fund size, our estimate indicates that the annual fund performance would decrease by 105.9 bps. Our estimate is slightly smaller than the estimated DRS in Chen et al. (2004) and Ferreira et al. (2013).

The estimated model for fund performance allows us to compute the posterior belief for each fund's skill in each period. In BG's theory, fund size is adjusted by investor in- or outflows in response to the most recent performance to the extent that the expected performance (based on the posterior belief and the impact of DRS) is equated to the fund's expense ratio. This relationship allows us to compute the fund size predicted by the rational (and frictionless) benchmark. We then construct a misallocation measure which is the difference between the model-predicted fund size and observed AUM. We find that this misallocation is large and persistent.[6] More importantly, this measure of misallocation strongly predicts subsequent fund performance, indicating that it is not simply reflecting mismeasurement of skill. A standard deviation increase in misallocation leads to a statistically significant 34 bps decrease in the current-year performance and a 25 bps decrease in next-year performance. In addition, fund flows also respond to misallocation but much slower than predicted by the frictionless model. This evidence indicates that

decision-making in finance and macroeconomics.

[3] In the absence of frictions and decreasing returns to scale the cross-section of fund sizes collapses to a point mass where the most skilled fund captures the entire market.

[4] Chen et al. (2004), Ferreira et al. (2013), Yan (2008) use OLS method and find a significant DRS at the fund-level. Pástor et al. (2015) use a recursive demeaning method and fail to find DRS at the fund-level.

[5] Including Chen et al. (2004); Ferreira et al. (2013); Yan (2008).

[6] Roussanov et al. (2018) document large *cross-sectional* misallocation but do not explore its variation over time. Pastor et al. (2017) measure misallocation by looking at properties of fund portfolio returns.

investors are either adjusting their fund investments slowly in response to new information, or systematically deviating from the rational beliefs, or both.

Motivated by the above findings, we estimate a model of fund flows that allows for all of these possibilities. Specifically, the model allows investors to use a belief updating process that deviates from the rational benchmark as measured by the performance data, at least in terms of the parameters underpinning the Kalman filter. In addition, we allow for a flexible elasticity of net flows to the (implied) beliefs about manager skill. In contrast to our first step, these parameters are estimated by matching the model to the *flows* data (again with MLE), instead of the performance data.

First, we find that, comparing with the rational benchmark, investors are more optimistic about fund managers' average skill. Investors' prior on the average skill of the funds is 6.1% compared to 1.3% estimated from the performance data. Intuitively, the stark difference is due to the fact that, despite the relatively poor record of average performance of mutual funds, they still enjoy significant inflows, particularly in the first years after inceptions.

Second, we find that investors tend to over-weight recent realized performance (a noisy signal of skill) against the prior. In other words, investors seem to think that a fund's recent performance reflects the manager's skill more than what it actually does, consistent with the "over-inference" implied by the representativeness heuristic of Kahnemann and Tversky (see Rabin (2002) and Benjamin (2018) for discussion). Moreover, we find that investors discount the more distant information more heavily than a rational Bayesian would. For example, the 5-years-ago performance is weighted 0.25 times less than the last-year performance in the rational benchmark, whereas it is weighted 0.89 times less by investors. This pattern of belief updating is also consistent with the representativeness heuristic, as argued by Gennaioli and Shleifer (2010) and Bordalo et al. (2017), Bordalo et al. (2018). The implied persistence of managers' skill is also substantially lower, at 0.77 in the flow model versus 0.96 estimated by the econometrician. This suggests that investors might "chase returns" or over-extrapolate from recent performance because of an implicit belief that manager ability (or at least that of a given fund) varies over time a lot more than it actually does.

Finally, we find that investors adjust their fund investments slowly in response to shifting posterior beliefs. Over time, fund sizes do tend towards the frictionless model-predicted allocations but at a speed much slower than predicted by the theory. In particular, it is expected to take 6 years for a typical fund to reach halfway of the adjustment towards its efficient allocation. This fact might be explained by information/search costs as in Hortaçsu and Syverson (2004) and Roussanov et al. (2018), or by models of limited attention and/or adjustment costs (see Gabaix (2017) for a survey).

To the extent that investors are heterogeneous in their beliefs about mutual fund performance, we also consider flows into retail and institutional share classes separately. We estimate beliefs implied by institutional fund flows to be closer to the rational beliefs than those of retail investors,

even though they are still quite optimistic about fund managers' skill, perceive skill as much less persistent than it appears to be in the historical data, and are also quite slow in adjusting towards efficient allocations.

1.1 Literature

To the best of our knowledge, our paper is the first to formally test the model of Berk and Green (2004) or to estimate the beliefs of mutual fund investors that are implied by the data. The closest to our paper is work by Baks et al. (2001), who develop a Bayesian method of performance evaluation and find that even extremely skeptical prior beliefs still lead to sizable allocation to active mutual funds. Pástor and Stambaugh (2012) show that uncertainty about decreasing returns to scale and the priors on average manager skill can lead to slow learning. In a related but distinct branch of literature focusing on asset prices, Pástor and Stambaugh (2002) develop an econometric framework that combines investor priors about asset pricing models and managerial skill with return data.

Our paper is related to a large literature on learning in finance. Learning has been applied in different areas in finance, such as volatility and predictability of asset returns, stock price bubbles, portfolio choice, IPO, trading volume, firm profitability and etc.[7]Among those, a closely related paper is Huang et al. (2012) which use mutual fund flows to infer how rational mutual fund investors are. They find empirical evidence consistent with rational learning. Based on our structural estimation approach, we find investors significantly deviating from the rational benchmark.

Our paper is also related to the vast literature on flow-performance relationship. Prior works show that fund flows respond to fund performance (Ippolito (1992), Chevalier and Ellison (1997), Sirri and Tufano (1998)). Our contribution to this literature is to empirically estimate the canonical BG model and explore its quantitative implications.

This paper also contributes to the large literature on how agents utilize historical information. Malmendier and Nagel (2011) show that investors who have experienced low stock-market returns throughout their lives so far report lower willingness to take financial risks. Greenwood and Nagel (2009) show that younger mutual fund managers invested more heavily in technology stocks than older mutual fund managers around the peak of the tech bubble. Our contribution to this literature is to show that mutual fund investors over-weight recent performance in a manner consistent with models based on the representativeness heuristic such as those in Bordalo et al. (2017) and Bordalo et al. (2018), which feature decision-makers putting excessive weight on the most recent observations (relative to the optimal Kalman filter).

The paper is organized as follows. In Section 2, we develop the model. In Section 3, we discuss the estimation methods. In Section 4, we describe the data used for the estimation. In

[7]For an extensive survey, please refer to Pástor and Veronesi (2009).

Section 5, we present our estimation results. In Section 6, we explore the implications of the estimates. In Section 7, we extend our model to account for investor heterogeneity. Section 8 summarizes our conclusions.

2 Model

The model of rational fund flows articulated by BG has two key components: (i) fund performance dynamics driven by unobservable skill and decreasing returns to scale and (ii) fund flows that are driven by learning about fund skill. To bring these components to data, we generalize them along a few dimensions. With the generalizations, our model still nests the rational benchmark as given by BG. In Section 3, the model is structurally estimated to reveal whether and to what extent our generalizations hold according to the data.

2.1 Fund performance

The realized gross alpha, denoted by $r_{j,t}$, for an active fund $j \in \{1, ..., N\}$ in a time period t is determined by three factors: (i) the fund manager's skill to generate expected returns in excess of those provided by a passive benchmark in that period, denoted by $a_{j,t}$, (ii) the impact of decreasing returns to scale, given by $D(Q_{j,t})$ where $Q_{j,t}$ is the fund's asset under management (AUM), and (iii) an idiosyncratic shock $\varepsilon_{j,t} \sim \mathcal{N}\left(0, \delta^2\right)$. Accordingly,

$$r_{j,t} = a_{j,t} - D\left(Q_{j,t}\right) + \varepsilon_{j,t}. \tag{1}$$

The above equation is the same as Equation (1) in BG. Next, we generalize slightly from BG by allowing the manager's skill to be time-varying. We assume manager's skill follows an AR(1) process:

$$a_{j,t} = (1-\rho)\mu + \rho a_{j,t-1} + \sqrt{1-\rho^2} \cdot v_{j,t}, \tag{2}$$

where $v_{j,t} \sim \mathcal{N}(0, \kappa^2)$. We specify that a fund's first-period skill is drawn from $\mathcal{N}(\mu, \kappa^2)$, the stationary distribution of the above process. Parameter ρ captures the persistence of the skill level. In the limiting case $\rho \to 1$, skill is fixed over time. The possibility of $\rho < 1$ captures the fact that fund managers and/or their strategies may change over time. More importantly, this possibility later allows us to examine whether investors' reaction to performance history (how they weight recent vs. earlier performance) aligns with the exact underlying degree of persistence of skill.

Following BG, we assume that the manager's skill is not observable to either the investors or the fund manager herself. Let $\widehat{a}_{j,t}$ be a rational investor's belief about $a_{j,t}$ in period t. More specifically, $\widehat{a}_{j,t}$ is the posterior mean of $a_{j,t}$ given all the historical information up to $t-1$ (not

including $r_{j,t}$ or $Q_{j,t}$). One can apply Kalman filter to derive an expression for $\widehat{a}_{j,t}$. Intuitively, equation (2) can be thought of as describing how the "hidden state", $a_{j,t}$, evolves over time, and equation (1) implies that a signal of this hidden state is $r_{j,t} + D\left(Q_{j,t}\right)$. Applying the Kalman filter gives us the following recursive formula for the posterior mean:

$$\widehat{a}_{j,t+1} = \rho\left[\widehat{a}_{j,t} + \frac{\widehat{\sigma}^2_{j,t}}{\widehat{\sigma}^2_{j,t} + \delta^2}\left(r_{j,t} + D(Q_{j,t}) - \widehat{a}_{j,t}\right)\right] + (1-\rho)\mu, \tag{3}$$

where the posterior variance satisfies:

$$\widehat{\sigma}^2_{j,t+1} = \rho^2\frac{\delta^2\widehat{\sigma}^2_{j,t}}{\widehat{\sigma}^2_{j,t} + \delta^2} + (1-\rho^2)\kappa^2. \tag{4}$$

For the initial period t when fund j was born, we simply have $\widehat{a}_{j,t} = \mu$ and $\widehat{\sigma}^2_{j,t} = \kappa^2$. In the special case of $\rho = 1$, these expressions coincide with BG's Proposition 1. For $\rho < 1$, the filtered posterior beliefs assign more weights to the realized performance in the more recent periods compared to earlier periods. Later, we will estimate δ, μ, κ, and ρ all from the performance data.

As to the functional form of $D(\cdot)$, that is, decreasing returns to scale, we assume the following parameterization:

$$D(Q_{j,t};\ \eta, \gamma) = \eta \cdot \frac{Q^{\gamma}_{j,t} - 1}{\gamma},\ \gamma \in [0, 1]. \tag{5}$$

This power function specification is fairly flexible. When $\gamma = 1$, it is linear in $Q_{j,t}$; when $\gamma \to 0$, it converges to $\log(Q_{j,t})$. For the intermediate values $\gamma \in (0, 1)$, the function is somewhere in between. One of the reasons to use this flexible parameterization is that the literature is inconclusive on the appropriate functional form for DRS.[8] Later, we estimate the exact value of η and γ from the performance data.

2.2 Fund flows (asset allocation)

In the model of BG, the impact of decreasing returns to scale (DRS) under the "efficient" fund size exactly offsets the difference between the posterior belief of fund skill ($\widehat{a}_{j,t}$) and the expense ratio ($p_{j,t}$). In other words, we have

$$D(\widehat{Q}^{BG}_{j,t};\ \eta, \gamma) = \widehat{a}_{j,t} - p_{j,t}.$$

Here, we generalize slightly by allowing investors to hold a belief that is different from $\widehat{a}_{j,t}$ (which represents the correct or "rational" belief). Let $\widetilde{a}_{j,t}$ denote the investor's belief. This belief might or might not coincide with $\widehat{a}_{j,t}$. To keep our model empirically tractable, we still impose a structure on $\widetilde{a}_{j,t}$, which we will make clear in just a bit. In addition, we also allow investors to use a set of DRS parameters that might be different from the underlying correct parameters, η

[8] Linear specifications were used in Pástor et al. (2015); logarithm specifications were used in Chen et al. (2004), Yan (2008), Elton et al. (2012), Ferreira et al. (2013), Reuter and Zitzewitz (2015).

and γ. Denote the investors' parameters as $\widetilde{\eta}$ and $\widetilde{\gamma}$. The BG-implied fund size, under investor's beliefs, is given by:

$$D(\widetilde{Q}_{j,t}^{BG};\ \widetilde{\eta}, \widetilde{\gamma}) = \widetilde{a}_{j,t} - p_{j,t}. \tag{6}$$

For ease of notation, in what follows, we use the lower-case q to denote the log transformation of any value represented by Q. Thus, $\widetilde{q}_{j,t}^{BG} \equiv \log(\widetilde{Q}_{j,t}^{BG})$. Using equation (6) and equation (5), we have

$$\widetilde{q}_{j,t}^{BG} = \frac{1}{\widetilde{\gamma}} \log\left[1 + \frac{\widetilde{\gamma}}{\widetilde{\eta}}(\widetilde{a}_{j,t} - p_{j,t})\right]. \tag{7}$$

Note that for $\widetilde{\gamma} \to 0$, the above equation reduces to

$$\widetilde{q}_{j,t}^{BG} = \frac{\widetilde{a}_{j,t} - p_{j,t}}{\widetilde{\eta}}.$$

Intuitively, $\widetilde{q}_{j,t}^{BG}$ is the log size for fund j that would be achieved if the capital allocations are fully adjusted to reflect investor beliefs within the period. For empirical application, we maintain the assumption that fund size adjusts towards $\widetilde{q}_{j,t}^{BG}$, albeit allowing a possibly slower rate:

$$q_{j,t} - q_{j,t-1} = \phi(\widetilde{q}_{j,t}^{BG} - q_{j,t-1}) + \xi_{j,t} \tag{8}$$

In the above, $q_{j,t} = \log Q_{j,t}$ is the log of fund size observed in the data, ϕ governs the rate of the convergence towards the efficient fund size, and $\xi_{j,t}$ is a shock term. If $\phi = 1$ then fund flows adjust immediately in response to performance information, as is the case of the BG model. Solving for $q_{j,t}$ in equation (8) gives

$$q_{j,t} = \phi\widetilde{q}_{j,t}^{BG} + (1-\phi)q_{j,t-1} + \xi_{j,t}.$$

We may assume that the error term $\xi_{j,t}$ is an independent innovation at time t. However, it is likely that ξ is serially correlated – a fund may carry on growing from one year to the next. So we allow serial correlation through an AR(1) process:

$$\xi_{j,t} = \beta\xi_{j,t-1} + \sqrt{1-\beta^2} \cdot \zeta_{j,t},$$

and $\zeta_{j,t} \sim \mathcal{N}(0, \omega^2)$ is an innovation at time t.

To complete the model, we need to specify $\widetilde{a}_{j,t}$. We assume that it follows the same structure as $\widehat{a}_{j,t}$ but under a potentially different set of performance-model parameters: $\widetilde{\mu}$, $\widetilde{\kappa}$, $\widetilde{\delta}$, and $\widetilde{\rho}$, as well as the DRS parameters: $\widetilde{\gamma}$ and $\widetilde{\eta}$. In other words, we assume that the skill-performance process as perceived by investors follows the same distributional family as we specified in Section 2.1. As a result, $\widetilde{a}_{j,t+1}$ and $\widetilde{\sigma}_{j,t+1}^2$ follow the same recursive formula as in equation (3) and (4) but with tilded parameters:

$$\widetilde{a}_{j,t+1} = \widetilde{\rho}\left[\widetilde{a}_{j,t} + \frac{\widetilde{\sigma}_{j,t}^2}{\widetilde{\sigma}_{j,t}^2 + \widetilde{\delta}^2}\left(r_{j,t} + D(Q_{j,t};\ \widetilde{\eta}, \widetilde{\gamma}) - \widetilde{a}_{j,t}\right)\right] + (1-\widetilde{\rho})\widetilde{\mu}, \tag{9}$$

$$\widetilde{\sigma}_{j,t+1}^2 = \widetilde{\rho}^2\frac{\widetilde{\delta}^2\widetilde{\sigma}_{j,t}^2}{\widetilde{\sigma}_{j,t}^2 + \widetilde{\delta}^2} + (1-\widetilde{\rho}^2)\widetilde{\kappa}^2. \tag{10}$$

For the initial period t when fund j was born, we have $\widetilde{a}_{j,t} = \widetilde{\mu}$ and $\widetilde{\sigma}^2_{j,t} = \widetilde{\kappa}^2$. Imposing a structure on $\widetilde{a}_{j,t}$ (instead of estimating it non-parametrically) keeps our model empirically tractable. Moreover, by keeping the rational belief and investor belief in the same distributional family, it facilitates a direct comparison between the two.

It is important to note that $q_{j,t}$ as specified in our model depends on the ratio $\widetilde{\kappa}/\widetilde{\delta}$ but not the absolute sizes of $\widetilde{\kappa}$ or $\widetilde{\delta}$. To see this, note that $q_{j,t}$ depends on the investor's posterior $\widetilde{a}_{j,t}$; however, in Bayesian updating, posterior means are not affected by an increase in the prior variance $\widetilde{\kappa}^2$ *and* a proportional increase in the signal variance $\widetilde{\delta}^2$. Hence, $\widetilde{\kappa}$ and $\widetilde{\delta}$ cannot be separately identified using only data on flows. Instead we can estimate the relative precision parameter $\lambda \equiv \widetilde{\kappa}/\widetilde{\delta}$, which captures how much of variation in the observed outperformance is perceived to be driven by skill (since $\widetilde{\kappa}$ controls the dispersion in the implied distribution of skill) relative to luck (since $\widetilde{\delta}$ is the implied volatility of random shocks to outperformance).

To summarize, our model allows the investor's behavior to deviate from the rational benchmark in several aspects. The first aspect is the way they form beliefs on the managers' skills. We assume investor beliefs to follow the same parametric family as the rational benchmark, but we allow their beliefs to follow a potentially different set of parameters from the "correct" parameters consistent with the performance data. Second and related, we allow investor behaviors to reflect a different degree of DRS from that consistent with the performance data. Third, we allow investors to adjust capital allocations slowly rather than immediately to the efficient benchmark.

At this stage, it is important to point out that the "learning" in our model happens on the same subject as in BG: the investors learn about the managers' skills. One can think of a further level of learning where investors also learn about the parameters (such as η and γ). However, this would lead to a much more complex model beyond the scope of this paper. We partially extend our analysis to this issue in Section 6.3.

Finally, we do not tempt to specify an explicit model for fund pricing, $p_{j,t}$. A benefit of this approach is to avoid biasing our estimation with a possibly misspecified pricing model. However, consistent with BG, we do assume that $p_{j,t}$ does not reveal about the underlying skill $a_{j,t}$ beyond $\widehat{a}_{j,t}$.

3 Estimation

There are three sets of parameters include to be estimated: (i) η, μ, κ, δ, ρ, γ, which together govern the evolution of fund performance, (ii) $\widetilde{\eta}$, $\widetilde{\mu}$, λ, $\widetilde{\rho}$, $\widetilde{\gamma}$, which together govern investor's beliefs, and (iii) ϕ, β, ω, which affect investor's choices. Below, we describe the estimation strategy for these parameters.

For ease of notation, let $Y_{j,t} = \{r_{j,t}, q_{j,t}, p_{j,t}\}$ denote the data about fund j from period t.

By equation (1), we can write down the conditional likelihood of observing $r_{j,t}$ as

$$\Pr(r_{j,t} \mid q_{j,t}, p_{j,t}, \boldsymbol{Y}_{j,t-1}, \boldsymbol{Y}_{j,t-2}, ...) \sim \mathcal{N}\left[\widehat{a}_{j,t} - D(Q_{j,t}; \eta, \gamma),\ \widehat{\sigma}^2_{j,t} + \delta^2\right]. \tag{11}$$

In the above, $\widehat{a}_{j,t}$ is the rational posterior on skill conditional on $\{\boldsymbol{Y}_{j,t-1}, \boldsymbol{Y}_{j,t-2}, ...\}$. Its recursive expression is derived in equation (3). Particularly, note that $\widehat{a}_{j,t}$ does not change upon observing the current-period fund size $q_{j,t}$ or price $p_{j,t}$. This is because: (i) price $p_{j,t}$ is assumed not to reveal about the underlying skill $a_{j,t}$ beyond $\widehat{a}_{j,t}$, and (ii) $q_{j,t}$ is a function of $\{\boldsymbol{Y}_{j,t-1}, \boldsymbol{Y}_{j,t-2}, ...\}$ and $p_{j,t}$, plus innovation $\zeta_{j,t}$ that does not hold information about the underlying skill.

From equation (8), we can write down the conditional likelihood of observing $q_{j,t}$:

$$\Pr(q_{j,t} \mid p_{j,t}, \boldsymbol{Y}_{j,t-1}, \boldsymbol{Y}_{j,t-2}, ...) \sim \mathcal{N}\left(\phi\widetilde{q}^{BG}_{j,t} + (1-\phi)q_{j,t-1} + \beta\xi_{j,t-1},\ (1-\beta^2)\omega^2\right), \tag{12}$$

where $\xi_{j,t-1}$ can be backed out using equation (8) from the observed previous-period fund sizes as follows.[9]

$$\xi_{j,t-1} = q_{j,t-1} - \left[\phi\widetilde{q}^{BG}_{j,t-1} + (1-\phi)q_{j,t-2}\right].$$

Combining equation (11) and (12), we can write the partial likelihood function[10] as

$$\prod_{j,t} \Pr(r_{j,t}, q_{j,t} \mid p_{j,t}, \boldsymbol{Y}_{j,t-1}, ...) = \prod_{j,t} \underbrace{\Pr(r_{j,t} \mid q_{j,t}, p_{j,t}, \boldsymbol{Y}_{j,t-1}, ...)}_{\text{Performance}} \cdot \underbrace{\Pr(q_{j,t} \mid p_{j,t}, \boldsymbol{Y}_{j,t-1}, ...)}_{\text{Flows}}. \tag{13}$$

In the above, the first part of the likelihood (labeled "performance") tries to fit the observed returns, particularly how returns correlate across periods. Note this part only relies on the performance model and does *not* make any assumptions on how fund sizes are determined (in other words, how investors choose funds). The observed fund sizes do enter this part of the likelihood, but only as conditional variables to account for the DRS. In contrast, the second part of the likelihood (labeled "flows") tries to fit the observed fund sizes.

Maximizing the likelihood in equation (13) estimates the performance and flow model jointly. It is important to point out that, instead of a joint estimation, we can also conduct our estimation in a separate manner. We can either: (i) maximize the performance part in (13) alone to estimate η, μ, κ, δ, ρ, γ (which together govern the evolution of fund performance), or (ii) maximize the flow part alone to estimate $\widetilde{\eta}$, $\widetilde{\mu}$, λ, $\widetilde{\rho}$, $\widetilde{\gamma}$ (which together govern investor's beliefs) and ϕ, β, ω (which affect investor's choices). Technically, maximizing either part amounts to a *partial* likelihood estimation.

In general, a joint estimation has advantages and disadvantages. It offers more efficiency, however, mis-specification in any part of the likelihood may "contaminate" the estimates in another part, if the two parts depend on some common parameters. Fortunately, in our context,

[9] For the likelihood at period t, our estimation needs to use the information of q_{t-1} and q_{t-2} so that we restrict the sample to periods of $t \geq 3$.

[10] For partial likelihood estimation, see Wooldridge (2001), chapter §13.8.

the performance model and flow model are allowed to use two completely different sets of parameters. As a result, a joint estimation of equation (13) is equivalent to separate estimations of the performance and flow parts.

3.1 Identification of decreasing returns to scale

BG model relies on the existence of fund-level DRS to maintain a non-degenerate cross-sectional distribution of funds.[11] However, previous findings on the fund-level DRS are mixed in the literature. Early studies employ the ordinary least square (OLS) method to quantify the magnitude of DRS by directly regressing fund returns on lagged fund sizes.[12] This method generates biased estimates due to the fact that fund sizes are not randomly assigned to mutual funds. There could exist omitted factors (such as fund skill) that affect both fund size and fund returns.

Recognizing this identification challenge, Reuter and Zitzewitz (2015) use the impact of Morningstar star change on inflow as an exogenous shock to fund size to gauge the causal impact of fund size on performance. Pástor et al. (2015) use a recursive demeaning method to remove the impact of skill on fund size and fund performance. Both studies failed to find statistically significant fund-level DRS.[13]

In this paper, we employ a different method which is to structurally estimate the BG model so that we can explicitly account for manager skills. While operationally this can be implemented through a partial MLE as described above, it is still important to understand, in theory, whether it is possible at all to identify DRS in the BG model from just fund performance and size data. To this end, we offer an argument on the identification of DRS in the BG model. With the generalizations of BG that we made in Section 2, the identification should be much more complex and we do not attempt it here. Nevertheless, our Monte Carlo experiments show that all the parameters in our generalized model can be recovered.

Specifically, let $D(Q_{j,t}) = \eta q_{j,t}$, where $q_{j,t}$ is the log size of the fund j in period t (in other words, $\gamma = 0$). Also, let a fund's skill be persistent over time, that is, $a_{j,t} = a_j$ for all t (in other words, $\rho \to 1$). For the ease of notation, here we will assume that every fund is born at $t = 1$. The performance of fund j in period t is given by

$$r_{j,t} = a_j - \eta q_{j,t} + \varepsilon_{j,t}.$$

So $r_{j,t} + \eta q_{j,t}$ can be regarded as a normally distributed signal centered around the underlying skill a_j. At period T, the posterior about the skill as seen by econometrician is given by

$$\mathbb{E}(a_j \mid Y_{j,T}, ..., Y_{j,1}) = \frac{\kappa^{-2}\mu + \delta^{-2}\sum_{t=1}^{T}(r_{j,t} + \eta q_{j,t})}{\kappa^{-2} + \delta^{-2}T},$$

[11] In the absence of frictions and decreasing returns to scale, the cross-section of fund sizes collapses to a point mass where the most skilled fund captures the entire market.

[12] Including Chen et al. (2004); Ferreira et al. (2013); Yan (2008).

[13] Pástor et al. (2015) find decreasing returns to scale at the industry level.

where $Y_{j,t} = \{r_{j,t}, q_{j,t}, p_{j,t}\}$ again denotes the data about fund j from period t.[14] We assume that $q_{j,T+1}$ does not provide more information about a_j beyond $\{Y_{j,T}, ..., Y_{j,1}\}$, which holds in both BG and our model. With this assumption, we have:

$$\begin{aligned}\mathsf{E}(r_{j,T+1} \mid q_{j,T+1}, Y_{j,T}, ..., Y_{j,1}) &= \mathsf{E}(a_j | Y_{j,T}, ..., Y_{j,1}) - \eta q_{j,T+1} \\ &= \frac{\kappa^{-2}\mu}{\kappa^{-2} + \delta^{-2}T} + \frac{\delta^{-2}\sum_{t=1}^{T} r_{j,t}}{\kappa^{-2} + \delta^{-2}T} + \frac{\delta^{-2}\eta\sum_{t=1}^{T} q_{j,t}}{\kappa^{-2} + \delta^{-2}T} - \eta q_{j,t+1}.\end{aligned}$$

As an identification argument, consider $T \to +\infty$, in which case the role of prior diminishes:

$$\mathsf{E}(r_{j,T+1} \mid q_{j,T+1}, Y_{j,T}, ..., Y_{j,1}) \to \overline{r}_{j,T} + \eta\left(\overline{q}_{j,T} - q_{j,T+1}\right),$$

where

$$\overline{r}_{j,T} \equiv \frac{1}{T}\sum_{t=1}^{T} r_{j,t},$$

$$\overline{q}_{j,T} \equiv \frac{1}{T}\sum_{t=1}^{T} q_{j,t}.$$

The identification of η should be clear from the above expression. The expression also offers an intuitive way to relate DRS to data pattern. For large T, the DRS η essentially manifests itself as the elasticity at which the deviation of return from historical average (i.e., $r_{j,T+1} - \overline{r}_{j,T}$) responds to a deviation of fund size from historical average (i.e., $\overline{q}_{j,T} - q_{j,T+1}$).

4 Data

We collect data from CRSP and Morningstar. Our sample contains 2,885 well-diversified actively managed domestic equity mutual funds from the United States between 1965 and 2014. Our sample has 31,098 fund-year observations. We closely follow data-cleaning procedures in Berk and van Binsbergen (2015) and Pástor et al. (2015).

There are three main data variables to be used in estimation: annual gross realized alpha (i.e. fund performance), fund size, and expense ratio. To compute the annual realized alpha $r_{j,t}$, we start with monthly return data. We first augment each fund's monthly net return with its monthly expense ratio (1/12th of the annual expense ratio) to get the monthly gross return. Then, we regress the excess gross return (over the 1-month U.S. T-bill rate) on the risk factors throughout the life of the fund to get the betas for each fund. We multiply betas with factor returns to get the benchmark returns for each fund at each point in time. We subtract the benchmark return from the excess gross return to get the monthly gross alpha. Last, we aggregate the monthly gross alpha to the annual realized alpha $r_{j,t}$. We use 5 different benchmark models: CAPM, the three-factor model of Fama and French (1993), the four-factor model of Carhart (1997), the five-factor model of Fama and French (2015), and a six-factor model that adds a momentum factor of Fama and French (2018). For our main results, we use the Fama-French

[14] See Proposition 1 in BG for more details about the derivation.

six-factor model as the benchmark. There is an alternative way of risk adjustment which is to use Morningstar benchmark index data. There are two shortcomings about this method: 1) the data is not freely available, 2) the researcher needs to assume the loading of each fund on the benchmark. Usually, the researcher will assume a loading equal to 1. But this doesn't have to be the case for each fund in a given category. In addition, Pástor et al. (2015) show that in terms of studying the impact of decreasing returns to scale and performance, the two methods yield similar results. Therefore, we take the first route.

Fund size for each year is the fund's AUM at the end of the previous year. To make fund size comparable across time, we inflate all the fund sizes to December 2011 dollars by following Pástor et al. (2015)'s method which is to use the ratio of the total market value of all CRSP stocks in December 2011 to its value at the end of the previous year. In our dataset, there is a huge skewness in funds' AUM. From the summary statistics, we can see that the mean of funds' AUM is much larger than the median. The funds at the 99 percentile are over 4,000 times larger than the funds at the 1 percentile. And the fund size at the third quartile is 13 times larger than the fund size at the first quartile. In the literature, to study the impact of decreasing returns of scale, dollar amount of the funds' AUM were used in Pástor et al. (2015), and the logarithm of the funds' AUM were used in Chen et al. (2004), Yan (2008), Elton et al. (2012), Ferreira et al. (2013), and Reuter and Zitzewitz (2015). Our flexible functional form of the DRS allows the data to inform us about the shape of the function. To lessen the effects of "incubation bias"[15], following Fama and French (2010), we limit the tests to funds that reach 15 million 2011 dollars in AUM. Once a fund passes the threshold, it is included for all subsequent periods, so this requirement does not create selection bias.

In the mutual fund industry, a single mutual fund may provide several share classes to investors that differ in their fees structures. Following much of the literature (with some exceptions, e.g., Bergstresser et al. (2009)), we conduct our analysis at the fund level instead of the share class level. We compute a fund's AUM by summing AUM across the fund's share classes, and compute the fund's realized alphas, expense ratios by using AUM weighted average across share classes.

[Table 1 about here.]

5 Parameter Estimates

The parameter estimates of the performance and flow models are presented in Table 2.

[Table 2 about here.]

[15]For details, please refer to Evans (2010).

5.1 Performance model parameters

Our estimate of the decreasing returns to scale (DRS), η, is 23 basis points with a t-statistic of 2.3. To see the economic magnitude of this estimate, consider an increase in fund size by \$100 million, which is about a 40% increase in the size for the median size fund. The estimate of η indicates that such an increase in size is associated with a decrease in expected annual fund performance of 105.9 bps. To compare, Chen et al. (2004) find that the same increase in the fund size leads to around 110 bps decrease in fund performance.[16] Ferreira et al. (2013) find that the same increase in the fund size leads to around 124 bps decrease in fund performance.[17] Note that both papers relied on OLS regressions. Pástor et al. (2015) overcomes the potential bias in OLS by a method of recursive demeaning. Their estimate of DRS is statistically insignificant, potentially due to a lack of statistical power.

To put the DRS estimate further in perspective, we note that the volatility of annual performance in our data is around 7 percent. Hence, a decrease of 105.9 bps of a fund's performance is approximately 15% of the annual volatility in mutual fund's performance (105.9 basis points divided by 7 percent).

Parameter γ, which measures the curvature of DRS, is estimated to be very close to zero. The estimate implies that a log specification of decreasing returns to scale is most in line with the data.

The mean of the prior distribution of managerial skill is 1.3% (per annum). This number is positive and statistically significant, which means that an average active mutual fund manager is skilled. This result is consistent with previous literature, for example, Berk and van Binsbergen (2015). Based on our estimates, for a fund with the average skill level, its size should be no larger than \$324 million, otherwise, DRS would make it produce negative expected performance. However, in the data, the average fund size is \$1.45 billion. This observation indicates that the mutual fund industry on average might be too large. Meanwhile, we do find a significant variation in skill across funds. For example, for a fund with the skill level one standard deviation (0.26%) higher than the average, its size can grow to around \$1 billion before generating negative expected performance.

Parameter ρ, which measures the persistence in fund manager's skill is estimated to be 0.96. The estimate indicates a fund's skill changes slowly over time and consequently, distant past performance are likely still relevant in predicting a manager's skill. This result of skill persistence is consistent with Berk and van Binsbergen (2015), who find that cross-sectional differences in value added persist for as long as 10 years.

[16] We use their coefficient of DRS for the monthly 4-factor alpha, 0.020. The result 110 bps equals 0.020 times the logarithm of 100 million dollars times 12.

[17] We use their coefficient of DRS for the quarterly 4-factor alpha, 0.0675. The result 124 bps equals 0.0675 times the logarithm of 100 million dollars times 4

5.2 Flow model parameters

First, we see that parameter $\widetilde{\mu}$, the prior mean of manager skills, is estimated to be 6.1%. Recall this estimate represents the investor belief, and is estimated from the flow data (instead of the performance data). This estimate is close to the value of the prior mean (6.5%) shown in BG's Table 1. In their paper, the value of the prior mean is picked using a calibration procedure to match the empirical relation between the flow of funds and performance, which is conceptually similar to our estimation based on flow data.

The difference between the mean of investor prior of skills ($\widetilde{\mu}$) and the rational benchmark (μ) is about 4.8%. Together with the estimate of DRS, this difference indicates that for an average fund ($1.45 billion), investors hold an optimism that the fund generates an extra annual performance of 1.38% compared to what the performance data tell us.[18] Intuitively, the stark difference is due to the fact that despite the relatively mediocre performance record of average mutual funds, they still enjoy significant inflows, especially in the early years after a fund's inception. In Figure 1, we plot the average annual net inflow of funds as a function of fund age. We can see that before the age of 6, the average inflows are statistically greater than zeros.

[Figure 1 about here.]

Parameter ϕ, which measures the adjustment rate of flow towards the efficient fund size, is estimated to be 0.068. Recall that for $\phi = 1$, we have $q_{j,t} = \widetilde{q}^{BG}_{j,t}$, that is, the market adjusts capital allocations completely in each period. The other polar case, $\phi = 0$, means the fund flows do not adjust towards the efficient allocations at all. Our point estimate of ϕ indicates that fund flows do respond to past performance, but slowly. Based on this estimate, it takes about 6 years for a typical fund to reach halfway of $\widetilde{Q}^{BG}_{j,t}$, the efficient fund size under investor belief.[19] This slow adjustment might be explained by information/search costs as in Hortaçsu and Syverson (2004) and Roussanov et al. (2018), or by models of limited attention and adjustment costs (Gabaix, 2017 and the references there within).

Parameter $\lambda \equiv \widetilde{\kappa}/\widetilde{\delta}$ is estimated to be 0.664, which is substantially larger than $\kappa/\delta = 0.208$. Recall that δ^2 is the variance of $\varepsilon_{j,t}$, the noise on fund performance, and κ^2 is the variance of the prior. So our estimates indicate an average investor regards the realized performance as less noisy signals of underlying skills than how noisy they really are. Put differently, investors seem to under-estimate the role of luck (against skill) in a fund's performance. Consequently, investors tend to over-react to fund performance. We will discuss more on this point later in Section 6.2.

[18] The average optimism is computed as follows: $0.061 - 0.007 \times \log(1453) - 0.013 + 0.0023 \times \log(1453) = 0.0138$.

[19] For this calculation, we start with a fund whose: (i) skill equals μ, (ii) initial fund size equals the median fund size at the age of 1 ($63 million), and (iii) fund price fixes at the median expense ratios (1.14%). We set all shocks in the model to zero and check how long it takes for the fund size to reach half of $\widetilde{Q}^{BG}_{j,t}$.

6 Implications

In this section, we use several exercises to illustrate the implications of our structural estimated model. First, we illustrate the persistence of misallocations by examining whether misallocation predicts performance. Second, we graphically illustrate how investors weight the historical performance of a fund when updating their beliefs about the fund's skill, and compare their weighting scheme to what a rational investor would do. Third, we compare the estimated rational beliefs and investor beliefs in terms of the ability to predict fund performance out-of-sample. We also compare them with simple prediction rules, such as 3-year and 10-year average past performance.

6.1 Misallocation and performance

According to BG's theory, if fund allocations are efficient (that is, the decreasing return offsets any positive skill), then fund performance will not be forecastable. Following this line of thought, we analyze whether misallocation, measured as the difference between fund's actual size and efficient size, can predict fund performance. The idea is as follows: if misallocation is non-existent or simply a noise independent across periods, then on average it should not predict fund performance. However, if fund flows adjust slowly, then misallocation will be able to predict performance. Funds that are "too small" relative to the benchmark would subsequently outperform due to DRS, and funds that are "too big" would subsequently underperform.

The results are provided in Panel A in Table 3. In Column (1), we regress realized performance $r_{j,t}$ onto misallocation, computed as the difference between the actual and efficient fund size, $q_{j,t} - \widehat{q}^{BG}_{j,t}$. Here, the efficient fund size is defined by (compare to equation 6)

$$D(\widehat{Q}^{BG}_{j,t};\ \eta, \gamma) = \widehat{a}_{j,t} - p_{j,t}.$$

We find a statistically significant coefficient in front of the misallocation measure. In terms of economic magnitude, a standard deviation increase in misallocation leads to 34 bps decrease in the expected performance. To check whether our misallocation measure is sensible, we break it up into $\widehat{q}^{BG}_{j,t}$ and $q_{j,t}$ separately. The results are provided in column (2). As intended, the coefficient in front of $\widehat{q}^{BG}_{j,t}$ is positive, while the coefficient in front of $q_{j,t}$ is negative.

As a robustness check, we repeat the above analysis but replace $\widehat{q}^{BG}_{j,t}$ with $\widetilde{q}^{BG}_{j,t}$, which is the efficient fund size computed using investor's belief (see equation 6). The results are provided in columns (3) and (4), which are not qualitatively different from columns (1) and (2). The magnitude of the coefficients are somewhat larger. This is due to the fact that $\widetilde{q}^{BG}_{j,t}$ is structurally estimated to fit $q_{j,t}$, and consequently, the variation of $q_{j,t} - \widetilde{q}^{BG}_{j,t}$ tends to be smaller than the variation of $q_{j,t} - \widehat{q}^{BG}_{j,t}$.

As a further robustness check, in Panel B of Table 3, we re-run the above regressions after changing the dependent variable to the next-period performance, $r_{j,t+1}$. The results stay qualita-

tively the same. One standard deviation increase in misallocation leads to about 25 bps decrease in next period's performance.

[Table 3 about here.]

Lastly, we test the prediction of the BG model that underpins the (lack of) persistence in performance: investor flows respond to misallocation of capital, whereby inflows inrease the size of funds that are "too small" given their updated skill, while outflows shrink the funds that are "too big." Thus, we regress flow $(q_{j,t+1} - q_{j,t})$ onto the misallocation measure and other controls. The results are provided in Table 4. If investors are quick to correct misallocation and push funds' sizes towards their efficient levels, the magnitude of the coefficient in front of the misallocation should be close to 1. Meanwhile, we find that the magnitude of the coefficient is significantly *smaller* than 1. This result is consistent with our estimate of ϕ in the model, meaning that fund flows respond to misallocation but the response is much weaker than predicted by the frictionless model.

[Table 4 about here.]

6.2 Weighting scheme

Our model allows the posterior of skills to flexibly weight the past performance of a fund. There are two important aspects in the investor's weighting scheme: (i) the extend to which more distant information is discounted, as measured by parameter $\widetilde{\rho}$, and (ii) how informative the realized performance is about the underlying skill, in comparison to the prior, which is measured by $\lambda = \widetilde{\kappa}/\widetilde{\delta}$.

[Figure 2 about here.]

In Figure 2, we plot the weighting schemes for four cases. Each weighting scheme shows how to weight the historical DRS-adjusted performance, $\{r_{j,t} + D(Q_{j,t}),\ t < T\}$, when forming the posterior about $a_{j,T}$.[20]

The blue curve with stars plots the weighting scheme under the estimated performance model (i.e., how the rational posterior $\widehat{a}_{j,t}$ weights historical information). The yellow curve with circles plots the weighting scheme under the estimated investor beliefs (i.e., how the $\widetilde{a}_{j,t}$ weights historical information). Comparing these two curves, we can clearly see that relative to the rational benchmark, investors over-weight recent performance (lag period 1 to 4) in a manner

[20]The weighting scheme can be easily computed by exploiting the fact that the posterior should be a linear function of the prior and signals. Technically, we first simulate the model for a number of funds and T periods, then regress $\widehat{a}_{j,T}$ or $\widetilde{a}_{j,T}$ on $\{r_{j,t} + D(Q_{j,t}),\ t < T\}$ and the prior. If implemented correctly, the regression will have a $R^2 = 1$ and the coefficients will sum up to 1.

consistent with models based on the "representativeness" heuristic, and under-weight distant performance information (lag period 5 onward).

To obtain more intuitions on the roles of different parameters in the weighting scheme, we plot two additional curves. The red curve with squares plots the same investor's weighting scheme as the yellow curve except that we impose $\widetilde{\rho} = 0.95$ (the estimated $\widetilde{\rho} = 0.766$). Comparing the red and yellow curves, we see that a larger $\widetilde{\rho}$ (red curve) implies more weight on the distant signals, which is intuitive because a larger $\widetilde{\rho}$ means manager skills are more persistent over time. Another useful result here is that the weight on prior for the red curve is 0.26; for the yellow curve is 0.55. Hence, when $\widetilde{\rho}$ is larger (red curve), the posterior puts less weight on the prior. Intuitively, this is because a smaller $\widetilde{\rho}$ means that the manager skill reverts to the stationary distribution faster, so that the posterior should rely more on the prior. In the opposite extreme case where $\widetilde{\rho} \to 1$, the weight on the prior will go to zero as $T \to \infty$.

The gray curve with asterisks plots the same investor's weighting scheme as the yellow curve except that we impose $\lambda = 0.45$ (the estimated $\lambda = 0.664$). The weight on prior for gray curve is 0.68; for the yellow curve is 0.55. We can see that as λ gets smaller (gray curve), the weight on the prior increases. Intuitively, this is because $\lambda = \widetilde{\kappa}/\widetilde{\delta}$ measures how one perceives the precision of signals against the precision of prior.

Finally, it is important to note that the weighting curve changes differently when we vary $\widetilde{\rho}$ and when we vary λ. Particularly, $\widetilde{\rho}$ mainly calibrates the relative weights of recent vs. older signals, while λ mainly calibrates the relative weights of signals vs. prior. Conceptually, this difference explains how the two parameters can be separately identified from the data.

6.3 Out-of-sample prediction

In this section, we explore the out-of-sample prediction of outperformance by the estimated rational belief ($\widehat{a}_{j,t}$), investor belief ($\widetilde{a}_{j,t}$), and some naive predictors such the moving averages. The goal is to see whether there is any gain, and if yes, how much gain we can get from using an econometrically estimated model, when predicting future performance out-of-sample.

To operationalize, we estimate our model parameters (both the performance and flow model) using data from 1965 up to 2009. Then, we use the estimated parameters to generate the rational posteriors and investor posteriors of fund skills from 2010 to 2014. The posteriors are computed using equation (3) and (9), respectively. Note, importantly, the performance data after 2009 are used in computing these posteriors. It is the parameter estimates that are obtained without using post-2009 data. Next, we subtract the impact of DRS (using last-period fund size q_{t-1}) from the posterior to construct the predictor for fund performance $r_{j,t}$. Aside from filtered posteriors, we also consider simple moving averages, such as $\sum_{k=1}^{5} r_{j,t-k}/5$, as predictors for $r_{j,t}$.

We compute the mean squared error (MSE) of the various predictors for $r_{j,t}$ from 2010 to 2014. The results are provided in Table 5. We find that the rational posterior has the smallest

MSE, which means that it performs the best in the out of sample prediction. Both rational posterior and the investor posterior outperform the naive predictors. The result suggests that it offers some advantage to use a properly specified econometric model to extract useful information from historical data.

[Table 5 about here.]

7 Extension: institutional vs. retail investors

Up to now, we have assumed a representative investor. While this is clearly a dramatic simplification, we cannot identify the heterogeneous beliefs of individual investors without account-level data, which is not available to us. We can, however, consider different groups of investors aggregated into (admittedly coarse) segments. In particular, it is natural to ask whether institutional investors are more sophisticated (and, by extension, more "rational") than households (see, e.g., Glode et al., 2017).

In order to explore this question, we exploit the fact that there are usually multiple share classes of the same fund, some are marketed to retail investors and others are only available to institutions. We extend our model of fund flows to allow for two types of investors, who hold different beliefs and invest in the two different classes of the same fund. We assume that each type of investors are dogmatic in their beliefs (ignoring the fact that their beliefs might differ from those of the other type of investors). While this is admittedly a simplification, we make this assumption for the sake of tractability.

7.1 Model and estimation

Let there be two different classes for each fund: (1) institutional and (2) retail. Then the sizes of the two share classes add up to the total net assets of the fund:

$$Q_{j,t} = Q_{j,t}^{(1)} + Q_{j,t}^{(2)}.$$

In the above, we use superscripts 1 and 2 to denote the two share classes. For the investors in class k, where $k \in \{1, 2\}$, we denote their posterior beliefs as $\widetilde{a}_{j,t}^{(k)}$ and $\widetilde{\sigma}_{j,t}^{(k)}$, which follow the same structure as in equation (9) and (10), but under a different set of parameters $\widetilde{\mu}^{(k)}$, $\widetilde{\kappa}^{(k)}$, $\widetilde{\delta}^{(k)}$, $\widetilde{\rho}^{(k)}$, $\widetilde{\eta}^{(k)}$, and $\widetilde{\gamma}^{(k)}$. So their perceived efficient fund size is given by (compare to equation 7)

$$\widetilde{q}_{j,t}^{BG(k)} = \frac{1}{\widetilde{\gamma}^{(k)}} \log\left[1 + \frac{\widetilde{\gamma}^{(k)}}{\widetilde{\eta}^{(k)}}(\widetilde{a}_{j,t}^{(k)} - p_{j,t}^{(k)})\right]. \tag{14}$$

Note that the above is the *fund* size, not the share class size, as perceived by class-k investors. Importantly, this is because DRS happens at the fund level, not the share-class level. Because the investment strategies are the same at the two share classes within the same fund, as the size of one share class increases, it should cause decreasing return on the other share class as well.

The flow for the share class k in fund j is specified as follows (compare to equation 8)

$$q_{j,t}^{(k)} - q_{j,t-1}^{(k)} = \phi^{(k)} \cdot \left(\psi^{(k)} \cdot \widetilde{q}_{j,t}^{BG(k)} - q_{j,t-1}^{(k)} \right) + \xi_{j,t}^{(k)}. \tag{15}$$

The new parameter $\psi^{(k)}$ presents the proportion of class k in the fund's efficient size. We will estimate the parameter from the data. Given that the sizes of the two share classes should add up to the fund size, one may impose that $\psi^{(1)} + \psi^{(2)} = 1$ in the estimation. However, from the standpoint of estimation, this restriction is not necessary.[21]

Again, we allow serial correlation in the residual $\xi_{j,t}^{(k)}$ as in Section 2.2, but with class-specific parameters: $\beta^{(k)}$ and $\omega^{(k)}$, $k \in \{1, 2\}$.

The estimation of the extended model follows Section 3, with a few differences. First, the data is expanded to included class-specific records:

$$Y_{j,t} \equiv \left\{ r_{j,t}, p_{j,t}^{(1)}, p_{j,t}^{(2)}, q_{j,t}^{(1)}, q_{j,t}^{(2)} \right\}.$$

Second, the likelihood function fits the sizes of share classes, instead of the fund size. The partial likelihood on the flow model is

$$\prod\nolimits_{j,t} \left[\prod\nolimits_{k \in K(j,t)} \Pr \left(q_{j,t}^{(k)} \mid p_{j,t}^{(1)}, p_{j,t}^{(2)}, Y_{j,t-1}, Y_{j,t-2}, \ldots \right) \right].$$

In the above $K(j,t) \subseteq \{1, 2\}$ denotes the share classes that fund j offers in period t (some funds did not have an institutional class in earlier years). Ideally, we want to estimate the model with the sample that has both share classes available at the same time so that the comparison between the estimated parameters for different share classes would be meaningful. To achieve that purpose, we focus on the sample period of 2000-2014, because before 2000, there were very few institutional share classes in the data. Finally, note that the partial likelihood on the performance model stays exactly the same as in Section 3, because we have made no changes to the performance model.

7.2 Results

The estimates of the two-class model are reported in Table 6. It reveals that there is indeed some heterogeneity in beliefs, at least between the (broadly defined) institutional and retail investors. Overall, compared to the retail classes, the beliefs implied by the flows at institutional-share classes are closer to the rational benchmark as estimated from the performance data (Table 2).

Specifically, first, institutional investors hold a prior about the skill of fund managers that is lower than retail investors and closer to the benchmark ($\widetilde{\mu}^{(1)} = 3.3\%$ vs. $\widetilde{\mu}^{(2)} = 5.3\%$, and $\mu = 1.3\%$). Second, compared to retail investors, institutional investors believe that realized performance are driven more by luck rather than skill, which is more aligned with the performance data ($\lambda^{(1)} = 0.423$ vs. $\lambda^{(2)} = 0.672$, and $\kappa/\delta = 0.208$). Third, institutional investors exhibits

[21] Without the restriction, the interpretation of $\psi^{(k)}$ is more akin to what class-k investors *perceive* as the proportion of fund size that they should be responsible for.

a less degree of "over-extrapolation" compared to retail investors, in the sense of believing that skill is more persistent; however, the degree of "over-extrapolation" is still very strong compared to the benchmark ($\widetilde{\rho}^{(1)} = 0.745$ vs. $\widetilde{\rho}^{(2)} = 0.705$, and $\rho = 0.958$). Fourth, institutional investors adjust capital allocations somewhat faster than retail investors ($\phi^{(1)} = 0.076$ vs. $\phi^{(1)} = 0.064$). Lastly, institutional investors perceive a degree of DRS that is smaller than retail investors and closer to the benchmark.

[Table 6 about here.]

8 Conclusion

We estimate a structural model of investor beliefs implicit in the mutual fund flows. We compare this estimated model with the rational Bayesian benchmark that is based on past performance. Our estimates imply that investors are more optimistic about fund manager's average skill than warranted by the historical data. They over-weight recent performance in a manner consistent with models based on the "representativeness" heuristic, yet respond slowly to changes in these beliefs, consistent with limited attention and/or informational frictions. These results offer new perspective on mutual fund investor's behaviors beyond the flow-performance relationship and pave roads for fruitful future research on household finance.

References

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

Benjamin, Daniel J, "Errors in probabilistic reasoning and judgment biases," Technical Report, National Bureau of Economic Research 2018.

Bergstresser, Daniel, John MR Chalmers, and Peter Tufano, "Assessing the costs and benefits of brokers in the mutual fund industry," *Review of financial studies*, 2009, *22* (10), 4129–4156.

Berk, Jonathan B and Jules H van Binsbergen, "Measuring skill in the mutual fund industry," *Journal of Financial Economics*, 2015, *118* (1), 1–20.

_ **and Richard C Green**, "Mutual fund flows and performance in rational markets," *Journal of political economy*, 2004, *112* (6), 1269–1295.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Diagnostic expectations and credit cycles," *The Journal of Finance*, 2018, *73* (1), 199–227.

_ , _ , **Rafael La Porta, and Andrei Shleifer**, "Diagnostic Expectations and Stock Returns," Working Paper 23863, National Bureau of Economic Research September 2017.

Carhart, Mark M, "On persistence in mutual fund performance," *The Journal of finance*, 1997, *52* (1), 57–82.

Chen, Joseph, Harrison Hong, Ming Huang, and Jeffrey D Kubik, "Does fund size erode mutual fund performance? The role of liquidity and organization," *The American Economic Review*, 2004, *94* (5), 1276–1302.

Chevalier, Judith and Glenn Ellison, "Risk taking by mutual funds as a response to incentives," *Journal of Political Economy*, 1997, *105* (6), 1167–1200.

Cochrane, John H., "Finance: Function Matters, Not Size," *Journal of Economic Perspectives*, May 2013, *27* (2), 29–50.

Elton, Edwin J, Martin J Gruber, and Christopher R Blake, "Does mutual fund size matter? The relationship between size and performance," *The Review of Asset Pricing Studies*, 2012, *2* (1), 31–55.

Evans, Richard B, "Mutual fund incubation," *The Journal of Finance*, 2010, *65* (4), 1581–1611.

Fama, Eugene F and Kenneth R French, "Common risk factors in the returns on stocks and bonds," *Journal of financial economics*, 1993, *33* (1), 3–56.

_ **and** _ , "Luck versus skill in the cross-section of mutual fund returns," *The journal of finance*, 2010, *65* (5), 1915–1947.

_ **and** _ , "A five-factor asset pricing model," *Journal of Financial Economics*, 2015, *116* (1), 1–22.

_ **and** _ , "Choosing factors," *Journal of Financial Economics*, 2018, *128* (2), 234–252.

Ferreira, Miguel A, Aneel Keswani, António F Miguel, and Sofia B Ramos, "The determinants of mutual fund performance: A cross-country study," *Review of Finance*, 2013, *17* (2), 483–525.

Gabaix, Xavier, "Behavioral inattention," Technical Report, National Bureau of Economic Research 2017.

Gennaioli, Nicola and Andrei Shleifer, "What comes to mind," *The Quarterly journal of economics*, 2010, *125* (4), 1399–1433.

Glode, Vincent, Burton Hollifield, Marcin Kacperczyk, and Shimon Kogan, "Is investor rationality time varying? Evidence from the mutual fund industry," in "Behavioral Finance: WHERE DO INVESTORS'BIASES COME FROM?," World Scientific, 2017, pp. 67–113.

Greenwood, Robin and Stefan Nagel, "Inexperienced investors and bubbles," *Journal of Financial Economics*, 2009, *93* (2), 239–258.

Hortaçsu, Ali and Chad Syverson, "Product differentiation, search costs, and competition in the mutual fund industry: A case study of S&P 500 index funds," *The Quarterly Journal of Economics*, 2004, *119* (2), 403–456.

Huang, Jennifer C, Kelsey D Wei, and Hong Yan, "Investor learning and mutual fund flows," *Working paper*, 2012.

Ippolito, Richard A, "Consumer reaction to measures of poor quality: Evidence from the mutual fund industry," *The Journal of Law and Economics*, 1992, *35* (1), 45–70.

Kahneman, Daniel and Amos Tversky, "Subjective probability: A judgment of representativeness," *Cognitive psychology*, 1972a, *3* (3), 430–454.

Malmendier, Ulrike and Stefan Nagel, "Depression babies: do macroeconomic experiences affect risk taking?," *The Quarterly Journal of Economics*, 2011, *126* (1), 373–416.

Pástor, Luboš and Pietro Veronesi, "Learning in financial markets," *Annual Review of Financial Economics*, 2009, *1* (1), 361–381.

_ **and Robert F Stambaugh**, "Investing in equity mutual funds," *Journal of Financial Economics*, 2002, *63* (3), 351–380.

_ **and** _ , "On the size of the active management industry," *Journal of Political Economy*, 2012, *120* (4), 740–781.

_ , _ , **and Lucian A Taylor**, "Scale and skill in active management," *Journal of Financial Economics*, 2015, *116* (1), 23–45.

Pastor, Lubos, Robert F Stambaugh, and Lucian A Taylor, "Fund tradeoffs," Technical Report, National Bureau of Economic Research 2017.

Rabin, Matthew, "Inference by believers in the law of small numbers," *The Quarterly Journal of Economics*, 2002, *117* (3), 775–816.

_ **and Dimitri Vayanos**, "The gambler's and hot-hand fallacies: Theory and applications," *The Review of Economic Studies*, 2010, *77* (2), 730–778.

Reuter, Jonathan and Eric Zitzewitz, "How much does size erode mutual fund performance? A regression discontinuity approach," Technical Report, National Bureau of Economic Research 2015.

Roussanov, Nikolai, Hongxun Ruan, and Yanhao Wei, "Marketing mutual funds," *NBER working paper*, 2018.

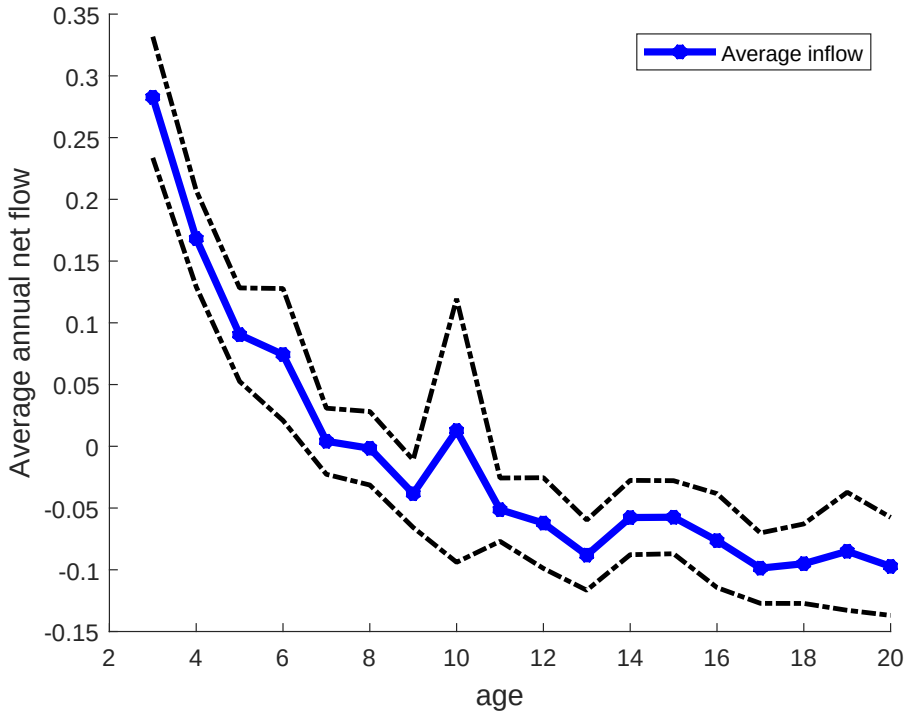
Sirri, Erik R. and Peter Tufano, "Costly Search and Mutual Fund Flows," *The Journal of Finance*, 1998, *53* (5), 1589–1622.

Tversky, Amos and Daniel Kahneman, "Judgment under uncertainty: Heuristics and biases," *science*, 1974, *185* (4157), 1124–1131.

_ **and** _ , "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment.," *Psychological review*, 1983, *90* (4), 293.

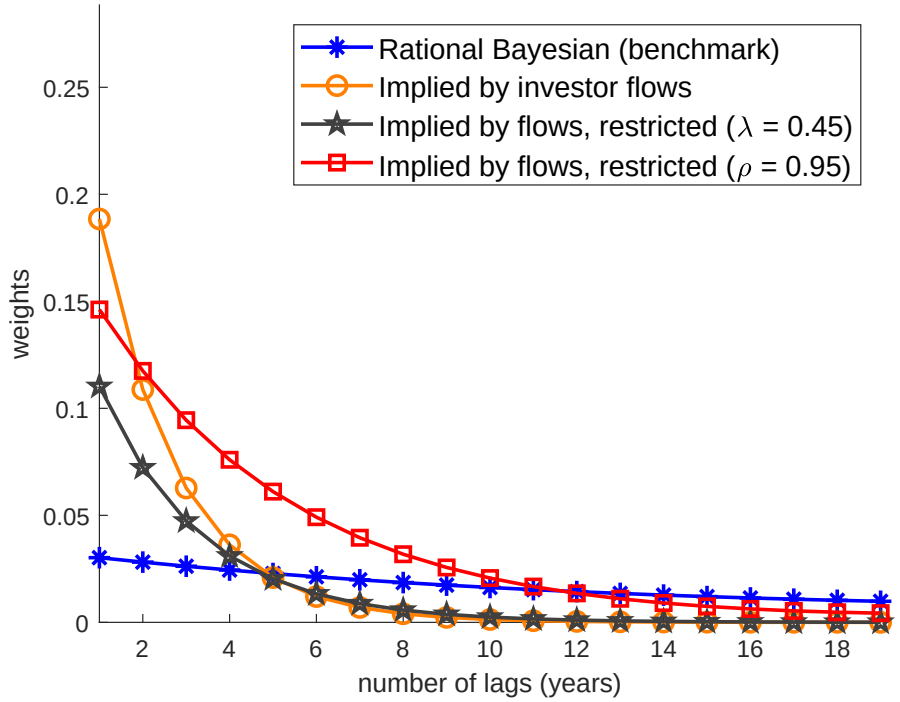
Yan, Xuemin, "Liquidity, investment style, and the relation between fund size and fund performance," *Journal of Financial and Quantitative Analysis*, 2008, pp. 741–767.

Figure 1: Average annual net flow as a function of age



This figure plots the average annual net flow against fund age. The dotted lines indicate the 95% confidence intervals.

Figure 2: Weights on historical information



A posterior mean on a fund's skill is a weighted sum of the historical DRS-adjusted performances of the fund. The weight decays with the lag between now and the year when the performance was realized. This plot displays how the weights decay in different posteriors. The blue curve with stars plots the weighting scheme under the estimated performance model (i.e., how the rational posterior $\widehat{a}_{j,t}$ weights historical information). The yellow curve with circles plots the weighting scheme under the estimated investor beliefs (i.e., how the $\widetilde{a}_{j,t}$ weights historical information). The red curve with squares plots the same investor's weighting scheme as the yellow curve except that we impose $\widetilde{\rho} = 0.95$ (the estimated $\widetilde{\rho} = 0.766$). The gray curve with asterisks plots the same investor's weighting scheme as the yellow curve except that we impose $\lambda = 0.45$ (the estimated $\lambda = 0.664$).

Table 1: Summary statistics

	Mean	SD	P1	P25	P50	P75	P99
Annual alpha (%)	0.24	7.07	-18.00	-3.22	0.04	3.26	22.28
Annual expense ratio (%)	1.18	0.50	0.14	0.90	1.14	1.43	2.55
Fund size ($million)	1,453	5,603	9	73	238	840	23,466

This table presents summary statistics for our sample of U.S. equity mutual funds. The sample period is from 1965 to 2014. Each observation is a fund-year combination. Annual alpha is computed using Fama-French six-factor model as in Fama and French (2018). Fund size is the fund's total AUM aggregated across share classes. Both annual expense ratio and annual alpha are computed as AUM-weighted averages across share classes.

Table 2: Parameter estimates

	Value	SE	Description
Panel A: Performance Model			
η	2.3e-3	(0.001)	size of DRS
γ	3.4e-4	(8.6e-4)	shape of DRS
μ	0.013	(0.003)	mean of skill prior
κ	0.014	(5.4e-4)	stdv. of skill prior
δ	0.067	(2.2e-4)	stdv. of return noise
ρ	0.958	(0.013)	skill persistence
Panel B: Flow Model			
$\widetilde{\eta}$	0.007	(0.001)	size of DRS
$\widetilde{\gamma}$	0.016	(0.002)	shape of DRS
$\widetilde{\mu}$	0.061	(0.005)	mean of skill prior
$\lambda = \widetilde{\kappa}/\widetilde{\delta}$	0.664	(0.033)	ratio of prior and noise stdv.
$\widetilde{\rho}$	0.766	(0.015)	skill persistence
ϕ	0.068	(0.005)	flow adjustment rate
β	0.344	(0.005)	serial corr. in flow residual
ω	0.340	(0.001)	stdv. of flow residual

This table reports the maximum likelihood estimates for our model. Panel A reports parameters in the model of fund performance; Panel B reports parameters in the model of fund flows. For more details about the definitions of the parameters, please refer to Section 2. The standard errors are in parentheses.

Table 3: Sensitivity of performance to misallocation

Panel A: $r_{j,t}$ as dependent variable

	(1)	(2)	(3)	(4)
$q_{j,t} - \widehat{q}^{BG}_{j,t}$	-0.11***			
	(5.54)			
$\widehat{q}^{BG}_{j,t}$		0.12***		
		(5.52)		
$q_{j,t}$		-0.20***		-0.21***
		(6.86)		(6.36)
$q_{j,t} - \widetilde{q}^{BG}_{j,t}$			-0.19***	
			(6.09)	
$\widetilde{q}^{BG}_{j,t}$				0.19***
				(5.53)
Constant	0.68***	1.14***	0.13**	0.23
	(6.72)	(6.90)	(2.99)	(1.22)
N	25,530	25,530	25,530	25,530
Adj R^2	0.003	0.003	0.004	0.004

Panel B: $r_{j,t+1}$ as dependent variable

	(1)	(2)	(3)	(4)
$q_{j,t} - \widehat{q}^{BG}_{j,t}$	-0.08***			
	(3.81)			
$\widehat{q}^{BG}_{j,t}$		0.08***		
		(3.78)		
$q_{j,t}$		-0.17***		-0.15***
		(5.62)		(4.83)
$q_{j,t} - \widetilde{q}^{BG}_{j,t}$			-0.11***	
			(4.06)	
$\widetilde{q}^{BG}_{j,t}$				0.10***
				(3.42)
Constant	0.49***	1.01***	0.11*	0.43*
	(4.89)	(5.74)	(2.47)	(2.13)
N	23,018	23,018	23,018	23,018
Adj R^2	0.001	0.002	0.002	0.002

This table reports the regressions of performance on misallocation. In Panel A, we regress the fund performance $r_{j,t}$ (in percentage) onto misallocation. In Panel B, we regress future performance $r_{j,t+1}$ (in percentage) onto misallocation. For the definitions of misallocations, please see Section 6.1. The t-statistics are in parentheses. Significance at the 1%, 5%, and 10% levels are indicated by ***, **, and *, respectively.

Table 4: Sensitivity of flows to misallocation

	(1)	(2)
$q_{j,t} - \widehat{q}_{j,t}^{BG}$	-0.03***	-0.03***
	(25.68)	(19.90)
Lag expense ratio		1.59*
		(2.20)
Lag load dummy		0.01*
		(2.08)
Lag flow		0.13***
		(15.12)
Lag annual alpha vol		-1.17***
		(9.08)
Lag log fundsize		-0.01***
		(5.40)
Lag age		2e-4
		(0.94)
Constant	5e-3*	0.07***
	(2.29)	(4.32)
N	23,018	20,716
Adj R^2	0.06	0.08

This table reports the regression of flow on misallocation. Flow is defined as $q_{j,t+1} - q_{j,t}$. The control variables include: lag expense ratio, lag load fund dummy (takes the value of 1 if the fund has front loads), lag flow, lag annual alpha vol measured as a fund's alpha's standard deviation over the prior year using monthly data, lag log of fund TNA, and lag fund age measured in years. The t-statistics are in parentheses. Significance at the 1%, 5%, and 10% levels are indicated by ***, **, and *, respectively.

Table 5: Out-of-sample MSE of various predictors for fund future performance

	(1)	(2)	(3)	(4)	(5)	(6)
	Rational posterior	Investor's posterior	Lag 1 year alpha	Lag 3 year alpha	Lag 5 year alpha	Lag 10 year alpha
MSE (%)	0.20	0.23	0.44	0.27	0.27	0.25

This table presents the mean squared errors of various predictors for realized alpha in 2010-2014. The first column uses the skill posteriors in the performance model, with the model parameters estimated from the performance data in 1965-2009. The second column uses the skill posteriors in the flow model, with the model parameters estimated from the fund size data in 1965-2009. For both columns, DRS based on last-period fund size is subtracted from the posterior for the current-period skill. The last four columns use moving averages.

Table 6: Parameter estimates: different share classes

	Inst Share		Retail Share			Benchmark	
	Value	SE	Value	SE		Value	SE
$\widetilde{\eta}$	3.7e-3	(0.001)	6.1e-3	(0.001)	η	2.3e-3	(0.001)
$\widetilde{\gamma}$	0.033	(0.015)	0.012	(0.002)	γ	3.4e-4	(8.6e-4)
$\widetilde{\mu}$	0.033	(0.004)	0.053	(0.003)	μ	0.013	(0.003)
$\lambda = \widetilde{\kappa}/\widetilde{\delta}$	0.423	(0.042)	0.672	(0.027)	κ/δ	0.208	(0.008)
$\widetilde{\rho}$	0.745	(0.024)	0.705	(0.014)	ρ	0.958	(0.013)
ϕ	0.076	(0.005)	0.064	(0.004)			
β	0.212	(0.006)	0.301	(0.005)			
ω	0.535	(0.002)	0.352	(0.001)			
$\psi^{Inst.}$	0.417	(0.055)					

This table reports parameter estimates for the extended flow model that accounts for institutional share class vs. retail share class. The sample period is from 2000 to 2014. The last two columns reproduce the rational benchmark estimates from panel A of Table 2. The standard errors are in parentheses.

Appendix

A Monte Carlo study

We conduct Monte Carlo experiments on our partial MLE described in Section 3. The first step is to simulate panel data on funds' return $r_{j,t}$, prices $p_{j,t}$, and size $q_{j,t}$. Our model specifies the data generating process for returns and sizes but not prices. Due to the partial MLE approach, we can generate $p_{j,t}$ as any function of $\{Y_{j,t-1}, Y_{j,t-2}, ...\}$. For the results shown below, we generate $p_{j,t}$ from a simple AR(1) process. In the simulated panel, we retain the same fund identities and years of existence for each fund as in the real data. As a result, the simulated panel data is of the same size as the real data.

The "true" parameters with which we generate the simulated data are set to the values in Table 2. We simulated 100 datasets, and for each dataset, we apply the partial MLE to recover the parameters. Table S1 shows the means and standard errors of the means of the recovered parameter values across the 100 datasets. As we can see, our partial MLE can recover all the parameters.

[Table S1 about here.]

B Out-of-sample estimation of fund performance

We re-conduct the out-of-sample exercises presented in Table 5 using rolling/expanding window generated alphas as our measure of fund performance. More specifically, we fix the starting point of the window at the birth time of the fund and the endpoint of the window progresses along time. The step size is one month since we are using monthly data to estimate alpha. To be able to estimate the beta for the initial periods of each fund, we set the initial size of the window as 24 months. That means the first 24 month's beta is the same. Then, from the 25th month onwards, we use last period estimated beta to compute the current period alpha. The endpoint of the window keeps on moving till the end of the fund's return data. The new results are presented in Table S2. We find that the results are quantitatively similar to the original results in Table 5.

[Table S2 about here.]

Table S1: Monte Carlo Results

	True value	Estimate mean	S.E. for mean
Performance model			
η	0.0023	0.0023	(1.95e-05)
γ	0.0001	0.0005	(3.26e-05)
μ	0.0133	0.0133	(8.44e-05)
κ	0.0139	0.0140	(7.73e-05)
δ	0.0672	0.0672	(3.49e-05)
ρ	0.9584	0.9538	(1.66e-03)
Flow model			
$\widetilde{\eta}$	0.0068	0.0068	(2.35e-05)
$\widetilde{\gamma}$	0.0162	0.0161	(1.40e-04)
$\widetilde{\mu}$	0.0606	0.0607	(1.49e-04)
$\lambda = \widetilde{\kappa}/\widetilde{\delta}$	0.6636	0.6629	(1.09e-03)
$\widetilde{\rho}$	0.7656	0.7666	(1.11e-03)
ϕ	0.0682	0.0686	(2.53e-04)
β	0.3442	0.3440	(7.12e-04)
ω	0.3396	0.3396	(2.03e-04)

This table reports the Monte Carlo results. The "true" parameters with which we generate the simulated datasets are set to the values in Table 2. We simulate 100 datasets, and for each dataset, we apply the partial MLE to estimate the parameters. This table reports the means and standard errors of the means of the parameter estimates across the 100 datasets.

Table S2: Out-of-sample MSE of various predictors for fund future performance (alternative alpha estimation)

	(1)	(2)	(3)	(4)	(5)	(6)
	Rational posterior	Investor's posterior	Lag 1 year alpha	Lag 3 year alpha	Lag 5 year alpha	Lag 10 year alpha
MSE (%)	0.22	0.27	0.55	0.32	0.31	0.29

This table presents the mean squared errors of various predictors for realized alpha in 2010-2014. The first column uses the skill posteriors in the performance model, with the model parameters estimated from the performance data in 1965-2009. The second column uses the skill posteriors in the flow model, with the model parameters estimated from the fund size data in 1965-2009. For both columns, DRS based on last-period fund size is subtracted from the posterior for the current-period skill. The last four columns use moving averages. In this table, we use rolling/expanding window generated alphas as our measure of fund performance. More specifically, we fix the starting point of the window at the birth time of the fund and the endpoint of the window progresses along time. The step size is one month since we are using monthly data to estimate alpha. To be able to estimate the beta for the initial periods of each fund, we set the initial size of the window as 24 months. That means the first 24 month's beta is the same. Then, from the 25th month onwards, we use last period estimated beta to compute the current period alpha. The endpoint of the window keeps on moving till the end of the fund's return data.